CONSERVATOIRE NATIONAL DES ARTS ET METIERS

PARIS

MEMOIRE D'ACTUARIAT

Présenté en vue d'obtenir

Le Diplôme d' ACTUARIAT du C.N.A.M.

Master Finances de marché, spécialité Actuariat

par

Jean-Marc AOUIZERATE

ALTERNATIVE NEURONALE EN TARIFICATION SANTE

Soutenu le 2 décembre 2010

MEMBRES DU JURY

M Michel FROMENTEAU

Mme Flor GABRIEL

M Pierre PETAUTON

M Dominique RIDET

M Vincent RUOL

M François WEISS

Résumé

Ces dernières années ont été marquées par une crise économique, une cohésion sociale fragile et un environnement législatif très fluctuant.

Dans le domaine de la protection sociale, le désengagement de la Sécurité sociale a reporté une partie de ses dépenses sur les régimes complémentaires dont le coût ne cesse d'augmenter.

Dans ce contexte, les entreprises sont soucieuses de maîtriser leurs budgets et d'anticiper l'évolution des montants de cotisations futures. Elles sont plus attentives au fait de payer au plus juste des garanties adaptées à leurs besoins. Gras Savoye se devait donc de répondre pleinement aux attentes de ses clients en développant des outils tarifaires performants.

Ce mémoire propose une alternative aux méthodes traditionnelles de tarification frais de santé, conçue sur la base d'un modèle de *Réseau de Neurones Artificiels*. Cette technique permet de capter automatiquement des dépendances non linéaires de haut niveau entre les variables explicatives. L'autre spécificité du modèle présenté est son approche globale du tarif qui n'est plus estimé pour chaque poste de dépenses comme dans le cadre classique.

Dans une première partie, un indicateur servant à estimer la performance des garanties a été défini. Pour un poste donné, il représente le taux de couverture obtenu en simulant la garantie correspondante sur l'ensemble du portefeuille. En neutralisant ainsi l'influence des tarifs pratiqués sur le taux de couverture, les garanties peuvent être comparées objectivement, quel que soit leur mode d'expression. Cet indice innovant servira ensuite pour intégrer la notion de garantie en entrée du modèle.

Une étude préliminaire a été menée sur le portefeuille Gras Savoye pour mettre en évidence les influences de chacun des facteurs sur la consommation médicale. Cette analyse a permis de faire une synthèse des ressources disponibles et de sélectionner les variables jugées pertinentes comme paramètres du modèle : le sexe, l'âge, le type de bénéficiaire, la C.S.P., le secteur d'activité, la localisation géographique, les garanties, le caractère obligatoire ou non du contrat, le régime (général ou Alsace Moselle) et la population (active ou inactive).

Dans une deuxième partie plus théorique, est exposé le principe général des réseaux de neurones ainsi que le panorama d'architectures et de techniques d'apprentissage existantes.

Enfin dans une dernière partie, le modèle tarifaire a été bâti, puis mis en œuvre sur le portefeuille Gras Savoye. L'ensemble de la consommation médicale de l'exercice de survenance 2007 (représentant 610 000 bénéficiaires) a été scindé en trois bases : apprentissage (40%), validation (30%) et test (30%). A partir de celles-ci, des simulations ont été réalisées selon différentes variantes pour tendre vers le réseau offrant les meilleurs résultats.

Le réseau de neurones obtenu précédemment a été confronté avec d'autres modèles avancés tels que les Modèles Linéaires Généralisés et Memory Based Reasoning. A l'issue de ce comparatif, le modèle neuronal a obtenu la plus faible erreur quadratique, ce qui a conforté le choix porté sur celui-ci.

Puis, le modèle a été appliqué dans des conditions réelles d'utilisation sur un panel de 13 nouveaux contrats entrés dans notre portefeuille après l'élaboration du modèle (60 000 bénéficiaires). Dans 79% des cas, les tarifs prédits par le modèle neuronal se sont avérés être plus proches de la consommation réelle que les taux de cotisations pratiqués issus d'outils de tarification standard.

La cohérence des résultats avec une logique économique a également été validée (croissance de la consommation avec l'âge, consommation plus élevée chez la population féminine,...).

L'expérimentation d'un modèle de réseau de neurones dans le domaine de la tarification santé a démontré ici son utilité en procurant des résultats plus probants qu'avec d'autres techniques de cotation.

Mots-clefs : Réseaux de Neurones - Perceptron Multicouches - Imputations Multiples - Augmentation de Données - Tarification - Assurance Frais de Santé

Abstract

Recent years have been marked by an economic crisis, a social cohesion difficulty and a legislative environment very changeable.

In social protection, the withdrawing of the French social security system has transfered a portion of its expenses to the supplementary insurance whose cost continues to rise.

In this context, firms are anxious to control their budgets and anticipate changes in the amounts of future contributions. They are more attentive to paying the true value of guarantees appropriate to their needs. Gras Savoye has to come up to expectations of its clients by developing powerful pricing tools.

This paper proposes an alternative to traditional methods of pricing for additional health insurance developed on the basis of a model of *Artificial Neural Networks*. This technology can automatically capture nonlinear dependencies between high-level explanatory variables. The other feature of the presented model is its overall approach to the tariff that is not valued for each item of expense, as it would have been in the classical framework.

In the first part, an indicator to estimate the performance of the guarantees has been defined. For a given medical expense item, it corresponds to the cover rate obtained by simulating the guarantee required to the entire portfolio. By neutralizing the influence of prices on the coverage, guarantees can be compared objectively regardless of their mode of expression. This innovative indicator is then used to integrate the concept of guaranteed into the model input.

A preliminary study has been conducted on the Gras Savoye portfolio to highlight the influences of each factor on medical consumption. This analysis led to a synthesis of available resources and to the selection of the relevant variables considered as parameters: sex, age, type of beneficiary, socio-professional group, sector, geographical location, guarantees, mandatory or voluntary contract, system (general or Alsace Moselle) and population (active or inactive).

In the second, more theoretical, is exposed the general principle of neural networks and the panorama of architectures and existing learning techniques.

Finally, in the last section, the princing model was built and implemented on the Gras Savoye's portofolio. All of the medical consumption occurred in 2007 (representing 610,000 beneficiaries) have been divided into three bases : training (40%), validation (30%) and test (30%). From these, simulations were performed by different variants to approach the network offering the best results.

The Neural Network obtained above was confronted with more advanced models such as Generalized Linear Models and Memory Based Reasoning. At the end of this comparison, the neural model obtained the lowest squared error, which has reinforced the decision to select it.

Then, the model has been applied to real conditions of use on a panel of 13 new contracts entered into our portfolio after the development of the model (60,000 beneficiaries). In 79% of cases, the tariff predicted by the neural networks were found to be closer to real medical consumption than insurance premiums rate from standard pricing tools.

The consistency of the results with an economic logic has also been validated (growth in consumption with age, higher consumption among the female population ,...).

The testing of a pricing model of neural networks applied to the sector of health insurance demonstrated its utility in providing better results than with other pricing techniques.

Key-words : Neural Networks - Multilayer Perceptron - Multiple Imputations - Data Augmentation - Pricing - Health Insurance

« Car la science en l'homme arrive la première.
Puis vient la liberté. »

Victor Hugo : « A qui la faute ? », *écrit le 25 juin 1871, L'Année terrible, VIII, 1872*

Je tiens tout d'abord à remercier les membres du corps professoral du Conservatoire National des Arts et Métiers pour la qualité de leur enseignement.

Mes remerciements vont également à ma famille et tout particulièrement à ma femme Claire, qui m'a toujours encouragé et soutenu tout au long de mon cursus et m'a permis d'aller jusqu'au bout et à mon fils Simon, pour sa patience face à un papa toujours étudiant.

Je tiens à exprimer toute ma gratitude à mon tuteur de mémoire Dominique Ridet, Directeur de l'Audit Interne de Gras Savoye, pour sa disponibilité, pour la pertinence de ses conseils qui ont su m'orienter efficacement dans mes réflexions ainsi que pour sa relecture.

Je remercie également Irène Granjon, responsable du service Actuariat, pour ses judicieux conseils et sa relecture.

Ma reconnaissance s'adresse aussi, pour leurs remarques pertinentes et leur relecture,

à mes collègues : Ahoua Assouan, Valérie Rosselin, Hélène Boucheau, Rozenn Treussier, Julien Rollet,

et amis : Chloë Marais, Colin Marais, Johan Kononovitch, Laurent Bris.

Enfin, je remercie aussi l'entreprise Gras Savoye qui m'a permis de suivre cette formation et donné les moyens de réaliser ce mémoire.

Table des matières

Introduction

La santé en France

En 2009, la part des dépenses courantes de santé des français représentait presque 12% de l'ensemble de leur consommation avec 223,1 milliards d'euros dont 175,7 milliards de C.S.B.M.[1] qui exclue, entre autres, les soins aux personnes âgées en établissement et les indemnités journalières[2].

D'après l'Enquête Santé Protection Sociale menée par l'I.R.D.E.S.[3], 93% des français sont couverts par une complémentaire santé. Parmi les personnes couvertes, environ quatre personnes sur dix bénéficient d'un contrat collectif et six sur dix d'un contrat individuel [4].

76% des dépenses de C.S.B.M. sont couverts par la Sécurité sociale, 14% sont financés par les organismes complémentaires d'assurance maladie et un peu plus de 9% restent à la charge des assurés. En valeur, la progression observée en 2009 (+3,3%) est moins marquée qu'en 2008 (+3,7%), mais rapportée au PIB, sa part a augmenté de 8,7% en à 9,2%. En volume, elle se maintient à +3% par rapport à l'an passé.

Après avoir subi une diminution entre 2005 et 2008 liée aux différentes mesures économiques, la part de l'assurance maladie obligatoire dans le financement des dépenses santé reste stable entre 2008 et 2009. Celle des organismes complémentaires est en légère progression de 13,7% à 13,8% et celle restant à la charge des ménages baisse légèrement de 9,5% à 9,4%.

Avec une progression de 6% en 2008, le « marché » de la complémentaire santé représentait un chiffre d'affaires de 29 milliards d'euros selon les déclarations des organismes complémentaires au Fonds de financement de la Couverture Maladie Universelle. Ce marché connait toujours une progression soutenue. De 2001 à 2008, il a progressé de 61% soit une croissance annuelle moyenne de 7,4%.

L'environnement de la santé en France évolue inexorablement vers une plus grande sollicitation des complémentaires santé. Les budgets alloués à la couverture maladie représentent une part de plus en plus importante pour les entreprises pour lesquelles la santé devient un réel enjeu économique. Le pilotage des régimes prend ainsi une dimension majeure au sein des sociétés qui souhaitent mieux maîtriser leurs dépenses tout en maintenant un niveau de protection convenable pour leurs salariés. Le contrôle des coûts, la mise en place de réformes, l'apparition de garanties innovantes ont donné une impulsion à l'activité tarification chez les différents acteurs de l'assurance et du courtage.

Le modèle proposé

Toujours dans l'optique de mieux répondre aux attentes des entreprises, Gras Savoye a mené une réflexion sur le pilotage des régime frais de santé. Celle-ci a abouti à l'élaboration d'une technique novatrice permettant d'appréhender plus précisément le risque et par conséquent le coût de leur régime.

[1] Consommation de Soins et de Biens Médicaux

[2] source : rapport de la cours des comptes de la Sécurité sociale

[3] Institut de Recherche et Documentation en Economie de la Santé

[4] source : D.R.E.E.S. (Direction de la Recherche, des Etudes de l'Evaluation et des Statistiques)

Sur de nombreux postes médicaux tels que l'optique ou le dentaire, les comportements des consommateurs sont étroitement liés aux niveaux de couverture dont ils disposent. De manière générale, la part de financement supportée directement par le salarié peut freiner sa consommation et par conséquent limiter les remboursements effectués par sa complémentaire. Cette influence du reste à charge intervient à différents niveaux. Premièrement, un effet dissuasif peut se produire, entraînant une diminution du nombre d'actes consommés. Deuxièmement, les assurés, davantage responsabilisés par leur contribution pécuniaire, peuvent être plus attentifs aux tarifs pratiqués, ce qui implique une baisse des coûts moyens. Troisièmement, les remboursements assureur peuvent être optimisés en ajustant les dépenses engagées sur les limitations contractuelles des garanties. Or, dans le cadre d'une méthode actuarielle de tarification classique d'un régime complémentaire santé, ces phénomènes ne sont pas recensés : pour une population donnée, les fréquences et distributions des coûts sont supposées constantes. C'est ce qui nous a motivé à expérimenter ici une modélisation tarifaire intégrant ces corrélations. Dans cette logique, nous avons pensé que la consommation médicale pouvait s'inscrire naturellement dans une approche neuronale et constituer ainsi une alternative aux méthodes plus conventionnelles de cotation.

Les techniques de modélisation non linéaires par apprentissage, développées récemment, suscitent beaucoup d'intérêt tant au sein du milieu académique que dans de nombreux secteurs d'activité tels que l'industrie, les sociétés de services ou encore les laboratoires de recherche. Les réseaux de neurones, inspirés directement du schéma de traitement de l'information des neurones biologiques, ont bénéficié du succès de la notion sous-jacente d'intelligence artificielle. Ces outils mathématiques performants permettent de classifier ou bien de régresser des fonctions complexes sans modélisation analytique particulière (aspect « boite noire »). Ce concept novateur, associé à l'évolution de la vitesse des traitements informatiques, ont fortement contribué à l'essor de ce modèle.

Dans un premier temps, nous allons créer un indicateur synthétique permettant d'effectuer des comparaisons homogènes de niveaux de garantie et ce quel que soit le mode d'expression de celle-ci. Nous porterons une attention toute particulière au traitement des données manquantes, notamment à l'aide de techniques d'*Imputations Multiples* et de l'emploi de méthodes de *Monte Carlo par Chaînes de Markov* (*M.C.M.C.*). Nous listerons ensuite l'ensemble des ressources disponibles afin de sélectionner les variables que nous utiliserons en entrée du modèle. Ce choix se fera notamment à l'aide d'analyses factorielles et de classifications pour mesurer l'incidence qu'elles peuvent avoir sur les dépenses et mettre en exergue leurs dépendances (Partie I).

Nous introduirons ensuite la définition d'un réseau de neurones, le mode de propagation de l'information, les différentes architectures possibles et les techniques d'apprentissage. Nous nous attarderons plus spécialement sur le modèle retenu, celui du *Perceptron Multi-Couches* avec l'*algorithme de rétropropagation de l'erreur*. La justification de la convergence du modèle ainsi que sa qualité d'approximateur universel seront évoquées succinctement (Partie II).

Enfin, nous appliquerons un modèle de réseau de neurones à la tarification d'un régime santé, après avoir rappelé brièvement le fonctionnement des méthodes plus traditionnelles et définit les postulats du modèle. Différentes techniques seront alors testées puis comparées pour sélectionner au final l'architecture et le mode d'apprentissage offrant les meilleurs résultats. Nous conclurons sur la mise en pratique du modèle dans un contexte réel d'utilisation en le confrontant à une méthode classique (Partie III).

Guide de lecture : Gras Savoye et son environnement

Gras Savoye est le premier courtier en assurances de France. En tant qu'intermédiaire, il met en relation son client et un organisme assureur. Il présente, propose et aide à conclure des contrats. Il a également un rôle de conseil dans le pilotage de régimes pour défendre au mieux les intérêts de l'assuré : évaluer les risques, proposer des garanties adéquates et innovantes, concevoir les contrats. Une grande partie de son activité concerne l'assurance collective frais de santé (ou assurance-groupe). Ces contrats interviennent en complément de la Sécurité sociale et éventuellement d'autres mutuelles. Gras Savoye dispose d'une délégation de gestion qui lui permet de rembourser directement les bénéficiaires à la place de l'assureur.

Dans tout ce qui suit, les notations suivantes seront utilisées :

- le montant des frais réels noté FR correspond aux dépenses engagées, c'est le tarif facturé par le praticien ou l'établissement de santé,
- le montant remboursé par la Sécurité sociale noté SS, est établi à partir d'une base tarifaire : la B.R. (Base de Remboursement) et d'un taux (généralement 70%),
- le montant remboursé par une première mutuelle noté $MUT1$ (souvent celle du conjoint),
- le montant remboursé par Gras Savoye noté GS,
- le reste à charge noté RAC, correspond à la part de dépenses non couverte par l'ensemble des intervenants.

Les garanties frais de santé sont définies par poste. Selon les contrats, elle peuvent être exprimées sous différentes formes et assiettes :

- en euro,
  par exemple, 200 €
- en pourcentage des frais réels (F.R.),
  par exemple, 90% des frais réels
- en pourcentage de la Base de Remboursement (B.R.) de la Sécurité sociale,
  par exemple, 400% de la B.R.
- en pourcentage du remboursement de la Sécurité sociale (R.S.S.),
  par exemple, 400% du R.S.S.
- en pourcentage du Ticket Modérateur (T.M.),
  par exemple, 100% du T.M.
- en pourcentage du Plafond Mensuel de la Sécurité sociale (P.M.S.S.),
  par exemple 10% du P.M.S.S.
- une combinaison d'assiette,
  par exemple 90% des F.R. limité à 400% de la B.R.
- y compris ou non le remboursement de la Sécurité sociale,
  par exemple 400% de la B.R. y compris Sécurité sociale
- avec des paliers,
  par exemple, 100% du T.M. puis 90% des F.R. au-delà.

Exemple de prise en charge d'une monture adulte ($BR = 2,84$ €) par un contrat prévoyant une garantie de 100 € :

- La monture est facturée 200 € par l'opticien ($FR = 200$),
- la Sécurité sociale rembourse 1,85 € ($SS = 65\% \times 2,84 = 1,85$),
- la mutuelle du conjoint rembourse 50 € ($MUT1 = 50$),
- le contrat géré par Gras Savoye rembourse 100 € ($GS = 100$),
- le reste à charge est de 48,15 € ($RAC = 200 - 1,85 - 50 - 100$).

Première partie

ANALYSE DESCRIPTIVE DES DONNEES SANTE

Cette première partie décrit un état des lieux précis des différentes données disponibles. L'objectif est double : fournir une aide dans la sélection des paramètres du modèle de tarification et compenser l'absence d'explication des réseaux de neurones sur leurs résultats.

Les influences sur la consommation médicale de facteurs tels que l'âge, le sexe, la localisation géographique, les secteur d'activité et les garanties seront mesurées.

Les garanties constituent l'un des paramètres clef en tarification. Elles seront converties en un indicateur synthétique, valant entre 0% et 100%, mesurant la performance de leur couverture. L'avantage de cet indice est de pouvoir quantifier, de manière homogène, des libellés exprimés différemment. L'intégration des garanties comme paramètres du modèle se fera au travers de cet indicateur.

La phase de codification des garanties fait appel, notamment, à des techniques d'*Imputations Multiples* utilisant des méthodes de *Monte Carlo par Chaînes de Markov*. La théorie sous-jacente sera préalablement décrite.

Chapitre 1

Base de données

1.1 Descriptif de l'infocentre

Le périmètre étudié recouvre l'intégralité du portefeuille de contrats d'assurance collective gérés par Gras Savoye (près de 700 entreprises).

En termes de volumétrie :

- Près de 610 000 bénéficiaires étaient recensées en 2007, soit plus de 270 000 familles assurées,
- Le nombre de lignes de décomptes[1] remboursées en 2007 avoisinait les 12 millions,
- Le montant des prestations était de l'ordre de 220 millions d'euros.

L'accès aux données de l'infocentre ainsi que l'ensemble des traitements statistiques seront effectués essentiellement sous *SAS*.

L'entrepôt des données consacré aux prestations, créé en 1994 est alimenté mensuellement. Il s'organise selon les quatre axes principaux suivants :

- la ligne de décompte :
  - *la date de soins*
  - *la date de règlement*
  - *le code de l'acte médical*
  - *le nombre d'actes (ou d'unités d'actes)*
  - *le montant des frais réels*
  - *le montant des remboursements effectués par la Sécurité sociale*
  - *le montant des remboursements effectués par une autre mutuelle*
  - *le montant des remboursements effectués par Gras Savoye*
  - *la base de remboursements de la Sécurité sociale*
  - *le taux de remboursement appliqué*

[1] Le décompte de remboursement de la Sécurité sociale et/ou complémentaire santé est un document papier ou dématérialisé. Il reprend les intitulés des actes médicaux pris en charge avec leur taux de remboursement. Ils précisent les dates et les montants des virements effectués

- les informations relatives au bénéficiaire :
  - *la date de naissance*
  - *le sexe*
  - *le lien familiale*
  - *l'adresse*

- l'adhésion à un contrat :
  - *l'alternative d'adhésion*
  - *l'établissement*
  - *le groupe*

- l'adhésion à l'entreprise :
  - *le nom*
  - *l'adresse*
  - *le code APE*
  - *l'activité*
  - *la catégorie INSEE*
  - *la classe INSEE,*

Les limitations de garanties ne sont pas accessibles en lecture directe dans les restitutions de notre système d'information. Nous ne pouvions nous passer de cette information capitale pour mener à bien notre étude. Deux possibilités s'offraient à nous : rechercher l'information manuellement ou estimer les garanties en vigueur. Nous avons retenu la deuxième solution qui sera développée dans une section ultérieure.

Dans notre étude, nous avons choisi d'utiliser les regroupements d'actes médicaux les plus significatifs suivants, dont le code abrégé est indiqué entre parenthèses :

- Verre (VER)
- Verre progressif de 0 à 2 dioptries (VP1)
- Verre progressif de 2,25 à 4 dioptries (VP2)
- Verre progressif de 4,25 à 6 dioptries (VP3)
- Verre progressif de plus de 6 dioptries (VP4)
- Verre simple de 0 à 2 dioptries (VS1)
- Verre simple de 2,25 à 4 dioptries (VS2)
- Verre simple de 4,25 à 6 dioptries (VS3)
- Verre simple de plus de 6 dioptries (VS4)
- Verre simple toutes dioptries (VS)
- Verre progressif toutes dioptries (VP)
- Monture (MON)
- Forfait optique (FOR)
- Chambre particulière (CHP)
- Honoraires en milieu hospitalier (HOH)
- Cure thermale (CTH)
- Forfait maternité (MAT)
- Consultation ou visite de généraliste (CGE)
- Consultation ou visite de spécialiste (CSP)
- Prothèses dentaires remboursées (PDA)
- Soins conservateurs (SCV)

1.2 Traitement des données manquantes et aberrantes

La base de données sur laquelle porte l'étude nous est transmise par le centre de gestion via une extraction de l'infocentre. Les lignes de décompte peuvent être directement intégrées par l'intermédiaire de différents flux télétransmis ou bien saisies manuellement par des gestionnaires. Il s'agit ici d'une base de gestion, en d'autres termes le principal objectif est de permettre de rembourser correctement chaque consommateur. Certaines pratiques employées pour y parvenir peuvent parfois générer une perte ou déformation de l'information.

En plus des variables présentes dans notre base, nous avons également passé en revue un certain nombre d'indicateurs moyens préalablement calculés qui seront repris dans la suite de l'étude :

- Les frais réels, Sécurité sociale, autres mutuelles, Gras Savoye moyens :
  Calcul par an par bénéficiaire effectué au prorata du temps de présence sur l'année,

- Les indicateurs de niveau de garantie (le calcul sera décrit dans une section ultérieure) :
  Indices de 0% à 100%, 100% étant une couverture intégrale des dépenses par le régime,

- Les fréquences annuelles par bénéficiaire :
  Nombre d'actes consommés annuellement,

- Les coûts moyens par acte :
  Dépenses rapportées au nombre d'actes,

- Les taux de couverture annuel par bénéficiaire :
  Dépenses remboursées rapportées au frais réels engagés.

Les fréquences indiquées pour les postes dentaires sont données en nombre d'unités d'actes (ou coefficients). Par exemple, une prothèse codifiée en "SPR 50" comporte 50 unités.

Les taux de couverture sont obtenus uniquement à partir des lignes de décomptes enregistrées en gestion, les actes non couverts par une garantie ne sont donc pas systématiquement pris en compte. Selon les règles d'alimentation de notre infocentre, un décompte n'apparaît que s'il y a au minimum une de ses lignes ayant un remboursement complémentaire non nul. Une partie des actes non couverts n'est alors pas intégrée, ce qui augmente artificiellement les taux de couverture.

Nous avons choisi de procéder au calcul des statistiques élémentaires suivantes sur ces indicateurs :
- Moyenne
- Ecart-type
- Minimum
- Maximum
- Quartile inférieur
- Médiane
- Quartile supérieur

Ces statistiques sont données, pour les actes les plus représentatifs, dans le tableau suivant :

Variable	Moyenne	Ecart-type	Min	Q. inf	Méd.	Q. sup	Max
Frais Réels/bénéf.	812 €	2 585 €	0 €	30 €	305 €	874 €	99 987 €
Gras Savoye/bénéf.	391 €	1 221 €	0 €	10 €	120 €	429 €	93 341 €
Sécurité soc./bénéf.	362 €	1 394 €	0 €	14 €	132 €	369 €	78 404 €
Autres mut./bénéf.	8 €	183 €	0 €	0 €	0 €	0 €	47 189 €
Indic. Garanties VER	74%	20%	5%	66%	73%	92%	100%
Indic. Garanties VP1	75%	14%	16%	67%	73%	91%	96%
Indic. Garanties VP2	77%	11%	12%	70%	80%	86%	96%
Indic. Garanties VP3	80%	9%	47%	77%	81%	85%	99%
Indic. Garanties VP4	84%	7%	56%	82%	83%	85%	100%
Indic. Garanties VS1	73%	16%	8%	55%	78%	86%	99%
Indic. Garanties VS2	80%	12%	30%	75%	84%	87%	99%
Indic. Garanties VS3	80%	9%	49%	75%	77%	86%	100%
Indic. Garanties VS4	88%	8%	64%	78%	92%	93%	100%
Indic. Garanties VS	79%	14%	41%	66%	76%	96%	97%
Indic. Garanties VP	83%	10%	39%	83%	83%	93%	97%
Indic. Garanties MON	79%	15%	3%	74%	80%	87%	100%
Indic. Garanties FOR	71%	24%	4%	65%	83%	87%	99%
Indic. Garanties CHP	95%	5%	45%	94%	96%	98%	100%
Indic. Garanties HOH	98%	2%	49%	98%	99%	99%	99%
Indic. Garanties CTH	82%	8%	54%	76%	81%	88%	98%
Indic. Garanties MAT	43%	21%	1%	29%	39%	57%	100%
Indic. Garanties CGE	97%	2%	73%	96%	98%	98%	100%
Indic. Garanties CSP	93%	7%	61%	93%	97%	98%	100%
Indic. Garanties PDA	81%	11%	25%	76%	85%	89%	100%
Indic. Garanties SCV	96%	2%	80%	94%	96%	97%	100%
Fréquence VER	0,27	0,7	0	0	0	0	9
Fréquence MON	0,13	0,34	0	0	0	0	4
Fréquence FOR	0,03	0,19	0	0	0	0	4
Fréquence CHP	0,009	0,586	0	0	0	0	320
Fréquence HOH	0,012	0,168	0	0	0	0	39
Fréquence CTH	0,002	0,061	0	0	0	0	4
Fréquence MAT	0,014	0,119	0	0	0	0	4
Fréquence CGE	2,23	3,23	0	0	1	3	96
Fréquence CSP	1,15	2,73	0	0	0	1	99
Fréquence PDA	7,29	34,99	0	0	0	0	950
Fréquence SCV	7,71	22,99	0	0	0	0	744
Coût moyen VER	156 €	112 €	1 €	80 €	120 €	204 €	998 €
Coût moyen MON	156 €	69 €	1 €	112 €	143 €	188 €	963 €
Coût moyen FOR	399 €	201 €	4 €	262 €	380 €	490 €	2 141 €
Coût moyen CHP	60 €	58 €	0 €	42 €	52 €	64 €	1 260 €
Coût moyen HOH	147 €	215 €	1 €	40 €	76 €	182 €	4 832 €
Coût moyen CTH	667 €	486 €	0 €	349 €	536 €	861 €	3 445 €
Coût moyen MAT	532 €	403 €	0 €	177 €	483 €	805 €	3 964 €
Coût moyen CGE	24 €	7 €	0 €	21 €	22 €	22 €	434 €
Coût moyen CSP	37 €	15 €	1 €	23 €	34 €	45 €	380 €
Coût moyen PDA	10,2 €	3,0 €	1,0 €	8,4 €	10,2 €	12,0 €	61,0 €
Coût moyen SCV	2,7 €	2,1 €	0,2 €	2,4 €	2,4 €	2,4 €	99,2 €
Couverture VER	86%	21%	1%	80%	95%	100%	100%
Couverture MON	88%	20%	1%	82%	100%	100%	100%
Couverture FOR	82%	28%	0%	72%	100%	100%	100%
Couverture CHP	97%	10%	20%	100%	100%	100%	100%
Couverture HOH	100%	3%	21%	100%	100%	100%	100%
Couverture CTH	88%	20%	8%	78%	100%	100%	100%
Couverture MAT	99%	6%	10%	100%	100%	100%	100%
Couverture CGE	98%	6%	1%	100%	100%	100%	100%
Couverture CSP	97%	8%	2%	99%	100%	100%	100%
Couverture PDA	86%	16%	3%	75%	91%	100%	100%
Couverture SCV	99%	46%	1%	100%	100%	100%	100%

Les indicateurs de niveaux de garantie, que nous détaillerons plus tard, peuvent s'apparenter à des taux de couvertures fictifs obtenus en simulant l'application de garanties à l'ensemble du portefeuille. Le fait que les taux de couverture soient systématiquement supérieurs aux indicateurs de niveau de garantie peut s'expliquer par l'influence que peuvent avoir les garanties sur les comportements de consommation médicale.

Les valeurs extrêmes que nous avons pu isoler trouvent leurs explications dans les deux pratiques suivantes :

- la présence de protocoles de gestion spécifiques : certains contrats peuvent avoir une partie de leur périmètre gérée par un autre système. Il se peut que des lignes de décomptes dont les dépenses sont globalisées soient affectées à une personne fictive pour que le montant de la consommation globale soit préservé. En revanche, cette pratique rend inexploitables les données de ces contrats qui n'indiquent pas le détail individuel.

- le mode de calcul au prorata du temps de présence : dans le cas d'une dépense très coûteuse sur une période très courte, le fait d'annualiser au prorata du temps de présence amplifie le phénomène. Bien qu'ayant un sens en termes de moyenne, il se peut que la valeur ainsi obtenue soit incohérente en tant que telle. Par exemple, dans le cas d'une personne présente un seul mois qui aurait eu une hospitalisation très onéreuse à 100 K€, le fait d'annualiser cette dépense porte celle-ci à 1 200 K€, montant exorbitant et irrationnel en pratique.

Afin de ne pas perturber notre analyse, nous avons effectué un retraitement sur l'ensemble des données servant à l'étude.

La question s'est posée du traitement à appliquer aux lignes révélant une anomalie (valeur manquante ou aberrante). Etait-il préférable de remplacer une valeur anormale par une estimation (valeur moyenne par exemple ou plus complexe encore avec la méthode de bootstrap) ou bien de la supprimer tout simplement ? Compte tenu du volume important de décomptes sur la période d'observation (près de 12 millions), nous avons privilégié la suppression de celles-ci qui ne représentaient que 1%.

Les principales corrections apportées sont les suivantes :

- Traitement des lignes de régularisation : afin de conserver l'historique des écritures passées, aucun décompte ne peut être supprimé. Pour l'annuler, il faut repasser une autre ligne signée négativement. Cet artifice de gestion ne reflète donc pas une réalité statistique. Nous avons donc effectué un traitement regroupant toutes les lignes de régularisation avec celles devant être rectifiées,

- Traitement des lignes ayant des montants de frais réels nuls : remplacement des frais réels par la somme des remboursements de la Sécurité sociale et des organismes assureurs,

- Traitement des lignes ayant un total des dépenses dépassant les frais réels : remplacement des frais réels par la somme des remboursements de la Sécurité sociale et des organismes assureurs.

Les principaux motifs de suppression de lignes sont les suivants :

- Suppression des lignes ayant au moins une valeur aberrante : certaines lignes dont les valeurs dépassaient des normes acceptables ont été ôtées de l'étude,

- Suppression des lignes ayant une valeur manquante,

- Suppression des lignes sans intervention de Gras Savoye,

- Suppression des lignes correspondant à un transfert de flux avec d'autres systèmes.

Cette première analyse nous a permis de mettre en évidence certaines valeurs hors normes telles que les dépenses annuelles moyennes par adhérent dont les valeurs maximales atteignent plusieurs millions d'euros (3,6 M€ de frais réels, 1,5 M€ de remboursements Gras Savoye, ...) ou encore des erreurs sur le nombre d'unités (200 verres consommés par une seule personne). Au global, 0,1% des bénéficiaires (742) ont été écartés de cette étude car au moins une de leurs caractéristiques était considérée comme hors normes.

Chapitre 2

Niveaux de garanties

2.1 Problématique

L'expression des garanties n'étant pas accessible via notre infocentre, nous avons choisi de définir une méthodologie permettant d'ordonner les différents niveaux de couverture en repartant des prestations déjà versées.

Dans le paramétrage d'un contrat, c'est l'alternative d'adhésion qui définit les garanties à appliquer. Un contrat peut être constitué de plusieurs alternatives, celles-ci couvrant différentes catégories de la population assurée. Elles peuvent dépendre des critères suivants : CSP, différentes entités, actif/inactifs, régime général/régime Alsace-Moselle, métropolitains/expatriés/détachés...

Certains postes tels que la pharmacie, les auxiliaires médicaux, les analyses et radios n'ont que très peu de dépassements. Les garanties y afférant n'apportent alors que très peu d'information. Pour cette raison, seuls les postes que nous estimions, a priori, comme étant les plus discriminants en terme de niveau de garantie ont été sélectionnés.

Les postes que nous avons retenus sont les suivants :
- les montures
- les verres (garanties indépendantes de la correction ou grilles optiques)
- les forfaits optiques
- les consultations/visites de généralistes
- les consultations/visites de spécialistes
- les prothèses dentaires remboursées
- les soins dentaires (soins conservateurs ou inlays/onlays)
- les chambres particulières
- les honoraires chirurgicaux
- les forfaits maternité
- les cures thermales

Cette codification des garanties devait répondre également à la problématique d'une classification ouverte à un large panel de garanties. Eu égard à la diversité des expressions de garanties, il nous a paru nécessaire de trouver un indicateur synthétique permettant d'effectuer un classement significatif entre des couvertures de différentes natures.

Par exemple, comment comparer une garantie à 90% des frais réels avec une à 400% de la BR (BR Base de Remboursement de la Sécurité sociale) ?

Prenons la cas d'une prothèse dentaire de type "SPR50" dont le montant des frais réels s'élève à 500 €. La Sécurité sociale rembourse : 70%×107,50 €, soit 75,25 €, la garantie à 90% des frais réels : 374,75 € , celle à 400% de la BR : 424,75 €.

Si le montant des frais réels s'élevait à 800 €, la première garantie rembourserait : 644,75 € et la deuxième : 430 €.

Lorsque le montant des frais réels dépasse $(430 + 75,25)/90\% = 561,40$ €, la garantie à 90% des frais réels offre un meilleur remboursement que celle à 400% de la BR.

Selon l'importance des montants des dépenses engagées, l'une ou l'autre des garanties peut être la plus avantageuse. De ce fait, il n'est pas possible d'effectuer une comparaison entre deux garanties, qui serait systématiquement vérifiée dans toutes les situations.

L'exemple suivant illustre la problématique de comparaison entre les deux garanties évoquées précédemment.

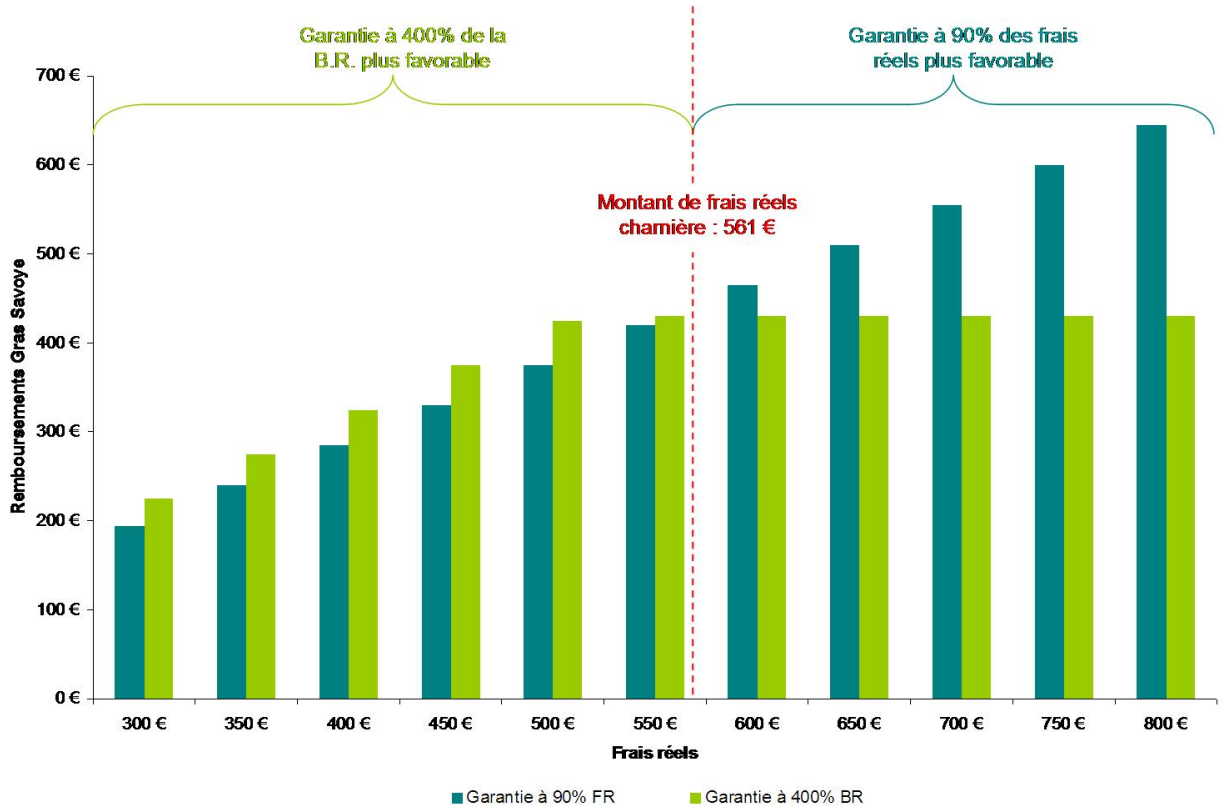


FIG. 2.1 – Comparatif de garanties

2.2 Création d'un indicateur du niveau de garantie

2.2.1 Principe

Nous avions initialement pensé à repérer les lignes de remboursements sur lesquelles il y avait un reste à charge afin d'en déduire le niveau de la garantie correspondante. Malheureusement, cette solution ne pouvait pas s'appliquer sur les garanties exprimées en pourcentage des frais réels pour lesquelles il y a systématiquement du reste à charge.

Le taux de couverture aurait pu être sélectionné comme indicateur du niveau de garantie. Ce choix, bien que naturel, présente un biais : les frais réels moyens par actes peuvent différer d'une alternative à une autre. Et ce, pour de multiples raisons : la localisation géographique, la CSP ou encore l'influence du niveau de garantie sur les dépenses engagées. En effet, le caractère haut de gamme d'un régime peut avoir un effet incitatif sur les frais réels. Des taux de couverture peuvent être identiques avec des niveaux de garanties différents tout simplement parce que les frais réels peuvent s'aligner sur les garanties.

Les dépenses engagées influent directement sur le taux de couverture. Une moyen de contourner cet écueil est de comparer les taux dans des mêmes conditions de frais réels. Pour y parvenir, il suffit de considérer la répartition des dépenses observée sur l'ensemble du portefeuille comme référentiel. De ce fait, l'indicateur retenu est le taux de couverture fictif obtenu en appliquant la garantie considérée à l'ensemble des prestations du portefeuille. L'indice est simulé en remplaçant la distribution des dépenses de l'alternative d'adhésion par celle portefeuille. En procédant ainsi, les impacts des variations de dépenses engagées sur les taux de couverture sont alors neutralisés. L'objectif recherché est bien atteint car seule la garantie a un effet sur le taux de couverture. Par ce biais, l'indicateur reflète bien le niveau des garanties. Il est alors possible de hiérarchiser des garanties de natures différentes.

Techniquement, ces taux sont calculés en procédant selon la procédure définie ci-après.

2.2.2 Méthode de calcul

Evacuons tout d'abord le cas simple des garanties exprimées sous forme de forfait comme les garanties maternité ou cure thermale, pour lesquelles la notion de frais réels n'existe pas. Pour cette raison, le calcul de l'indicateur n'est pas possible. Nous avons fixé arbitrairement le montant maximum de la garantie à 1 200 € et calculé l'indicateur comme suit : $CV_{Alt,acte} = \frac{GS_{Alt,acte}}{1200}$.

Pour les postes, la distribution des frais réels notés FR est discrétisée par tranches de coûts : T_k avec $k \in [1, K]$ où K est le nombre de tranches. Les pas des tranches sont calibrés en fonction des actes médicaux et sont exprimés selon l'assiette de la garantie la plus usité pour chacun des postes : en euros, en % du PMSS ou en % de la BR (exemples : tranches de 50 € pour les montures ou les verres, de 50% de la BR pour les consultations ou les prothèses dentaires ...)

Soient les lignes de décomptes indicées : $i \in [1 : n]$, $\mathbf{D}$ étant l'ensemble des lignes de décomptes du portefeuille et $\mathcal{D}$ l'ensemble de ses parties et Alt un sous-ensemble tel que $Alt \in \mathcal{D}$. Les parties Alt symbolisent chacune des alternatives d'adhésions. $nb\,actes_i$ et FR_i représentent respectivement le nombre d'actes et le montant des frais réels correspondant à la ligne i.

Afin d'alléger les notations de cette section, nous supposons implicitement que toutes les expressions sont indiquées pour un acte donné.

Pour une alternative *Alt*, la répartition des frais réels dans chacune des tranches de dépenses peut être définie comme étant le rapport entre le montant des frais réels dans la tranche k et le montant total des frais réels.

En d'autres termes, la répartition des frais réels par tranches de coûts notée $r_{k,Alt}$ pour la tranche k peut être retranscrite comme suit :

$$\boxed{r_{k,Alt} = \frac{\sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times FR_i}{\sum_{i \in Alt} FR_i}}$$

Notons $R_{Alt,acte}$ le vecteur colonne associé à l'acte *acte* reprenant le poids des frais réels dans chacune des tranches pour une alternative *Alt* :

$$\overrightarrow{R_{Alt,acte}} = \begin{bmatrix} r_{1,Alt} \\ r_{2,Alt} \\ \vdots \\ r_{K,Alt} \end{bmatrix}$$

Les types de dépenses associés à la ligne i sont définis de la façon suivante :

- les remboursements de la Sécurité sociale : SS_i
- les remboursements d'une première mutuelle : $MUT1_i$
- les remboursements assureur : GS_i
- le reste à charge RAC_i tels que : $FR_i = SS_i + MUT1_i + GS_i + RAC_i$

Le taux de couverture de la ligne i noté : CV_i est donné par la relation suivante :

$$CV_i = \frac{SS_i + MUT1_i + GS_i}{FR_i}$$

Le taux de couverture associé à *Alt* noté : CV_{Alt} est donné par la relation suivante :

$$CV_{Alt} = \frac{\sum_{i \in Alt} SS_i + MUT1_i + GS_i}{\sum_{i \in Alt} FR_i}$$

Le taux de couverture associé à la tranche k de Alt noté : $CV_{k,Alt}$ est donné par la relation suivante :

$$\boxed{CV_{k,Alt} = \frac{\sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times (SS_i + MUT1_i + GS_i)}{\sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times FR_i}}$$

En développant CV_{Alt} :

$$CV_{Alt} = \frac{\sum_{k=1}^{K} \sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times (SS_i + MUT1_i + GS_i)}{\sum_{i \in Alt} FR_i}$$

$$CV_{Alt} = \frac{\sum_{k=1}^{K} \left[\sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times FR_i \right] \times \frac{\sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times (SS_i + MUT1_i + GS_i)}{\left[\sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times FR_i \right]}}{\sum_{i \in Alt} FR_i}$$

$$CV_{Alt} = \frac{\sum_{k=1}^{K} \sum_{i \in Alt} \mathbb{1}_{\{T_{k-1} < FR_i \leq T_k\}} \times FR_i \times CV_{k,Alt}}{\sum_{i \in Alt} FR_i}$$

$$CV_{Alt} = \sum_{k=1}^{K} r_{k,Alt} \times CV_{k,Alt}$$

Notons $CV_{Alt,acte}$ le vecteur colonne associé à l'acte *acte* indiquant le taux de couverture de chaque tranche pour le sous-ensemble *Alt* :

$$\overrightarrow{CV_{Alt,acte}} = \begin{bmatrix} CV_{1,Alt} \\ CV_{2,Alt} \\ \vdots \\ CV_{K,Alt} \end{bmatrix}$$

Le taux de couverture associé à l'acte *acte* de la partie *Alt* est le suivant :

$$CV_{Alt,acte} = \begin{bmatrix} r_{1,Alt} & r_{2,Alt} & \dots & r_{K,Alt} \end{bmatrix} \times \begin{bmatrix} CV_{1,Alt} \\ CV_{2,Alt} \\ \vdots \\ CV_{K,Alt} \end{bmatrix}$$

ou vectoriellement :

$$\boxed{CV_{Alt,acte} = {}^t\overrightarrow{R_{Alt,acte}} \times \overrightarrow{CV_{Alt,acte}}}$$

L'indicateur de niveau de garantie recherché noté $GAR_{Alt,acte}$ est donné par la formulation suivante, en remplaçant $\overrightarrow{R_{Alt,acte}}$ par $\overrightarrow{R_{Referentiel,acte}}$:

$$\boxed{GAR_{Alt,acte} = {}^t\overrightarrow{R_{Referentiel,acte}} \times \overrightarrow{CV_{Alt,acte}}}$$

Le référentiel choisi est l'ensemble des contrats de notre portefeuille.

En résumé, l'indicateur de garantie indique le taux de couverture obtenu en simulant l'application des garanties de l'alternative considérée à l'ensemble du portefeuille. En le définissant ainsi, il neutralise l'effet que peuvent avoir les frais réels sur le taux de couverture et permet de hiérarchiser de manière homogène des garanties de natures différentes.

L'exemple suivant est celui du poste consultation de spécialiste avec une garantie à 400% de la BR. Pour chaque tranche de dépenses, sont représentés le taux de couverture et le poids des dépenses. Jusqu'à la limite de garantie, les dépenses sont couvertes intégralement. Au-delà, du reste à charge apparaît.

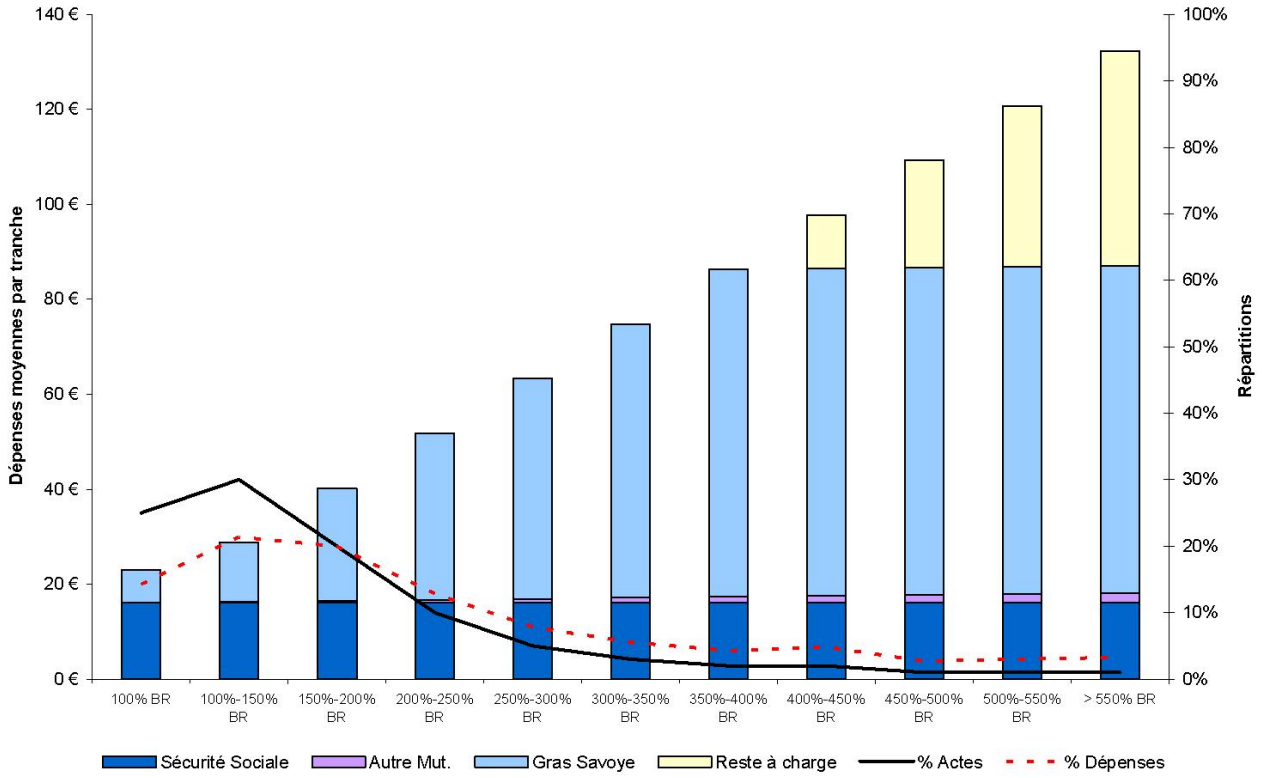


FIG. 2.2 – Répartition par tranche de dépenses - Exemple : consultations de spécialistes avec une garantie à 400% de la BR

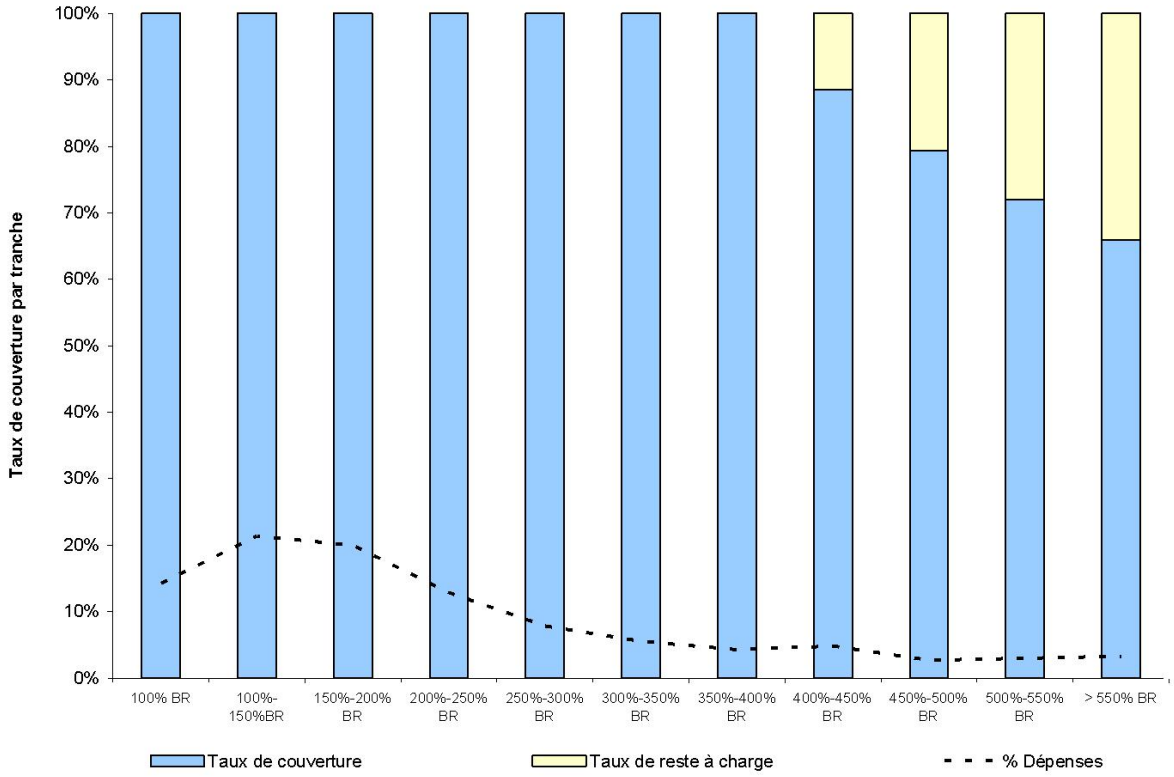


FIG. 2.3 – Taux de couverture par tranche - Exemple : consultations de spécialistes avec une garantie à 400% de la BR

Chapitre 3

Traitement des données manquantes de l'indicateur de garantie

3.1 Différentes approches du manque d'information

Pour être calculé, l'indicateur de garantie défini dans la section précédente nécessite qu'il y ait eu au moins un acte dans chaque tranche de dépenses. Or, en fonction du volume d'actes dans le périmètre observé, il se peut qu'il y ait des tranches de dépenses dans lesquelles il n'y ait aucun acte. Le calcul du taux de couverture de la tranche correspondante devient ainsi impossible, et par suite celui de l'indicateur de garantie également.

Une tranche de dépenses dépourvue d'acte ne permet donc pas de calcul du taux de couverture de celle-ci puisque les notions de frais réels, de remboursements de la Sécurité sociale, d'autres mutuelles et de Gras Savoye sont inconnues.

Deux alternatives de traitement des données manquantes pouvaient être envisagées : soit intervenir directement sur les taux de couverture manquants, soit en amont sur les différentes dépenses sous-jacentes (frais réels, Sécurité sociale, autres mutuelles et Gras Savoye) pour ensuite calculer celui-ci.

Nous avons opté pour la deuxième solution en appliquant un traitement différent adapté à chaque type de dépenses. Ce choix nous permet de tirer profit au maximum de la connaissance que nous avons des données.

3.2 Caractéristiques des données manquantes

Pour nous orienter dans le choix de méthodes à appliquer pour appréhender au mieux le manque d'information, nous avons choisi de lister les spécificités propres à chaque type de dépenses après avoir déterminé au préalable la nature de ces données manquantes.

Une tranche de dépenses sans décompte nécessite d'évaluer les données moyennes suivantes :
- Frais réels
- Remboursements Sécurité sociale
- Remboursements d'autres mutuelles
- Remboursements Gras Savoye

Les données manquantes peuvent appartenir à l'une des trois catégories suivantes :

- *Missing Completely At Random* (*MCAR*) : La probabilité qu'une observation soit manquante est constante. Elle ne dépend donc ni des valeurs des variables observées ni de la valeur non observée.
- *Missing At Random* (*MAR*) : La probabilité qu'une observation soit manquante ne dépend que des valeurs observées.
- *Missing Not At Random* (*MNAR*) : La probabilité qu'une observation soit manquante dépend de la valeur de la variable non observée.

Dans le cas présent, les quatre informations manquantes sont quasiment de type *MAR*. En effet, la valeur elle-même n'influe pas sur le fait que l'observation soit manquante ou non. Nous employons le terme "quasiment" car nous négligeons le fait qu'une faible intervention de la complémentaire santé puisse freiner la consommation d'actes coûteux. Dans ce cas précis, la non observation serait liée à la valeur des remboursements Gras Savoye. La proportion de données manquantes est essentiellement liée à l'alternative d'adhésion observée : plus le périmètre de bénéficiaires associé est restreint, plus il est probable qu'il y ait des observations manquantes.

Par exemple, il est fort probable que dans une alternative d'une dizaine de personnes, il n'y ait pas eu un acte de consommé dans chacune des tranches. En revanche, il n'y a aucune raison particulière pour que la valeur du remboursement Gras Savoye moyen sur une tranche influe sur son observation (en excluant l'hypothèse qu'il pourrait y avoir des différences significatives de niveaux de garanties liées à la taille de l'alternative).

Les différentes tranches de dépenses sont ordonnées et constituent ainsi des séries chronologiques pour chacune des dépenses. Une autre caractéristique des données manquantes est qu'elles ne sont pas monotones, c'est à dire qu'une valeur manquante sur une tranche n'entraîne pas une valeur manquante pour toutes les tranches suivantes.

D'une manière générale, les tendances attendues pour un même acte et une même alternative selon les tranches sont les suivantes :

- les frais réels moyens par tranche devraient être strictement croissants de par la définition même de la tranche,
- les remboursements de la Sécurité sociale moyens devraient être sensiblement identiques (sauf pour les verres, où les corrections les plus fortes sont généralement les plus chères mais aussi les mieux remboursées par la Sécurité sociale),
- les remboursements Gras Savoye devraient être croissants puis constants une fois la limite de garantie atteinte,
- les taux de couverture devraient décroître une fois la limite de garantie atteinte,

3.3 Méthodologie appliquée à chaque type de dépense

Parmi les différentes dépenses, seuls les remboursements Gras Savoye contiennent une information propre liée à l'alternative d'adhésion. En effet, le montant remboursé étant calculé en fonction de la garantie associée, celui-ci dépend donc intégralement de l'alternative. Contrairement aux autres dépenses, pour une donnée manquante, seul le remboursement Gras Savoye peut tirer de l'information sur les valeurs des autres tranches. Pour cette raison, nous avons choisi de traiter différemment les remboursements Gras Savoye et les autres dépenses.

Pour les frais réels, les remboursements de la Sécurité sociale et les autres mutuelles, nous avons remplacé la donnée manquante par la donnée moyenne de la tranche correspondante. Pour les remboursements Gras Savoye qui comportent l'information majeure, nous avons choisi de remplacer les données manquantes par une méthode plus adaptée développée dans le paragraphe suivant : la méthode des *Imputations Multiples avec augmentation de données.*

3.4 Imputations Multiples

La solution envisagée pour les remboursements Gras Savoye est d'estimer les valeurs manquantes par la méthode des "Imputations Multiples" qui permet de générer plusieurs jeux de données où les valeurs manquantes sont complétées par plusieurs valeurs plausibles. L'intérêt de cette méthode est de refléter correctement l'incertitude des valeurs manquantes tout en préservant les aspects importants des distributions ainsi que les relations entre les variables.

Cette technique intuitive constitue une alternative intéressante pour contrer le problème des données manquantes. Développés en premier par Rubin [RUB87] dans un contexte de sondage avec non-réponse puis par Schaffer [SCH97], ces modèles d'imputation multiple s'inscrivent dans un cadre bayésien visant à générer des données manquantes.

La méthode nécessite la vérification d'hypothèses et notamment celle d'"imputation multiple propre", c'est à dire que la variabilité introduite puisse garantir que les inférences valident le modèle.

Ce processus décrit par Rubin s'organise en trois étapes :
- Imputation
- Analyse
- Combiner les résultats obtenus

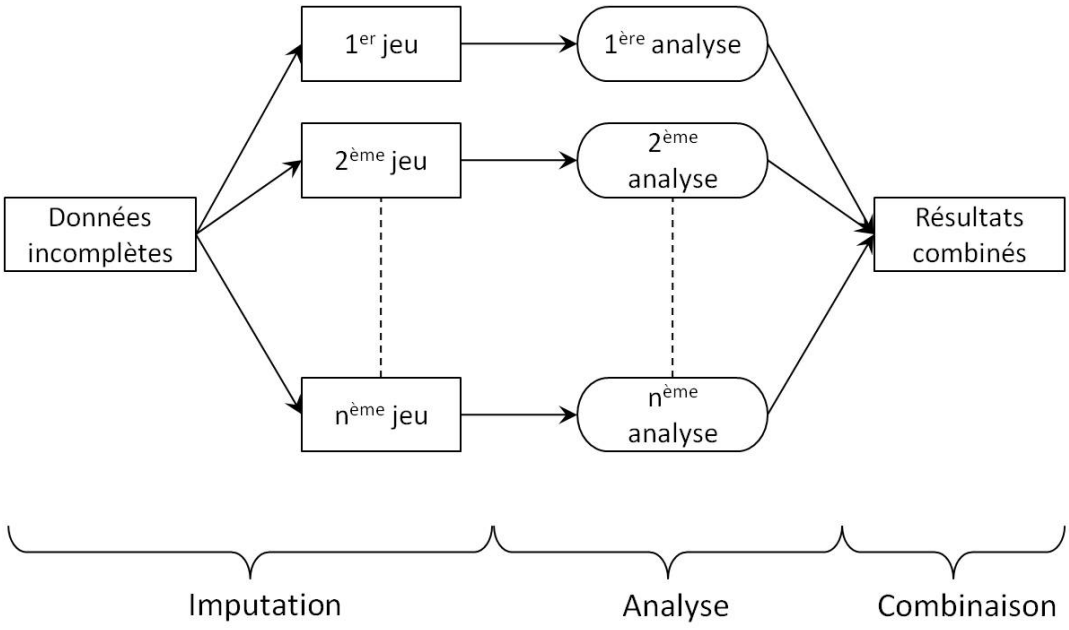


FIG. 3.1 – Méthodologie de l'imputation multiple

Phase Imputation

Chaque donnée manquante est complétée par $m > 1$ valeurs simulées afin de générer m jeux de données. Différentes méthodes peuvent être utilisées pour le remplacement des données manquantes : score de propension, régression (linéaire, log-linéaire ou logistique) ou encore *MCMC* (partiellement pour rendre les données monotones afin de pouvoir employer un autre modèle ou bien en complétant entièrement par *MCMC*). Rubin préconise d'utiliser seulement 3 ou 4 jeux de données.

Phase Analyse

Les m jeux de données complétées peuvent ainsi être analysés indépendamment par des techniques de régression classiques pour obtenir une estimation des paramètres d'intérêt de la modélisation. Un grand avantage de cette méthode réside dans la faculté de pouvoir utiliser des méthodes statistiques standard sur des ensembles complets de données, après imputation.

Phase Combinaison des résultats

Les m jeux de données sont ensuite combinés pour estimer les paramètres d'intérêt du modèle. L'assemblage des résultats des m analyses permet ainsi de refléter la variabilité supplémentaire due aux données manquantes. Une étude de la variance peut ensuite être menée pour mesurer l'influence de la non-observation des données sur les estimations.

Nous allons développer les règles d'induction définies par Rubin permettant de regrouper les informations de chaque jeux de données complétées.

Dans sa méthodologie, le paramètre d'intérêt du modèle de régression recherché noté Q est estimé simplement par la moyenne de ceux trouvés dans chacune des imputations :

$$\overline{Q} = \frac{1}{m} \sum_{i=1}^{m} \widehat{Q_i}$$

La variance de l'estimé combiné $\overline{Q}$ se décompose en deux parties :

La variance intra-imputation obtenue par la moyenne des variances :

$$U = \frac{1}{m} \sum_{i=1}^{m} \widehat{U}_i$$

et la variance inter-imputation :

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\widehat{Q}_i - Q) \times (\widehat{Q}_i - Q)^T$$

La variance totale s'obtient en sommant les deux variances à un coefficient de correction près :

$$T = \overline{U} + (1 + \frac{1}{m}) \times B$$

La statistique $(Q - \overline{Q}) \times T^{-\frac{1}{2}}$ est approximativement distribuée comme une loi de Student t avec v_m degrés de liberté :

$$(Q - \overline{Q}) \times T^{-\frac{1}{2}} \sim t_{v_m}$$

où v_m est donné par :

$$v_m = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1}) \times B}\right]^2$$

Un intervalle de confiance de niveau α pour Q est défini par :

$$Q \in [\widehat{Q} - t_{v_m,1-\frac{\alpha}{2}} \times \sqrt{T}, \widehat{Q} + t_{v_m,1-\frac{\alpha}{2}} \times \sqrt{T}]$$

Deux statistiques aident à diagnostiquer la manière dont les données non observées contribuent à l'incertitude sur le paramètre estimé :

La première appelée l'augmentation de variance due à la non observation de donnée, notée r :

$$r = \frac{(1+m^{-1}) \times B}{U}$$

et la seconde appelée la part d'information manquante, notée λ :

$$\lambda = \frac{(r+2)/(v+3)}{r+1}$$

Rubin définit également l'efficience relative :

$$RE = (1 + \frac{\lambda}{m})^{-1}$$

Cet indicateur permet de mesurer en unité de variance, selon la part d'information manquante, le gain en terme de précision en fonction du nombre d'imputations.

Efficience relative :

	λ				
m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

Avec des proportions d'informations manquantes raisonnables, seulement quelques imputations donnent déjà des résultats de bonne efficacité.

La méthode choisie pour simuler les valeurs manquantes est une méthode de *Monte Carlo par Chaîne de Markov* (*M.C.M.C*) dite d'"*augmentation de données*" ou "*Data Augmentation*" (*D.A.*). Les paragraphes suivants esquissent une ébauche succincte des notions sous-jacentes dont elle découle. Ces techniques assez pointues seront détaillées en annexes.

3.5 Méthodes de Monte Carlo par Chaînes de Markov (M.C.M.C.)

3.5.1 Généralités

Dans ce paragraphe, nous allons aborder succinctement les principes de base des méthodes de *Monte Carlo par Chaînes de Markov* dites *M.C.M.C.* (*Monte Carlo Markov Chain*). Nous vous proposons de vous référer au cours de Sandrine Vaton [VAT02] ou à l'ouvrage de Christian Robert [ROB96] pour approfondir le sujet de manière plus détaillée.

Ces méthodes ont été initialement élaborées pour répondre à des problématiques physiques telles que l'exploration des distributions d'équilibre des interactions moléculaires. Leur essor depuis le début des années 90 est essentiellement dû au développement des techniques informatiques et notamment grâce aux vitesses de calcul qui n'ont cessé de s'accroître. Elles sont utilisées principalement pour l'échantillonnage de variables aléatoires, le calcul d'intégrales et l'optimisation de fonctions lorsqu'il n'existe pas de solutions analytiques, ou pour résoudre des problèmes en grande dimension. Les méthodes les plus connues sont : l'*algorithme de Hastings-Métropolis* (cf. *Annexe B*), l'*échantillonneur de Gibbs* (cf. *Annexes C et D*) et la *Data Augmentation* développée plus loin.

Une chaine de Markov est une séquence de variables aléatoires dont la probabilité de passage d'un état au suivant ne dépend que du précédent : $p(x_i|x_{i-1}, ..., x_0) = p(x_i|x_{i-1})$

Elle se définit selon deux composantes :
- la distribution initiale : $p(x_0)$
- le noyau de transition : $T(x, A) = p(x_{i+1} \in A|x_i = x)$

L'idée sous-jacente des méthodes *MCMC* est de construire une chaine de Markov prenant la forme d'une marche guidée pour explorer l'espace multidimensionnel des paramètres pour parvenir à estimer une distribution de probabilité en les échantillonnant périodiquement.

Les méthodes *MCMC* s'inscrivent dans une approche d'inférence bayésienne où l'information des paramètres inconnus est contenue dans la loi de distribution a posteriori $p(\theta|y)$ qui est obtenue en utilisant le Théorème de Bayes :

Théorème 3.5.1 (Théorème de Bayes)**.**

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

La distribution a priori de ce paramètre aléatoire $p(\theta)$ et la vraisemblance $p(y|\theta)$ des observations y définissent ainsi le cadre du modèle. Le facteur de normalisation $p(y)$ est une constante indépendante de θ ce qui réduit le théorème de la façon suivante :

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Plusieurs finalités de l'analyse bayésienne peuvent se résumer à calculer, quelle que soit la fonction g où l'espérance existe, l'expression suivante :

$$\mathbb{E}(g(\theta)|y) = \int_\theta g(\theta)p(\theta|y)dy$$

Or, il n'est quasiment jamais possible d'obtenir l'expression précédente analytiquement, d'où la nécessité de recourir à des simulations.

Le principe général des méthodes de Monte Carlo est de tirer aléatoirement un échantillon $\theta^{(m)}$ selon $p(\theta|y)$ pour chaque valeur de m jusqu'à M, pour ensuite utiliser la relation de convergence suivante :

$$\frac{1}{M}\left(\sum_{m=1}^{M} g(\theta^{(m)})\right) \xrightarrow{p.s} \mathbb{E}(g(\theta)|y)$$

Les méthodes de *Monte Carlo par Chaînes de Markov* permettent de générer une chaîne de Markov $\theta^{(m)}$ dont les échantillons sont distribués asymptotiquement selon la loi a posteriori $p(\theta|y)$.

3.5.2 Algorithme Data Augmentation (D.A.)

L'algorithme *Data Augmentation* est une méthode de *Monte Carlo par Chaînes de Markov* particulièrement adapté pour traiter les problèmes de données manquantes. Cette technique proposée par Tanner et Wong [TAN87] est une forme particulière de l'*échantillonneur de Gibbs*, elle fut développée ensuite par Schaffer dans les ouvrages suivants [SCH97] [SCH98].

Le principe de la méthode d'augmentation des données est d'assimiler les observations manquantes comme des paramètres supplémentaires à estimer pour traiter l'échantillon augmenté comme s'il était complet afin d'exploiter l'*échantillonnage de Gibbs*.

Sous hypothèse que les données suivent une distribution normale multivariée, l'augmentation de données peut être appliquée en répétant les deux étapes suivantes :

- Etape "I-step" : Imputation des données manquantes

  $$Y_{mqt}^{(t+1)} \sim P(Y_{mqt}/Y_{obs}, \theta^{(t)})$$

  Partant d'une estimation du vecteur moyenne et de la matrice de covariance, les données manquantes sont simulées pour chaque observation indépendamment. En notant Y_{mqt} les variables ayant des données manquantes et Y_{obs} celles ayant des valeurs observées, les données manquantes sont alors tirées selon la distribution conditionnelle aux données observées.

- Etape "P-step" : Distribution des paramètres θ a posteriori

  $$\theta^{(t+1)} \sim P(\theta/Y_{obs}, Y_{mqt}^{(t+1)})$$

  Partant d'une base complétée, cette étape consiste en la simulation a posteriori du vecteur moyenne et de la matrice de covariance. Cette estimation sera ensuite utilisée dans l'étape "I-step". Le déroulement de cette étape va dépendre de l'information a priori dont on dispose sur les paramètres.

La chaîne de Markov ainsi créée : $(Y_{mqt}^1, \theta^{(1)}), (Y_{mqt}^2, \theta^{(2)}), ..., (Y_{mqt}^t, \theta^{(t)})$ converge en probabilité vers la distribution prédictive a posteriori des données $P(Y_{mqt}, \theta/Y_{obs})$. Par ce biais,

sont simulés alternativement données manquantes et paramètres (les deux étapes sont développées en *Annexe E*). Cette méthode nécessite des valeurs de départ pour les paramètres, une bonne pratique peut consister à les initialiser en appliquant des algorithmes $Esprance - Maximisation(E.M.)$. (cf. *Annexe A*).

3.6 Application aux indicateurs de garantie

3.6.1 Proportion de données manquantes

Afin de nous aider dans le choix du traitement à appliquer aux données manquantes, il nous paraissait important de pouvoir quantifier leur proportion pour chaque acte et dans chacune des tranches.

Toutes les personnes qui adhérent à la même alternative bénéficient des mêmes garanties. Les tranches de dépenses d'une alternative s'appliquent alors à chaque bénéficiaire rattaché à celle-ci. Par conséquent, il semble intéressant de conserver pour chaque alternative la notion du volume des effectifs qu'elle couvre. Pour cette raison, nous avons jugé pertinent d'indiquer également la proportion de ces informations manquantes pour chaque bénéficiaire. Ce qui revient à rattacher les tranches de garanties à l'ensemble des bénéficiaires couvertis par une même alternative. En d'autres termes, il suffit de pondérer les alternatives en fonction du nombre de bénéficiaires.

Le poids de l'alternative *Alt* est évalué comme suit : $P_{Alt} = \frac{Nb\ Beneficiaires\ \in\ Alt}{Nb\ Beneficiaires\ Totaux}$

Les taux de données manquantes sont indiqués selon les deux approches :
- les taux de données manquantes par alternative,
- les taux de données manquantes par bénéficiaire

La seconde version du taux de données manquantes trouve tout son intérêt dans notre approche de modélisation tarifaire qui sera réalisée individu par individu.

Il n'est pas obligatoire qu'une alternative comporte des garanties sur l'ensemble des postes. Nous avons posé comme hypothèse que s'il n'y avait eu aucun règlement d'effectué sur un acte, il n'était pas couvert par l'alternative (sauf exception pour l'optique). Cette supposition n'est pas systématiquement vérifiée car il n'est pas impossible qu'une alternative possède des actes sur lesquels aucune consommation ne soit observée.

Le poste optique présente une spécificité dans la formulation de ses garanties. Une alternative peut proposer des garanties optiques exprimées sous plusieurs formes :

- une garantie unique pour l'ensemble de verres sans distinction
- une garantie différente pour les verres simples et pour les verres complexes
- une grille optique avec des remboursements différents selon la correction des verres
- un forfait global annuel alloué à l'ensemble des dépenses optiques (verres, montures, lentilles,...)

Le paramétrage d'une garantie repose sur une catégorie d'actes plus ou moins fine selon le degré d'information nécessaire au paiement des prestations. Pour des motivations évidentes de productivité, les gestionnaires ne saisissent la codification la plus précise que si la garantie l'exige. Dans le cadre d'un forfait optique, une seule ligne de décompte est renseignée sous une codification unique regroupant l'ensemble des dépenses optiques. Dans ce cas, la répartition des frais entre verres et monture n'est pas disponible. A l'opposé, une grille optique bénéficiera d'une gamme de 8 codes pour chaque correction de verres.

Le versement d'une prestation optique permet de savoir, selon la précision du code acte utilisé, la nature de la garantie employée : forfait, garantie identique sur tous les verres, correction simple/complexe ou grille. Il suffit qu'une seule prestation issue d'une garantie paramétrée sur plusieurs actes soit observée, pour en déduire que la couverture de l'alternative s'étend également à tous les autres actes. Par conséquent, sur une garantie sous forme de grille optique ou simple/complexe, il est possible de savoir s'il y a des codes actes bénéficiant d'une garantie sur lesquels il n'y a eu aucune dépense d'observée. Par exemple, si des remboursements sont effectués sur le code acte "verre simple", une garantie doit être également prévue pour les verres complexes. Si aucun verre complexe n'a été recensé, cette absence d'information constitue une donnée manquante.

En résumé, sur le poste optique une donnée peut être considérée comme manquante sur une alternative :

- si des prestations ont été versées sur un acte mais que toutes les tranches de dépenses n'ont pas pu être observées.

- si sur une codification reposant sur plusieurs actes (correction simple/complexe ou bien grille), certains n'ont pas pu être observés

Dans la plupart des cas, sur un exercice de survenance donné, les différentes formules optiques ne peuvent pas coexister au sein d'une même alternative. La codification permet donc de connaître la nature de la garantie proposée : unique, simple/complexe, grille ou forfait.

Cependant il peut y avoir quelques exceptions :

- Première exception, dans le cas d'une mise en place en cours d'année d'une grille optique en remplacement d'une garantie indépendante de la correction,

- Deuxième exception, dans les contrats disposant d'une grille optique avec la possibilité de passer dans un réseau d'opticiens partenaires. La garantie est la même quelque soit l'accès choisi par le bénéficiaire. Elle sera codifiée en grille en dehors du réseau et en verre toute correction dans le réseau,

- Troisième exception, des lignes dont la saisie a été forcée manuellement.

Dans tous ces cas de figure, plusieurs types de garantie peuvent être présents sur la même année, ce qui remet en cause l'unicité de la garantie rattachée à une alternative. Si différentes expressions de la garantie optique coexistent (par exemple, des forfaits et des verres simples/complexes), l'hypothèse choisie est de retenir la plus représentative.

	TRANCHES											
Acte	N°1	N°2	N°3	N°4	N°5	N°6	N°7	N°8	N°9	N°10	N°11	**Total**
Verre	49%	25%	29%	39%	46%	49%	59%	70%	79%	87%	89%	**56%**
Verre simple	83%	65%	57%	41%	37%	43%	43%	48%	59%	65%	48%	**54%**
Verre progressif	98%	93%	76%	76%	80%	83%	76%	72%	76%	72%	46%	**77%**
Verre simple 0-2 Dioptries	44%	27%	34%	62%	76%	80%	89%	93%	95%	98%	98%	**72%**
Verre simple 2,25-4 Dioptries	72%	45%	51%	69%	84%	89%	95%	98%	98%	100%	100%	**82%**
Verre simple 4,25-6 Dioptries	93%	67%	65%	70%	85%	89%	95%	98%	100%	99%	100%	**87%**
Verre simple >6 Dioptries	97%	90%	78%	77%	84%	86%	93%	95%	98%	99%	100%	**91%**
Verre progressif 0-2 Dioptries	92%	73%	66%	49%	49%	49%	56%	67%	83%	92%	93%	**70%**
Verre progressif 2,25-4 Dioptries	97%	91%	83%	73%	68%	67%	66%	78%	89%	91%	94%	**82%**
Verre progressif 4,25-6 Dioptries	99%	97%	95%	94%	88%	82%	82%	86%	93%	97%	98%	**92%**
Verre progressif >6 Dioptries	100%	98%	96%	96%	92%	87%	90%	91%	94%	94%	95%	**94%**
Forfait optique	53%	48%	39%	42%	45%	52%	61%	67%	73%	82%	84%	**59%**
Monture	68%	39%	20%	27%	40%	56%	76%	83%	93%	95%	95%	**63%**
Prothèses dentaires remboursées	97%	90%	49%	45%	34%	35%	58%	82%	92%	95%	96%	**70%**
Soins conservateurs	2%	49%	81%	84%	90%	91%	94%	95%	95%	96%	80%	**78%**
Consultations généralistes	2%	20%	38%	53%	69%	80%	87%	92%	95%	98%	95%	**66%**
Consultations spécialistes	8%	27%	14%	26%	45%	56%	79%	78%	89%	89%	89%	**55%**
Chambre particulière	87%	60%	32%	54%	76%	91%	94%	90%	97%	95%	92%	**79%**
Honoraires chirurgicaux hospi	25%	34%	71%	83%	88%	93%	94%	97%	97%	97%	94%	**79%**
Cure thermale	66%	89%	89%	93%	70%	73%	84%	89%	90%	94%	65%	**82%**
Total	**44%**	**47%**	**45%**	**53%**	**62%**	**69%**	**80%**	**85%**	**91%**	**94%**	**91%**	**69%**

FIG. 3.2 – Proportion de données manquantes par alternative d'adhésion

	TRANCHES											
Acte	N°1	N°2	N°3	N°4	N°5	N°6	N°7	N°8	N°9	N°10	N°11	**Total**
Verre	8%	2%	3%	6%	10%	12%	20%	28%	39%	51%	57%	**21%**
Verre simple	23%	8%	8%	3%	1%	3%	8%	3%	10%	14%	9%	**8%**
Verre progressif	96%	93%	22%	22%	49%	23%	17%	12%	18%	14%	9%	**34%**
Verre simple 0-2 Dioptries	4%	1%	1%	8%	19%	20%	30%	43%	43%	70%	69%	**28%**
Verre simple 2,25-4 Dioptries	13%	2%	5%	10%	20%	25%	37%	71%	95%	99%	99%	**43%**
Verre simple 4,25-6 Dioptries	35%	10%	9%	11%	23%	28%	61%	46%	79%	71%	79%	**41%**
Verre simple >6 Dioptries	44%	31%	16%	15%	20%	21%	37%	38%	49%	71%	100%	**40%**
Verre progressif 0-2 Dioptries	31%	13%	10%	6%	5%	6%	7%	12%	22%	40%	36%	**17%**
Verre progressif 2,25-4 Dioptries	68%	30%	23%	12%	12%	11%	11%	22%	30%	36%	41%	**27%**
Verre progressif 4,25-6 Dioptries	97%	43%	42%	34%	24%	20%	20%	22%	36%	47%	49%	**40%**
Verre progressif >6 Dioptries	100%	77%	59%	44%	34%	25%	32%	34%	42%	38%	61%	**50%**
Forfait optique	6%	63%	62%	63%	63%	64%	66%	67%	69%	75%	78%	**61%**
Monture	15%	4%	2%	3%	5%	10%	24%	32%	51%	59%	55%	**24%**
Prothèses dentaires remboursées	75%	48%	11%	8%	5%	6%	17%	38%	57%	64%	71%	**36%**
Soins conservateurs	0%	8%	30%	35%	46%	48%	54%	58%	53%	56%	30%	**38%**
Consultations généralistes	1%	1%	4%	9%	17%	29%	41%	53%	59%	71%	65%	**32%**
Consultations spécialistes	0%	1%	0%	2%	7%	12%	32%	28%	44%	46%	45%	**20%**
Chambre particulière	53%	24%	11%	27%	45%	63%	67%	67%	81%	72%	77%	**53%**
Honoraires chirurgicaux hospi	5%	7%	28%	36%	44%	54%	59%	67%	72%	83%	57%	**47%**
Cure thermale	60%	71%	72%	78%	56%	42%	80%	74%	75%	94%	38%	**67%**
Total	**31%**	**22%**	**19%**	**22%**	**27%**	**31%**	**42%**	**48%**	**58%**	**65%**	**59%**	**39%**

FIG. 3.3 – Proportion de données manquantes par bénéficiaire

Globalement, tous postes confondus, 69% des remboursements Gras Savoye moyens par tranches sont manquants par alternative, ce qui représente 39% par bénéficiaire. Mécaniquement, plus la population couverte par une alternative est nombreuse, plus il est probable que beaucoup de tranches de garanties puissent être observées. De même, les taux de données manquantes les plus élevés sont sur les tranches de dépenses les moins fréquentes. Autre élément majeur, 99% des alternatives ont au moins une tranche manquante, soit 81% des bénéficiaires.

Ce volume non négligeable d'informations manquantes ne nous permet pas d'exclure systématiquement les lignes ayant des données non observées.

3.6.2 Organigramme du traitement des données

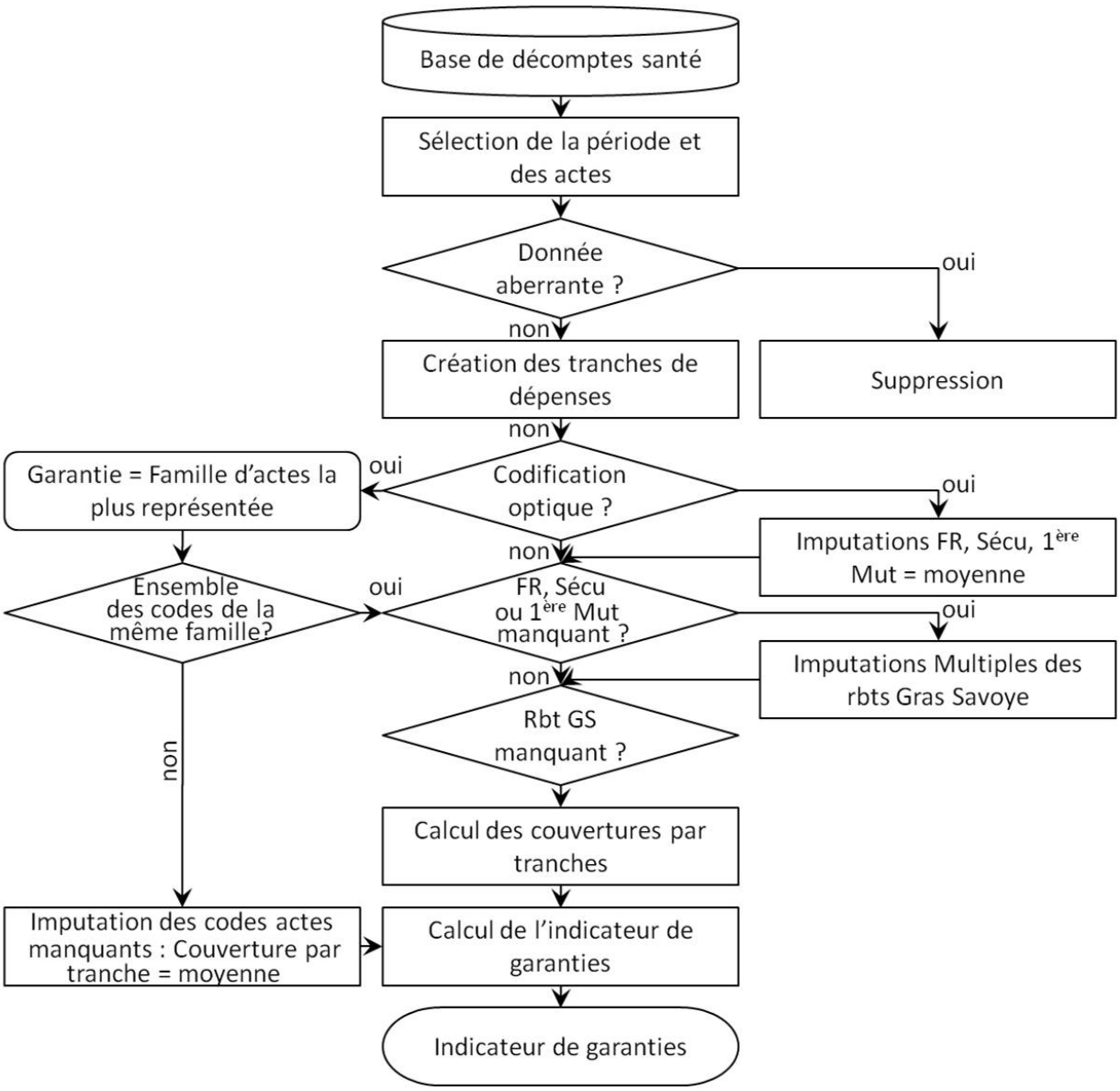


FIG. 3.4 – Traitement des données

3.6.3 Sélection de la Méthode

Comme nous l'avions indiqué, l'ensemble des données manquantes peuvent être globalement considérées comme *MAR*, le postulat que les données sont manquantes selon un mécanisme aléatoire étant validé, les méthodes d'imputation multiples sont donc applicables dans le cas présent.

Nous avons vu précédemment que les valeurs manquantes des remboursements Gras Savoye moyen pour une tranche de dépenses $GS_{i,Alt}$ n'étaient pas monotones, pour cette raison nous avons préféré écarter les méthodes de régression ou du score de propension qui nécessitent de remplir cette condition.

Nous nous sommes donc orientés vers une méthode d'imputation multiple par le biais des techniques d'augmentation de données qui requiert que les variables soient distribuées selon une loi normale multivariée. Cette condition à valider sera étudiée dans la section suivante.

3.6.4 Postulat de multinormalité requis pour la Data Augmentation

Validité de l'hypothèse de normalité

L'application de la méthode de *Data Augmentation* suppose que l'hypothèse de normalité soit vérifiée. Nous avons dû étudier dans quelles mesures les remboursements Gras Savoye sur les différentes tranches de dépenses notés $GS_{i,Alt}$ se rapprochaient d'une distribution normale multivariée (ou multinormale).

Bien que la normalité de toutes les distributions marginales soit une condition nécessaire, elle n'est pas suffisante. Des distributions multidimensionnelles peuvent très bien vérifier la normalité dans chaque loi marginale sans pour autant suivre une loi multinormale.

Définition 3.6.1 (Distribution normale multivariée)**.** *Un vecteur aléatoire $\overrightarrow{X}$ à n composantes $(X_1, ..., X_n)$ est gaussien, c'est à dire qu'il suit une loi normale de dimension n (appelée également loi multinormale ou encore loi normale multivariée) sur $\mathbb{R}^n$ si et seulement si toutes les combinaisons linéaires des composantes $(X_1, ..., X_n)$ de $\overrightarrow{X}$ sont des variables normales sur $\mathbb{R}$ (univariées).*

La définition ci-dessus, (cf. le livre de J.-L. Ouvrard [OUVOO]) peut laisser penser qu'un test de multinormalité peut se ramener à l'étude de plusieurs cas univariés. En pratique, ceci est bien souvent inutilisable pour deux raisons : quand le nombre de dimensions augmente, la quantité de combinaisons linéaires à tester croît rapidement et le risque de première espèce globale converge assez vite vers 100% et ce quelque soit le risque de première espèce individuel : $\alpha_{global} = lim_{n\to\infty}(1 - (1 - \alpha_{ind})^n)$. Par exemple, avec $\alpha_{ind} = 5\%$ le risque de première espèce globale vaut $\alpha_{global} = 40\%$ pour 11 dimensions.

Les remboursements Gras Savoye ne peuvent pas être négatifs et sont bornés par définition de la tranche. L'hypothèse qu'ils suivent strictement une loi normale est donc infirmée puisque elle possède une queue de distribution dans les valeurs négatives. De même, les montants arrondis dans le libellé des limites de garantie excluent également cette hypothèse (intuitivement, il est plus fréquent d'avoir une garantie à hauteur de 400% de la BR qu'à 402,34%).

« Tout ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable. »
([VAL42])

Dans la plupart des cas, les données économiques réelles ne suivent pas exactement des lois théoriques. Les modélisations sont des visions simplifiées de la réalité ne prenant pas l'intégralité des paramètres. Il n'est alors pas surprenant qu'à partir d'un volume de données suffisant, le moindre écart à la loi théorique devienne significatif. Quel que soit le test statistique utilisé, il est toujours possible de trouver un échantillon suffisamment grand à partir duquel l'hypothèse selon laquelle les données suivent une lois théorique soit invalidée. [SAP06].

Nous allons donc effectuer une série de tests afin de déterminer si la loi de distribution, bien que ne suivant pas une loi normale, ne s'en écarte pas trop. Les principes des différents

tests sont décrits dans le cours de Master 2ème année intitulé "Mesures d'association" de Frédéric et Myriam Bertrand [BERO6]. Nos différents indicateurs de garantie portent sur 20 actes comportant chacun 11 tranches de dépenses, soit un total de 220 tranches. Nous n'étudierons donc pas dans le détail l'hypothèse de normalité de chacun de ces actes. Nous avons essentiellement procédé par sondage, en effectuant des tests classiques de normalité univariés : *Shapiro-Wilk*, *Kolmogorov-Smirnov*, *Cramer-Von Mises* et *Anderson-Darling*, tout en privilégiant l'apparence visuelle des distributions. Dans un second temps, nous avons utilisé des tests de normalité multivariés : une extension du test de Shapiro-Wilk à la multinormalité proposée par Royston et les tests de *Mardia-Kurtosis* et *Mardia-Skewness*.

Le test de Shapiro and Wilk (1965) est reconnu comme étant l'un des tests omnibus de normalité univariée les plus puissants. Le terme "omnibus" caractérise les tests de significativité globale où l'apport des facteurs explicatifs est testé globalement par rapport à un modèle de référence naïf. Royston (1983) a étendu son application pour tester la multinormalité, mais la procédure implique cependant des approximations asymptotiques devant être justifiées pour éviter d'avoir de mauvais comportements dans des échantillons finis. [ROY82] [ROY95].

Les tests de multinormalité de *Mardia* reposent sur le calcul de deux statistiques : le degré d'asymétrie (*MKU : Mardia KUrtosis*) d'une part et le degré d'aplatissement (*MSK : Mardia SKewness*) d'autre part [MAR70] [MAR80].

Exemples de tests de normalité univariés des remboursements Gras Savoye sur les consultations de spécialistes

Voici des exemples de tests univariés appliqués sur deux tranches de remboursements Gras Savoye sur le poste consultations de spécialistes : la tranche de 100% à 150% de la BR et de 250% à 300% de la BR.

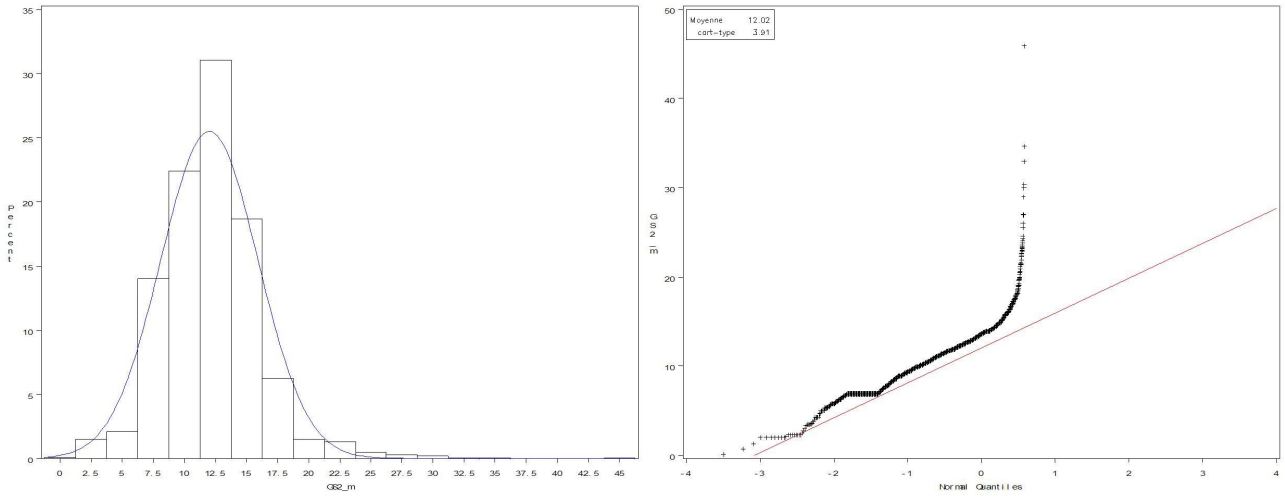


FIG. 3.5 – Répartition - Droite de Henry (Tranche de 100% à 150% de la BR)

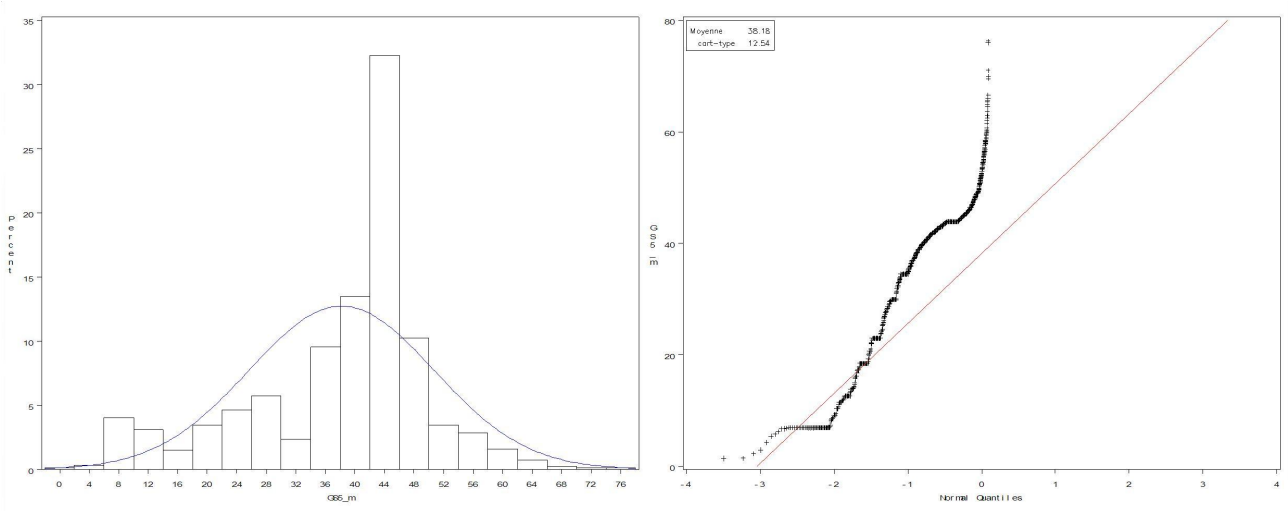


FIG. 3.6 – Répartition - Droite de Henry (Tranche de 250% à 300% de la BR)

Nombres d'occurrences

Valeur	Comptage	Pourcentages Cellule	Pourcentages Cum.	Valeur	Comptage	Pourcentages Cellule	Pourcentages Cum.	Valeur	Comptage	Pourcentages Cellule	Pourcentages Cum.
0.10000000	1	0.1	0.1	6.57500000	1	0.1	4.3	7.76444444	1	0.1	13.8
0.70000000	1	0.1	0.1	6.68000000	3	0.2	4.5	7.77500000	1	0.1	13.9
1.27000000	1	0.1	0.2	6.69625000	1	0.1	4.5	7.77545455	1	0.1	13.9
2.00000000	7	0.4	0.5	6.71125000	1	0.1	4.6	7.78137931	1	0.1	14.0
2.08333333	1	0.1	0.6	6.73571429	1	0.1	4.6	7.79421053	1	0.1	14.0
2.30000000	8	0.4	1.0	6.77285714	1	0.1	4.7	7.82000000	1	0.1	14.1
2.50000000	1	0.1	1.1	6.85000000	1	0.1	4.7	7.84866667	1	0.1	14.1
2.75000000	1	0.1	1.1	6.88333333	1	0.1	4.8	7.86666667	1	0.1	14.2
3.00000000	1	0.1	1.2	6.90000000	1	0.1	4.8	7.90000000	8	0.4	14.6
3.31578947	1	0.1	1.2	6.90000000	5	0.3	5.1	7.92000000	1	0.1	14.7
3.43000000	2	0.1	1.3	6.90000000	97	5.2	10.3	7.93333333	1	0.1	14.7
3.44000000	1	0.1	1.4	6.90000000	9	0.5	10.8	7.94560606	1	0.1	14.8
3.50000000	1	0.1	1.5	6.90000000	1	0.1	10.9	7.96666667	1	0.1	14.8
3.54000000	1	0.1	1.5	6.90000000	1	0.1	10.9	8.03000000	1	0.1	14.9
3.63333333	1	0.1	1.6	6.90000000	1	0.1	11.0	8.04400000	1	0.1	14.9
3.73333333	1	0.1	1.6	6.94900000	1	0.1	11.0	8.06000000	1	0.1	15.0
4.00000000	1	0.1	1.7	6.95000000	2	0.1	11.1	8.08333333	1	0.1	15.1
4.22500000	1	0.1	1.7	6.95769231	1	0.1	11.2	8.09090909	1	0.1	15.1
4.23666667	1	0.1	1.8	6.96600000	1	0.1	11.2	8.10200000	1	0.1	15.2
4.25000000	1	0.1	1.8	6.96666667	1	0.1	11.3	8.15000000	1	0.1	15.2
4.27000000	1	0.1	1.9	7.00000000	7	0.4	11.7	8.15714286	1	0.1	15.3
4.30000000	1	0.1	1.9	7.06468750	1	0.1	11.7	8.20000000	1	0.1	15.3
4.65000000	1	0.1	2.0	7.07170732	1	0.1	11.8	8.22285714	1	0.1	15.4
4.69090909	1	0.1	2.0	7.08533333	1	0.1	11.8	8.24875000	1	0.1	15.4
5.00000000	3	0.2	2.2	7.09450000	1	0.1	11.9	8.24909091	1	0.1	15.5

FIG. 3.7 – Nombre d'occurrences (Tranche de 100% à 150% de la BR)

Les résultats des tests sont les suivants :

Tranche de 100% à 150% de la BR

Test	Statistique		Seuil de significativité	
Shapiro-Wilk	W	0.946813	Pr<W	<0.0001
Kolmogorov-Smirnov	D	0.71319	Pr>D	<0.0100
Cramer-Von Mises	W-Sq	1.984601	Pr>W-Sq	<0.0050
Anderson-Darling	A-Sq	13.38224	Pr>A-Sq	<0.0050

Tranche de 250% à 300% de la BR

Test	Statistique		Seuil de significativité	
Shapiro-Wilk	W	0.91194	Pr<W	<0.0001
Kolmogorov-Smirnov	D	0.163365	Pr>D	<0.0100
Cramer-Von Mises	W-Sq	10.49012	Pr>W-Sq	<0.0050
Anderson-Darling	A-Sq	54.31624	Pr>A-Sq	<0.0050

Prenons par exemple la tranche de 100% à 150% de la BR. Avec un seuil à 5%, l'hypothèse de normalité est rejetée dans les quatre tests. Le test de Kolmogorov-Smirnov est celui qui réfute le moins cette hypothèse avec une P-value à 1%. Ce résultat isolé est assez représentatif de l'ensemble des tranches que nous avons testées.

Tout d'abord, nous devons tenir compte du nombre élevé d'observations sur lesquelles portent nos tests : 1 860 valeurs renseignées sur cette tranche, ce qui explique en partie le rejet de

l'hypothèse de normalité car les différences avec la loi théorique sont rapidement mises en évidence. Toujours sur cette même tranche, la valeur 6,90 Euros apparaît davantage que celles avoisinantes car elle correspond à des consultations dont le tarif pratiqué est exactement celui de la base de remboursement conventionnée, soit 23 Euros (6,90 Euros = 23 Euros - 70% x 23 Euros).

La représentation des fréquences réelles cumulées n'épouse pas la trajectoire de la droite de Henry. La valeur de 6,90 Euros indiquée précédemment génère un "aplatissement" sur la représentation des fréquences réelles cumulées et les queues de distributions n'ont pas les mêmes formes que celles de la loi normale. En conclusion, la forme de distribution de cette variable ne décrit pas exactement une courbe Gaussienne, cependant malgré ces différences, l'allure générale de sa distribution est en forme de "cloche".

Exemples de tests de normalité multivariés des remboursements Gras Savoye sur les consultations de spécialistes

Bien que la normalité univariée n'ait pas pu être vérifiée, des tests de multinormalité ont été réalisés sur l'ensemble des tranches de remboursements Gras Savoye de l'exemple précédent, à savoir le poste consultations de spécialistes. La plupart des tests proposés par les différents logiciels de statistiques ne prennent pas en compte les lignes comportant des données manquantes. Les informations sur chaque tranche de dépenses pour une alternative et un acte donnés ne sont que très rarement complètes, ce qui réduit considérablement le volume de lignes pouvant être testées.

Les résultats des tests sont les suivants :

Ensemble des tranches de consultations de spécialistes

Variable	Test	Skewness / Kurtosis	Statistique	Seuil de significativité
Tranche 1	Shapiro-Wilk	-	0.90	<0.0001
Tranche 2	Shapiro-Wilk	-	0.83	<0.0001
Tranche 3	Shapiro-Wilk	-	0.62	<0.0001
Tranche 4	Shapiro-Wilk	-	0.76	<0.0001
Tranche 5	Shapiro-Wilk	-	0.92	<0.0001
Tranche 6	Shapiro-Wilk	-	0.91	<0.0001
Tranche 7	Shapiro-Wilk	-	0.92	<0.0001
Tranche 8	Shapiro-Wilk	-	0.90	<0.0001
Tranche 9	Shapiro-Wilk	-	0.88	<0.0001
Tranche 10	Shapiro-Wilk	-	0.93	<0.0001
Tranche 11	Shapiro-Wilk	-	0.95	0.0023
Toutes tranches	Shapiro-Wilk multivarié	-	0.72	<0.0001
Toutes tranches	Mardia Skewness	65.081	1057.71	<0.0001
Toutes tranches	Mardia Kurtosis	210.045	19.22	<0.0001

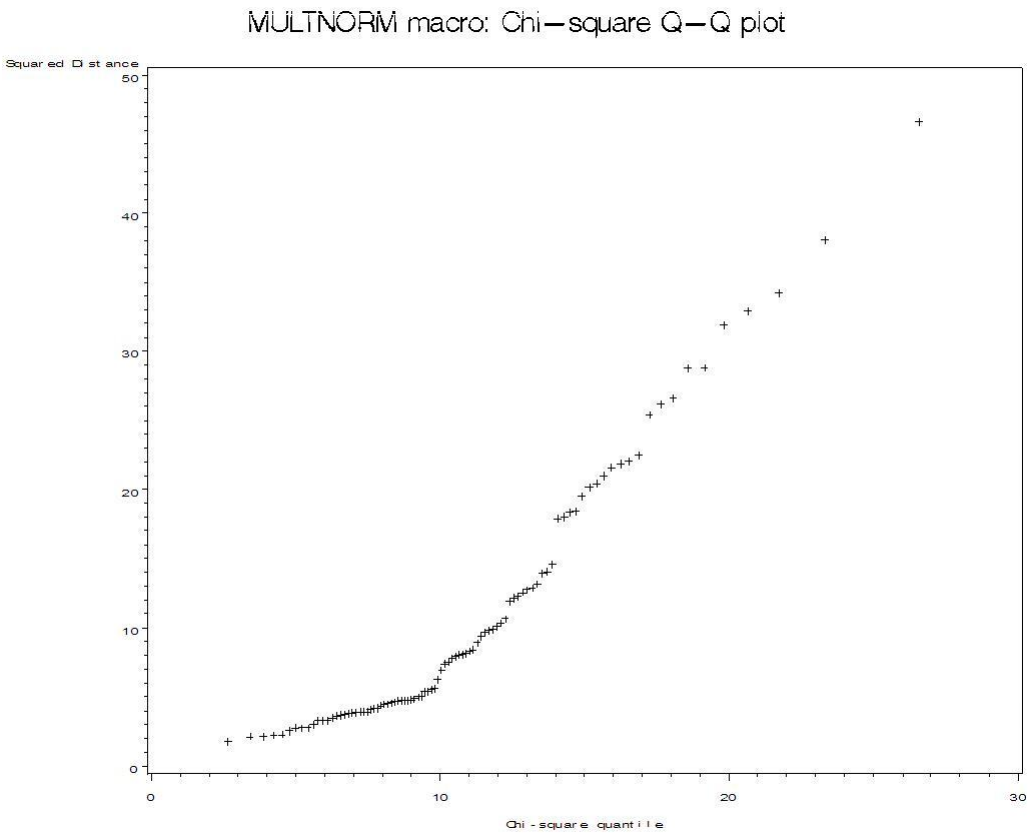


FIG. 3.8 – Droite de Henry multivariée (Consultations de spécialistes)

L'hypothèse de normalité n'est pas validée par cette série de tests. La représentation des fréquences réelles cumulées multivariées ne suit pas exactement la tendance de la droite de Henry, mais ne s'en éloigne pas non plus de manière importante.

Nous avions également envisagé d'utiliser un changement de variable du type : $Y = ln(X)$, mais les résultats obtenus n'ont pas été plus probants. En procédant à ce changement de variable, le test de Shapiro-Wilk multivarié indiquait une p-value de $8,025 \times 10^{-12}$ contre $6,672 \times 10^{-12}$ sans celui-ci.

Conclusions

Les tests statistiques n'ont pas pu valider que les remboursements Gras Savoye suivaient approximativement une loi multinormale. Pour autant, la forme générale des tranches suit grossièrement une distribution en forme de "cloche" malgré quelques irrégularités.

Schafer et Graham [SCH02] précisent que la méthode reste efficace même dans les cas où les données ne se conforment pas complètement à ce postulat de normalité et qu'elle fournit d'excellents résultats dans un bon nombre de situations.

La robustesse de la méthode à cette hypothèse nous permet donc de ne pas l'exclure directement sachant que nos données ne suivent pas strictement une loi normale. Nous choisissons donc d'employer cette méthode.

3.6.5 Imputation multiple des remboursements Gras Savoye

Dans cette partie, nous exposerons les hypothèses que nous avons prises lors de la mise en application de la méthode d'Imputation Multiples aux montants de remboursements Gras Savoye moyens par tranche. Les résultats seront illustrés toujours sur le même exemple que précédemment, à savoir les consultations de spécialistes.

Phase Imputation

Nous avons opté pour simuler 5 jeux de données en suivant les préconisations de Rubin, avec imputation par la méthode *MCMC* de *Data Augmentation*.

Nous avons choisi d'appliquer la méthode *MCMC* pour imputer l'intégralité des données. Cette méthode aurait pu être utilisée partiellement pour chercher à rendre monotone les données afin d'utiliser ensuite d'autres modélisations. Arbitrairemant, cette alternative n'a pas été retenue.

Les valeurs de départ des paramètres sont initialisées par un algorithme *EM* (*Espérance-Maximisation*).

Nous avons pondéré chacune des observations par le montant des frais réels correspondant pour que la prise en compte des tranches soit fonction de leur volume de dépenses.

Comme chaque état d'une chaîne de Markov a une influence sur le suivant, nous avons laissé 200 itérations avant la première imputation et 100 entre chaque afin d'éliminer les séries de dépendance sur les valeurs de départ pour tendre vers une distribution stationnaire.

Nous avons supposé que nous ne disposions pas d'information a priori sur l'estimation des moyennes et des covariances en utilisant la distribution de Jeffreys ([SCH97], page 154)

Les 10 premiers motifs de donnés manquantes sur les consultations de spécialistes sont les suivants (les "X" symbolisent les valeurs connues et les "." celles manquantes) :

Groupe	Tr.1	Tr.2	Tr.3	Tr.4	Tr.5	Tr.6	Tr.7	Tr.8	Tr.9	Tr.10	Tr.11	Pourcentage
1	X	X	X	X	X	X	X	X	X	X	X	53.08%
2	X	X	X	X	X	X	.	.	.	.	.	6.25%
3	X	X	X	X	X	X	.	X	.	.	.	3.62%
4	X	X	X	X	X	X	X	X	.	.	.	3.53%
5	X	X	X	X	X	X	X	X	.	X	X	3.49%
6	X	X	X	X	.	.	.	.	.	.	.	3.31%
7	X	X	X	X	X	X	X	X	X	.	X	3.15%
8	X	X	X	X	X	X	X	X	X	X	.	2.79%
9	X	X	X	X	X	X	X	X	X	.	.	2.36%
10	X	X	X	X	X	X	X	.	.	.	.	2.34%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

53.08% des motifs sont entièrement connus (groupe 1), les autres 46.92% des motifs ont au moins une tranche manquante.

Globalement, les données manquantes sont essentiellement situées sur les tranches de dépenses élevées.

Les moyennes par groupes correspondantes sont les suivantes :

Groupe	Tr.1	Tr.2	Tr.3	Tr.4	Tr.5	Tr.6	Tr.7	Tr.8	Tr.9	Tr.10	Tr.11
1	7.46	13.04	24.35	34.90	45.91	53.54	60.98	64.59	67.76	70.35	73.89
2	7.35	12.37	21.29	27.56	31.95	36.96	.	.	.	.	.
3	7.29	12.28	23.03	29.86	37.49	41.62	.	53.11	.	.	.
4	7.43	12.15	22.01	30.20	37.76	44.82	50.85	51.47	.	.	.
5	7.48	12.75	24.91	35.07	44.23	51.19	60.50	63.77	.	69.13	78.57
6	7.01	12.06	21.01	27.37	.	.	.	.	.	.	.
7	7.29	12.33	24.07	31.63	40.17	44.97	54.50	55.20	55.78	.	63.00
8	7.47	12.85	24.39	33.10	40.75	44.28	48.73	52.65	55.77	59.46	.
9	7.31	12.14	24.50	34.84	44.99	50.94	57.66	66.24	65.52	.	.
10	7.49	12.77	22.77	30.36	36.68	42.55	48.68	.	.	.	.
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Les estimations produites par l'*algorithme EM* sont les suivantes :

Moyennes :

Variable	Tr.1	Tr.2	Tr.3	Tr.4	Tr.5	Tr.6	Tr.7	Tr.8	Tr.9	Tr.10	Tr.11
Moyenne	7.40	12.76	23.59	32.78	41.90	48.39	55.02	57.95	60.41	62.78	66.89

Covariances :

Variable	Tr.1	Tr.2	Tr.3	Tr.4	Tr.5	Tr.6	Tr.7	Tr.8	Tr.9	Tr.10	Tr.11
Tr.1	0.31	0.27	0.32	0.51	0.56	0.61	0.37	0.65	1.12	0.78	-0.68
Tr.2	0.27	4.33	3.53	4.78	5.6	5.29	6.75	6.28	8.35	7.17	4.18
Tr.3	0.32	3.53	12.93	18.9	27.11	30.9	38.03	41.87	44.11	48.68	48.36
Tr.4	0.51	4.78	18.9	39.77	56.19	67.94	84.77	91.91	98.7	107.01	106.3
Tr.5	0.56	5.6	27.11	56.19	105.14	121.95	160.53	175.03	190.44	204.93	210.84
Tr.6	0.61	5.29	30.9	67.94	121.95	180.45	231.1	266.97	296.66	320.23	336.69
Tr.7	0.37	6.75	38.03	84.77	160.53	231.1	362.05	403.84	459.74	492.95	536.7
Tr.8	0.65	6.28	41.87	91.91	175.03	266.97	403.84	507.66	584.88	631.89	706.47
Tr.9	1.12	8.35	44.11	98.7	190.44	296.66	459.74	584.88	721.06	762.58	873.41
Tr.10	0.78	7.17	48.68	107.01	204.93	320.23	492.95	631.89	762.58	859.9	979.85
Tr.11	-0.68	4.18	48.36	106.3	210.84	336.69	536.7	706.47	873.41	979.85	1238.94

Nous avons ensuite appliqué la méthode *MCMC* de *Data Augmentation* sur nos cinq jeux de données afin qu'ils deviennent entièrement complétés.

La connaissance de toutes les dépenses moyennes sur chaque jeux pour chacune des alternatives va permettre de calculer l'indicateur de garantie. Préalablement, une phase d'analyse doit être effectuée sur les cinq jeux pour vérifier la cohérence des données imputées et éventuellement agir sur celles qui ne sont pas conformes à une certaine réalité.

Phase Analyse

La deuxième partie analyse statistique de la méthodologie générale d'imputation multiple, telle qu'elle est décrite par Rubin, n'a pas réellement lieu d'être dans le cas présent. Cette partie a pour objectif d'appliquer des modèles de régression standard sur chaque jeux de données (modèles linéaires généralisés, logistiques,...), pour en estimer ensuite ses paramètres tout en testant l'impact de la non observation de données.

L'objectif de cette étude est de générer artificiellement un indicateur de niveau de garantie. Sa création ne nécessite pas de modélisation particulière car nous connaissons déjà de façon exacte son lien avec les valeurs imputées.

Au lieu de chercher à modéliser des paramètres dans nos jeux de données nous allons tester la cohérence des valeurs imputées, pour les modifier si besoin.

Les critères que nous avons testés sont les suivants :

- Les remboursements Gras Savoye moyens doivent être positifs
- Les remboursements Gras Savoye moyens additionnés aux autres dépenses ne peuvent pas dépasser le montant des frais réels
- Les remboursements Gras Savoye moyens imputés doivent être supérieurs à ceux de la tranche précédente

Ce test nous a permis de mettre en évidence que 3% des tranches imputées avaient des remboursements Gras Savoye hors normes. Nous avons ensuite corrigé ces valeurs de telle sorte à ce qu'elles respectent la logique liée aux dépenses et aux tranches.

Une fois les corrections apportées sur nos cinq jeux de données imputées, nous avons pu bâtir notre indicateur de niveau de garantie comme nous l'avions décrit précédemment :

$$Gar_{Alt,Specialistes} = \left[\begin{array}{cccc} r_{1,Ref.} & r_{2,Ref.} & \dots & r_{11,Ref.} \end{array} \right] \times \left[\begin{array}{c} CV_{1,Alt} \\ CV_{2,Alt} \\ \vdots \\ CV_{11,Alt} \end{array} \right]$$

avec les $r_{i,Alt}$ représentant la répartition moyenne dans chacune des tranches observée sur l'ensemble de notre portefeuille.

A ce stade, nous disposons de cinq jeux de données complètes et cohérentes bénéficiant pour chaque alternative d'un indicateur de niveau de garantie synthétique. Il ne nous reste plus qu'à combiner ces jeux en intégrant la variance liée à la non observation de données pour en obtenir un seul.

Exemple de calcul de l'indicateur de niveau de garantie sur une alternative *Alt* :

Tranche i	Répartition Référentiel $r_{i,Ref.}$	Taux de couverture $CV_{i,Alt}$	Produit $r_{i,Ref.} \times CV_{i,Alt}$
Tr.1	28.9%	100.0%	28.9%
Tr.2	7.1%	95.0%	6.8%
Tr.3	26.5%	90.0%	23.8%
Tr.4	18.1%	85.0%	15.4%
Tr.5	8.9%	80.0%	7.1%
Tr.6	6.1%	75.0%	4.6%
Tr.7	1.5%	70.0%	1.0%
Tr.8	1.5%	65.0%	1.0%
Tr.9	0.4%	60.0%	0.2%
Tr.10	0.5%	55.0%	0.3%
Tr.11	0.5%	50.0%	0.3%
Indicateur de garantie			89.3%

Phase Combinaison des résultats

La phase combinaison des résultats décrite par Rubin, a été conçue pour estimer des paramètres de modélisation. Ici, l'objectif est de créer un indicateur de garantie résultant de plusieurs variables imputées (les 11 tranches de coûts). Il est calculé pour chacune des alternatives d'adhésion et non pas en moyenne sur l'ensemble du portefeuille. L'approche n'est donc plus globale mais individuelle.

En adaptant la méthodologie proposée par Rubin, pour obtenir l'estimation de l'indicateur de niveau de garantie intégrant la volatilité liée à la non observation, il suffit de prendre la moyenne observée sur les cinq jeux d'imputation pour chaque alternative :

$$\overline{Gar_{Alt}} = \frac{1}{5} \sum_{Imp=1}^{5} \hat{Gar}_{Imp.,Alt.}$$

Dans un second temps, nous avons cherché à quantifier l'impact que pouvait avoir la non observation de données sur notre indicateur.

En s'inspirant de la démarche proposée par Rubin, les variances intra-imputation (U) et inter-imputation (B) ont été estimées comme suit :

La variance intra-imputation (U) a été obtenue en réalisant la moyenne des variances observées sur les cinq jeux :

$$U = \hat{Var}_{Intra} = \frac{1}{5} \sum_{Imp.=1}^{5} \hat{Var}_{Imp.}$$

avec

$$\hat{Var}_{Imp.} = \frac{1}{Nb_{Alt} - 1} \sum_{Alt.=1}^{Nb_{Alt}} \left(\hat{Gar}_{Imp.,Alt.} - \overline{Gar_{Imp.}} \right)^2$$

La variance inter-imputation (B) a été approchée par la moyenne des variances obtenues sur chaque alternative :

$$B = \hat{Var}_{Inter} \simeq \frac{1}{Nb_{Alt}} \sum_{Alt=1}^{Nb_{Alt}} \hat{Var}_{Alt.}$$

avec

$$\hat{Var}_{Alt.} = \frac{1}{4} \sum_{Imp.=1}^{5} \left(\hat{Gar}_{Imp.,Alt.} - \overline{Gar_{Alt.}} \right)^2$$

N°Alternative	GAR Imp.1	GAR Imp.2	GAR Imp.3	GAR Imp.4	GAR Imp.5	Moyenne	Var_{Alt} (Inter)
1	91.11%	90.91%	90.95%	90.69%	91.18%	90.97%	3.64E-06
2	93.20%	93.20%	93.20%	93.20%	93.20%	93.20%	6.53E-12
3	90.93%	90.89%	90.91%	90.88%	90.93%	90.91%	5.91E-08
4	95.06%	95.05%	95.06%	95.09%	95.12%	95.07%	8.30E-08
5	92.80%	92.80%	92.80%	92.80%	92.80%	92.80%	0
6	93.55%	93.55%	93.55%	93.55%	93.55%	93.55%	0
7	98.40%	98.40%	98.40%	98.40%	98.40%	98.40%	4.02E-11
8	98.59%	98.58%	98.59%	98.67%	98.59%	98.60%	1.25E-07
9	93.11%	93.06%	93.37%	93.06%	93.32%	93.19%	2.17E-06
10	98.66%	98.66%	98.66%	98.66%	98.66%	98.66%	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$Var_{Imp.}$ (Intra)	0,0078	0,0077	0,0077	0,0078	0,0078	94.82%	

La valeur moyenne de l'indicateur de niveau de garantie sur l'ensemble des alternative est de 94.82%, la variance intra-imputation vaut $0,0078$ (soit un écart-type de près de 9%) et celle inter-imputation seulement $0,00001$ (soit un écart-type de 0,3%).

La faible valeur de la variance inter-imputation par rapport à celle intra-imputation tend à prouver que l'inférence liée à la non observation des données manquantes préserve l'information contenue dans le calcul de notre indicateur. L'efficience relative, telle que la définit Rubin, vaut quasiment de 1 ($RE = 0,9999998$), ce qui signifit que le rajout d'imputations supplémentaires n'aurait pas grand intérêt.

Ces résultats peuvent s'expliquer pour deux raisons :

- premièrement, la part des données manquantes reste raisonnable,
- deuxièmement, les tranches les plus rarement observées sont aussi celles qui jouent le moins dans le calcul de l'indicateur.

Conclusions

Nous avons pu construire notre indicateur de niveau de garantie sur l'ensemble des alternatives. Grâce notamment aux méthodes *M.C.M.C.*, ce calcul a été rendu possible même sur celles dont la totalité des tranches de dépenses n'avaient pu être observée. Ces indices vont pouvoir nous permettre de quantifier les niveaux de couverture de toutes les garanties et ce, quelque soit leur mode d'expression. Les alternatives pourront ensuite être comparées entre elles selon un critère objectif et nous permettre ainsi de poursuivre notre étude.

Chapitre 4

Analyse de la consommation médicale

4.1 Périmètre de l'étude

Notre étude porte sur l'ensemble des prestations versées au titre de l'exercice de survenance 2007 (date de soin) et réglées jusqu'au 31/12/08. Notons qu'à cette date d'observation la quasi-totalité des sinistres de l'exercice est connue (99,6% soit un taux de P.S.A.P. (Provisions pour Sinistres A Payer) inférieur à 0,4%).

Nous avons restreint volontairement le périmètre de notre étude aux personnels actifs métropolitains pour rester sur une base de comparaison homogène.

4.2 Composition de notre portefeuille

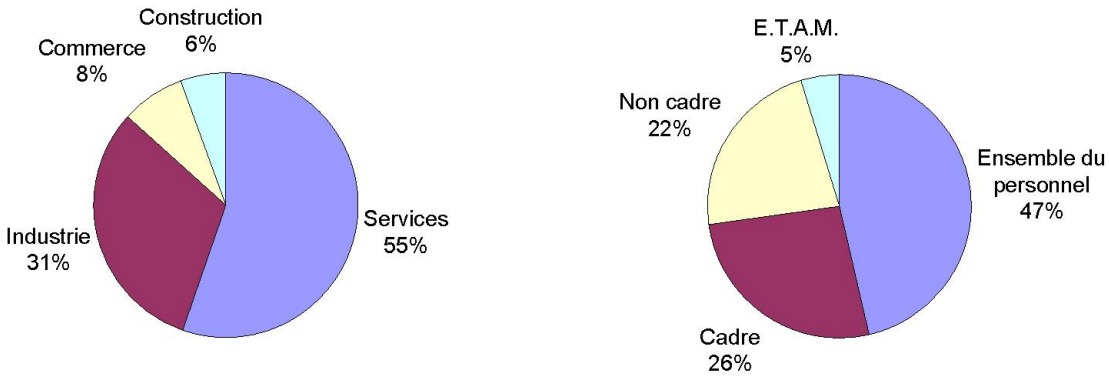


FIG. 4.1 – Répartition des bénéficiaires par secteurs d'activité et type de contrats

La majeure partie des entreprises référencées dans notre portefeuille sont issues du secteur d'activité des Services.

Un peu plus de la moitié de nos contrats sont paramétrés en fonction de la catégorie socio-professionnelle des adhérents. Dans ces contrats plus de la moitié des effectifs sont des cadres. D'expérience, les contrats ensemble du personnel respectent aussi globalement ces proportions.

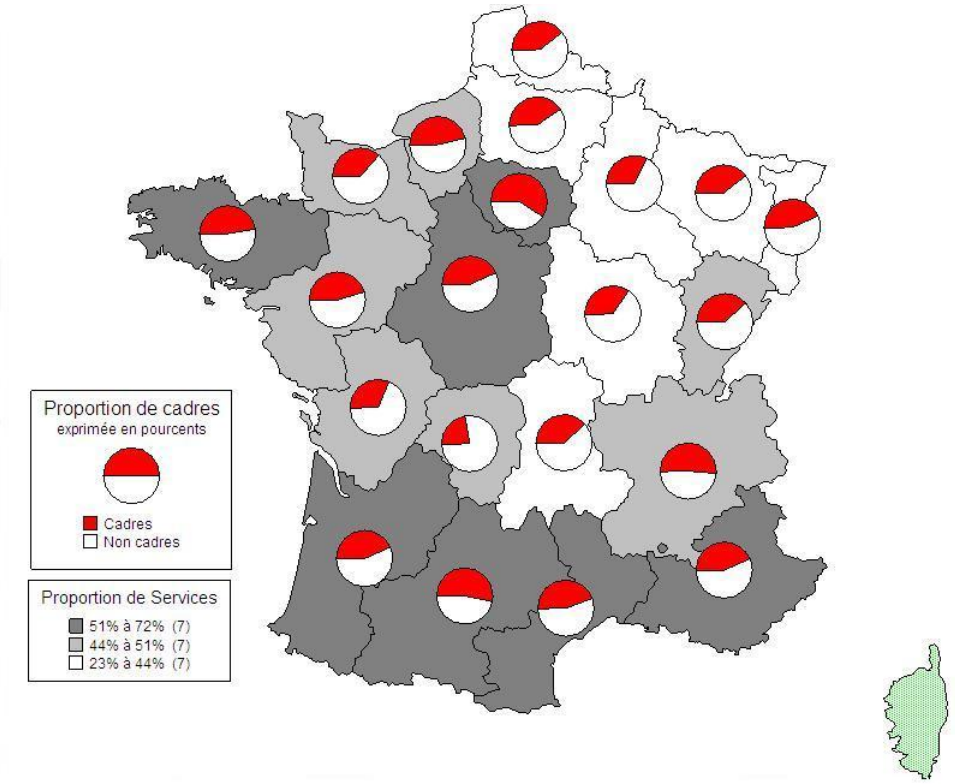


FIG. 4.2 – Répartition des bénéficiaires par secteurs d'activité et type de contrats

Catégorie INSEE	Classe INSEE	Répartition des effectifs
Services	Services aux entreprises	40,6%
	Transports	6%
	Services aux particuliers	3,9%
	Activités financières	2,5%
	Activités immobilières	2,3%
Industrie	Industries des biens d'équipement	8,3%
	Industries des biens intermédiaires	6,9%
	Energie	6,5%
	Industries agricoles et alimentaires	5,1%
	Industrie des biens de consommation	3,9%
	Industrie automobile	0,8%
Commerce	Commerce de gros	4,8%
	Commerce et réparation automobile	2,2%
	Commerce de détail divers	0,4%
	Commerce de détail alimentaire	0,1%
	Commerce de détail équipt personne	0,1%
	VPC	<0,1%
	Commerce de détail de santé beauté	<0,1%
Construction	Construction	5,6%

La forte proportion de cadres, caractéristique de notre portefeuille, est étroitement liée aux principaux secteurs d'activités tels que les services.

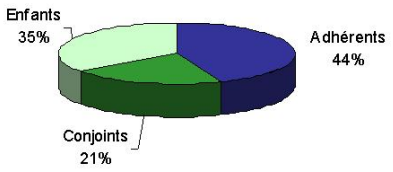
Schématiquement, les entreprises de service sont essentiellement localisées en Ile-de-France et dans la moitié sud. La situation géographique des cadres est concentrée dans les mêmes zones que celles des entreprises de service. En Ile-de-France, les entreprises de services représentent 67% et 59% des populations assurées sont cadres.

Taille de l'entreprise (en nombre d'adhérents)	Nombre d'entreprises
plus de 20 000	2
de 10 000 à 20 000	1
de 5 000 à 10 000	5
de 4 000 à 5 000	1
de 3 000 à 4 000	6
de 2 000 à 3 000	8
de 1 000 à 2 000	24
de 500 à 1 000	48
moins de 500	570
Total	655

En terme de taille, 47 entreprises ont plus de 1 000 adhérents, 3 d'entre elles sont supérieures à 10 000. L'entreprise moyenne couverte est de 500 têtes.

4.3 Etude Démographique

Eléments Démographiques	sur 2006		sur 2007		Evolution
	Effectifs	Répartition	Effectifs	Répartition	
Adhérents	228 347	44%	268 221	44%	17%
Conjoints	112 511	22%	130 087	21%	16%
Enfants	181 986	35%	214 066	35%	18%
Cumul des bénéficiaires	522 844	100%	612 375	100%	17%
Poids des ayants droit	56%		56%		



Familles au 31/12/2007	Type	Effectifs	Répartition
Adhérent sans enfant	C0	116 224	43%
Adhérent avec enfants	C1+	24 258	9%
Couple sans enfant	M0	36 200	13%
Couple avec enfants	M1+	91 539	34%
	Total	268 221	100%

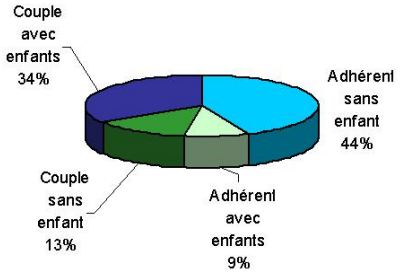
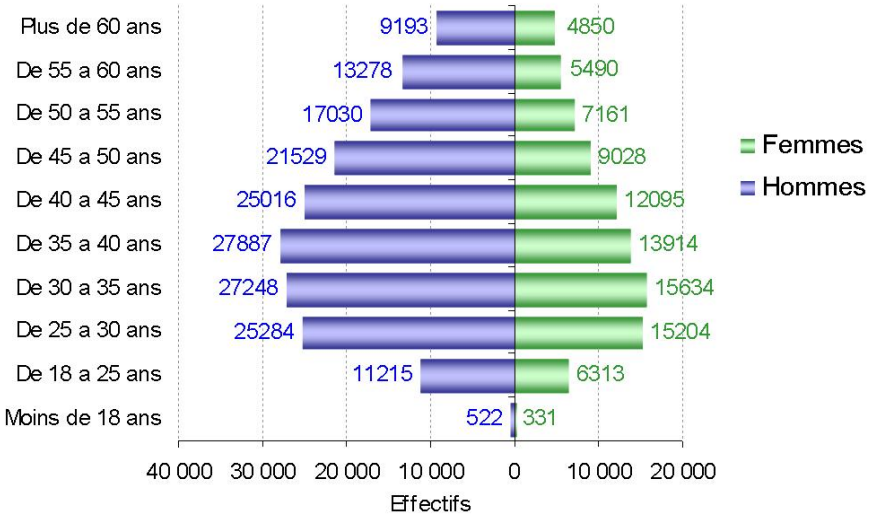


FIG. 4.3 – Eléments démographiques

En 2007, notre portefeuille comporte près de 270 00 adhérents et 610 000 bénéficiaires, ces effectifs sont en progression de 17% par rapport à 2006.

48% des adhérents sont mariés, 43% ont des enfants et en ont en moyenne 0,80.



	Hommes	Femmes	Global
âge moyen :	40,4 ans	39,2 ans	40,0 ans
Répartition H/F :	66,4%	33,6%	

FIG. 4.4 – Pyramide des âges des adhérents

L'âge moyen des adhérents est de 40 ans, une population majoritairement masculine avec les deux tiers d'adhérents hommes.

4.4 Cadence de règlements

	Années de Survenance								
	2004		2005		2006		2007		2008
Périodes de règlement	Montants	Répartition	Montants	Répartition	Montants	Répartition	Montants	Répartition	Montants
de jan à déc 2004	135 698 040 €	87,7%							
de jan à déc 2005	18 364 391 €	11,9%	139 207 935 €	86,1%					
de jan à déc 2006	607 208 €	0,4%	21 620 408 €	13,4%	156 872 151 €	86,0%			
de jan à déc 2007			772 802 €	0,5%	24 732 808 €	13,6%	189 843 224 €	87,3%	
de jan à déc 2008					801 406 €	0,4%	27 730 191 €	12,7%	147 679 531 €
Total	154 669 639 €	100%	161 601 146 €	100%	182 406 365 €	100%	217 573 415 €	100%	147 679 531 €
PSAP	0 €		0 €		0 €		846 697 €		
Total (yc PSAP)	154 669 639 €		161 601 146 €		182 406 365 €		218 420 111 €		

PSAP (*Provision pour sinistres à payer*) : - pour 2007 0,39%

- pour 2006 0,00%

Statistiques arrêtées le 31/12/2007

Observées le 31/12/2008

FIG. 4.5 – Cadence de règlements

En 2007, la vitesse de liquidation des prestations s'est accélérée par rapport à 2006 : au 31/12/06 il restait 14% de prestations à verser sur l'exercice de survenance 2006, alors qu'au 31/12/07 pour la survenance 2007 il ne restait que 13,1% à régler.

4.5 Ventilation des remboursements Gras Savoye

Année de survenance	2006 Montants	2006 Répartition	2007 Montants	2007 Répartition	Evolution (yc PSAP)
Chirurgie	4 541 061 €		5 474 526 €		
Chambre particulière	4 384 413 €		5 096 055 €		
Frais de séjour	5 977 822 €		7 770 853 €		
Forfait journalier	3 419 025 €		4 157 003 €		
Autres actes	1 215 905 €		1 536 172 €		
Hospitalisation	**19 538 226 €**	**11%**	**24 034 610 €**	**11%**	**23%**
Actes de Spécialité	**3 647 855 €**	**2%**	**4 338 601 €**	**2%**	**19%**
Généraliste	13 191 617 €		15 871 503 €		
Spécialistes	13 736 145 €		16 211 880 €		
Cumul Consultations, Visites	**26 927 761 €**	**15%**	**32 083 383 €**	**15%**	**19%**
Radios, Analyses, AM	**18 334 506 €**	**10%**	**22 308 948 €**	**10%**	**22%**
Pharmacie	**26 502 070 €**	**15%**	**30 950 690 €**	**14%**	**17%**
Soins dentaires	5 194 220 €		6 142 307 €		
Prothèses dentaires	24 059 162 €		28 228 757 €		
Orthodontie	6 982 141 €		8 750 516 €		
Cumul Dentaire	**36 235 524 €**	**20%**	**43 121 579 €**	**20%**	**19%**
Optique verres	18 485 591 €		22 531 234 €		
Optique monture	14 949 164 €		18 155 022 €		
Optique lentilles	3 341 175 €		4 284 920 €		
Opération de la myopie	187 492 €		286 543 €		
Cumul Optique	**36 963 422 €**	**20%**	**45 257 720 €**	**21%**	**22%**
Maternité	5 807 206 €		6 103 483 €		
Frais d'obsèques	412 903 €		312 893 €		
Divers	8 036 892 €		9 908 204 €		
Cumul Autres Postes	**14 257 001 €**	**8%**	**16 324 580 €**	**7%**	**15%**
Total (y compris PSAP)	**182 406 365 €**		**218 420 111 €**		**20%**
Nombre d'adhérents	**228 347**		**268 221**		**17%**
Remboursements moyens par adhérent	**799 €**		**814 €**		**2%**

PSAP : Provision pour sinistres à payer

Provisions :	0,000%	0,389%

FIG. 4.6 – Ventilation des remboursements Gras Savoye

Entre l'exercice 2006 et 2007, les remboursements ont progressé de 20%. Dans le même temps le nombre d'adhérents s'est accru de 17%. Au final, les montants remboursés en moyenne par adhérents ont augmenté de 2%. Les postes optique et dentaire sont ceux qui ont le plus contribué à cette dérive.

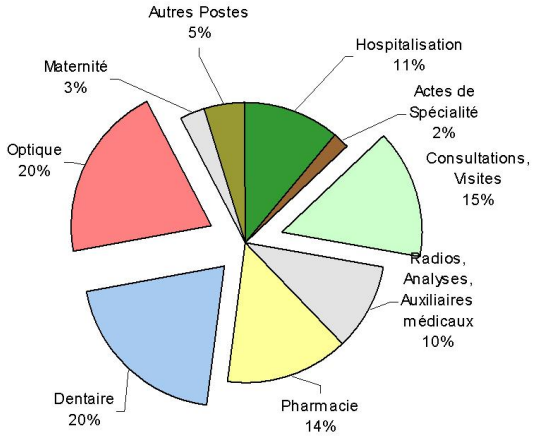


FIG. 4.7 – Ventilation des remboursements Gras Savoye

Les principaux postes de dépenses sont l'optique, le dentaire et les consultations / visites.

4.6 Taux de couverture

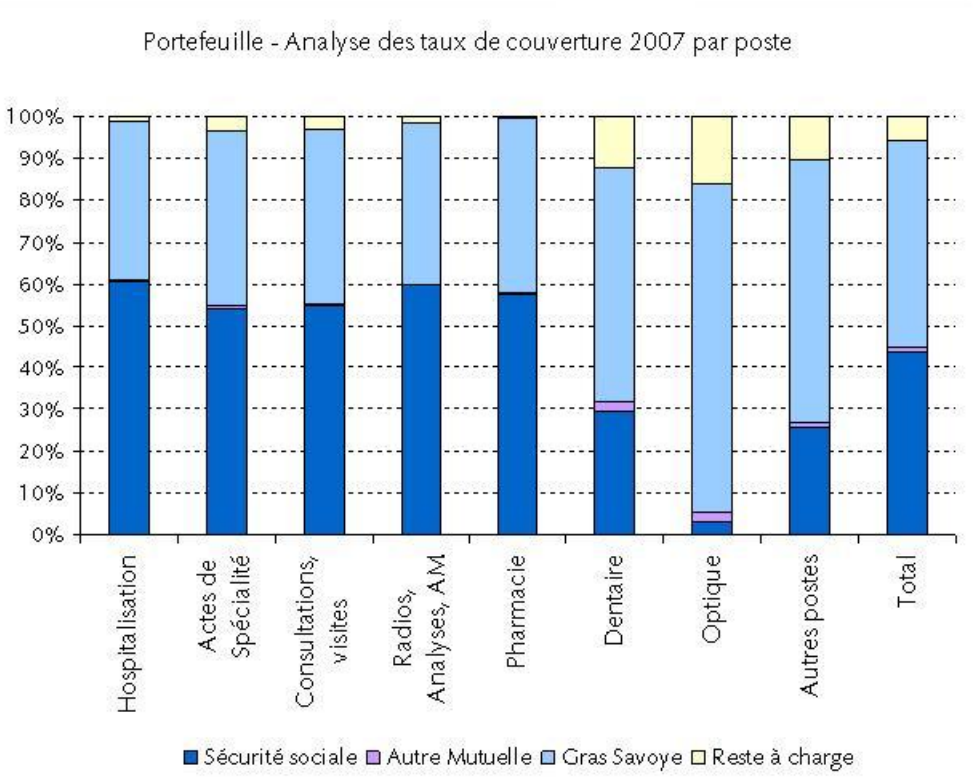


FIG. 4.8 – Taux de couverture par poste

Le taux de couverture, défini comme étant le rapport entre les dépenses prises en charge (Sécurité sociale, Gras Savoye et autres mutuelles) et les frais réels engagés, atteint les 94%, tous postes confondus.

Certains postes tels que l'optique ne bénéficient que d'une contribution minime de la part de la Sécurité sociale (quelques euros seulement), le régime prend alors en charge la part la plus importante des dépenses.

Le taux de couverture réel d'une famille n'est jamais parfaitement connu au travers d'une base de gestion. Seules les dépenses dont une demande de remboursement a été faite sont prises en compte. Celles qui ne passent pas par ce circuit sont alors inconnues et génèrent un léger biais. A titre d'exemple, il est possible de diminuer facialement le taux de couverture d'un régime en rajoutant une garantie telle que la chirurgie de l'œil. Avant que la garantie ne soit mise en place, cet acte n'apparaissait pas. Une fois la garantie effective, l'acte est intégré dans le calcul, il suffit alors qu'il ne soit pas remboursé au même niveau que le reste des autres garanties pour que le taux de couverture soit abaissé.

Par ailleurs, le calcul du taux de couverture inclut les lignes de décomptes sans intervention de Gras Savoye. A priori, cette approche peut contribuer dans les deux sens : les lignes de décomptes remboursées intégralement par la Sécurité sociale améliorent la couverture alors que celles des actes non pris en charge par les garanties la détériorent. Nous avons fait le choix arbitraire de ne pas prendre en compte les lignes sans remboursement de la part de Gras Savoye (pour information, le taux de couverture global aurait été de 91% sinon).

Chapitre 5

Analyses préliminaires des données

5.1 Méthodologie

Les analyses univariées, puis multivariées constituent la première étape d'une étude statistique exploratoire dans laquelle chacune des variables de recherche doit faire l'objet d'une étude synthétique visant à résumer l'information disponible. Afin de mener à bien une étude statistique, il est toujours intéressant de procéder ainsi pour avoir un ordre d'idée du contenu des différentes variables. Cela permet de repérer d'éventuelles incohérences mais également d'avoir une première idée de l'influence des variables par rapport à un critère donné.

Ces analyses sont établies à partir de l'ensemble des prestations versées au titre de l'exercice de survenance 2007 et réglées jusqu'au 31/12/08. Comme nous l'avions déjà indiqué précédemment, les PSAP sur l'exercice 2007 sont relativement faibles. Dans toute la suite de l'étude, nous négligerons cette provision compte tenu de son faible impact. Nous avons préféré l'exercice 2007 plutôt que 2008 car, au 31/12/08, sa connaissance est quasiment complète.

En moyenne, les frais réels moyens annuels par bénéficiaire sont de 720 € avec 319 € de remboursements de la Sécurité sociale, 7 € provenant d'autres mutuelles, 350 € de remboursements Gras Savoye. Le reste à charge de 44 € porte le taux de couverture moyen à 94%, valeur qui reflète le caractère haut de gamme de nos contrats.

Nous avons continué à appliquer la même règle que celle indiquée précédemment : les lignes de décomptes sans intervention de Gras Savoye ne sont pas prises en compte.

Nous avons décliné des analyses univariées des différentes caractéristiques pouvant influer sur la consommation médicale :

- le type de bénéficiaire
- l'âge,
- le sexe,
- la catégorie socio-professionnelle,
- la localisation géographique,
- le type d'activité.

Ces analyses préliminaires permettront de compenser le fait que les modélisations neuronales ne donnent pas d'explication sur les résultats obtenus.

5.2 L'effet type de bénéficiaire

Trois types de bénéficiaires existent : adhérent, conjoint et enfant. Les remboursements Gras Savoye annuels moyens par bénéficiaire sont de 406 € pour les adhérents, de 484 € pour les conjoints et de 206 € pour les enfants.

L'écart entre les adultes (adhérents et conjoints) d'une part et les enfants d'autre part, est lié essentiellement à l'âge.

L'écart entre les adhérents et les conjoints provient notamment du mode de consommation des conjoints qui peut différer de celui des adhérents. En effet, les conjoints peuvent déjà bénéficier d'une autre couverture par ailleurs et donc ne pas utiliser systématiquement le régime de leur époux (se). Cette remarque reste valable également pour les enfants.

Dans la suite de l'étude, nous nous intéresserons uniquement à la consommation médicale des adhérents afin de neutraliser ce biais.

5.3 L'effet homme/femme

Le montant de remboursements Gras Savoye moyen des adhérents de sexe masculin est de 343 € alors que celui des adhérents de sexe feminin est à 533 €, soit près de 55% en plus. Ce facteur présente donc une très forte influence sur la consommation médicale.

La disparité de la consommation par type de bénéficiaire au global indiquée précédemment est atténuée par le fait que les adhérents des contrats de notre portefeuille sont majoritairement des hommes.

En effet, le fait que les femmes consomment davantage que les hommes compense celui que les conjoints consomment moins que les adhérents.

Il est alors intéressant de croiser le facteur sexe avec le type de bénéficiaires. Un conjoint de sexe féminin consomme 22% de moins que s'il était adhérent principal, et 12% de moins pour les conjoints de sexe masculin.

La différenciation par sexe se fait moins ressentir chez les enfants.

	Sexe	Homme	Femme	Ecarts Femme/Homme
Type de bénéficiaire	Adhérent	343 €	533 €	+55%
	Conjoint	302 €	415 €	+37%
	Enfant	201 €	210 €	+4%
Ecarts	Conjoint/Adhérent	-12%	-22%	
	Enfant/Adhérent	-41%	-61%	

5.4 L'effet âge

Globalement la consommation médicale croît avec l'âge. Cette évolution peut être décomposée en trois phases :

- de 18 à 30 ans : La croissance au fil des ans est très forte et l'écart par sexe se creuse nettement avec +18% en moyenne par an pour les hommes et +28% pour les femmes.

- de 31 à 40 ans : La consommation est quasi stationnaire avec +1% en moyenne par an pour les hommes et +2% pour les femmes.

- de 41 à 60 ans : La croissance reprend doucement de 3% aussi bien pour les hommes que pour les femmes.

Attention, l'évolution mesurée ici est celle des remboursements Gras Savoye moyen par an et non pas celle des frais réels engagés.

La tendance de la courbe des frais réels engagés aurait été plus marquée car le poids des postes de consommation fluctue en fonction de l'âge et le régime n'intervient pas au même niveau sur tous les postes.

Par exemple, le passage de verres simples en verres progressifs fréquent autour de la quarantaine génère un pic de consommation sur le poste optique fortement sollicité par le régime, alors que les pathologies plus lourdes et plus coûteuses, relativement bien prises en charge par le régime général, touchent plus souvent des personnes plus âgées.

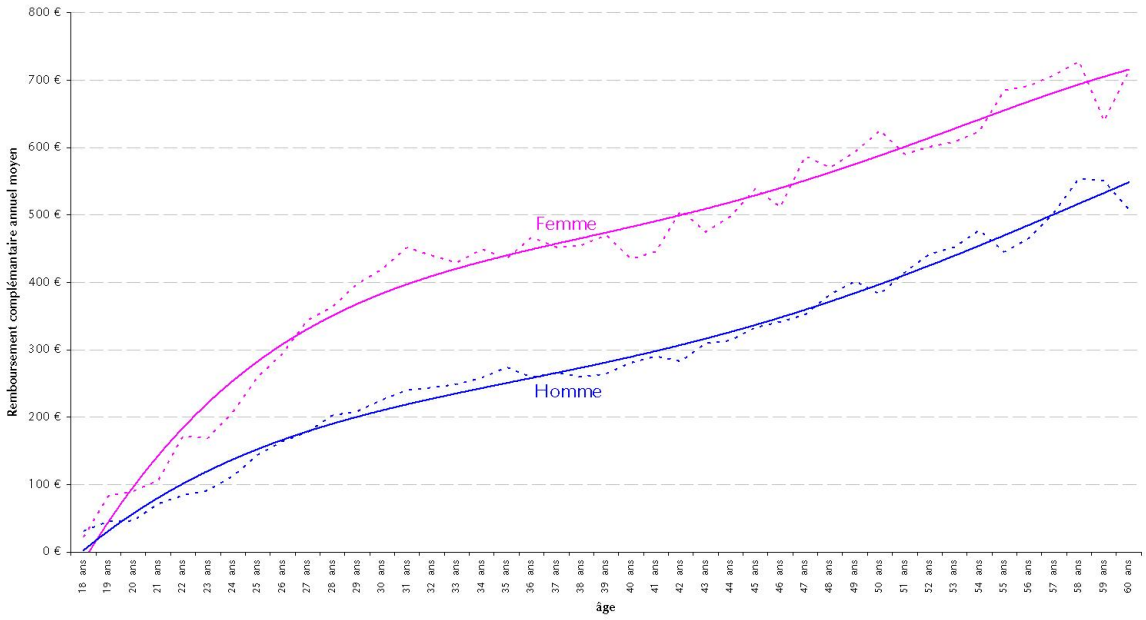


FIG. 5.1 – Consommation médicale d'un adhérent par âge et sexe

5.5 L'effet catégorie socio-professionnelle

Seuls les contrats bénéficiant d'un régime distinct pour chaque collège peuvent nous permettre d'étudier cet effet. Les contrats "ensemble du personnel" n'indiquent aucune information sur les catégories de populations assurées.

Les populations cadres sont les plus consommatrices avec en moyenne 374 € de remboursements Gras Savoye par adhérent contre 311 € pour les E.T.A.M. et 296 € pour les non cadres.

D'une manière générale, plus le niveau social est élevé plus les garanties proposées sont haut de gamme. Dans le même temps, l'impact financier du reste à charge pour l'assuré se fait d'autant plus ressentir que le niveau de CSP est bas. Ce phénomène est mis en évidence en pratique par des taux de PSAP différents selon la CSP : les taux de PSAP sont plus importants pour les classes élevés pour qui les délais de remboursements portent moins à conséquence.

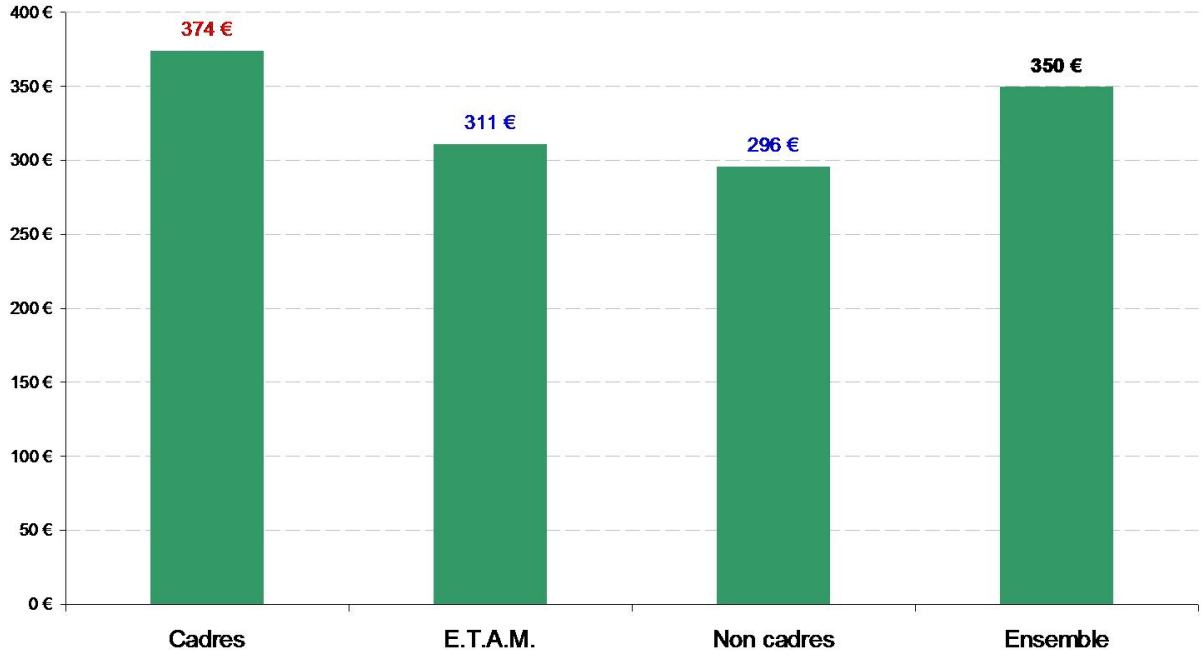


FIG. 5.2 – Influence de la CSP sur les remboursements complémentaires

5.6 L'effet localisation géographique

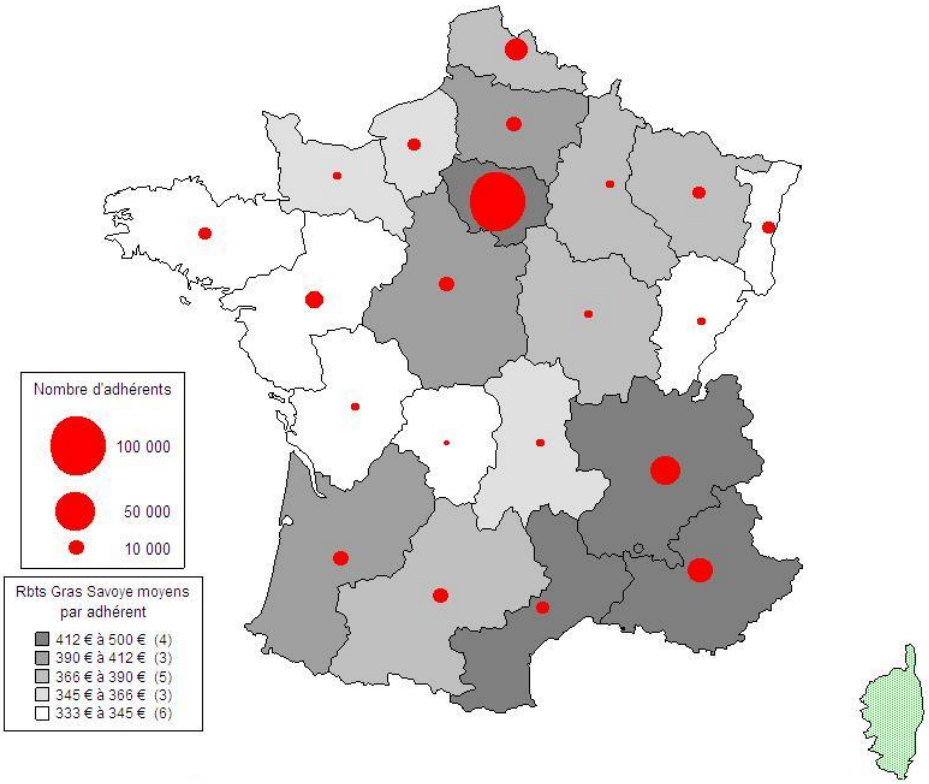


FIG. 5.3 – Remboursements Gras Savoye moyens par adhérent selon la région

Les zones où les montants remboursés sont les plus élevés sont celles où le nombre d'adhérents est aussi le plus concentrés (Ile-de-France et le quart sud est). Dans les grandes agglomérations, les tarifs pratiqués par les professionnels de santé y sont souvent plus élevés que dans les zones moins peuplées. Une des explications de ces différences de coûts peut provenir, entre autres, d'une répercussion des écarts de frais de fonctionnement selon la région et de la bonne solvabilité des patients. Typiquement, un dentiste dans le seizième arrondissement de Paris ne paiera pas le même loyer que s'il était installé dans la Creuse.

La région Alsace-Lorraine bénéficie d'un régime général particulier. Le taux de remboursement de la Sécurité sociale pour la plupart des postes est de 90% au lieu de 70% dans le reste de la France. La Sécurité sociale intervenant davantage, le régime complémentaire frais de santé est donc moins sollicité, ce qui explique la faiblesse des montants remboursés par Gras Savoye dans cette région.

5.7 L'effet activité

Catégorie INSEE	Classe INSEE	Rbts GS moyens
Services	Services aux entreprises	450 €
	Transports	375 €
	Services aux particuliers	562 €
	Activités financières	524 €
	Activités immobilières	492 €
Total Services		457 €
Industrie	Industries des biens d'équipement	398 €
	Industries des biens intermédiaires	388€
	Energie	359 €
	Industries agricoles et alimentaires	428 €
	Industrie des biens de consommation	519 €
	Industrie automobile	319 €
Total Industrie		409 €
Commerce	Commerce de gros	435 €
	Commerce et réparation automobile	340 €
	Commerce de détail divers	272 €
	Commerce de détail alimentaire	319 €
	Commerce de détail équipt personne	388 €
	VPC	634 €
	Commerce de détail de santé beauté	643 €
Total Commerce		396 €
Construction	Construction	309 €
Total		429 €

Le secteur d'activité de Services, qui représente plus de la moitié des adhérents couverts par un contrat de notre portefeuille, détient le montant de remboursement moyen le plus élevé.

Comme nous l'avions déjà évoqué précédemment, ce phénomène est étroitement lié à la forte proportion de cadres le caractérisant.

5.8 L'effet garanties

Les garanties ont une influence directe sur la consommation médicale. Plus les niveaux de garanties sont élevés plus les fréquences et les coûts moyens par acte le sont également.

L'exemple suivant représentant le coût moyen d'une monture en fonction de la valeur de l'indicateur de garantie illustre parfaitement ce phénomène. Le tarif moyen d'une monture dans le cadre d'un régime haut de gamme (indicateur entre 90% et 100%) est supérieur de près de la moitié à celui d'un régime de milieu de gamme (indicateur entre 60% et 70%).

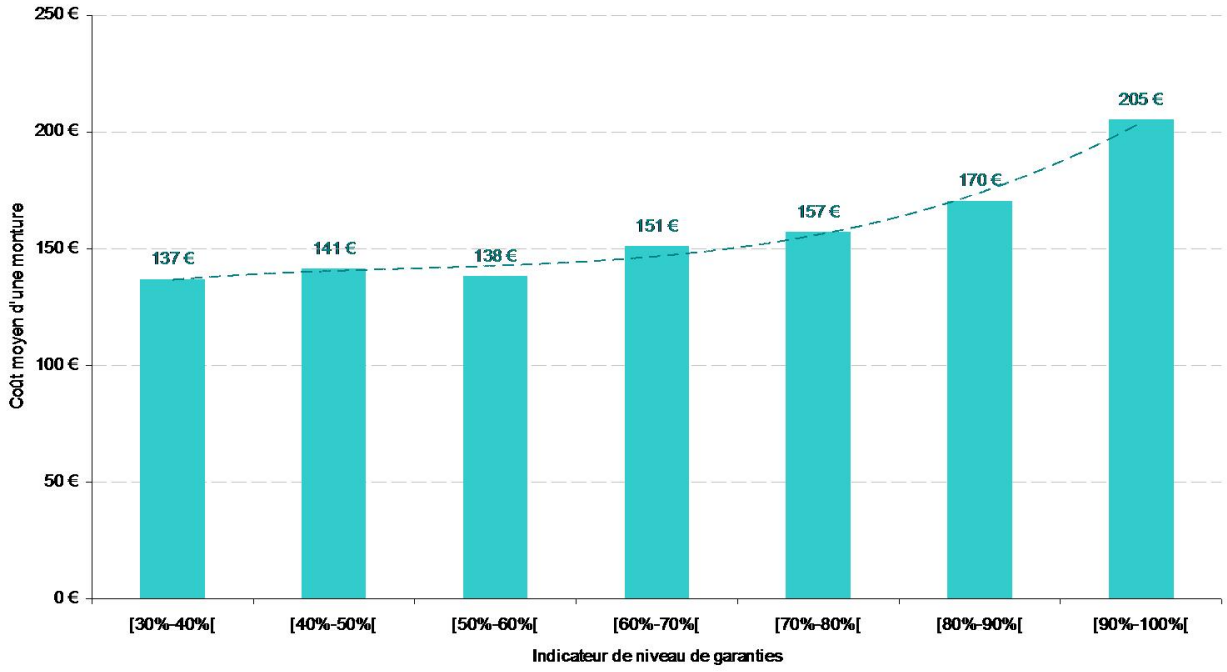


FIG. 5.4 – Influence de la garantie sur les tarifs pratiqués

Chapitre 6

Transfert de dépenses entre verres et montures

6.1 Problématique

Une des motivations qui nous a poussé à porter notre choix sur une modélisation de réseau de neurones plutôt que sur une approche tarifaire classique était, entre autres, la possibilité de prendre en compte les interactions existantes entre les différentes garanties. A titre d'exemple, nous traiterons, ici, le cas des équipements optiques (monture + 2 verres) dans lequel les dépenses sont étroitement liées aux niveaux de garanties proposées.

Premièrement, le fait que les assurés puissent conserver une partie des dépenses à leur charge peut freiner la consommation médicale. Ils sont alors davantage responsabilisés et deviennent plus vigilants sur la comparaison des différents tarifs. Réciproquement, des garanties couvrant l'intégralité des dépenses peuvent inciter des personnes à consommer sans se préoccuper des coûts. Typiquement, les bénéficiaires peuvent être moins hésitants à prendre une monture coûteuse de marque de luxe dans le cadre d'un contrat aux frais réels, que s'ils doivent en prendre une partie importante à leur charge.

Deuxièmement, certains professionnels de santé peu scrupuleux peuvent parfois ajuster leurs tarifs sur les limites de garanties contractuelles. Ces pratiques s'inscrivent dans une volonté d'optimisation des garanties visant à minimiser le reste à charge de l'assuré au détriment du régime de frais de santé.

Que ce soit à l'initiative du bénéficiaire ou bien à celle de l'opticien, ces ajustements peuvent se faire à deux niveaux : soit au sein d'un même acte (verre ou monture), soit entre les verres et la monture.

Dans le premier cas, le prix de chaque acte est adapté avec le plafond de sa propre garantie. Dans le second, la répartition des prix entre les verres et la monture est modifiée en fonction des garanties des deux actes, de manière à conserver le coût global de l'équipement. Ce procédé contribue alors à un transfert de coûts entre les deux postes : une partie du prix de l'acte le moins bien couvert est reporté vers celui qui est le mieux remboursé dont le plafond de garantie n'est pas encore atteint.

6.2 Indépendance du prix des verres avec les garanties des montures (ANOVA)

Dans le cadre du transfert de budget des verres vers la monture, nous avons tenté de mettre en évidence un début de preuve de cette dépendance entre les tarifs appliqués par les praticiens et les garanties.

La démarche que nous avons choisie est bâtie sur le postulat suivant : si ce transfert n'existait pas, le tarif moyen d'un verre devrait suivre la même distribution quelle que soit le niveau de garantie de la monture.

Tout d'abord, afin de neutraliser certains biais liés au type de bénéficiaire ou à la localisation géographique, nous avons restreint le périmètre de notre étude à la consommation des adhérents (hors conjoints et enfants) résidant en Ile-de-France et ayant été remboursés d'un équipement complet (1 monture + 2 verres).

Ensuite, nous avons défini arbitrairement 3 groupes, selon le niveau de l'indicateur de garantie de la monture :

- Garantie Basse : indicateur de niveau de garantie inférieur à 75%
- Garantie Moyenne : indicateur de niveau de garantie comprise entre 75% et 90%
- Garantie Haute : indicateur de niveau de garantie supérieur à 90%

Nous avons donc choisi de tester l'hypothèse nulle :

« $H_0 : \mu_1 = \mu_2 = \mu_3$, le coût moyen d'un verre est identique dans chaque classe de niveau de garantie. »

Vs

« H_1 : au moins deux de ces moyennes sont différentes. »

Afin de tester ces dépendances, nous avons opté pour un test *ANOVA* (*ANalyse Of VAriance*), cette méthode permet d'étudier la différence de moyennes de populations selon un facteur qualitatif. Dans le cas présent, l'identité des moyennes des tarifs des verres est testée dans nos trois groupes ayant des niveaux de garanties des montures différents.

L'idée directrice de cette méthode est fondée sur la décomposition de la variation totale des données en deux composantes :

Variation totale (autour de la moyenne) = Variation dûe au facteur (intergroupe) + Variation dûe à l'erreur d'échantillonnage.

Les sommes de carrés servent à mesurer ces variations :

Variation totale : $SCT = \sum_{i,j}(X_{i,j} - \overline{X})^2$

Variation due au facteur : $SCF = \sum_i(n_i\overline{X}_i - \overline{X})^2$

Variation due à l'erreur : $SCE = \sum_{i,j}(X_{i,j} - \overline{X}_i)^2$

La relation suivante en découle : $SCT = SCF + SCE$

Le principe du test repose sur le fait que si l'hypothèse nulle est vraie, la variation dûe au facteur est censée être faible comparativement à celle dûe à l'erreur d'échantillonnage. La statistique F, si H_o est vraie obéit à une loi Fischer à $k-1$ et $n-1$ degrés de liberté. Cette loi sert alors à l'obtention de la valeur critique du test et du seuil expérimental.

L'utilisation du test F de Fisher suppose de vérifier les trois conditions suivantes :

- Les X_i suivent une loi normale dans les populations de référence ou les échantillons sont de taille assez grande,
- Les populations ont la même variance,
- Les populations sont prélevées de manière indépendante.

La normalité de frais réels a déjà été traitée précédemment. De plus, le nombre de données observées avoisine les 6 000, nous supposons le premier point vérifié. Le troisième point se déduit naturellement de sa définition. Les variances des trois groupes sont assez proches : 12 615 pour le premier groupe, 10 909 pour le second et 11 329 pour le dernier. Sans pour autant avoir testé rigoureusement cette condition, nous la supposons vérifiée. Nous considérons donc que l'ensemble des conditions d'application du *test F* de Fischer sont réunies.

Le tableau d'analyse de variance et test F de Fisher se présente de la façon suivante :

Source	Degrés de Liberté	Somme des carrés	Carré moyen	Valeur F
Modèle	SCF	k-1	CMF=SCF/(k-1)	F=CMF/CME
Erreur	SCE	n-k	CME=SCE/(n-k)	
Total	SCT	n-1		

Source	Degrés de Liberté	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	2	89 764	128 901	5,61	0,0037
Erreur	5 983	68 680 623	5 503		
Total	5 985	68 809 525			

La p-value obtenue de $0,37\%$ permet d'infirmer, avec un niveau de confiance raisonnable, que le prix des verres est indépendant du niveau de garantie des montures, ce qui revient à dire que le facteur niveau de garantie des montures a bien un effet sur le prix des verres.

Les modélisations de phénomènes liés au comportement humain, restent bien souvent difficiles à interpréter avec certitude, du simple fait de la multitude de paramètres pouvant entrer en jeux. Neanmoins, l'issue de ce test abonde plutôt dans le sens d'une influence des garanties du poste monture sur les tarifs des verres.

6.3 Report de budget des verres sur les montures

L'analyse de la variance que nous avons réalisée a permis de tester l'égalité des trois moyennes, mais n'a fourni aucune information expliquant les éventuels écarts. Les coûts moyens des montures et de l'équipement ont été rajoutés afin de tenter comprendre le phénomène.

Garantie monture	Coût moyen monture	Coût moyen verre	Coût moyen équipement
Basse	145 €	200 €	544 €
Moyenne	167 €	189 €	545 €
Haute	205 €	191 €	587 €
Total	171 €	192 €	556 €

Premier constat, le coût moyen de la monture croît avec le niveau de garantie. D'expérience, nous anticipions ce résultat car comme nous l'avions déjà évoqué précédemment, les garanties influent sur le comportement de consommation et notamment sur les prix.

Second constat, le coût moyen de l'équipement croît entre des garanties moyennes et hautes. Toutes choses étant égales par ailleurs, une augmentation de garantie ne peut que profiter à la hausse des tarifs.

Troisième constat important : en passant d'une garantie moyenne à basse sur les montures, alors que le coût global de l'équipement reste stable, le coût des verres augmente. Pour un équipement complet de même coût, une diminution des garanties sur les montures entraîne une augmentation du poids des verres de 69% à 74% dans les dépenses totales. Ces effets, peuvent très bien matérialiser le transfert de coûts de la monture vers les verres lorsque les garanties des montures sont faibles.

Il est important de souligner que notre démarche ne constitue en aucun cas une preuve irréfutable de cette pratique. Comme dans toute démarche bayésienne, elle permet juste de ne pas exclure son éventualité et de la considérer comme vraie jusqu'à preuve de son contraire.

En synthèse de l'ensemble de ces éléments, l'hypothèse suivante ne semble pas être contredite :

Lorsque les garanties proposées sur les montures sont faibles, le prix des verres semble être artificiellement augmenté pour compenser une baisse de prix appliquée à la monture. Les remboursements de la complémentaire santé sont alors optimisés en restant dans une enveloppe budgétaire globale équivalente. Le bénéficiaire est alors mieux couvert et l'opticien peut augmenter sa facturation.

Chapitre 7

Analyses factorielles

7.1 Analyse selon le secteur d'activité

7.1.1 Analyse en Composantes Principales

Dans cette partie, nous avons choisi d'explorer plus finement les différentes spécificités liées aux secteurs d'activité. Plusieurs dizaines de variables les caractérisent, ce qui rend difficile la visualisation de l'espace de points et le repérage des différentes typologies de comportement. L'*Analyse en Composantes Principales* (*A.C.P.*) est une technique synthétisant de l'information contenue dans un grand nombre de variables. Elle permet de trouver le meilleur sous-espace de faible dimension (plan ou volume) grâce auquel on aura la meilleure représentation du nuage.

L'*A.C.P.* construit de nouvelles variables artificielles et les représente graphiquement. Les relations peuvent être visualisées, ainsi que l'existence éventuelle de groupes d'éléments ou de variables.

Techniquement, cela revient à maximiser la variance des observations projetées (problème d'optimisation sous contrainte) afin de déterminer les axes qui expliquent le mieux la dispersion des points disponibles.

L'*Analyse en Composantes Principales* peut donc être vue comme une technique de réduction de dimensionalité de façon à conserver le maximum d'information.

Dans les représentations graphiques des variables et des individus, les corrélations sont symbolisées par une proximité dans la projection. L'interprétation de ces représentations est délicate et doit respecter une démarche rigoureuse pour éviter de tirer des conclusions trop hâtives.

Les projections des variables sur un plan factoriel s'inscrivent dans un cercle de rayon unitaire, appelé cercle des corrélations. La qualité de leur représentation d'autant plus importante qu'elles se situent en bordure du cercle. Deux variables proches et positionnées au centre du cercle peuvent très bien avoir des caractéristiques diamétralement opposées.

La représentation d'un point individu notée M sur un plan factoriel de centre O et d'axes U et V se mesure grâce à l'indicateur suivant : $cos^2(\overline{OM}, \overline{U}) + cos^2(\overline{OM}, \overline{V})$ (Pythagore). Les points sont bien représentés dès lors que la valeur est proche de 1.

Pour chaque secteur d'activité, nous avons sélectionné les variables suivantes :

- Démographie : âge moyen, prop. d'hommes, prop. de mariés, nombre d'enfants, prop. vivant en Ile-de-France,
- Garanties : consultations de spécialistes, montures, verres, prothèses dentaires, chambres particulières
- Fréquences : consultations de spécialistes, montures, verres, prothèses dentaires, chambres particulières
- Coûts moyens : consultations de spécialistes, montures, verres, prothèses dentaires, chambres particulières
- Consommation des adhérents : Remboursements GS annuels, taux de couverture

Avant de commencer l'analyse, il est important de s'assurer que le pourcentage de variance expliquée n'est pas insignifiant, auquel cas il faudrait vérifier qu'il n'y ait pas trop de variables redondantes.

Dans un second temps, le choix optimum du nombre de plans peut être repéré graphiquement en observant le pourcentage de variance expliquée pour chaque axe afin de repérer une cassure.

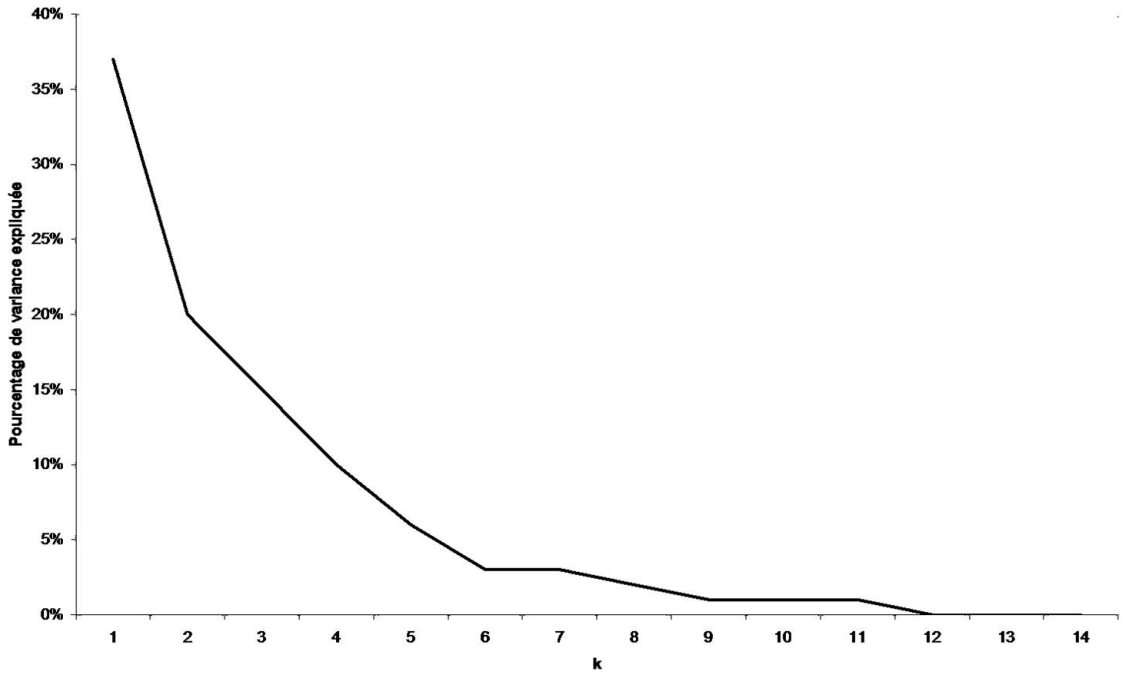


FIG. 7.1 – Pourcentage de variance expliquée pour chaque axe

Les deux premiers axes représentent 57% de la variance (71% avec trois). Il n'y a pas de cassure très nettement marquée, une première se situe entre le deuxième et le troisième axe la suivante entre le sixième et le septième.

Nous choisissons donc de projeter sur les trois premiers plans en privilégiant l'étude du premier.

Représentation des variables

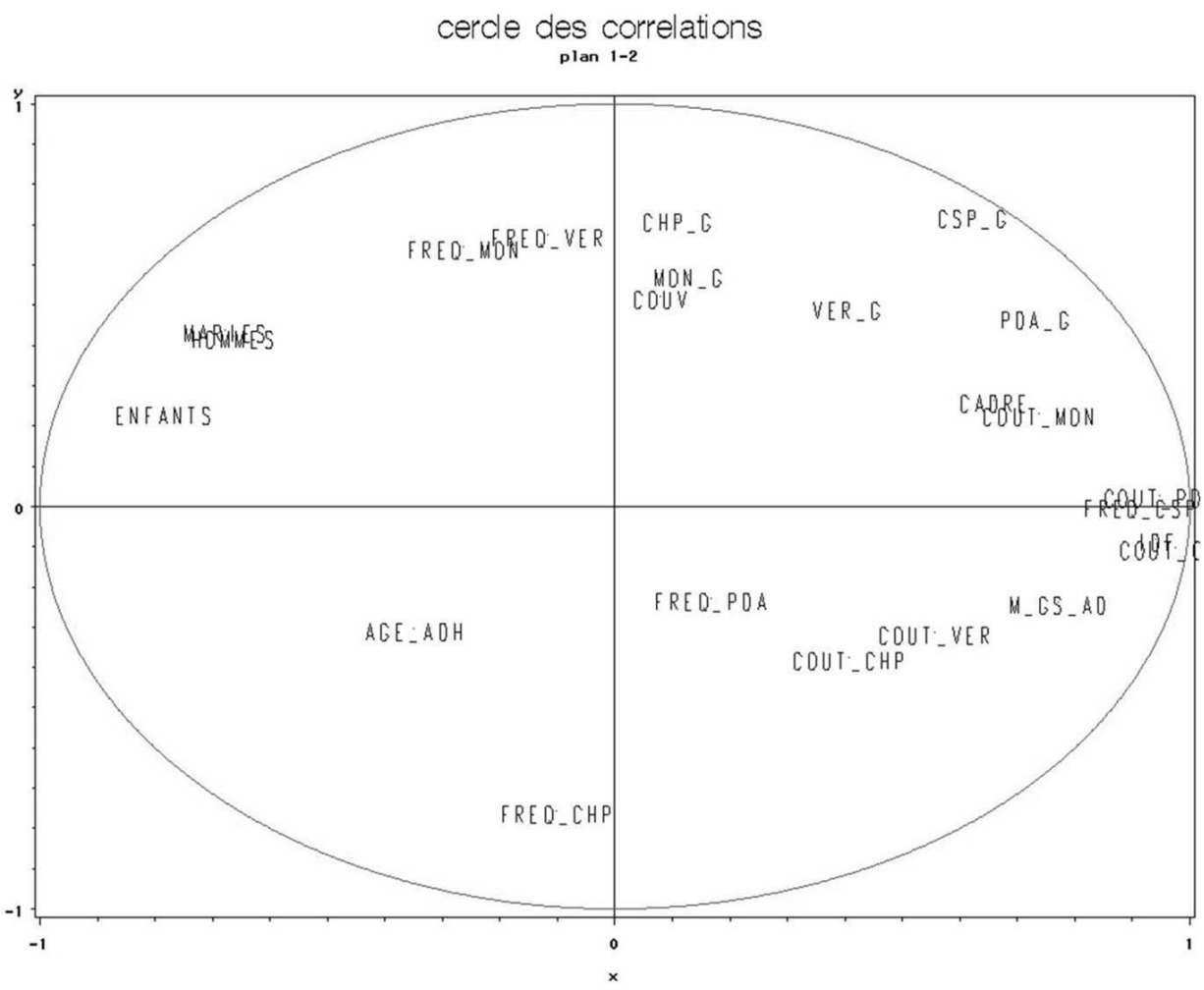


FIG. 7.2 – Représentation des variables

La plupart des variables, à l'exception de la fréquence des prothèses dentaires, ne sont pas trop éloignées de la bordure du cercle de corrélation et bénéficient ainsi d'un bon niveau de représentation.

Plusieurs relations (parfois évidentes) apparaissent :

- Le fait d'être un homme est anticorrélé avec les montants Gras Savoye,
- Les mariés ont plus d'enfants que les célibataires,
- Les populations cadre ont des coûts de montures élevées,
- Les fréquence de consommation de verres et de monture sont fortement corrélées,
- Le niveau de couverture global est étroitement lié aux niveaux de garanties,
- Les coûts élevés des prothèses dentaires et des spécialistes est typique de l'Ile-de-France,
- Des montants de remboursement Gras Savoye élevés sont liés à l'Ile-de-France et aux coûts élevés.

Représentation des individus sur le premier axe

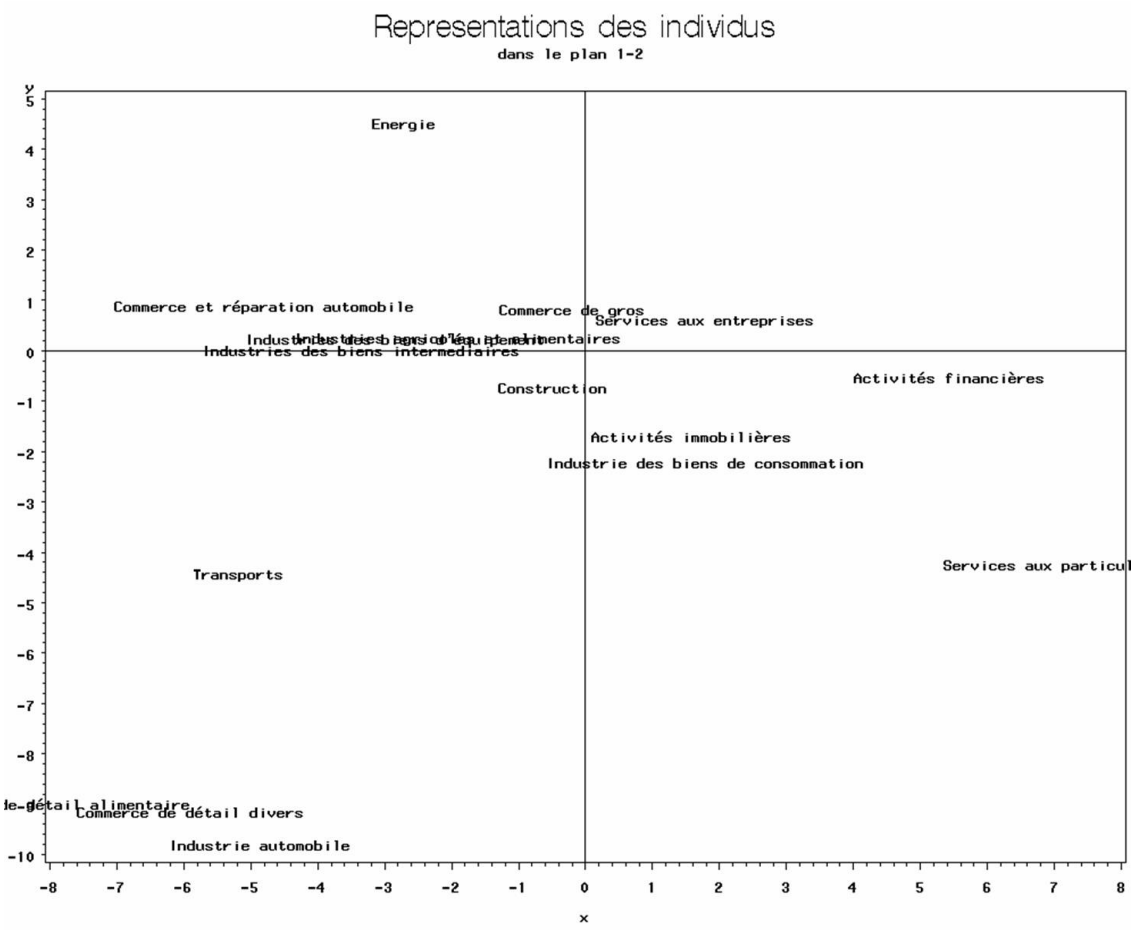


FIG. 7.3 – Représentation des individus sur le premier plan

Le critère somme des cos^2 sur les deux premiers axes pour vérifier la bonne représentation des individus n'est applicable que pour ceux éloignés du centre de gravité. Il ne peut donc pas être utilisé pour les secteurs construction, commerce de gros, services aux entreprises et industries agricoles et alimentaires. Pour les secteurs suffisamment éloignés du centre, l'activité immobilière et l'industrie de biens de consommation, la somme des cos^2 sur les deux premiers axe vaut respectivement 0,13 et 0,20. Leur représentation sur le premier plan factoriel n'est donc pas de bonne qualité.

En observant les coordonnées des vecteurs propres, les deux axes sont principalement caractérisés dans le sens positif comme suit :

- Le premier axe est essentiellement défini par un coût et une fréquence élevés de consultations de spécialistes, des prothèses dentaires onéreuses, une population féminine, une localisation en Ile-de-France et un montant de remboursements Gras Savoye élevé,

- Le second indique un niveau de couverture élevé associé à des garanties haut de gamme.

Les services aux particuliers, les transports et les services aux entreprises contribuent le plus à l'explication de l'axe 1 (resp. 21%, 19% et 16%) l'énergie et les transports pour l'axe 2 (resp. 30% et 26%).

Représentation des individus sur le deuxième axe

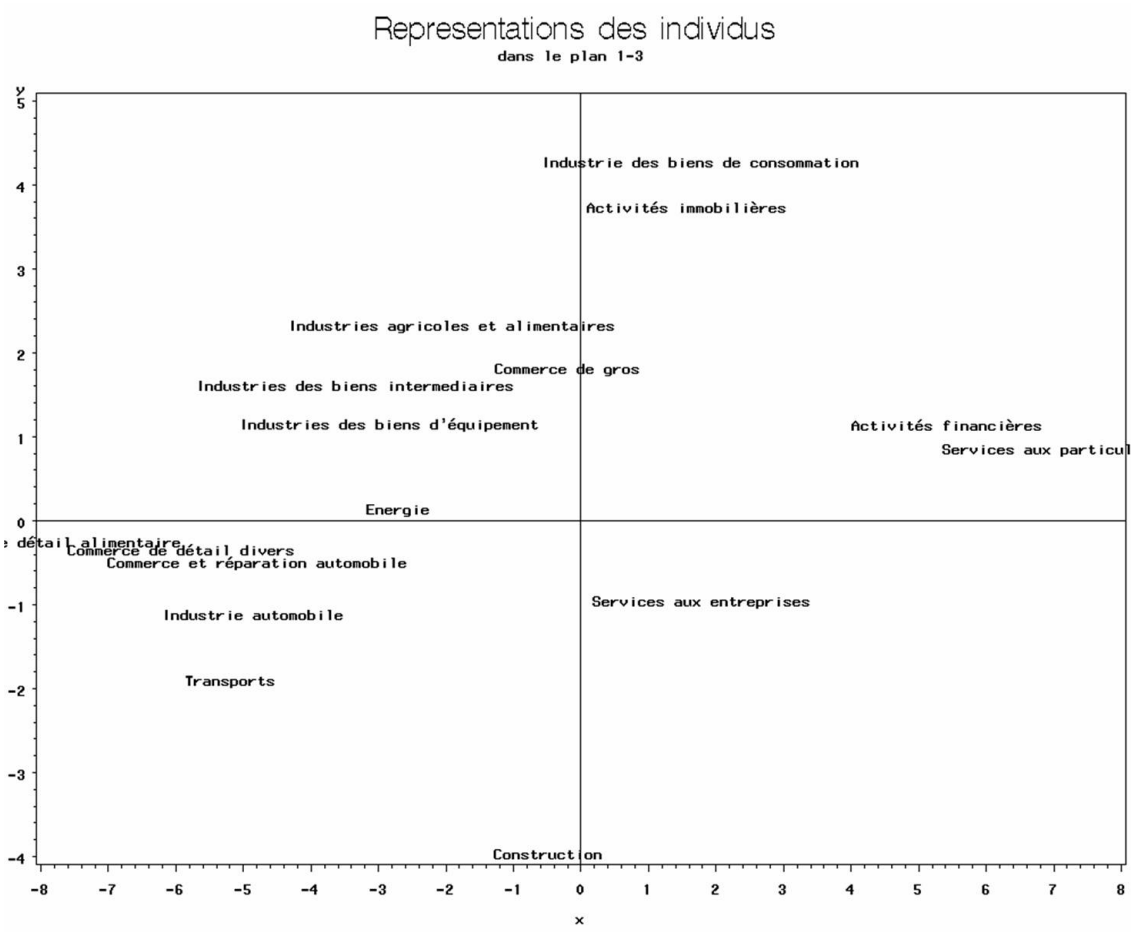


FIG. 7.4 – Représentation des individus sur le deuxième plan

Le second plan est constitué des axes 1 et 3.

L'axe 3 se caractérise par un âge et des fréquences élevés.

La construction est le secteur qui contribue le plus à son explication.

Les secteurs d'activités sont assez regroupés en grandes classes que sont les services, l'industrie, le commerce et la construction.

Représentation des individus sur le troisième axe

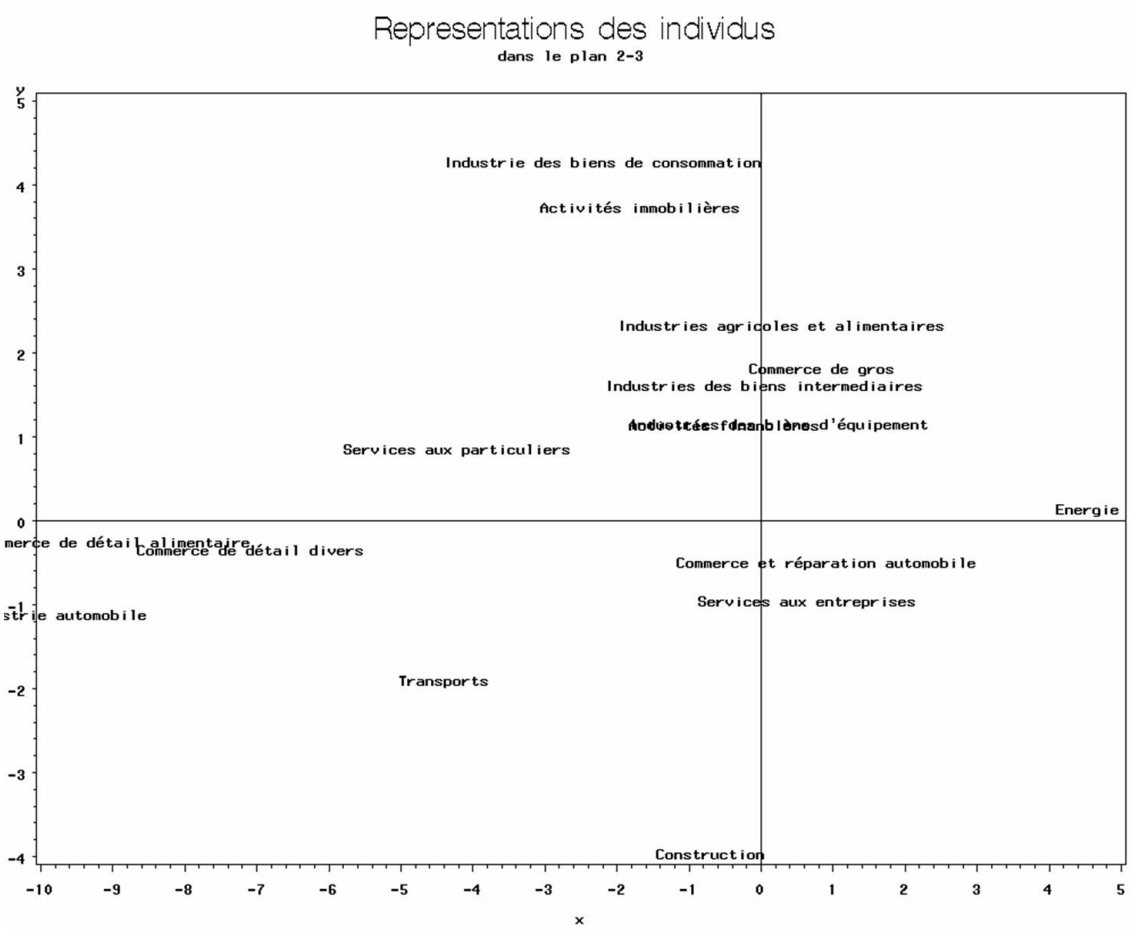


FIG. 7.5 – Représentation des individus sur le troisième plan

Le troisième plan est constitué des axes 2 et 3 définis précédemment.

Nous ne l'évoquons ici qu'à titre indicatif car l'information apportée ne semble pas suffisante pour être exploitée ultérieurement.

7.1.2 Classification hiérarchique

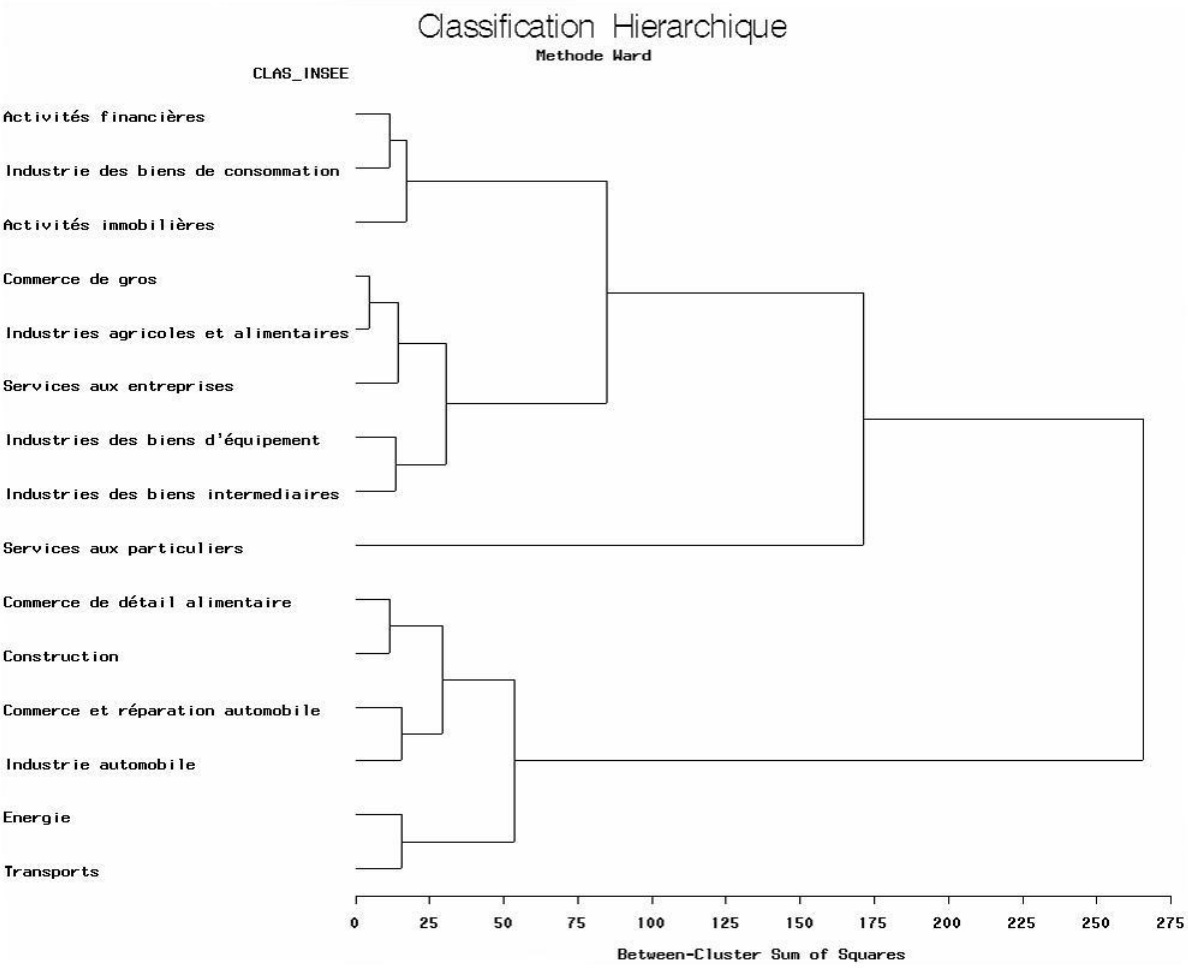


FIG. 7.6 – Classification hiérarchique

La classification hiérarchique permet de regrouper le plus efficacement possible les secteurs d'activités ayant des similitudes de comportements. Cette méthode permet de partitionner en classes homogènes (les plus compactes possibles) et distinctes (les plus séparées possibles). La visualisation de l'ensemble des éléments est organisée en arbres hiérarchiques où toute coupure de l'arbre par une droite horizontale fournit une partition de l'ensemble. Nous avons choisi de créer une rupture à partir de quatre classes.

La classification dite hiérarchique ascendante ou par agrégation procède par fusions successives de classes déjà existantes. A chaque étape, les classes dont la "distance" est la plus faible, vont fusionner. La distance entre deux groupes de points la plus couramment utilisée est celle de Ward [WAR63] :

$$d(A,B) = \frac{n_A n_B}{n(n_A + n_B)}.d^2(g_A, g_B)$$

avec d la distance euclidienne, g la gravité et n le nombre d'individus.
A chaque étape, le principe consiste à sélectionner le regroupement de classes tel que l'augmentation de l'inertie intra classe, utilisée comme indice de niveau, soit minimum.

7.1.3 Synthèse

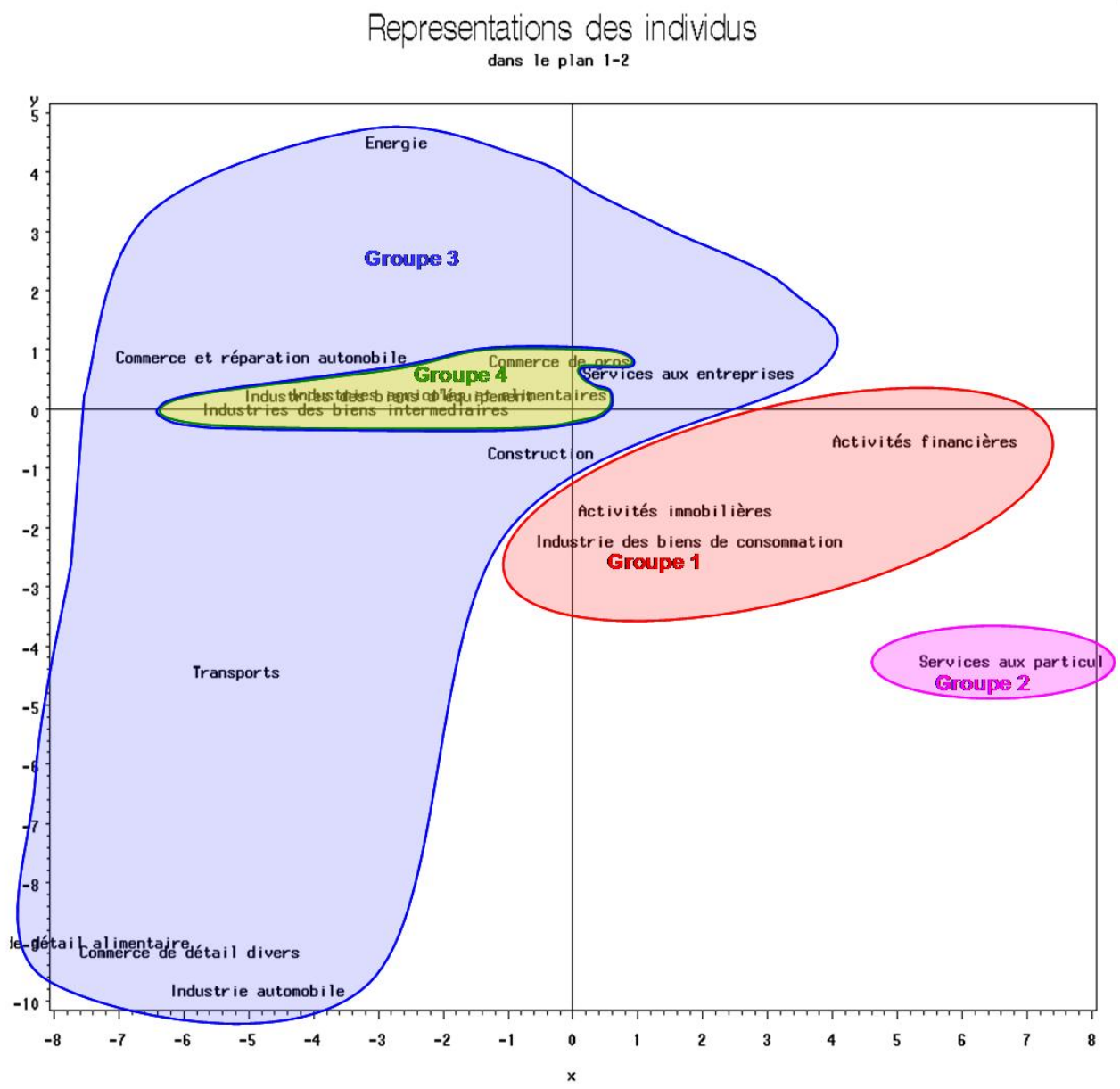


FIG. 7.7 – Représentation des groupes sur l'ACP

Les données moyennes, obtenues par la classification en quatre groupes, en termes de garanties, de coûts moyens par acte, de fréquences et de consommation médicale annuelles sont les suivantes :

Indicateurs de niveaux de garanties					
Classes	Spécialistes	Monture	Verre	Prothèse dent.	Chambre part.
1	94%	81%	67%	81%	92%
2	95%	77%	68%	88%	93%
3	92%	75%	66%	81%	94%
4	95%	81%	82%	83%	92%

Coûts moyens par acte					
Classes	Spécialistes	Monture	Verre	Prothèse dent.	Chambre part.
1	40,13 €	161,86 €	155,14 €	10,94 €	57,59 €
2	43,38 €	169,57 €	371,7 €	11,65 €	62,99 €
3	34,52 €	143,67 €	133,14 €	9,65 €	53,09 €
4	36,84 €	158,8 €	156,99 €	10,39 €	59,37 €

Fréquences moyennes par bénéficiaire					
Classes	Spécialistes	Monture	Verre	Prothèse dent.	Chambre part.
1	1,37	0,16	0,32	8,63	0,01
2	1,5	0,05	0,19	9,36	0,02
3	0,93	0,12	0,26	6,88	0,01
4	1,22	0,13	0,28	7,33	0,01

Eléments démographiques						
Classes	Age	Prop. Hommes	Prop. Mariés	Nb Enfants	Prop. Cadre	Prop. IDF
1	41,5 ans	47%	33%	0,71	66%	58%
2	38,2 ans	53%	37%	0,66	68%	75%
3	40,5 ans	79%	54%	0,87	56%	30%
4	39,6 ans	70%	52%	0,85	55%	41%

Consommation globale		
Classes	Remboursements Gras Savoye annuels des adhérents	Taux de couverture
1	513 €	92,2%
2	618 €	94,4%
3	349 €	93,5%
4	442 €	94,1%

Les quatre groupes sont principalement définis par les caractéristiques suivantes :

- Groupe n°1 - Activités financières, activités immobilières et industrie de biens de consommation : Représente près de 9% des effectifs, leur moyenne d'âge est élevée, peu de mariés, une part importante sont des cadres, résident majoritairement en Ile-de-France et sont essentiellement des femmes. Leurs niveaux de garanties, fréquences et coûts moyens sont au dessus de la moyenne et ont une consommation annuelle importante.

- Groupe n°2 - Service aux particulier : Constitué que d'un seul secteur, il représente près de 4% des effectifs, leur moyenne d'âge est en dessous de la moyenne, peu de mariés, la proportion de cadre et de résidants en Ile-de-France est encore plus marquée que dans le premier groupe. Leurs niveaux de garanties, fréquences et coûts moyens sont très élevés et leur consommation annuelle est très importante.

- Groupe n°3 - Service aux entreprises, construction, transports, commerce de détail alimentaire, commerce de détail divers, industrie automobile, commerce et réparation automobile et énergie : Représente près de 22% des effectifs, leur moyenne d'âge est légèrement au dessus de la moyenne, majoritairement des mariés de sexe masculin, la proportion de cadre est en dessous de la moyenne et résident en dehors en Ile-de-France. Leurs niveaux de garanties, fréquences et coûts moyens sont faibles et leur consommation annuelle reste faible.

- Groupe n°4 - Commerce de gros, industries de biens intermédiaires, industries agricoles et alimentaires, industries des biens d'équipement et services aux entreprises : Représente la majorité des effectifs, leur moyenne d'âge est dans la moyenne, majoritairement des mariés, la proportion de cadre est faible et résident majoritairement en dehors de l'Ile-de-France. Leurs niveaux de garanties sont élevés, les fréquences, coûts moyens sont dans la moyenne et leur consommation annuelle est également moyenne.

7.2 Analyse selon la localisation géographique

7.2.1 Analyse en Composantes Principales

De la même manière que nous avons exploré les spécificités relatives aux secteurs d'activité, nous avons reproduit la même trame d'étude aux régions. L'objectif étant de vérifier si la localisation géographique pouvait avoir un impact notable sur les attitudes de consommation médicale et de faire ressortir des groupes ayant des similarités comportementales.

Pour chaque région, nous avons sélectionné les mêmes variables que dans la précédente étude à l'exception de la proportion d'adhérents vivants en Ile-de-France, que nous avons remplacée par la proportion d'adhérents appartenant au secteur d'activité de services.

Les deux premiers axes expliquent, à eux seuls, un peu plus des deux tiers de la variance. Pour cette raison, nous nous consacrons à l'étude uniquement du premier plan factoriel.

Représentation des variables

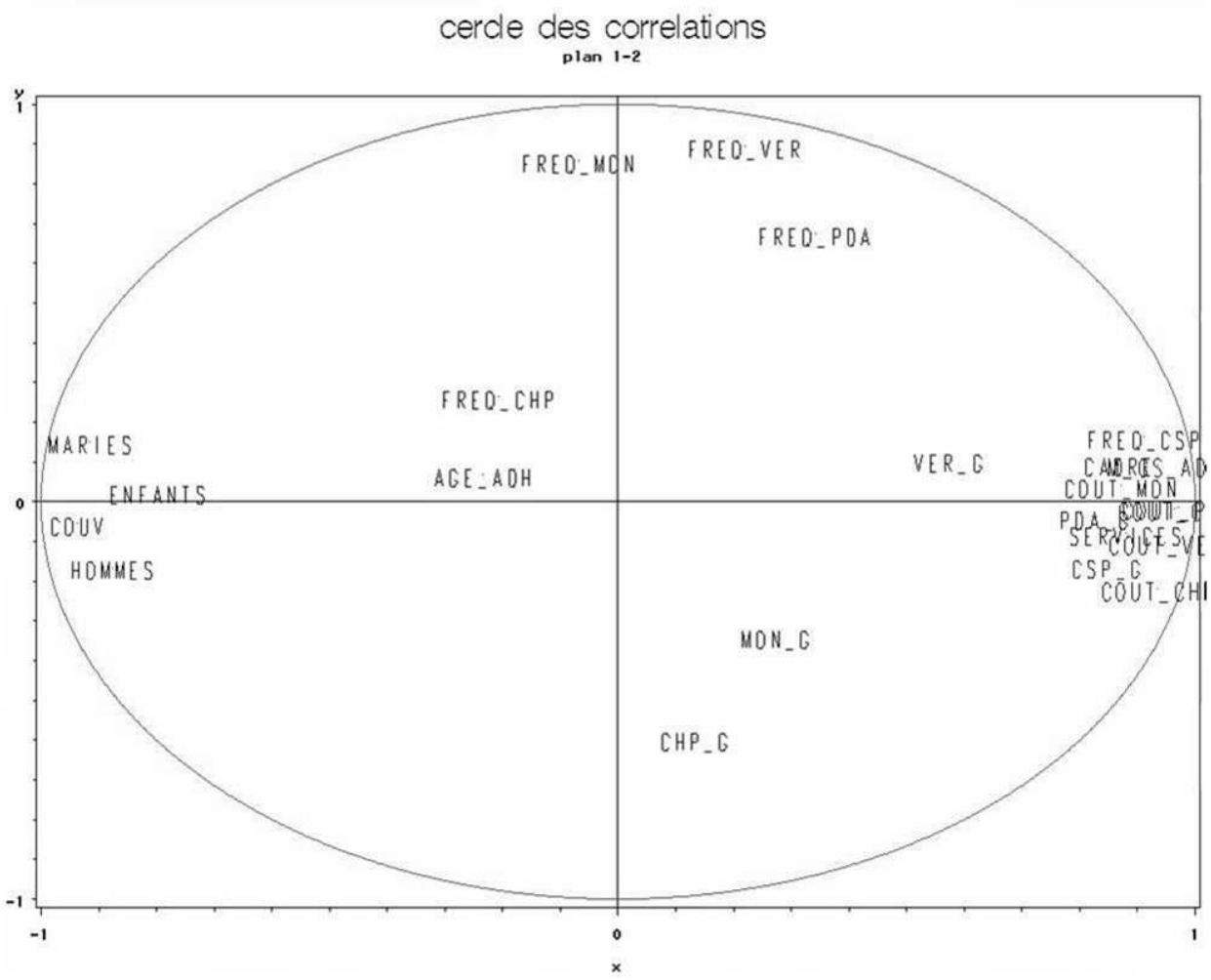


FIG. 7.8 – Représentation des variables

La majorité des variables sont bien représentées, à l'exception de l'âge et de la fréquence de chambres particulières.

Les relations évoquées lors de l'analyse par secteur d'activité se retrouvent pour la plupart ici, ce qui permet de valider ce que nous avions déjà remarqué : le sexe féminin, la proportion de cadres, des garanties élevées favorisent des coûts et fréquences élevés, ce qui accroît l'intervention de Gras Savoye.

La nouvelle variable introduite ici, à savoir la proportion d'adhérents issus du secteur de services, se caractérise notamment par des coûts et des fréquences de consommation élevés.

Représentation des individus sur le premier axe

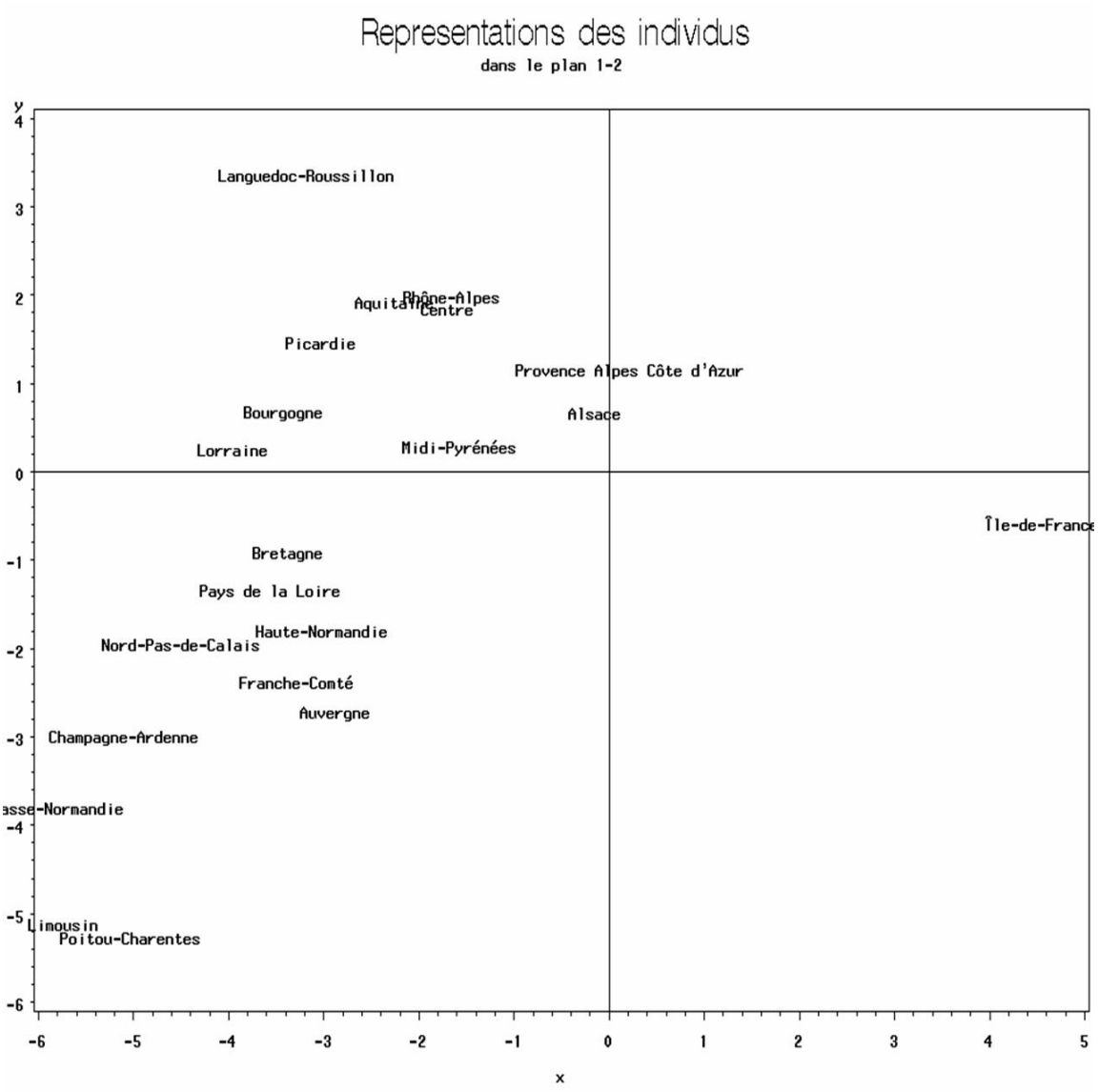


FIG. 7.9 – Représentation des individus sur le premier plan

Les trois régions Alsace, le Midi-Pyrénées et la région PACA étant trop proches du centre de gravité, le critère : somme des cos^2 sur le premier plan factoriel, n'a pas de signification. En revanche, les régions étant suffisamment éloignées du centre de gravité : Picardie, Lorraine et Limousin ne sont pas très bien représentées. Leur valeur de la somme des cos^2 sur le premier plan factoriel sont basses, elles valent respectivement : 0.24, 0.38 et 0.46. L'Ile-de-France, à l'inverse, est très bien représentée avec une valeur proche de 1. Ceci s'explique par le poids important de la région dans l'analyse.

Le premier axe est caractéristique des coûts élevés et de remboursements Gras Savoye importants dans le sens positif. Dans le sens négatif, il représente essentiellement une proportion d'hommes et couverture élevées. Le second symbolise des fortes fréquences dans le sens positif et de garanties importantes dans l'autre sens.

L'Ile-de-France contribue à 57% de l'axe 1 et la région Rhône-Alpes à 19% de l'axe 2.

Représentation des individus sur le premier axe hors Ile-de-France

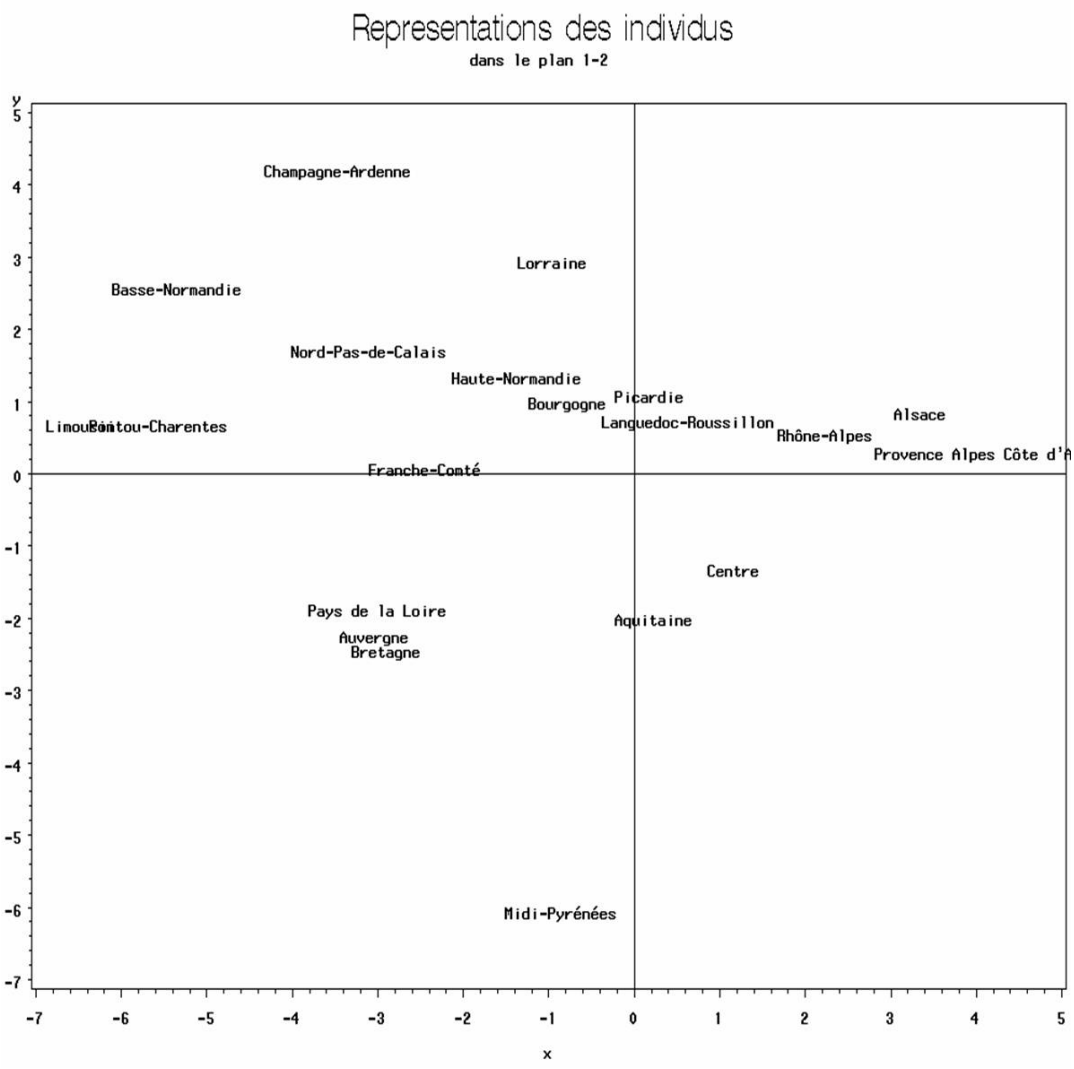


FIG. 7.10 – Représentation des individus sur le premier plan

La sur-pondération de l'Ile-de-France nous a poussé à réaliser le même exercice en l'ôtant de l'étude pour mettre plus en avant les différences entre les autres régions. Cette analyse n'est donnée qu'à titre indicatif car elle n'a pas apporté d'éléments majeurs si ce n'est plus de lisibilité. L'éclatement des régions est plus marqué, mais leurs positionnements relatifs ne sont pas fondamentalement modifiés.

7.2.2 Classification hiérarchique

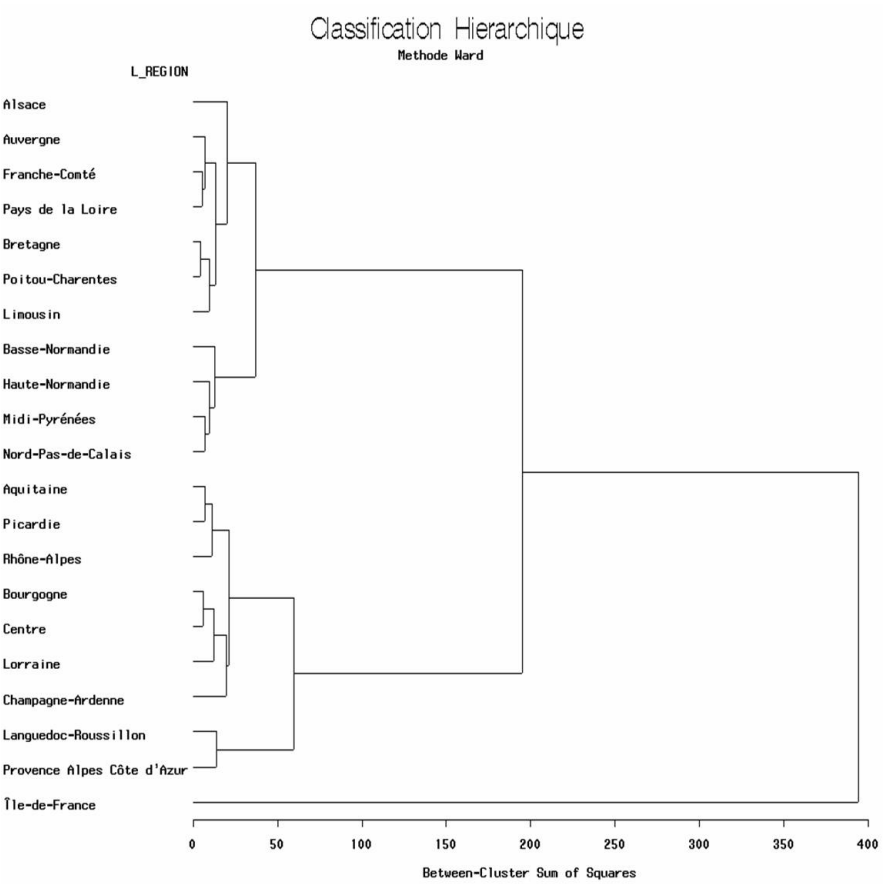


FIG. 7.11 – Classification hiérarchique

Nous avons repris la même méthode de classification que pour les secteurs d'activité.

L'Ile-de-France, qui est la région qui pèse le plus dans l'analyse, constitue une classe à elle seule.

La proximité géographique de certaines régions se retranscrit parfois sur le plan factoriel : la Basse et la Haute Normandie, le Poitou-Charentes et le Limousin, la Lorraine et la Champagne-Ardenne, le Languedoc-Roussillon et la Provence Alpes Côte d'Azur.

Nous avons choisi de sectionner l'arbre hiérarchique en quatre classes.

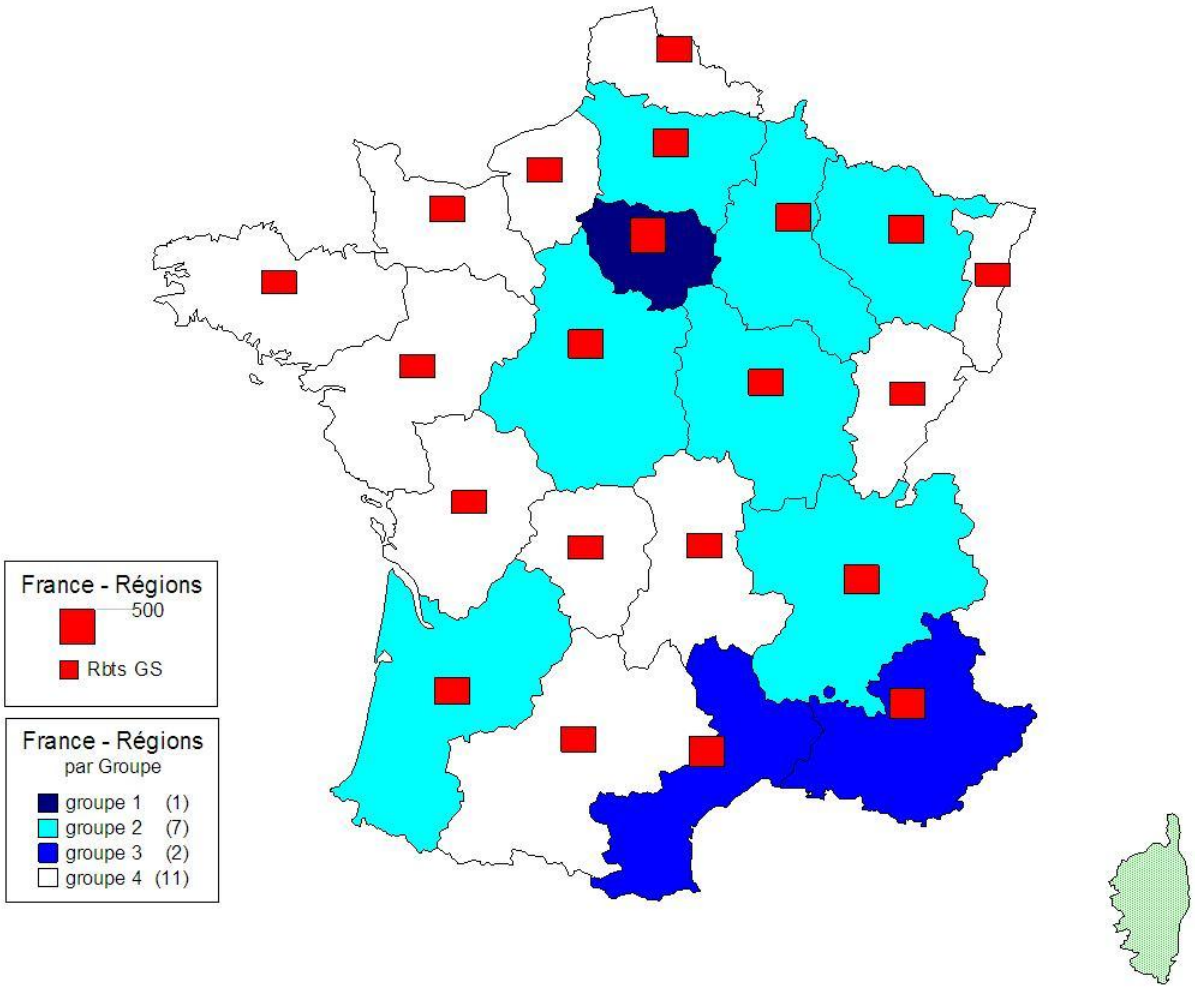


FIG. 7.12 – Représentation des groupes sur une carte de France

En représentant les quatre groupes sur une carte de France, la visualisation de leurs zones géographiques correspondantes s'avèrent être bien marquées. Le premier groupe correspond toujours à l'Ile-de-France, le deuxième au centre-est et à l'Aquitaine, le troisième au sud-est, le quatrième à l'ouest. La classification de la région Alsace-Lorraine, qui bénéficie d'un régime spécifique, est difficilement comparable aux autres régions.

7.2.3 Synthèse

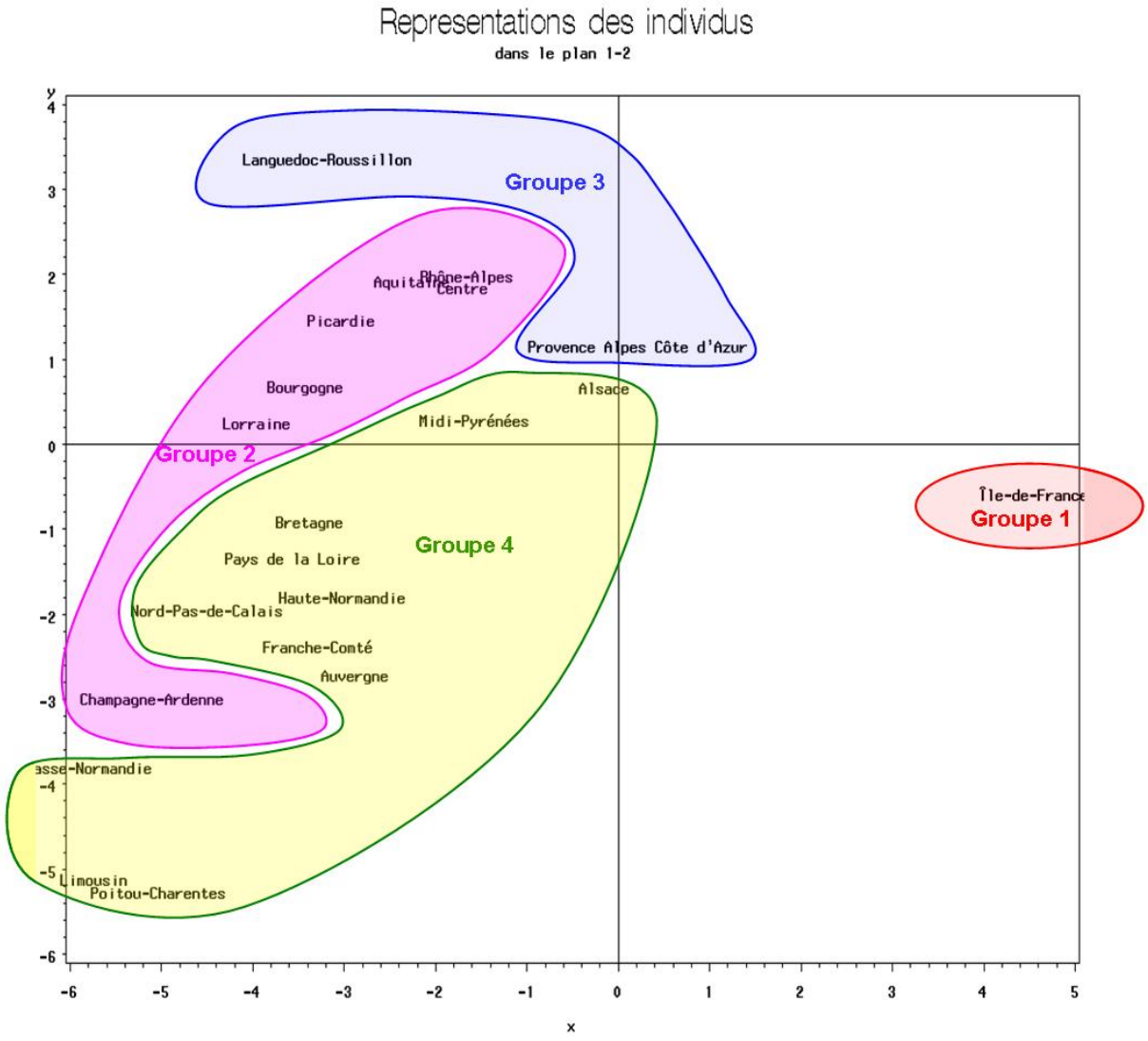


FIG. 7.13 – Représentation des groupes sur l'ACP

Les quatre groupes qui découlent de notre classification, sont caractérisés par les données moyennes suivantes concernant leurs garanties, leurs coûts moyens par acte, leurs fréquences et leur consommation médicale annuelle :

Indicateurs de niveaux de garanties					
Classes	Spécialistes	Monture	Verre	Prothèse dent.	Chambre part.
1	95%	80%	76%	84%	93%
2	93%	79%	74%	80%	92%
3	93%	80%	73%	82%	92%
4	93%	80%	74%	81%	93%

Coûts moyens par acte					
Classes	Spécialistes	Monture	Verre	Prothèse dent.	Chambre part.
1	43,68 €	165,55 €	182,21 €	11,83 €	87,53 €
2	33,12 €	150,24 €	146,52 €	9,39 €	47,27 €
3	34,87 €	159,74 €	141,06 €	10,11 €	57,72 €
4	29,75 €	147,49 €	137,34 €	8,90 €	45,03 €

Fréquences moyennes par bénéficiaire					
Classes	Spécialistes	Monture	Verre	Prothèse dent.	Chambre part.
1	1,44	0,13	0,27	7,53	0,01
2	1,02	0,14	0,28	7,65	0,03
3	1,33	0,14	0,29	8,85	0,01
4	0,91	0,12	0,25	6,53	0,01

Eléments démographiques						
Classes	Age	Prop. Hommes	Prop. Mariés	Nb Enfants	Prop. Cadre	Prop. Services
1	40,0 ans	62%	44%	0,75	68%	59%
2	40,4 ans	68%	54%	0,87	52%	33%
3	40,7 ans	69%	54%	0,85	49%	39%
4	39,9 ans	72%	55%	0,90	48%	35%

Consommation globale		
Classes	Remboursements Gras Savoye annuels des adhérents	Taux de couverture
1	510 €	92,4%
2	406 €	94,5%
3	451 €	94,1%
4	358 €	95,4%

Les principales propriétés des quatre groupes ainsi formés peuvent se résumer de la façon suivante :

- Groupe n°1 - Région Ile-de-France : Constitué d'une seule région, il représente plus de deux tiers des effectifs (35%), peu de mariés, moins d'enfants et une proportion d'hommes moins importante que dans les autres groupes. Une part importante sont des cadres et travaillent majoritairement dans les sociétés de services. Leurs niveaux de garanties, fréquences et coûts moyens sont les plus élevés. Leur consommation annuelle est importante. Leur taux de couverture reste faible car les tarifs pratiqués sont tellement élevés que les garanties, bien que haut de gamme, ne suffisent pas à couvrir les dépenses dans les mêmes proportions que dans les autres régions.

- Groupe n°2 - Régions Aquitaine, Bourgogne, Centre, Champagne-Ardenne, Lorraine, Picardie et Rhône-Alpes : Constitué de sept régions, il représente près de 28% des effectifs dont 13% pour la région Rhône-Alpes. Un tiers seulement travaillent dans des sociétés de services. Leurs garanties sont les plus faibles, en adéquation avec des fréquences et des coûts moyens également faibles.

- Groupe n°3 - Régions Languedoc-Roussillon et Provence-Alpes-Côte d'Azur : Représente près de 11% des effectifs, dont 8% pour la région PACA. Les proportions de cadres et de société de services sont dans la moyenne. Les fréquences, les tarifs pratiqués et leur consommation sont élevés. Ils restent cependant en dessous du niveau de l'Ile-de-France.

- Groupe n°4 - Autres régions : Représente le quart des effectifs, avec onze régions. Ils sont en grande majorité des hommes, mariés avec un nombre important d'enfants. La proportion de sociétés de services est moins marquée, la proportion de cadres est la plus faible. Les fréquences et les tarifs sont les plus faibles, leurs niveaux de garantie sont peu élevés mais suffisent à maintenir une bonne couverture globale. Leur consommation est la plus basse.

Chapitre 8

Synthèse des ressources disponibles

Cette étude préliminaire nous a permis d'explorer notre infocentre pour établir un bilan exhaustif des ressources disponibles.

Nous avons porté une attention toute particulière aux traitements des données manquantes ou aberrantes afin de parasiter le moins possible nos résultats.

Nous avons pu ensuite construire des indicateurs de niveaux de garantie grâce à la méthode d'*Imputation Multiple* par chaînes de Markov de *Data Augmentation*. N'importe quel type de garantie peut alors être hiérarchisé de manière homogène en lui affectant une valeur comprise entre 0 et 100%. Celle-ci correspond à une estimation du taux de couverture obtenu en simulant la garantie concernée à l'ensemble de notre portefeuille. Ces outils trouveront tout leur intérêt dans la modélisation neuronale. La prise en compte des garanties se fera par leur intermédiaire.

Les effets des différentes variables sur la consommation médicale ont été passés en revu afin d'en mesurer l'impact.

Parallèlement, des groupes de comportements homogènes ont été constitués selon deux approches : le secteur d'activité et la région géographique.

Ces analyses sont d'autant plus utiles, que les modélisations neuronales n'apportent aucune explication sur résultats obtenus. Cette première phase était donc essentielle pour pallier cette carence d'informations.

Cette étape préliminaire nous a aiguillé dans la sélection des variables pertinentes. Celles qui présentaient une réelle influence sur la consommation et, par conséquent, sur la tarification ont été retenues :

- nos indicateurs de garanties
- la localisation géographique
- le secteur d'activité
- le type de bénéficiaire
- le sexe
- l'âge
- la catégorie socio-professionnelle

Les effets des différentes variables peuvent être synthétisés dans le tableau récapitulatif suivant :

Critère	Consommation faible	Consommation élevée
Type de bénéficiaires	Enfants	Adhérents et Conjoints
Sexe	Homme	Femme
Age	Jeunes	Moins jeunes
Garanties	Basses	Hautes
Secteur d'activité	Construction	Services
C.S.P.	Ouvriers	Cadres supérieurs
Région	Petites agglomérations	Ile-de-France, Rhône Alpes, PACA

En résumé, voici quelques ordres de grandeurs des principaux effets qui ressortent de cette première étude :

- La consommation d'un adhérent de sexe féminin est supérieure de plus de la moité que celle d'un adhérent de sexe masculin (55%).

- Pour les hommes, la consommation d'un conjoint représente 88% de celle d'un adhérent (78% pour les femmes), celle d'un enfant représente 67% de celle d'un adhérent (51% pour les femmes).

- Dans les tranches d'âge des actifs, l'effet lié au vieillissement est de l'ordre de 1 à 3% par an (son les âges et le sexe).

- Le secteur d'activité ou bien la localisation géographique, souvent liés à d'autres facteurs, peuvent avoir un impact sur la consommation médicale allant jusqu'à plus 50% supplémentaire.

Deuxième partie

THEORIE DES RESEAUX DE NEURONES

Dans cette seconde partie, nous allons esquisser un bref aperçu des modèles de réseaux de neurones, de leur origine à leur applications en passant par les variantes possibles.

Nous nous intéresserons plus particulièrement au modèle que nous allons employer, à savoir le *Perceptron Multicouche* (*P.M.C.*).

Chapitre 9

Principes et fondements biologiques

9.1 Historique

Les prémices du concept de réseau de neurones sont apparues en 1943 avec la présentation par W. Mc Culloch et W. Pitts du neurone formel [CUL43]. (philosophie qu'il redéveloppera plus tard en 1959 [LET59]). Dans cet article fondateur, le neurone formel est défini comme une abstraction du neurone physiologique. L'idée directrice est de démontrer que le cerveau humain peut être assimilé à une machine de Turing, ce qui revient à dire que la pensée est régie par des mécanismes logiques et matériels.

Une machine de Turing est un modèle abstrait du fonctionnement d'un ordinateur et de sa mémoire, créé par Alan Turing. Elle se compose d'une tête de lecture comportant un nombre fini d'états internes et d'un ruban infini. La tête se déplace sur la bande en lisant les symboles un à un. Elle se réfère à une table d'états indiquant le changement de symbole à effectuer s'il y a lieu ainsi que le nouvel état à prendre en compte.

Par ce mécanisme relevant de l'imaginaire, elle parvient à reproduire n'importe quel processus. Tout problème calculable devient alors solvable par une machine. Par extension, une machine qui reprendrait les règles de toutes les autres serait universelle.

En 1949, D. Hebb propose une formulation du mécanisme d'apprentissage en modifiant des connexions synaptiques : si des neurones sont activés simultanément de façon répétée, leur liaison va s'intensifier. [HEB49]

En 1958, F. Rosenbalt invente le perceptron en s'inspirant du système visuel [ROS57] [ROS58]. C'est le premier modèle de réseau de neurones avec algorithme d'apprentissage. Il est composé d'une couche de perception et d'une couche décisionnaire.

Le modèle de l'*Adaline* (*ADAptative LINear Element*), présenté par B. Widrow et Hoff, fait son apparition dans la même période [WID60]. Son réseau est constitué d'un unique neurone combinant linéairement ses entrées et son algorithme neuronal optimise le critère des moindres carrés en minimisant l'erreur quadratique. Proche du Perceptron, il diffère de celui-ci par sa loi d'apprentissage qui est à l'origine de l'algorithme de rétropropagation du gradient.

En 1969, une critique des propriétés du *Perceptron* est publiée par M. Minsky et S. Papert [MIN69]. Celle-ci va avoir de lourdes retombées sur le développement du modèle qui

va voir ses activités de recherche s'arrêter. Les modèles de réseaux de neurones artificiels ne peuvent pas traiter efficacement des problèmes non linéaires. Cette limitation est le principale reproche qui leur est fait.

Ce n'est qu'en 1982, que ce domaine retrouve de l'intérêt grâce à un article publié par le physicien John Joseph Hopfield qui introduisit un nouveau modèle de réseau de neurones inspiré des verres de spin (systèmes magnétiques) [HOP82]. Il parvient à justifier l'utilité des réseaux de neurones au travers d'une théorie décrivant précisément et quantitativement le mécanisme d'un réseau de neurones formels. En fixant préalablement les objectifs de son modèle, il définit la structure ainsi que la loi d'apprentissage permettant d'y parvenir. A ce stade, la limitation aux cas linéaires n'est pas encore levée.

Apparue en 1983, la machine de Boltzmann est le premier modèle capable de traiter les limitations rencontrées dans le cas du Perceptron [HIN83a] [HIN83b], cependant la difficulté d'utilisation ainsi que la longueur des temps de calculs sont encore trés pénalisants. Un an plus tard, le système de rétropropagation du gradient de l'erreur attire l'attention des chercheurs.

L'année 1986 est marquée par une réelle avancée en la matière, la rétropropagation est alors adaptée dans le cadre des *Perceptrons-Multicouches*. Proposés initialement par Werbos [WER74], ils sont popularisés par Rumelhart [RUM86] et suscitent un réel engouement. Le modèle est enfin capable de traiter les phénomènes non linéaires. Cette révolution associée aux progrès informatiques ont participé à l'essor de ce modèle qui rencontra un succès notamment en raison de ses capacités d'apprentissage et de généralisation.

9.2 Neurone biologique

Les modèles de réseaux de neurones artificiels trouvent directement leur inspiration dans le mécanisme biologique de la pensée du cerveau humain.

Les neurones biologiques forment des réseaux de communication complexes où ils établissent entre eux de très nombreuses connexions. Pour avoir un ordre d'idée de la puissance de la machine humaine : le cerveau humain est constitué d'environ 10^{11} neurones. En moyenne chaque neurone a de l'ordre de 10^4 connexions.

Un neurone biologique est composé de quatre parties :

- le corps cellulaire : Il se compose du noyau et du mécanisme biochimique nécessaire à la synthèse des enzymes. De forme sphérique ou pyramidale, sa taille est de l'ordre de quelques microns de diamètre. Ce corps cellulaire assure la vie de la cellule grâce à ses molécules essentielles.

- les dendrites : D'aspect tubulaire, ces fines extensions de quelques dizaines de microns de longueur, se ramifient autour du neurone pour former une arborescence. Elles captent les signaux envoyés au neurone.

- l'axone : Beaucoup plus long que les dendrites, de quelques millimètres à plusieurs

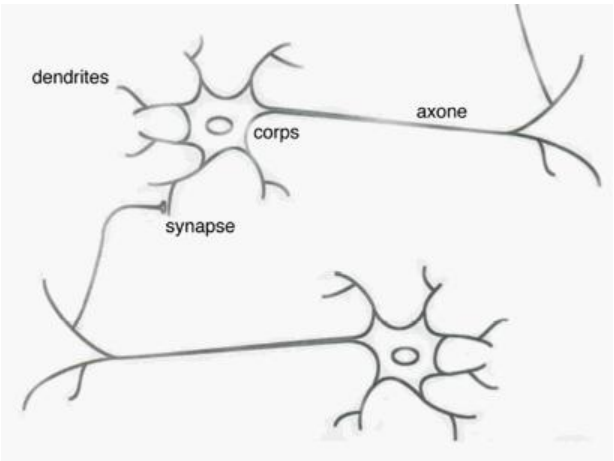


FIG. 9.1 – Neurone biologique

mètres, il se connecte aux dendrites des autres neurones pour véhiculer les signaux émis par le neurone.

– la synapse : C'est la jonction entre deux neurones, le plus souvent entre l'axone d'un neurone et le dendrite d'un autre.

D'un point de vu fonctionnel, l'information circule toujours dans le même sens : des dendrites vers l'axone. Les dendrites captent les signaux venant d'autres neurones. Le corps cellulaire renvoit un influx par l'intermédiaire de l'axone si la somme des influx qui lui arrivent dépassent un certain seuil via un potentiel d'action (signal électrique). Lorsque le signal arrive aux terminaisons synaptiques, des vésicules synaptiques vont venir fusionner avec la membrane synaptique. Des neurotransmetteurs sont ensuite utilisés pour permettre le passage d'informations d'un neurone à l'autre. Ils excitent ou inhibent le neurone suivant pour propager ou non un nouvel influx nerveux.

Les synapses, en fonction de leur historique, mémorisent les activations répétées ou non entre deux neurones et ajustent leur fonctionnement en conséquence pour faciliter ou non le passage des influx nerveux. Les mécanismes d'apprentissage s'inspirent de ce schéma.

9.3 Neurone formel

Un neurone formel est une modélisation mathématique du neurone biologique. Il est généralement constitué de plusieurs entrées et d'une sortie par analogie aux dendrites et à l'axone. Les grands principes y sont repris, notamment la sommation des entrées. A chacune de ces entrées, est associé un poids qui correspond aux actions excitatrices et inhibitrices des synapses. Une phase d'apprentissage est alors utilisée pour ajuster ces coefficients.

Schématiquement, le neurone calcule la somme pondérée des entrées qu'il perçoit auquel se rajoute un seuil, applique à la valeur obtenue une fonction dite d'« activation » classiquement non linéaire. Puis, la valeur finale est affectée en sortie du neurone.

Considérons le cas d'un neurone formel à m entrées auxquelles sont affectées m grandeurs numériques notées x_1 à x_m, pour définir la règle de calcul.

Un poids synaptique noté w_1 à w_m est associé à chaque entrée, la sommation pondérée est obtenue de la façon suivante :

$$w_1.x_1 + ... + w_m.x_m = \sum_{i=1}^{m} w_i.x_i$$

Le seuil noté w_0 est additionné à cette grandeur, pour ensuite être transformé par la fonction d'activation notée ϕ :

$$\phi\left(w_0 + \sum_{i=1}^{m} w_i.x_i\right)$$

En général, l'expression de la valeur de sortie est simplifiée en ajoutant une entée fictive x_0 fixée à la valeur 1 :

$$\phi\left(\sum_{i=0}^{m} w_i.x_i\right)$$

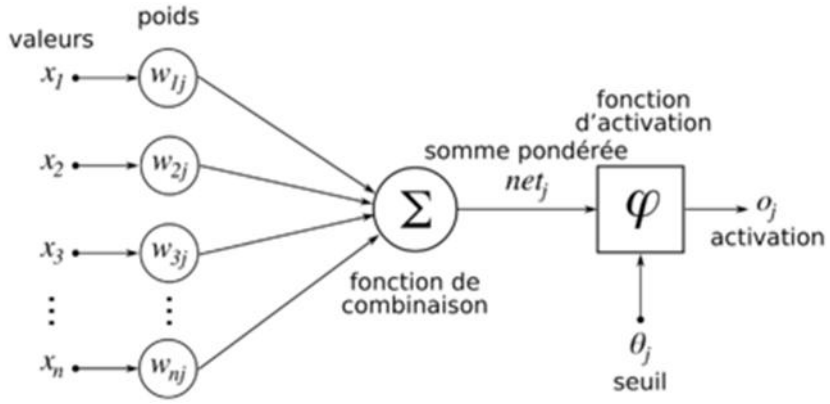


FIG. 9.2 – Neurone formel

9.4 Fonction d´activation

Le choix de la fonction d'activation est important pour obtenir un modèle utile en pratique car il influe sur les propriétés du neurone formel. De préférence strictement croissante et bornée, une fonction d'activation se doit de respecter une certaine forme de régularité et des conditions de dérivabilité.

Les plus fréquemment utilisées sont des « sigmoïdes » ayant une forme de « S » et symétrique par rapport à l'origine comme la fonction logistique, la tangente hyperbolique ou l'arc-tangente. Il existe cependant d'autres fonctions : le pas unitaire, la linéaire seuillée, la gaussienne, l'identité.

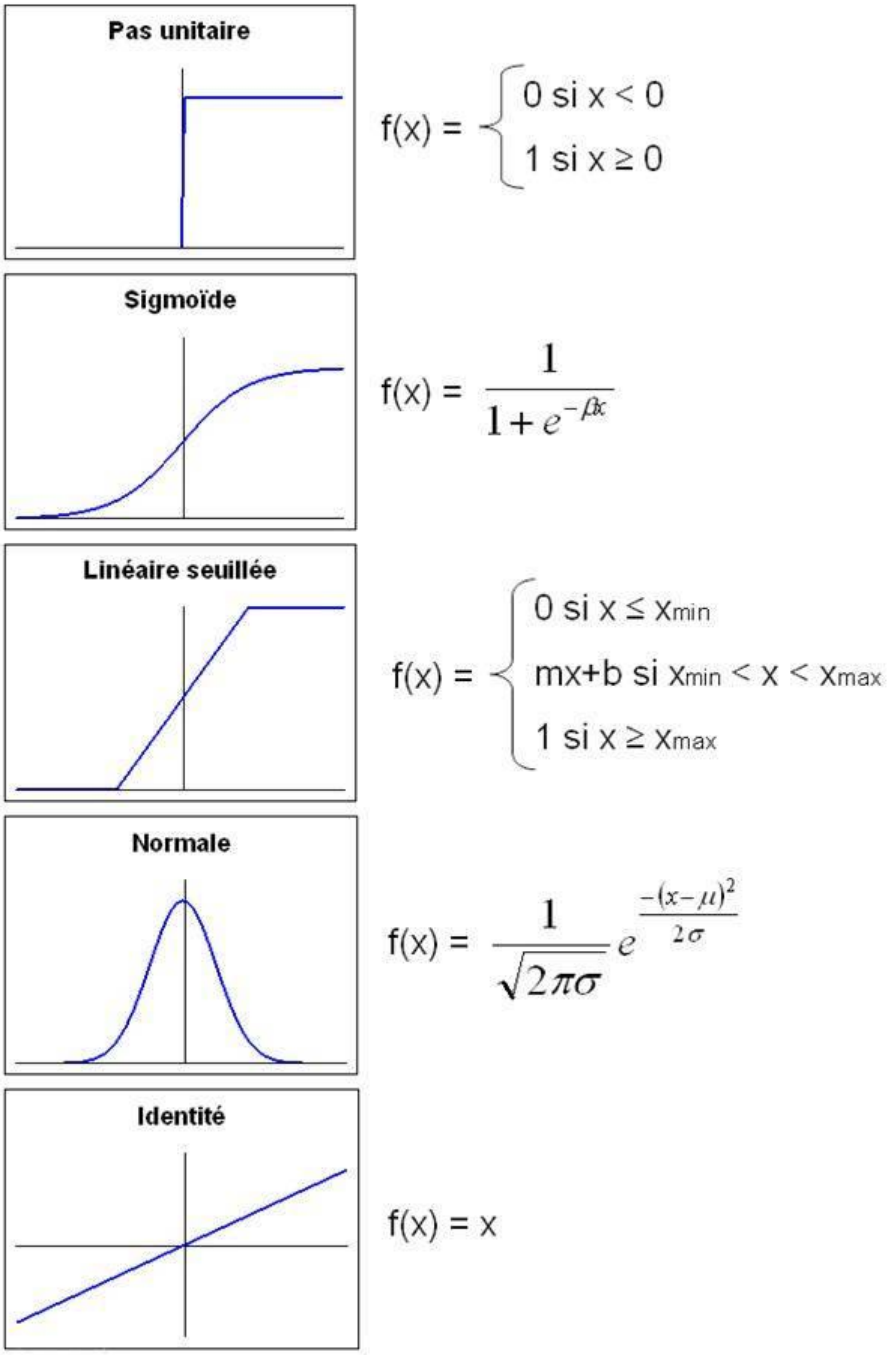


FIG. 9.3 – Fonctions d'activations

9.5 Propagation de l´information

Un réseau de neurones artificiels est composé d'un ensemble de neurones formels qui, associés en couches, fonctionnent en parallèle. Chacune des couches effectue un traitement indépendamment des autres pour transférer le résultat de son analyse à la couche suivante. L'information se propage de couche en couche, de l'entrée vers la sortie. Le nombre de couches cachées est variable selon la nature du modèle.

En fonction du sens et de l'orientation des différentes connexions définissant la structure du réseau, la valeur obtenue en sortie du neurone par la fonction d'activation est alors trans-

mise comme entrée aux neurones suivants qui lui sont reliés. Par ce mécanisme, l'information peut alors circuler de neurone en neurone et de couche en couche au travers du réseau.

Les couches dites d'entrée et de sortie constituent l'interface avec l'extérieur. La couche d'entrée reçoit les variables en entrée du modèle et la couche de sortie renvoie les résultats attendus. Les neurones des autres couches, internes au réseau, sont appelées neurones cachés.

Conventionnellement les neurones d'entrée ont une fonction d'activation identité pour ne pas modifier l'information.

En fonction de l'algorithme d'apprentissage, l'information peut également être propagée en arrière ("*back propagation*"). De manière générale, chaque neurone, à l'exception de ceux des couches d'entrée et de sortie, est connecté à tous les neurones de la couche précédente et de la couche suivante.

Un réseau est caractérisé par trois composantes :
- le type d'interconnexion
- le choix des fonctions d'activation
- le mode d'apprentissage (ou comment estimer les poids)

Les deux premières définissent l'architecture du réseau.

9.6 Architecture réseau

Les réseaux de neurones artificiels peuvent être regroupés en deux grandes catégories : les réseaux dits "non bouclés" (ou statiques ou *Feed-forward*) et ceux dits "bouclés" (ou dynamiques ou récurrents ou *Feed-back*). La différence entre les deux réside dans le mode de propagation de l'information. Dans les réseaux non bouclés, l'information se propage de couche en couche, sans retour en arrière possible, alors que dans ceux bouclés, il y a retour en arrière de l'information.

Les boucles de rétroaction, qui peuvent être mise en place soit au niveau des neurones soit au niveau des couches, permettent de prendre en compte des aspects temporels et de mémorisation. La contrepartie réside dans la difficulté à les mettre en œuvre, notamment en terme de convergence. Dans les réseaux non bouclés, les données peuvent être présentées dans n'importe quel ordre sans qu'il y ait d'influence sur l'évolution des poids lors de la phase d'apprentissage. En revanche, dans le cas d'un réseau bouclé simulant un processus dépendant du temps, l'ordre de présentation du jeu de données au réseau est primordial.

9.6.1 Les réseaux non bouclés

Le Perceptron monocouche

Historiquement, c'est le premier réseau de neurones artificiels : le *Perceptron de Rosenblatt*. Sa conception est relativement simple puisqu'il n'est composé que de deux couches : une entrée et une sortie. Conçu initialement pour la reconnaissance des formes, il est calqué sur le système visuel. Il peut cependant traiter d'autres problématiques comme la classification ou la résolution d'opérations logiques simples telles que le "ET" ou le "OU". Ce modèle a néanmoins une contrainte forte, il résout uniquement des problèmes séparables linéairement.

Classiquement, il suit un apprentissage supervisé selon la règle de correction de l'erreur (ou celle de Hebb).

Le Perceptron multicouches (P.M.C.)

Ce modèle n'est autre qu'une évolution du précédent, intégrant une ou plusieurs couches cachées entre l'entrée et la sortie. Hormis pour les couches d'entrée et de sortie, chaque neurone est connecté avec tous ceux de la couche suivante et tous ceux de la précédente. Par contre, il n'existe pas de liaison entre des neurones issus de la même couche. Les fonctions d'activation principalement utilisées sont des fonctions à seuil ou bien des sigmoïdes. Comparativement au Perceptron monocouche, il a l'avantage de résoudre des problèmes non linéairement séparables ainsi que des problématiques logiques plus complexes comme le "OU" exclusif. La phase d'apprentissage est supervisée et régie par la règle de correction de l'erreur.

Les réseaux à fonction radiale

Appelés aussi *R.B.F.* ("*Radial Basic Functions*"), ces modèles conservent la même architecture que les *P.M.C.* à la variante près que les fonctions d'activation sont des lois normales. Comme les *P.M.C.*, ils sont employés en classification et plus spécifiquement en approximation de fonctions. En terme d'apprentissage, le mode le plus fréquent est hybride avec soit la règle de correction de l'erreur soit celle d'apprentissage par compétition.

9.6.2 Les réseaux bouclés

Les réseaux de neurones compétitifs et cartes auto organisatrices

Ces types de réseaux détectent les corrélations et les régularités présentes dans les données en entrée pour adapter les réponses en conséquence afin d'apprendre à catégoriser des vecteurs d'entrée.

Les réseaux compétitifs ajustent les poids des neurones vainqueurs suivant la loi de Kohonen. Ils possèdent un couche dite compétitive dans laquelle les neurones réagissent différemment aux entrées. Un neurone est élu vainqueur et le gagnant a le droit de modifier ses poids de connexion. L'apprentissage peut être supervisé ou non.

Les cartes auto organisatrices (*Self Organizing Map*, *S.O.M.*), qui appartiennent à la classe des réseaux compétitifs, permettent d'établir une carte discrète ordonnée topologiquement à partir de patterns en entrée. L'architecture du réseau peut s'apparenter à un treillis dont chacun des nœuds correspond à un neurone associé à une vecteur poids. Pour chaque entrée, la correspondance entre chaque vecteur poids est calculée. Le principe est d'effectuer une sélection du vecteur de poids bénéficiant de la meilleure corrélation ainsi que certains de ses voisins. Ce process permet de les modifier pour accroître davantage cette corrélation. Le mode d'apprentissage est non-supervisé.

Les réseaux de Hopfield

La caractéristique de ces réseaux est qu'ils sont récurrents et entièrement connectés. En d'autres termes, chaque neurone est connecté à tous les autres sans faire de distinction entre les neurones d'entrée et de sortie. Ils peuvent être assimilés à une mémoire associative non-linéaire. En fonction des représentations bruitées ou partielles, ils sont capables de trouver un

objet stocké. Ces réseaux trouvent leur utilité principalement dans la résolution de problèmes d'optimisation ainsi que l'entrepôt de connaissances. Le mode d'apprentissage est également non-supervisé.

Les ART

Les réseaux de type *A.R.T.* (*Adaptative Resonance Theorie*) sont caractérisés par un mode d'apprentissage par compétition. Le point crucial est de trouver un bon compromis entre la stabilité et la plasticité. Dans ce mode d'apprentissage, la stabilité des catégories formées n'est pas garantie sauf en limitant la capacité d'apprentissage, ce qui serait au détriment de la plasticité. L'intérêt de cette architecture est d'éviter de se retrouver confronté à ce dilemme grâce à un procédé dit de résonance. Ici, la proximité des entrées avec les prototypes déjà existants déterminera si les vecteurs de poids devront être adaptés ou bien si une nouvelle catégorie doit être créée car trop éloignée des autres.

Récapitulatif

Les architectures réseaux peuvent être résumées dans le classement suivant :

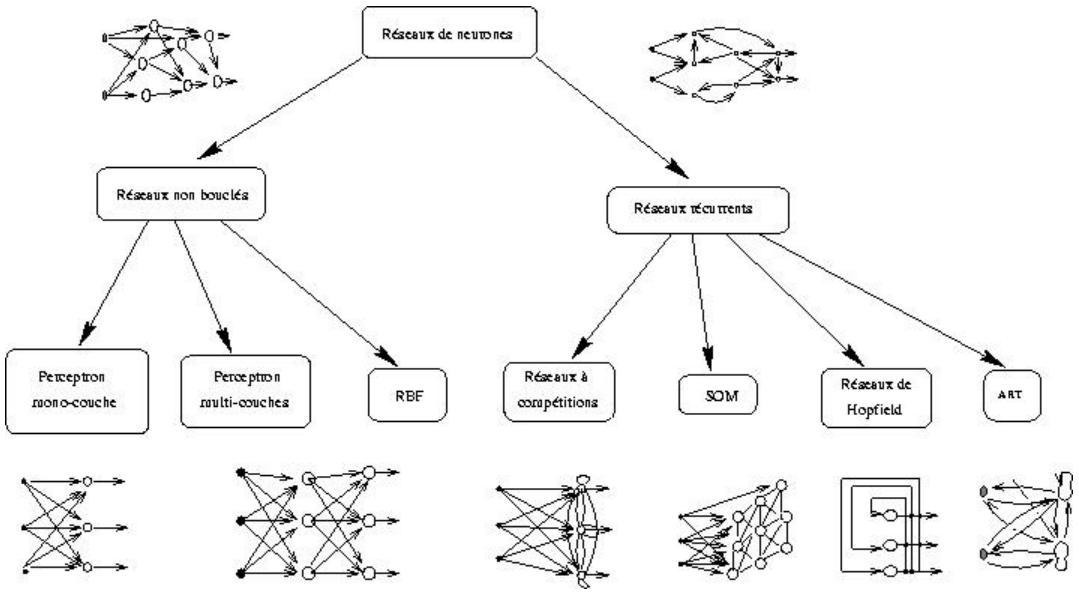


FIG. 9.4 – Architecture

9.7 Les types d'apprentissage

9.7.1 Le mode supervisé

Dans ce mode, les entrées et les sorties du jeu de données sont présentées au réseaux simultanément. Les résultats, calculés par les réseaux en fonction des données fournies en entrée, sont comparés à ceux attendus. Le réseaux va s'adapter progressivement en modifiant ses poids jusqu'à ce que les sorties calculées se rapprochent de celles espérées. Il faut alors définir un critère d'arrêt à partir duquel l'écart entre les deux soit jugé satisfaisant.

9.7.2 Le renforcement

Dans ce modèle parfois considéré comme un apprentissage supervisé, le système a un retour sur ses actions sans avoir accès à des exemples. L'action est évaluée par le biais de l'attribution d'une récompense ou d'une punition. Son objectif est d'apprendre la corrélation entrée/sortie par l'intermédiaire de son erreur. Ce qui revient à étudier le rapport échec/succès en maximisant un indice de performance appelé signal de renforcement. Sans pour autant connaître la bonne réponse, le modèle parvient à déterminer si la réponse calculée est correcte ou non.

9.7.3 Le mode non-supervisé

La spécificité du mode d'apprentissage non supervisé est que seules les entrées sont disponibles. Le réseau se modifie selon les régularités statistiques perçues en entrée pour établir des catégories. Basé sur des probabilités, le principe est d'optimiser une valeur de qualité attribuée aux catégories reconnues.

9.7.4 Le mode hybride

Ce mode est un compromis entre le mode supervisé et non supervisé. Une partie des poids est déterminée par apprentissage supervisé et l'autre par apprentissage non supervisé.

9.8 Les règles d'apprentissage

9.8.1 Règle de correction d'erreur

Cette règle s'inscrit dans le paradigme de l'apprentissage supervisé. L'écart entre la valeur de la sortie calculée et celle espérée est utilisé pour réduire l'erreur globale du système en modifiant les poids.

Règle de Widrow-Hoff

Proposée en 1960, elle repose sur un principe de minimisation de l'erreur quadratique où l'erreur d'une couche est la différence entre la réponse fournie et celle attendue. Les poids sont modifiés itérativement de manière à se rapprocher de la valeur attendue en sortie. Cette règle permet à un neurone de corriger ses différents poids comme suit :

Notons c^s la sortie attendue et o^s la sortie calculée pour l'exemple s.

$$\Delta W_{ij} = \delta(c^s - o^s)x_i^s$$

A noter que les poids sont modifiés après chaque exemple et non pas une fois qu'ils sont tous présentés.

Loi de rétropropagation du gradient de l'erreur

Dans cette règle utilisée pour les réseaux multicouches, l'idée est de minimiser une fonction erreur en appliquant une descente de gradient. Cette méthode est celle que nous avons retenue pour notre sujet, elle fera l'objet d'un chapitre particulier.

Dans sa conception initiale, la rétropropagation de l'erreur (*Backpropagation*) telle que décrite par Rumelhart est une technique efficace permettant de calculer le gradient de l'erreur en partant de la dernière couche du réseau vers la première pour un *Perceptron Multi Couches*.

Par extension, l'appellation courante d'« *algorithme de rétropropagation du gradient de l'erreur* » fait référence à l'algorithme classique de correction des erreurs basé sur le calcul du gradient grâce à la rétropropagation. Il convient cependant de distinguer ces deux notions qui sont la méthode utilisée pour le calcul du gradient de l'erreur et la technique servant à corriger les erreurs.

9.8.2 Règle de Hebb

Cette loi, qui s'inspire de données biologiques, est la suivante :

Définition 9.8.1 (Postulat de Hebb)**.** *"Si des neurones, de part et d'autre d'une synapse, sont activés de manière synchrone et répétée, la force de connexion synaptique va aller croissant."*

A noter que l'apprentissage est localisé et que la modification des poids n'est fonction que de l'activation de deux neurones.

$$\Delta W_{ij} = \delta x_i^s x_j^s$$

δ est une constante positive représentant la force d'apprentissage.

9.8.3 Apprentissage de Boltzmann

Cette règle trouve son utilité dans les réseaux de Boltzmann. Ces réseaux symétriques et récurrents, se composent de deux sous-groupes de neurones : des cellules visibles liées à l'environnement et d'autres cachées qui ne le sont pas. Les poids des connexions sont ajustés afin que les cellules visibles satisfassent une distribution probabiliste souhaitée. C'est une règle stochastique.

9.8.4 Apprentissage par compétitions

L'objectif de cette règle est de catégoriser des données. En fonction des corrélations des données, les neurones sont mis en compétition et le plus représentatif d'une classe devient son représentant. Seuls les poids de connexions du neurone gagnant peuvent alors être ajustés.

9.9 Domaines d´application

Les secteurs d'activité dans lesquels ils peuvent être appliqués sont assez vastes : l'aérospatial, l'automobile, la défense, l'électronique, la finance, le secteur médical ou bien encore les télécommunications...

Les réseaux de neurones peuvent être employés pour de nombreuses problématiques :

- En classification, pour répartir de objets en plusieurs classes, pour obtenir une information qualitative à partir de données quantitatives ou encore en reconnaissance de formes,

- En recherche opérationnelle, afin de résoudre des problèmes dont la solution est inconnue,

- En mémoire associative, afin de restituer à partir d'informations incomplètes ou bruitées une donnée.

A titre d'exemple, voici deux applications concrètes de ces modèles :

- Premier exemple, dans un centre de tri postal, les codes postaux peuvent être reconnus à partir d'une modélisation par apprentissage statistique. Chaque chiffre doit être catégorisé parmi 11 classes : les dix chiffres auxquels s'ajoute une autre classe pour les chiffres trop mal écrits et devant être traités manuellement.

- Deuxième exemple, dans un service technique d'E.D.F., la demande d'électricité future peut être prédite au moyen de l'extrapolation des valeurs de consommation antérieures. La prédiction est basée sur les valeurs de consommation passées ainsi que des variables exogènes comme la température, le degré de nébulosité, la vitesse du vent, le degré d'humidité, le niveau de pluie, la saison, le fait que ce soit un jour ouvrable ou férié, le jour de la semaine, une période spéciale, ...

Voici également les noms de quelques entreprises faisant appel à cette technique : EasyReader, Mimetics, Thomson, Silac, Renault, Canon, VLSI, IBM, JVC, Lyonnaise, CGE, EDF, Cofiroute, Air Liquide, Elf, Atochem, Lafarge, Sagem,...

9.10 Spécificités du modèle

Les grands atouts de ce modèle sont principalement sa capacité de généralisation, sa bonne résistance à l'imprécision des données ainsi que sa capacité à être mis en œuvre sur de grandes bases. A l'inverse, ce modèle ne fournit aucune explication sur l'obtention du résultat (aspect "boite noire"), le paramétrage est parfois délicat (nombre de neurones dans la couche cachée), un fichier de validation doit être utilisé pour éviter un sur-apprentissage.

Chapitre 10

L´algorithme de rétropropagation du gradient de l´erreur

10.1 Méthode du gradient

La méthode du gradient a pour objectif de trouver le minimum d'une fonction notée f d'une variable réelle à valeurs réelles, remplissant des conditions de dérivabilité. Le principe est de construire une suite notée x_n de telle sorte à ce qu'elle converge vers ce minimum.

Cette méthode utilise la direction de descente de la plus grande pente, c'est à dire la direction dans laquelle f décroît le plus vite.

En partant d'une valeur choisie arbitrairement notée x_0, la relation de récurrence est définie comme suit :

$$\forall n > 0, x_{n+1} = x_n + \Delta x_n$$

avec $\Delta x_n = \varepsilon f'(x_n)$ où ε est une valeur correctement sélectionnée (si ε est trop petit, le nombre d'itérations peu être élevé, s'il est trop grand la suite de valeurs peut osciller autour du minimum sans converger).

En prenant une valeur de $\varepsilon f'(x)$ suffisamment petite, le théorème des approximations finies donne la relation suivante :

$$f(x_{n+1}) = f(x_n + \varepsilon f'(x_n)) \approx f(x_n) - \varepsilon(f'(x_n))^2$$

Sous réserve que l'approximation faite précédemment ne s'écarte pas trop de la réalité, la fonction $f(x_{n+1})$ est inférieure à $f(x_n)$, ce qui prouve la décroissance de celle-ci.

Plus la pente est grande, plus l'écart entre x_n et x_{n+1} est important. Le critère d'arrêt généralement retenu est une pente suffisamment faible.

Les inconvénients majeurs de cette méthode sont le choix empirique de ε et le fait que le minimum obtenu ne soit pas systématiquement global.

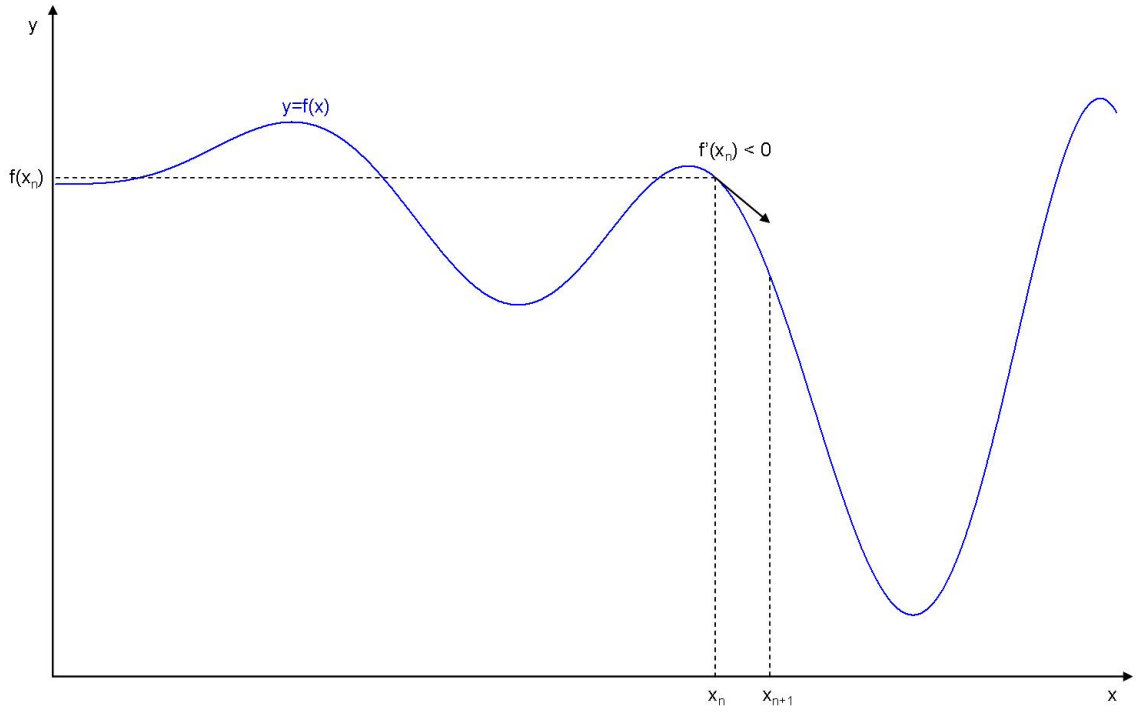


FIG. 10.1 – Méthode du gradient

10.2 Définition du Perceptron Multi-Couches

Définissons la fonction d'activation comme étant une sigmoïde de la manière suivante :

$$\phi(x) = \frac{1}{1+e^{-x}}$$

Celle-ci est une approximation indéfiniment dérivable de la fonction à seuil de Heaviside dont la dérivée est relativement simple à calculer :

$$\phi(x)' = \frac{e^x}{(1+e^x)^2} = \phi(x)(1-\phi(x))$$

Dans cette section, la difficulté majeure réside dans la complexité des notations que nous définirons de la manière suivante :

- l'échantillon d'apprentissage comprend S exemples
- chaque exemple d'apprentissage est noté par un indice s
- chaque cellule est définie par un indice noté i
- le réseau comprend p cellules de sortie
- la sortie attendue pour une cellule de sortie est notée c_i
- le poids synaptique de la cellule j vers la cellule i est noté w_{ij}
- l'ensemble des cellules en entrée de la cellule i est noté $Pred(i)$ (prédécesseur)
- l'ensemble des cellules en sortie de la cellule i est noté $Succ(i)$ (successeur)
- l'entrée totale de la cellule i est notée y_i de telle sorte que : $y_i = \sum_{j \in Pred(i)} w_{ij} x_{ij}$
- la sortie de la cellule i est notée o_i de telle sorte que : $o_i = \phi(y_i)$
- l'erreur d'apprentissage est notée E
- le nombre d'entrées est noté n
- le nombre de sorties est noté p
- le vecteur des poids synaptiques associé à tous les liens du réseau est noté $\overrightarrow{w}$

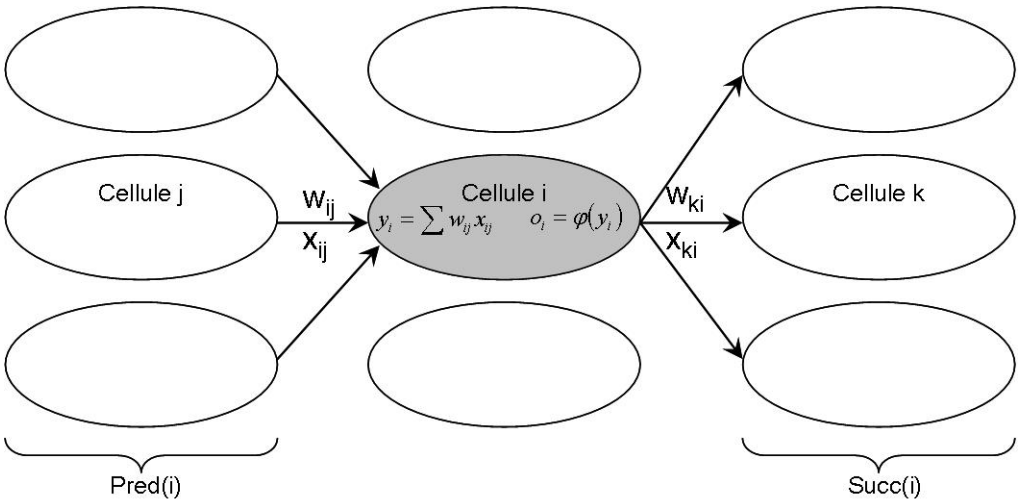


FIG. 10.2 – Notations utilisées

10.3 Justification de l'algorithme

Cet algorithme itératif a pour objectif de faire converger le modèle vers un minimum d'erreur (appelée aussi fonction coût) entre les valeur attendues et celles calculées. Pour chaque neurone du réseau, le gradient de l'erreur est calculé afin de mesurer la contribution de chacun des poids synaptiques à l'erreur commise. Par ce biais, les corrections sont effectuées au fur et à mesure que les exemples d'apprentissage sont présentés. Comme la modification d'un poids influe sur tous ceux des neurones des couches suivantes, les corrections d'erreur doivent être propagées de la dernière couche vers la première.

Sur un échantillon complet d'apprentissage de S exemples, l'erreur du PMC est définie comme suit :

$$E(\overrightarrow{w}) = \sum_{s \in S} \frac{1}{2} \sum_{k=1}^{p} (c_k^s - o_k^s)^2$$

En se restreignant à un seul exemple s, l'erreur devient :

$$E_s(\overrightarrow{w}) = \frac{1}{2} \sum_{k=1}^{p} (c_k - o_k)^2$$

Comme pour le perceptron avec la règle du Widrow-Hoff, la procédure à suivre est de minimiser l'erreur sur chaque exemple présenté et non pas directement l'erreur globale sur l'échantillon complet.

Appliquons ensuite la méthode du gradient en calculant les dérivées partielles par rapport aux poids synaptiques de la fonction erreur sur un exemple, que nous noterons $\frac{\partial E}{\partial w_{ij}} = \frac{\partial E_s(\overrightarrow{w})}{\partial w_{ij}}$:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} x_{ij}$$

Cette décomposition de la dérivée partielle en deux termes est permise puisque les poids

synaptiques n'ont d'impact sur la sortie du réseau qu'au travers de l'entrée totale de la cellule.

Il reste à calculer $\frac{\partial E}{\partial y_i}$, deux cas sont à distinguer : celui où la cellule i est une cellule de sortie ou bien interne.

Traitons dans un premier temps le cas où la cellule est de sortie :

Cellule de sortie

Comme y_i n'impacte la sortie du réseau que par l'intermédiaire de o_i, la règle de chaînage des dérivées partielles nous permet d'écrire la décomposition suivante :

$$\frac{\partial E}{\partial y_i} = \underbrace{\frac{\partial E}{\partial o_i}}_{(1)} \times \underbrace{\frac{\partial o_i}{\partial y_i}}_{(2)}$$

La première dérivée peut s'écrire de la manière suivante :

$$(1) = \frac{\partial E}{\partial o_i} = \frac{\partial \frac{1}{2} \sum_{k=1}^{p} (c_k - o_k)^2}{\partial o_i}$$

Tous les termes ont une dérivée nulle à l'exception du i^{me} terme :

$$(1) = \frac{\partial E}{\partial o_i} = \frac{\partial \frac{1}{2}(c_i - o_i)^2}{\partial o_i} = -(c_i - o_i)$$

La deuxième dérivée s'obtient en partant de la définition d'une cellule élémentaire $o_i = \sigma(y_i)$ et en dérivant la fonction sigmoïde :

$$(2) = \frac{\partial o_i}{\partial y_i} = \frac{\partial \sigma(y_i)}{\partial y_i} = \sigma(y_i)(1 - \sigma(y_i)) = o_i(1 - o_i)$$

La formule de départ devient après remplacement des deux dérivées par leurs expressions :

$$\frac{\partial E}{\partial y_i} = (1) \times (2) = -(c_i - o_i)o_i(1 - o_i)$$

d'où :

$$\frac{\partial E}{\partial w_{ij}} = -(c_i - o_i)o_i(1 - o_i) \times x_{ij}$$

La modification des poids synaptiques s'effectue ensuite en appliquant la méthode du gradient :

$$\Delta w_{ij} = -\varepsilon \frac{\partial E}{\partial w_{ij}} = \varepsilon \delta_i x_{ij}$$

avec :

$$\delta_i = o_i(1 - o_i)(c_i - o_i)$$

Passons au cas où i est une cellule interne :

Cellule interne

Contrairement aux cellules de sortie, les variations de la quantité y_i se répercutent dans tous les calculs issus des cellules suivantes :

$$\frac{\partial E}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial o_i} \frac{\partial o_i}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \times w_{ki} \times o_i(1 - o_i)$$

En réorganisant la formule :

$$\frac{\partial E}{\partial y_i} = o_i(1 - o_i) \times \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \times w_{ki}$$

d'où :

$$\frac{\partial E}{\partial w_{ij}} = o_i(1 - o_i) \times \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} \times w_{ki} \times x_{ij}$$

La modification des poids synaptiques s'effectue ensuite en appliquant la méthode du gradient :

$$\Delta w_{ij} = -\varepsilon \frac{\partial E}{\partial w_{ij}} = \varepsilon \delta_i x_{ij}$$

avec :

$$\delta_i = o_i(1 - o_i) \times \sum_{k \in Succ(i)} \delta_k w_{ki}$$

Synthèse

En résumé, nous avons calculé l'expression des dérivées partielles $\frac{E(\vec{w})}{\partial w_{ij}}$ quelle que soit la position de la cellule i, en commençant par les cellules de sorties, puis cachées pour finir par celles d'entrées.

La modification des poids synaptiques d'effectue ensuite en appliquant la méthode du gradient :

$$\boxed{\Delta w_{ij} = -\varepsilon \frac{\partial E}{\partial w_{ij}} = \varepsilon \delta_i x_{ij}}$$

Dans le cas d'une cellule de sortie :

$$\boxed{\delta_i = o_i(1 - o_i)(c_i - o_i)}$$

Dans le cas d'une cellule interne :

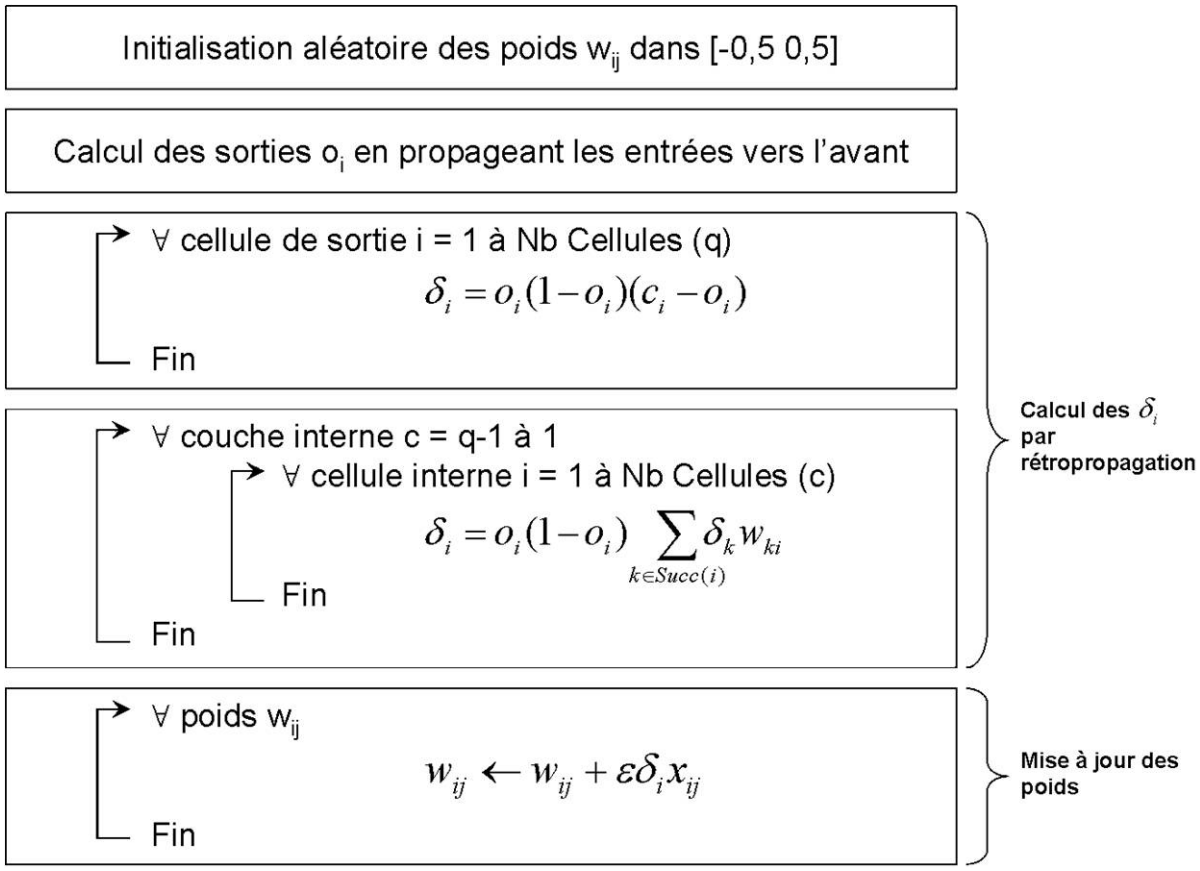
$$\boxed{\delta_i = o_i(1 - o_i) \times \sum_{k \in Succ(i)} \delta_k w_{ki}}$$

Nous disposons maintenant de tous les éléments pour mettre en place l'algorithme de rétro-propagation du gradient de l'erreur.

10.4 L'algorithme de rétropropagation du gradient l'erreur

Soit en entrée un échantillon S et un PMC de couche d'entrée C_0, de couches $q-1$ cachées $C_1, ..., C_{q-1}$ et de couche de sortie C_q.

L'algorithme de rétropropagation du gradient de l'erreur peut s'énoncer de la façon suivante :

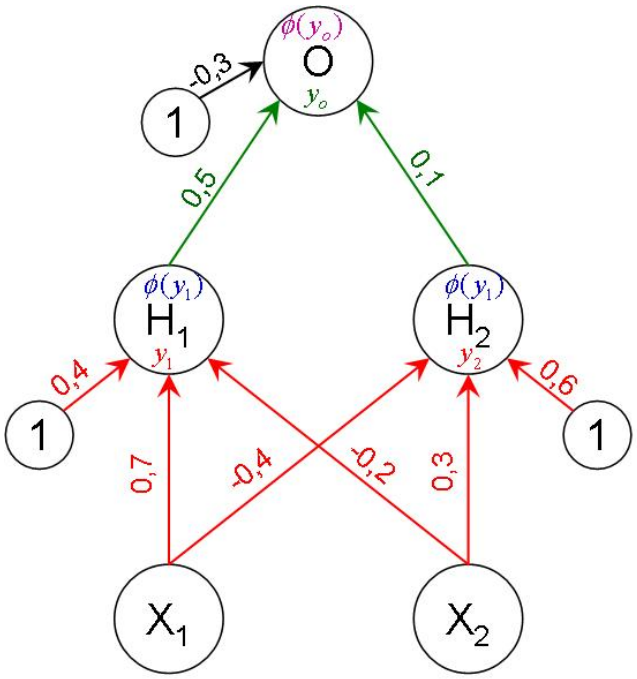


Les étapes de calcul de δ_i par rétropropagation et de mise à jour des poids sont à répéter jusqu'à ce que le critère d'arrêt soit satisfait. Celui-ci peut être un nombre maximum d'itérations, le moment où la racine de l'Erreur Quadratique Moyenne (E.Q.M.) devient inférieure à un certain seuil ou bien encore la méthode dite d'« arrêt prématuré »(early stopping). Cette dernière méthode permet d'éviter le *sur-apprentissage* en arrêtant l'algorithme avant d'avoir atteint le minimum. Un ensemble d'exemples dit de validation sert à stopper l'algorithme lorsque l'erreur remonte trop sur cet ensemble.

Cet algorithme peut être conçu de deux manières. La version "séquentielle" ou "on-line" consiste à actualiser les poids à chaque présentation d'exemple au modèle. Dans la version dite "batch", les erreurs sont calculées pour tous les échantillons sans modifier les poids, les erreurs sont uniquement additionnées, les poids ne sont ensuite mis à jour qu'après le passage des toutes les données dans le réseau. La méthode batch est généralement plus rapide mais peut être contraignante en terme de stockage si le nombre d'exemples est important. La version séquentielle, quant à elle, peut être confrontée à des problèmes de convergence.

10.5 Exemple

Voici un exemple issu de l'ouvrage de Larene Fausett [FAU94] dans lequel l'échantillon d'apprentissage vaut : $(X_1 = 0, X_2 = 1, Y = 1)$ et $\varepsilon = 0,25$. Le réseau comporte une couche d'entrée composée de deux neurones, une couche cachée composée également de deux neurones et une couche de sortie comprenant un seul neurone.



Entrée de la couche cachée :

$$y_1 = 0,4 + 0 \times 0,7 + 1 \times -0,2 = 0,2$$

$$y_2 = 0,6 + 1 \times -0,4 + 1 \times 0,3 = 0,9$$

Sortie de la couche cachée :

$$\phi(y_1) = \frac{1}{1 + e^{-0,2}} = 0,550$$

$$\phi(y_2) = \frac{1}{1 + e^{-0,9}} = 0,711$$

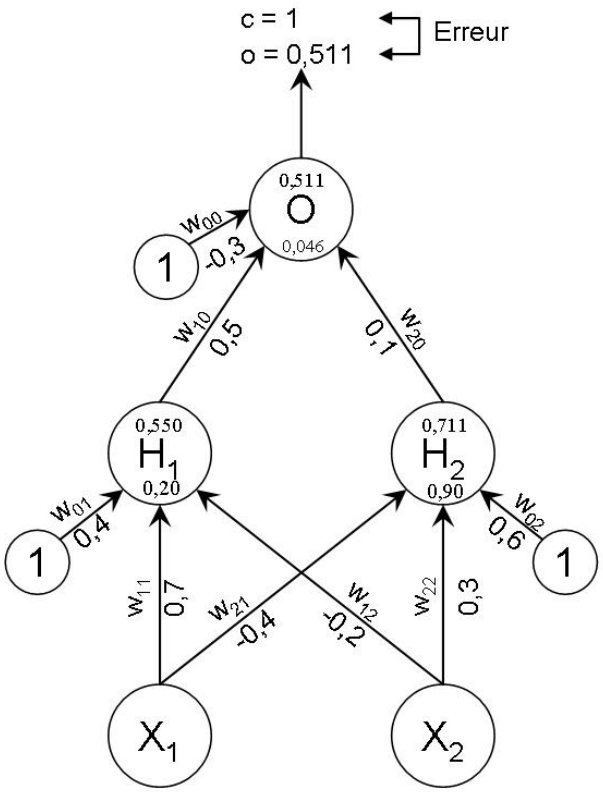
Entrée de la couche de sortie :

$$y_o = -0,3 + 0,550 \times 0,5 + 0,711 \times 0,1 = 0,046$$

Sortie de la couche de sortie :

$$\phi(y_0) = \frac{1}{1 + e^{-0,046}} = 0,511$$

FIG. 10.3 – Exemple : Propagation avant



Erreur :

$$c - o = 1 - 0{,}511 = 0{,}488$$

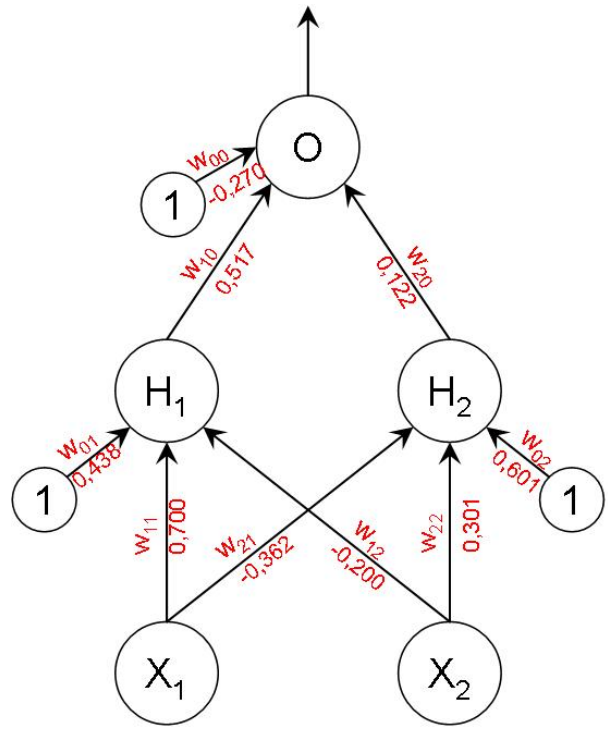
Calcul des δ :

$$\delta_k = o(1-o)(c-o)$$
$$\delta_k = 0{,}511(1-0{,}511)(1-0{,}511)$$
$$\delta_k = 0{,}122$$

$$\delta_{j_1} = o_1(1-o_1)\delta_k w_1$$
$$\delta_{j_1} = 0{,}550 \times (1-0{,}550) \times 0{,}122 \times 0{,}5$$
$$\delta_{j_1} = 0{,}015$$

$$\delta_{j_2} = o_2(1-o_2)\delta_k w_2$$
$$\delta_{j_2} = 0{,}711 \times (1-0{,}711) \times 0{,}122 \times 0{,}1$$
$$\delta_{j_2} = 0{,}0025$$

FIG. 10.4 – Exemple : Calcul des delta



Calcul des Δw_{ij} :

$$\Delta w_{00} = \varepsilon\delta_k = 0{,}25 \times 0{,}122 = 0{,}0305$$
$$\Delta w_{10} = \varepsilon\delta_k o_1 = 0{,}25 \times 0{,}122 \times 0{,}550 = 0{,}0168$$
$$\Delta w_{20} = \varepsilon\delta_k o_2 = 0{,}25 \times 0{,}122 \times 0{,}711 = 0{,}0217$$

$$\Delta w_{01} = \varepsilon\delta_{j_1} = 0{,}25 \times 0{,}015 = 0{,}038$$
$$\Delta w_{11} = \varepsilon\delta_{j_1} x_1 = 0{,}25 \times 0{,}015 \times 0 = 0$$
$$\Delta w_{21} = \varepsilon\delta_{j_1} x_2 = 0{,}25 \times 0{,}015 \times 1 = 0{,}038$$

$$\Delta w_{02} = \varepsilon\delta_{j_2} = 0{,}25 \times 0{,}0025 = 0{,}0006$$
$$\Delta w_{12} = \varepsilon\delta_{j_2} x_1 = 0{,}25 \times 0{,}0025 \times 0 = 0$$
$$\Delta w_{22} = \varepsilon\delta_{j_2} x_2 = 0{,}25 \times 0{,}0025 \times 1 = 0{,}0006$$

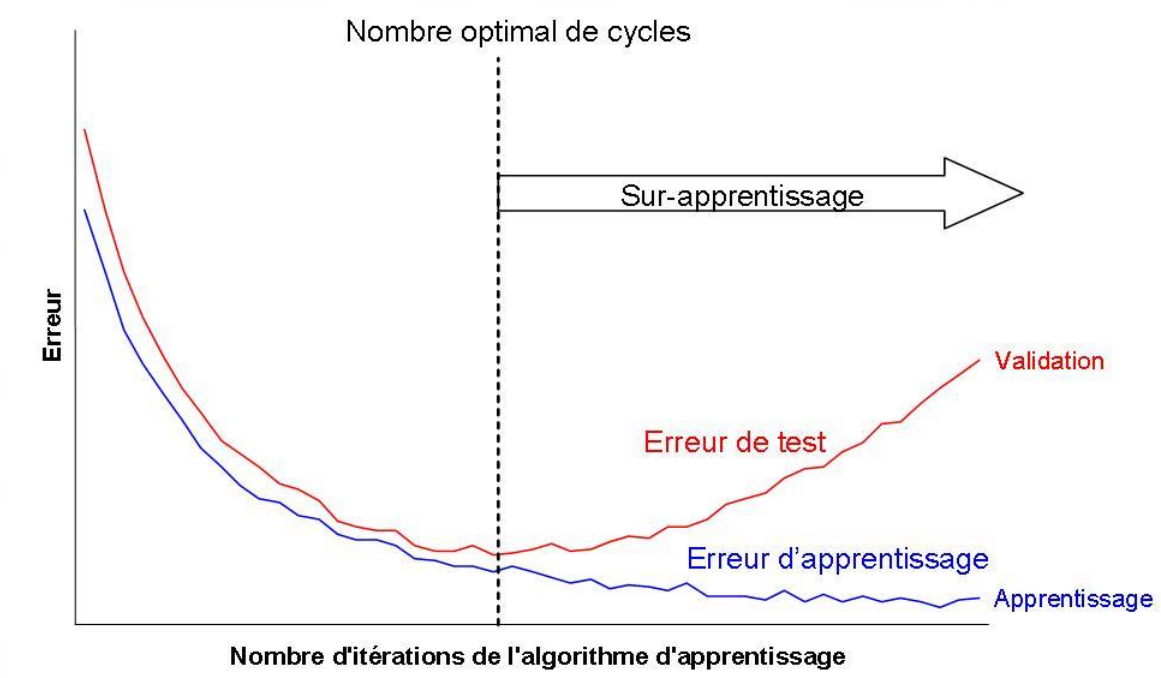
FIG. 10.5 – Exemple : Modification des poids

10.6 Early stopping

Cette technique permet d'anticiper le phénomène de sur-apprentissage en arrêtant prématurément la présentation des exemples. Un premier cycle est réalisé en présentant aléatoirement tous les exemples de la base d'apprentissage. Un second est réalisé sur la base de validation. Le processus s'arrête dès lors que l'erreur simulée sur la base de validation a atteint son minimum.

Soient le nombre d'exemples de la base d'apprentissage N, la k-ième valeur attendue y_k et la k-ième valeur prédite $f(x_k)$, l'Erreur Quadratique Moyenne (E.Q.M.) peut être définie comme suit :

$$EQM(i) = \frac{1}{N}\sum_{k=1}^{N}(y_k - g(x_k))^2$$



Une bonne pratique consiste à séparer la base de données en trois parties, à condition bien sûr que la taille le permette :

- une base d'entraînement
- une base de validation
- une base de test

Le base d'entraînement sert à l'apprentissage des coefficients, celle de validation est utilisée pour déterminer le nombre d'itérations optimal et celle de test permet de vérifier les performances du réseau ainsi que sa capacité de généralisation. Une répartition entre ces trois bases peut être 60% pour l'apprentissage, 20% pour la validation et 20% pour les tests.

Chapitre 11

Propriétés des réseaux de neurones

11.1 Propriété d´approximation universelle

Les approximateurs conventionnels servent à modéliser une fonction non linéaire inconnue à partir des entrées/sorties, qui elles, sont connues.

La base fondamentale des réseaux de neurones repose sur le théorème de superposition de Kolmogorov [BOR07] que nous nous contenterons d'énoncer sans démonstration :

Théorème 11.1.1 (Théorème de superposition de Kolmogorov)**.** *Une fonction continue, non linéaire, réelle $y = f(x)$ à n variables $x = (x_1, x_2, ..., x_n)$ peut être approximée par une somme de $2n + 1$ fonctions continues d'une seule variable :*

$$\widehat{y} = F(x_1, x_2, ..., x_n) = \sum_{i=1}^{2n+1} g_i(z_i) = g_1(z_1) + g_2(z_2) + ... + g_{2n+1}(z_{2n+1})$$

où g_i est une fonction dépendant d'une unique variable z_i telle que :

$$z_i = \sum_{j=1}^{n} h_{ij}(x_j) = h_{1i}(x_1) + h_{2i}(x_2) + ... + h_{ni}(x_n)$$

Il en résulte que :

$$\begin{aligned} g_1(z_1) &= g_1(h_{11}(x_1) + h_{21}(x_2) + ... + h_{n1}(x_n)) \\ g_2(z_2) &= g_2(h_{12}(x_1) + h_{22}(x_2) + ... + h_{n2}(x_n)) \\ &\vdots \\ g_{2n+1}(z_{2n+1}) &= g_{2n+1}(h_{1,2n+1}(x_1) + h_{2,2n+1}(x_2) + ... + h_{n,2n+1}(x_n)) \end{aligned}$$

En condensant l'écriture, le théorème de Kolmogorov permet d'affirmer que la fonction $f(x_1, x_2, ..., x_n)$ peut être décrite par l'approximateur non linéaire suivant :

$$\widehat{y} = F(x_1, x_2, ..., x_n) = \sum_{i=1}^{2n+1} g_i \left(\sum_{j=1}^{n} h_{ji}(x_j) \right)$$

Une variante de ce théorème introduite par Sprecher consiste à remplacer h_{ji} par $\lambda_i h_i$ avec λ_i constantes et h_i des fonctions continues strictement croissantes :

$$\widehat{y} = F(x_1, x_2, ..., x_n) = \sum_{i=1}^{2n+1} g_i \left(\sum_{j=1}^{n} h_{ji}(x_j) \right) = \sum_{i=1}^{2n+1} g_i \left(\lambda_i \sum_{j=1}^{n} h_j(x_j) \right)$$

En développant la théorie de Kolmogorov et en s'appuyant sur des approximations polynomiales, Hornik [HOR89] et Funahashi [FUN89] ou encore Cybenko [CYB89] ont démontré en 1989 que les réseaux de neurones étaient une classe d'approximateurs universels. En d'autres termes, un Perceptron Multi-Couches doté d'une unique couche cachée, avec un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision voulue.

L'aboutissement de ces travaux conduit au théorème d'approximation universelle applicable aux réseaux de neurones qui s'énonce comme suit :

Théorème 11.1.2 (Théorème d'approximation universelle)**.** *Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire.*

Cette propriété est un théorème d'existence et ne donne pas d'indication sur le nombre optimal de neurones dans la couche cachée pour un type de fonction donné. Elle justifiera par la suite notre choix d'un réseau de neurones à une seule couche cachée.

Pour approfondir davantage les différentes approches de démonstration de cette propriété (réseau infini ou approximation), vous pouvez vous référer à l'article de P. Common sur la classification supervisée par réseaux multicouches [COM91].

11.2 Propriété d´approximation parcimonieuse

Une autre propriété importante, spécifique aux réseaux de neurones, qui fait tout leur intérêt est donnée par le théorème suivant :

Théorème 11.2.1 (Théorème d'approximation parcimonieuse)**.** *Si le résultat de l'approximation est une fonction non linéaire des paramètres ajustables, elle est plus parcimonieuse que si elle était une fonction linéaire de ses paramètres. De plus, pour des réseaux de neurones à fonction d'activation sigmoïdale, l'erreur commise dans l'approximation varie comme l'inverse de neurones cachés, et elle est indépendante du nombre de variables de la fonction à approcher. Par conséquent, pour une précision fixée (un nombre de neurones donnés), le nombre de paramètres du réseau est proportionnel au nombre de variables de la fonction à approcher.*

Les réseaux à une couche cachée forment une famille d'approximateurs parcimonieux, ce qui signifie qu'à nombre égal de paramètres, il est possible d'approximer correctement davantage de fonctions qu'avec des polynômes. Le nombre de paramètres croît linéairement avec le nombre de variables et non pas exponentiellement avec des approximateurs linéaires.

L'intérêt de cette propriété est de limiter le nombre d'exemples nécessaires pour arriver à une estimation satisfaisante de la fonction de régression.

En général, l'approximation se fait à partir d'un nombre fini de données observées expérimentalement qui sont le plus souvent entachées de bruit. La propriété fondamentale de parcimonie des réseaux de neurones leur procure une flexibilité, une bonne capacité de généralisation ainsi qu'une robustesse à des données dégradées par le bruit.

Dans notre cas, le comportement de consommation médicale d'individus ne peut être résumé en un nombre fini de paramètres ce qui lui confère un caractère aléatoire assimilable à un bruit. Pour cette raison, la vertu d'approximation parcimonieuse des réseaux de neurones a orienté notre choix vers ce modèle. Une fois les fonctions d'activation choisies, le seul paramètre à moduler sera le nombre de neurones cachés dans l'architecture du réseau. Celui-ci permet de donner plus ou moins de souplesse afin de gérer au mieux l'aléa contenu dans les données pour ajuster le modèle.

11.3 Introduction à la théorie statistique d'apprentissage de Vapnik

11.3.1 Minimisation du risque empirique

L'objectif de l'apprentissage est de trouver le paramètre θ minimisant le taux d'erreur moyen sur l'ensemble des formes possibles appelé risque réel. Dans le cas de la régression, il est peut être défini comme suit :

$$R = \int (y - f(x, \theta))^2 \, dP(x, y)$$

où $f(x, \theta)$ est la valeur prédite, y est la, valeur attendue, et dP la loi de probabilité jointe.

Puisque la loi de probabilité P est inconnue, le risque réel ne peut être obtenu. L'idée est de mesurer, à la place du risque réel, le risque empirique sur l'échantillon :

$$R_{Emp} = \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2$$

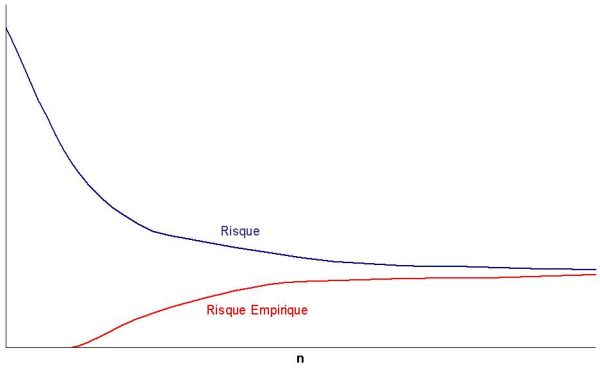


FIG. 11.1 – Risque et risque empirique

En moyenne, le risque est supérieur au risque empirique. Moins il y a d'exemples, plus le modèle arrive facilement à les approcher, ce qui se traduit par un risque empirique faible voire nul au début (surparamétrage). Par contre, le même modèle n'a que très peu de chance d'approximer correctement de nouvelles données d'où un risque réel fort au début. Ces deux risques sont censés converger, quand la taille de l'échantillon augmente, vers des limites assez proches si le processus d'apprentissage est bien modélisé.

11.3.2 Dilemme biais / variance

L'erreur de généralisation se définit comme étant l'erreur commise en appliquant le modèle sur un nouveau jeux de données.

Soient la prédiction g au point x paramétrée par w notée $g(x, w)$ et $G(x, W)$ la variable aléatoire correspondante. L'erreur de prédiction théorique notée P^2 estimant l'erreur de généralisation en un point x est donnée est la relation suivante :

$$P^2 = \underbrace{\sigma^2}_{bruit} + \underbrace{var\left(G(x,W)\right)}_{variance} + \underbrace{\left(E\left(f(x) - G(x,W)\right)\right)^2}_{biais}$$

Le biais et la variance varient en sens inverse de la complexité du modèle. Or l'erreur de généralisation fait intervenir la somme de ces deux termes. La conception du modèle est alors confrontée au dilemme biais / variance.

De manière générale, plus un modèle est complexe, plus le biais est faible, la contrepartie est l'augmentation de la variance. Le biais caractérise l'ajustement du modèle sur les données d'apprentissage qui s'améliore avec la complexité. La variance représente la variabilité en appliquant le modèle à de nouvelles données. Il est donc important de trouver un bon compromis entre ajustement et robustesse. Plus la taille de l'échantillon est importante, plus les procédures envisagées peuvent être complexes.

11.3.3 Consistance

La théorie statistique de l'apprentissage développée par Vapnik apporte des éléments de reponse concernant l'erreur de généralisation en vérifiant le bon comportement de prédiction d'un modèle sur de nouvelles données. Au travers de la minimisation du risque empirique, la question se pose de la convergence vers une même limite du risque et du risque empirique. Cette notion définit alors la consistance d'un modèle en représentant sa capacité à généraliser correctement sur de nouvelles données. Un modèle est dit consistant si l'erreur obtenue sur des données d'apprentissage converge vers l'erreur de généralisation lorsque la taille de l'échantillon d'apprentissage augmente.

Dans le cas des PMC, la vitesse de convergence est liée avec le nombre de neurones cachés. L'objectif est d'établir une relation entre le nombre de neurones cachés et nombre d'observations pour que la consistance soit vérifiée.

11.3.4 Dimension de Vapnik-Chervonenkis

La dimension de Vapnik-Chervonenkis est un outil mathématique mesurant la complexité d'un modèle, elle permet notamment de relier le risque au risque empirique mesuré sur l'ensemble d'apprentissage.

Soit un échantillon $(x_1, ..., x_n) \in \mathbf{R}^m$, il peut être séparé en deux sous-échantillons de 2^n façons différentes.

Définition 11.3.1 (Pulvérisation). *Soit un ensemble F de fonctions $f :\rightarrow \mathbf{R}^m \rightarrow \{-1, 1\}$, il pulvérise $(x_1, ..., x_n)$ si toutes les 2^n séparations peuvent être construites avec des représentants de F.*

Définition 11.3.2 (Capacité). *La capacité d'une famille de fonction est définie comme le plus grand ensemble de points pouvant être pulvérisés (dans le cas des réseaux de neurones, une famille est caractérisée par l'ensemble des différents réseaux obtenus en faisant varier les poids d'une même architecture).*

Dans le cas de classifications binaires, la dimension de Vapnik-Chervonenkis (appelée VC-dimension) correspond à la capacité d'une famille de fonctions. A titre d'exemple, la VC-dimension des hyperplans de $\mathbf{R}^m$ est $m + 1$. La VC-dimension de $\mathbf{R}^2$ est 3 car 4 points ne peuvent pas forcément être séparés par une droite.

11.3.5 Théorème de Vapnik-Chervonenkis

La force de la théorie de Vapnik est de parvenir à majorer le risque par la somme du risque empirique mesuré sur l'échantillon d'apprentissage et d'une somme déterministe.

Ce théorème majeur s'énonce comme suit :

Théorème 11.3.1 (Vapnik-Chervonenkis). *Si la VC-dimension notée d_{vc} est finie, alors $\forall f \in F$, avec une probabilité au moins égale à $1 - \delta$, pour $n > d_{vc}$*

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{d_{vc}\left(log(\frac{2n}{d_{vc}}) + 1\right) - log\frac{\delta}{4}}{n}}$$

Une condition de consistance des PMC à une seule couche cachée a pu être formulée à partir de ce théorème grâce aux travaux de Karpinski et Macintyre en 1995 [KAR95] en bornant la VC-dimension des PMC à fonctions sigmoïdes en $O(W^4)$.

Il suffit pour cela que le nombre de cellules cachées notées M_n satisfasse la relation suivante :

$$M_n < O\left(\sqrt[4]{(n)}\right)$$

Ces bornes théoriques sont souvent trop larges pour être réellement applicables en pratique.

Elles permettent néanmoins de montrer que l'erreur d'un modèle appliqué à de nouvelles données peut être contrôlée. De plus, cette approche théorique présente l'avantage d'être universelle et ne nécessite aucune hypothèse sur la distribution des données.

Chapitre 12

Conclusions

Au travers de ce chapitre, nous avons balayé les différentes perspectives qu'offrent les réseaux de neurones, après avoir défini les principes généraux. Nous nous sommes attardés plus particulièrement sur le modèle utilisé, à savoir le Perceptron Multicouche avec l'algorithme de rétropropagation du gradient l'erreur.

Pour finir, nous avons introduit très brièvement leur propriété fondamentale d'approximateurs parcimonieux ainsi que la théorie statistique d'apprentissage développée par Vapnik.

Troisième partie

UNE ALTERNATIVE AUX METHODES CLASSIQUES DE TARIFICATION SANTE

Chapitre 13

Méthodes classiques de tarification

13.1 Modèles de type fréquence × coût moyen

Les techniques classiques de tarification santé s'inpirent souvent de la méthode élaborée en 1968 par le B.C.A.C.[1], actualisée en 1981, puis en 2001. Elles reposent sur le principe de décomposition du risque entre la fréquence et le coût moyen. Ainsi, pour un acte médical donné, le montant de la prime peut être déterminé comme étant égale au produit de la fréquence f_{acte} par le remboursement complémentaire moyen $\bar{C}_{acte}$.

En d'autres termes, pour un individu :

$$\Pi = \sum_{acte \in Actes} \Pi_{acte} = \sum_{acte \in Actes} f_{acte} \times \bar{C}_{acte}$$

La fréquence par acte est obtenue en observant la consommation réalisée sur les exercices précédents. En général, elle est décomposée comme étant le produit d'une fréquence "centrale" par différents correctifs associés à l'âge, au sexe, à la C.S.P, au type de bénéficiaire, au caractère obligatoire ou non du régime, à la situation géographique, au secteur d'activité,...

La fréquence par acte peut alors s'écrire de la façon suivante :

$$f_{acte} = \bar{f}_{acte} \times C_{sexe} \times C_{age} \times C_{C.S.P.} \times C_{Benef.} \times C_{Oblig} \times C_{Region} \times C_{Activite}$$

La fréquence centrale correspond à celle d'un individu médian, par exemple, un adhérent de sexe masculin âgé de 40 ans. Pour une femme, conjointe, âgée de 50 ans, le tarif pourrait être le suivant :

$$f_{F,50,conjoint,acte} = \bar{f}_{acte} \times C_{sexe} \times C_{age} \times C_{Benef.}$$

$$f_{F,50,conjoint,acte} = f_{H,40,adherent,acte} \times 1,5 \times 1,2 \times 0,8$$

Une fois la fréquence déterminée, le montant du remboursement complémentaire moyen est évalué en fonction de la garantie associée à l'acte. Comme pour la fréquence, les frais réels sont observés sur un historique de consommation du portefeuille. L'information de la moyenne ne suffit généralement pas, celle de la répartition est plus riche et plus adaptée. Elle permet de prendre en compte la volatilité des tarifs qui a des effets mécaniques sur les remboursements complémentaires. A partir du moment où les frais réels sont modélisés, il ne reste plus qu'à en déduire le montant des remboursements complémentaires en fonction de la garantie.

[1] Bureau Commun d'Assurances Collectives

Prenons par exemple la répartition en déciles des prothèses dentaires remboursées. Le montant de remboursements complémentaires moyens pour une garantie à 400% de la BR sera obtenu sommant les remboursements complémentaires moyens simulés sur chaque décile :

$$\bar{C} = \sum_{i=1}^{10} 10\% \times min(400\% \times BR, FR_i - Secu)$$

Avec une BR à 107,50 € pour une prothèse de type "SPR50", la Sécurité sociale rembourse à hauteur de 70% soit 75,25 €

Déciles	1er	2ème	3ème	4ème	5ème	6ème	7ème	8ème	9ème	10ème
Frais réels moyens	370 €	370 €	470 €	540 €	580 €	600 €	640 €	710 €	800 €	800 €
Sécu. sociale	75 €	75 €	75 €	75 €	75 €	75 €	75 €	75 €	75 €	75 €
Rbts compl.	295 €	295 €	395 €	430 €	430 €	430 €	430 €	430 €	430 €	430 €

Le montant de remboursements complémentaires ainsi obtenu est de 399 €. Si la fréquence de consommation d'un bénéficiaire sur ce poste est de 0,4 prothèse par an, la prime correspondante serait de $0,4\times$ 399 € = 160 € par an.

En pratiquant de la sorte sur l'ensemble des postes, la prime pure est donc évaluée. Le passage à la prime commerciale se fait naturellement en rajoutant les chargements et taxes.

$$\Pi_{commerciale} = \Pi_{pure} \times \frac{1 + taxes}{1 - chargements}$$

13.2 Avantages et limites du modèle

Ce modèle a le mérite d'être assez simple d'utilisation et d'être employé par une majorité d'entreprises d'assurance. Dans le cadre de l'activité de courtage, il est relativement fréquent d'être en négociation avec un assureur, pour obtenir un tarif le plus proche possible du risque réel afin de piloter au mieux le régime. En ce sens, le fait de pouvoir discuter sur la base du même modèle est un atout en terme de communication. Le fait d'avoir un coût par poste est également appréciable.

La contrepartie de cette simplicité est le manque de précision des résultats obtenus. Implicitement dans cette approche, les dépendances entre les fréquences, les frais réels et le garanties ne sont pas supposées exister. Or dans les faits, ces grandeurs sont étroitement liées.

« Toutes les innovations en cours sont condamnées à l'échec si les systèmes d'information statistiques, qui permettent d'analyser en temps réels les résultats des mesures mises en oeuvres, ne suivent pas. Les assureurs santé, qui ont dans ce domaine un retard certain, devront donc s'atteler dans l'avenir à construire des bases de données fiables qui leur permettrons de créer les tableaux de bord et les outils de tarification indispensables. »[2]

[2]Jalma, Panorama de l'assurance santé 2000, Dossier technique : Gérer le risque santé

Chapitre 14

Méthode neuronale

14.1 Postulat du modèle

La part de la santé en assurance collective dans le budget des ménages a augmenté de plus de 40% en huit ans en passant de 2,75% en 2001 à 3,89% en 2009[1]. Les régimes complémentaires santé sont donc de plus en plus sollicités. Les enjeux financiers grandissant, les entreprises se préoccupent davantage du pilotage de leurs régimes dans l'optique de maîtriser au mieux leurs contraintes budgétaires.

Face à cette évolution, il nous est paru important d'évaluer le coût du risque le plus finement possible, d'où notre volonté de concevoir un modèle tarifaire plus précis qu'à l'ordinaire.

A l'origine de cette réflexion, quelques observations montraient que les méthodes traditionnelles n'utilisaient pas suffisamment les corrélations entre les différents facteurs. Ce constat nous a amené à penser qu'elles pouvaient être sensiblement améliorées :

- **La part des dépenses restant à la charge du bénéficiaire peut constituer un réel frein modifiant ainsi son comportement.** Dans ce cas, la fréquence de consommation de même que les montants de frais réels sont d'autant plus faibles que la contribution financière supportée par le salarié est importante. Cette influence peut être plus ou moins prononcée selon la Catégorie Socio Professionelle pour des motivations purement économiques.
- **Les bénéficiaires et les professionnels de santé peuvent être parfois tentés d'optimiser les remboursements complémentaires en fonction des garanties dont ils disposent.** Sur l'optique et le dentaire, il n'est pas rare de constater un ajustement des prix sur les limitations de garanties ou bien un transfert de budget des montures sur les verres.
- **Certains postes de dépenses, souvent consommés simultanément, peuvent être étroitement liés.** Par exemple, un contrat ne proposant qu'un faible remboursement des montures pourrait limiter la consommation sur ce poste et entrainer indirectement une diminution de la consommation de verres.
- **L'aspect préventif d'une garantie peut aussi limiter la consommation future sur un autre poste.** Par exemple, la mise en place de garanties couvrant des soins dentaires tels que la parodontologie peut à terme diminuer la consommation de couronnes, même principe avec la chirurgie de l'œil et les équipements optiques.

[1] source : "Quel avenir pour l'assurance maladie", Jalma

Toutes ces pratiques ont donc bien un impact direct sur le coût du régime et ne sont pas intégrées dans les méthodes de base.

Afin de pallier les carences dont souffrent les modèles classiques, il nous fallait trouver un modèle capable d'intercepter des liaisons de haut niveau entre les variables. Séduits par leurs propriétés d'approximateurs universels, notre choix s'est porté instinctivement vers un modèle de réseau de neurones.

Dans cette partie, nous allons bâtir une méthodologie de tarification globale, et non pas poste par poste, qui prendra en compte implicitement les corrélations entre les différents paramètres.

14.2 Méthodologie

Après avoir défini un processus de base, nous allons ensuite décliner notre modèle selon différentes variantes. Après comparaison, le choix se portera sur celle qui procurera les meilleurs résultats. Un fois le modèle défini précisément, nous nous assurerons de la cohérence des résultats face à des exemples fictifs. Enfin, il sera comparé à d'autres méthodes pour vérifier sa validité.

Les différentes phases du modèle peuvent se résumer de la façon suivante :

- codification des niveaux de garantie
- transformation des variables en entrée
- séparation de la base de donnée en trois : entraînement, validation et test
- entraînement du modèle jusqu'au critère d'arrêt

Les différentes variantes seront testées selon les paramètres suivants :

- le nombre de couches cachées
- le nombre de neurones cachés
- les paramètres pour la méthode de rétropropagation du gradient (pas, momentum d'inertie)
- les méthodes d'apprentissage du premier ordre (rétropropagation du gradient, Rprop)
- les méthodes d'apprentissage du second ordre (Levenberg-Marquardt, Gradient Conjugué, Quasi Newton)
- la fonction d'activation
- la fonction objective

Le réseau de neurones sélectionné sera comparé aux modèles suivants :

- Modèle Linéaire Généralisé (G.L.M.)
- Memory Base Reasoning (M.B.R.)

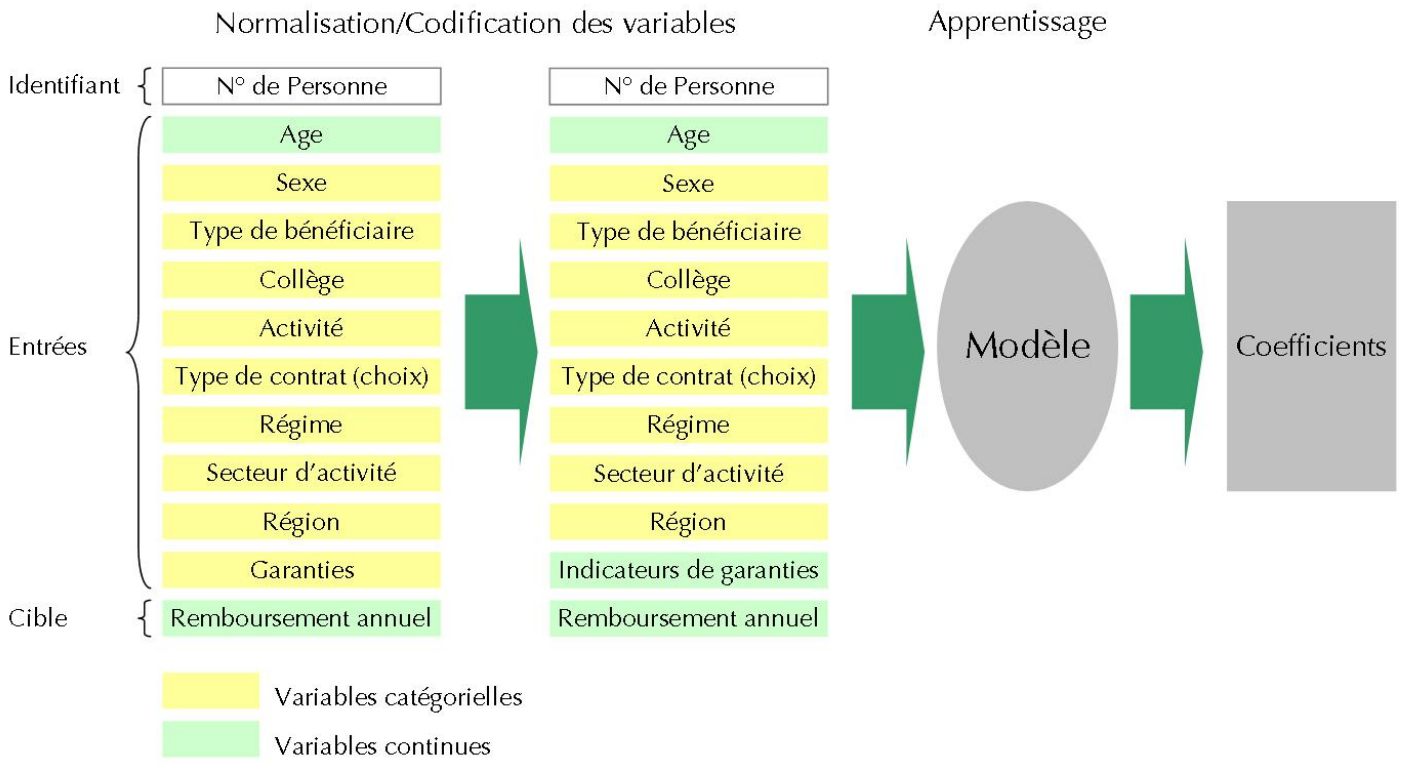


FIG. 14.1 – Process

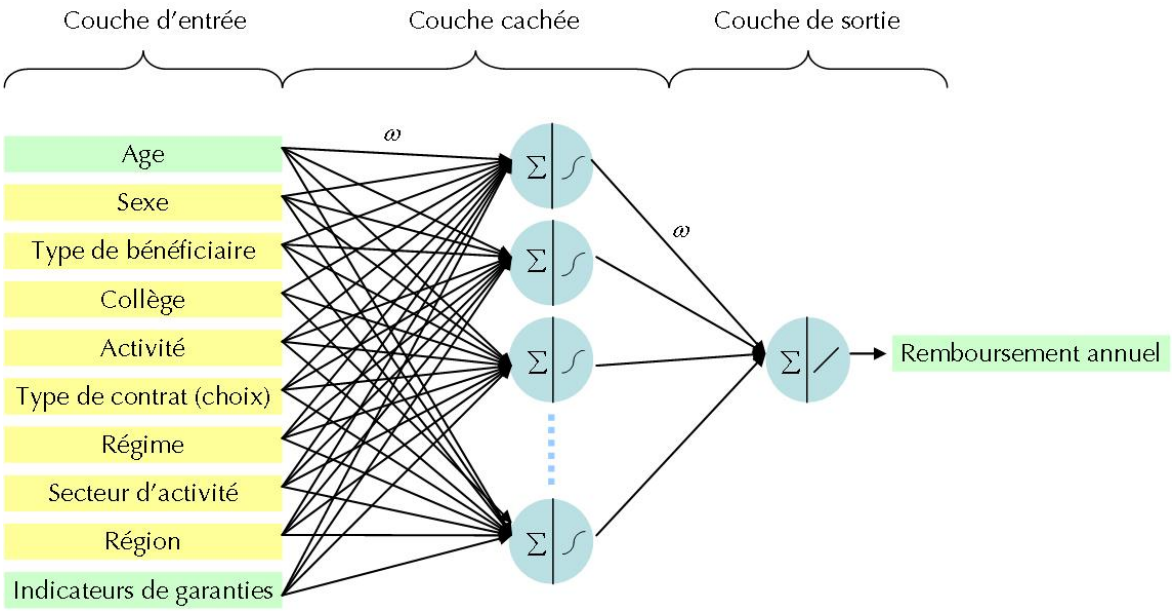


FIG. 14.2 – Schéma du PMC utilisé

14.3 Choix des variables discriminantes

Les informations que nous avons retenues en entrée du modèle sont indiquées par bénéficiaire ayant consommé on non sur une année. En effet, les personnes n'ayant pas bénéficié de soins doivent bien être prises en compte dans le coût du régime.

Par souci d'homogénéité, la consommation annuelle est corrigée du temps de présence sur l'année, en l'occurrence sur l'exercice 2007. A priori, il paraît raisonnable de considérer que la consommation est proportionnelle au temps de présence sur la période d'observation.

L'information sur la consommation du personnel expatrié est partielle. Elle correspond uniquement à leur consommation en France et non à l'étranger. Pour cette raison, nous avons ôté du périmètre de l'étude cette population.

Les variables jugées pertinentes que nous avons sélectionnées sont les suivantes :

- l'identifiant du bénéficiaire (N° interne)
- le sexe (variable binaire : Homme ou Femme)
- l'âge (variable continue)
- le type de bénéficiaire (variable catégorielle : Adhérent, Conjoint ou Enfant)
- le collège (variable catégorielle : Ensemble, Cadre, ETAM, Non cadre)
- l'activité (variable binaire : actif ou inactif)
- le choix, type de contrat (variable catégorielle : Obligatoire, Base, Option)
- le régime (variable catégorielle : Général, Alsace-Mos elle)
- le secteur d'activité (variable catégorielle : 24 secteurs recensés)
- la région (variable catégorielle : 22 régions recensées)
- les indicateurs de garanties (21 variables continues)
- le remboursement moyen (variable continue)

Le remboursement moyen correspond à la variable cible de notre modèle.

14.4 Codification/transformation des variables

14.4.1 Codification des variables catégorielles

Les variables catégorielles ont dû subir un retraitement afin qu'elles puissent être prises en compte dans le réseau. Chacune de ces variables disposant de n modalités possibles est dupliquée en $n-1$ variables artificielles reprenant ses modalités. Par convention, la dernière modalité (par ordre alphabétique) n'est pas reprise compte tenu du fait qu'elle est directement liée aux autres.

Le mode de conversion retenu est le suivant :
Dans les cas où la valeur de variable à coder est différente de la dernière modalité, les $n-2$ variables artificielles n'indiquant pas la bonne modalité se voient attribuer la valeur 0 et celle qui correspond prend la valeur 1. Dans le cas où la variable à coder prend la valeur de la dernière modalité, toutes les variables artificielles valent -1.

Par exemple, la variable type de bénéficiaire qui comporte 3 modalités distinctes (adhérent, conjoint ou enfant) a été convertie en 2 variables (bénéficiaire=adhérent, bénéficiaire=conjoint).

La codification est la suivante :

Type de bénéficiaire	Variables artificielles	
	bénéficiaire=adhérent	bénéficiaire=conjoint
Adhérent	1	0
Conjoint	0	1
Enfant	-1	-1

14.4.2 Normalisation des variables

Les variables continues utilisées n'ont pas forcément des valeurs du même ordre de grandeur. Des écarts importants peuvent nuire aux algorithmes d'apprentissage en les rendant inefficaces. Un prétraitement peut consister à les normaliser en effectuant un changement de variable remplaçant chacune d'entre elles par leur transformation en variable centrée réduite.

Nous avons donc opté pour cette solution, la transformation pour la variable x de l'exemple i, notée x_i, est la suivante :

$x_i \rightarrow \frac{x_i - \bar{x}}{\sigma(x)}$ avec $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ et $\sigma(x) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

Les variables catégorielles, quant à elles, ne subissent aucune normalisation.

14.4.3 Initialisation des poids

Avant de lancer l'algorithme, les valeurs initiales des poids et des biais ont besoin d'être définies. Si elles ne sont pas correctement fixées, il peut y avoir d'importantes répercussions sur la vitesse d'apprentissage ainsi que sur la capacité de généralisation.

La stratégie retenue s'inspire de celle décrite par Smieja [SMI91]. Les poids des connexions de la couche cachée vers celle de sortie sont initialisés à zéro, le biais de la couche de sortie est initialisé à la valeur moyenne de la cible, soit 392,14. Pour les autres, un tirage aléatoire est effectué selon une loi normale centrée et de faible écart-type (0,1), dans un intervalle $[-\frac{2}{\sqrt{d_{in}}}, \frac{2}{\sqrt{d_{in}}}]$ où d_{in} est le nombre de connexions entrantes.

Pour une architecture donnée, l'initialisation aléatoire des poids est conservée pour toutes les simulations afin qu'elles puissent être réalisées dans les mêmes conditions.

Voici un extrait de l'initialisation des 1561 poids (avec 76 variables en entrée, 20 neurones cachés et une valeur de sortie il y a : $76 \times 20 + 20 + 20 + 1 = 1561$ poids). La notation "Hxy" indique le y$^{\text{ème}}$ neurone de la x$^{\text{ème}}$ couche cachée. Par exemple, "H12" représente le deuxième neurone de la première couche cachée.

Connexions		Poids
Entrée	Sortie	
age	H11	-0,0005
CGE	H11	-0,0011
VS4	H11	0,0071
ACTIVITE=Actifs	H11	0,0375
CHOIX=Base	H11	0,0229
CHOIX=Option	H11	0,0626
CLASSE=Activité indéterminée	H11	-0,1603
CLASSE=Activités financières	H11	-0,0509
⋮	⋮	⋮
CLASSE=Transports	H11	0,1558
COLLEGE=Cadre	H11	0,0695
COLLEGE=Ens. Pers	H11	-0,2457
COLLEGE=ETAM	H11	0,1443
BENEF=AD	H11	0,0202
BENEF=CJ	H11	-0,0618
REGION=Alsace	H11	-0,1296
REGION=Aquitaine	H11	-0,0548
⋮	⋮	⋮
REGION=Rhône-Alpes	H11	-0,0024
REGIME=Alsace	H11	0,0162
SEXE=0	H11	-0,1091
⋮	⋮	⋮
Biais	H11	0,2033
Biais	H12	-0,0227
⋮	⋮	⋮
Biais	H120	-0,0766
H11	M GS moy	0
H12	M GS moy	0
⋮	⋮	⋮
H120	M GS moy	0
Biais	M GS moy	392,1378

14.4.4 Entraînements préliminaires

Les différents poids et biais finaux du réseau entraîné dépendent de leurs valeurs initiales. Pour réduire le risque de convergence vers un mauvais optimum local, plusieurs entraînements préliminaires peuvent être répétés sur quelques itérations en partant de différentes valeurs initiales. Celui offrant les meilleurs résultats est alors retenu. Les poids correspondants servent alors, à leur tour, pour initialiser la phase d'entraînement.

Pour ces raisons, une phase d'entraînement préliminaire de 50 lancés sur 20 itérations sera appliquée dans toutes nos simulations avec la méthode d'apprentissage de Levenberg-Marquardt (développée plus loin).

14.5 Définition du modèle de base

Nous avons choisi arbitrairement une première modélisation de réseau de neurones de base servant de référence. Partant de celle-ci, nous allons étudier un par un ses différents paramètres afin de les ajuster au mieux selon les résultats obtenus.

Ce premier réseau possède les caractéristiques d'architecture et d'apprentissage suivantes :
- une répartition de la base aléatoire : 40% entraînement, 30% validation et 30% test
- poids initiaux aléatoires (sauf pour la couche de sortie)
- une phase préliminaire (50 lancés sur 20 itérations avec la méthode de Levenberg-Marquardt)
- 1 seule couche cachée
- 20 neurones cachés
- entre 2 couches, tous les neurones sont liés
- pas de liaisons directes
- une fonction d'activation sigmoïde de type tangente hyperbolique
- fonction de coût à minimiser : erreur quadratique
- technique d'apprentissage : rétropropagation version batch avec un pas de 0,01 et sans momentum
- critère d'arrêt : Early stopping avec un maximum de 1 000 itérations

La base de test sera mise de côté est ne servira qu'une fois le modèle ajusté pour estimer au plus juste l'erreur réelle à la fin de l'apprentissage.

Les deux grandes classes d'indicateurs qui peuvent être utilisées pour apprécier la qualité de la modélisation sont l'erreur quadratique moyenne (A.S.E. Average Square of Error) et les critères d'informations classiques.

La minimisation de l'erreur quadratique est issue du principe de maximum de vraisemblance reposant sur une hypothèse de distribution approximativement normale des sorties avec une variance constante.

En notant k le nombre de paramètres, n le nombre d'exemples et $L(\hat{\theta})$ la vraisemblance, les principaux critères d'informations sont :

- le critère A.I.C. (Akaïké Information Criterion) d'Akaïké :
  $AIC = -2lnL(\hat{\theta}) + 2k$

- le critère B.I.C. (Bayesian Information Criterion) de Schwarz :
  $BIC = -2LnL(\hat{\theta}) + log(n)k$

Ces derniers pénalisent la log-vraissemblance en prenant en considération le nombre de paramètres (la plus faible valeur de ce critère sera préférée). Pour autant, le nombre de paramètres ne reflète pas systématiquement la complexité du modèle et notamment en ce qui concerne le cas des réseaux de neurones (cf [SAP06]). Pour cette raison, ces deux critères ne seront pas retenus dans le choix de nos critères de décision, seule l'erreur quadratique moyenne sera prise en compte.

Afin de pouvoir comparer les différentes variantes de modèles le plus objectivement possible, les simulations sont établies à partir des mêmes bases d'entraînement, de validation et de test. La présentation aléatoire des exemples est définie une fois pour toute selon le même ordre lors de chaque apprentissage (même graine de départ).

Le processus d'apprentissage est le suivant :

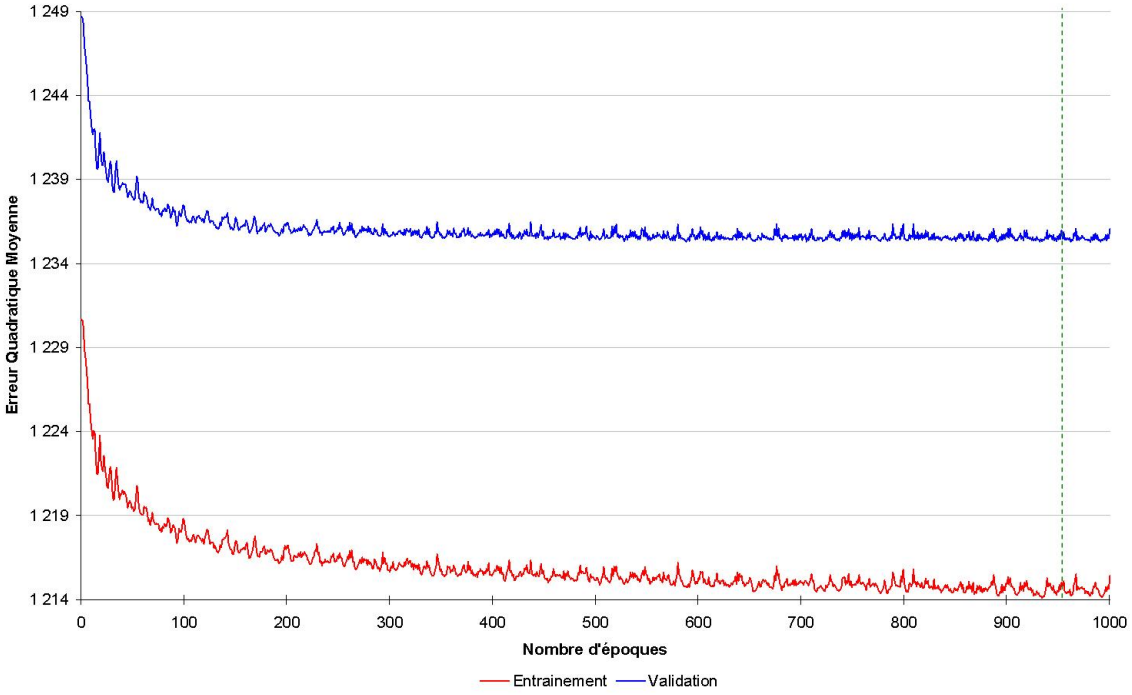


FIG. 14.3 – 1 couche cachée, 20 neurones, Backprop batch, pas de 0.01 et momentum nul

L'algorithme semble avoir convergé vers un minimum, l'erreur sur la base d'entraînement ne diminue quasiment plus, de même que celle de la base de validation. Le fait que nos données soient fortement entachées d'aléa explique l'absence de surapprentissage sur la base de validation (l'erreur ne remonte pas clairement après avoir atteint son minimum). Le minimum sur la base de validation est atteint à la 962 ème itération, la valeur de l'erreur quadratique moyenne sur la base d'apprentissage est de 1 214,25 et 1 235,27 sur la base de validation. En stoppant l'apprentissage à ce minimum d'erreur sur la base de validation, nous définissons ainsi les coefficients de notre premier modèle. En appliquant celui-ci à la base de test, l'erreur obtenue est de 1 286,03. Cette valeur reflète la qualité du modèle "lâché dans la nature" sur des données inconnues.

Pour une meilleure lisibilité, les courbes d'apprentissage et de validation seront parfois lissées en appliquant la moyenne par pas de 100 itérations.

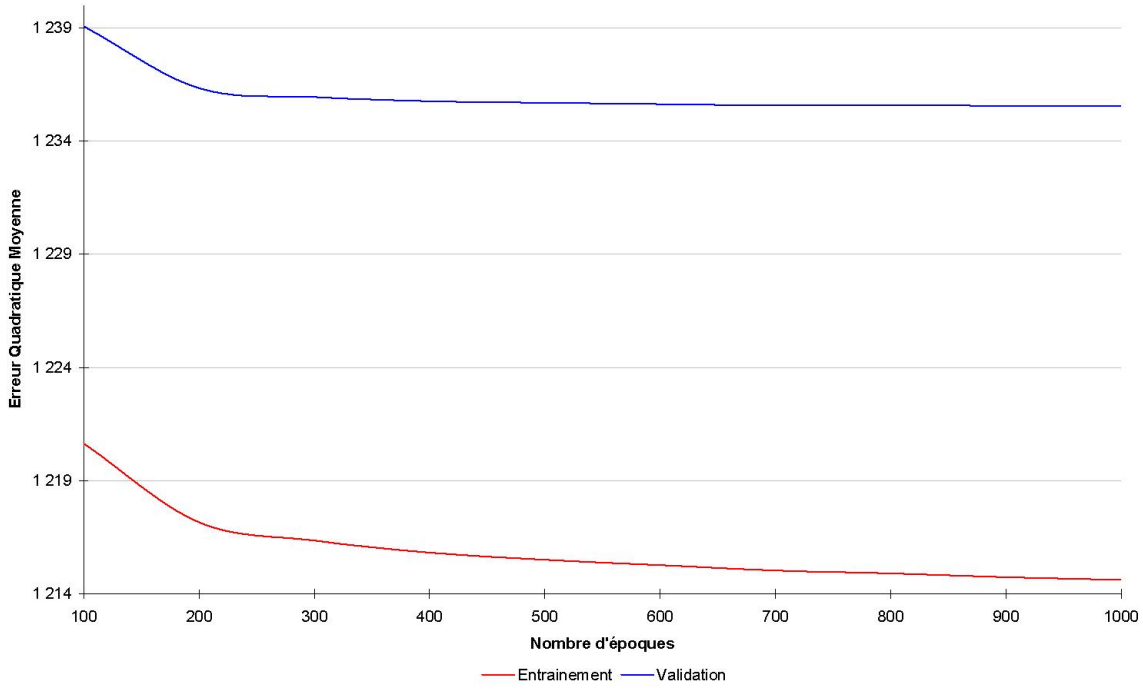


FIG. 14.4 – Lissage : moyenne sur 100 époques

Ce modèle va servir de référentiel pour la suite de l'étude.

Chapitre 15

Sélection du modèle

15.1 Nombre de neurones dans la couche cachée

Pour déterminer, dans un premier temps, le nombre de neurones à insérer dans la couche cachée, cinq simulations sont réalisées en partant de notre modèle de base.

Dans les cinq simulations, le seul paramètre qui varie est le nombre de neurones cachés :
- une simulation avec 10 neurones cachés
- une simulation avec 20 neurones cachés
- une simulation avec 30 neurones cachés
- une simulation avec 40 neurones cachés
- une simulation avec 50 neurones cachés

Plus le nombre de neurones est important plus l'erreur diminue rapidement sur la base d'apprentissage, ce premier constat est conforme sur le plan théorique.

Avec les 1 000 époques sur la base de validation, l'erreur semble s'être presque stabilisée dans les 5 scénarii. Sur cette même base, ce sont les modèles avec 10 ou 20 neurones cachés qui donnent les moins bons résultats. Entre 30, 40 et 50 neurones cachés, l'écart est moins accentué. Pour l'instant, notre choix se porte sur le réseau comportant 40 neurones cachés, comme étant la valeur moyenne entre 30, 40 et 50 neurones cachés.

Cette sélection est non définitive, elle pourra être réétudiée ultérieurement selon les résultats obtenus dans la suite de notre démarche. En l'occurrence, la convergence n'ayant pas été clairement observée (absence de remontée de l'erreur sur la base de validation), une fois les autres paramètres du réseau définis, nous retesterons le modèle ainsi obtenu sur un plus grand nombre d'itérations. L'objectif étant de s'assurer que l'erreur a bien atteint son minimum.

	Nombre de neurones cachés				
	10	20	30	40	50
EQM Min (validation)	1 235,72	1 235,27	1 234,94	1 234,88	1 234,76
Epoque	$679^{\text{ème}}$	$962^{\text{ème}}$	$775^{\text{ème}}$	$910^{\text{ème}}$	$942^{\text{ème}}$
EQM (test)	1 286,72	1 286,03	1 286,10	1 285,75	1 285,83

Sur la base de validation, l'Erreur Quadratique Moyenne (E.Q.M.) atteint son minimum avec une valeur de : 1 234,88 à la $910^{\text{ème}}$ itération.

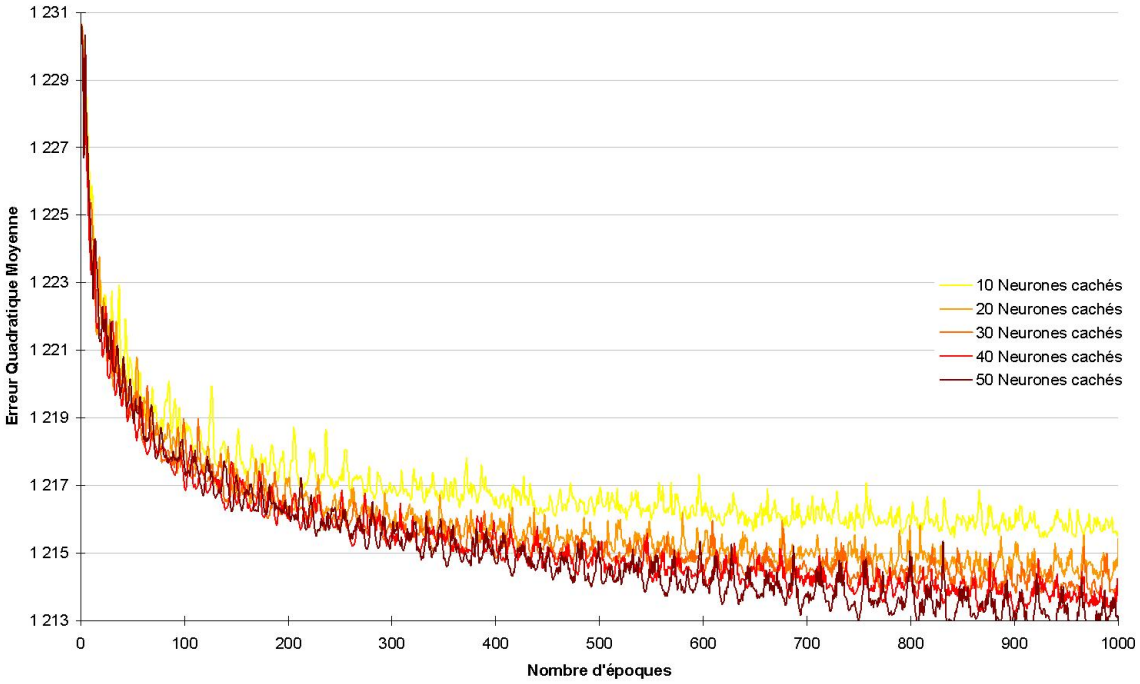


FIG. 15.1 – Entraînement (brut) - Nombre de neurones cachés

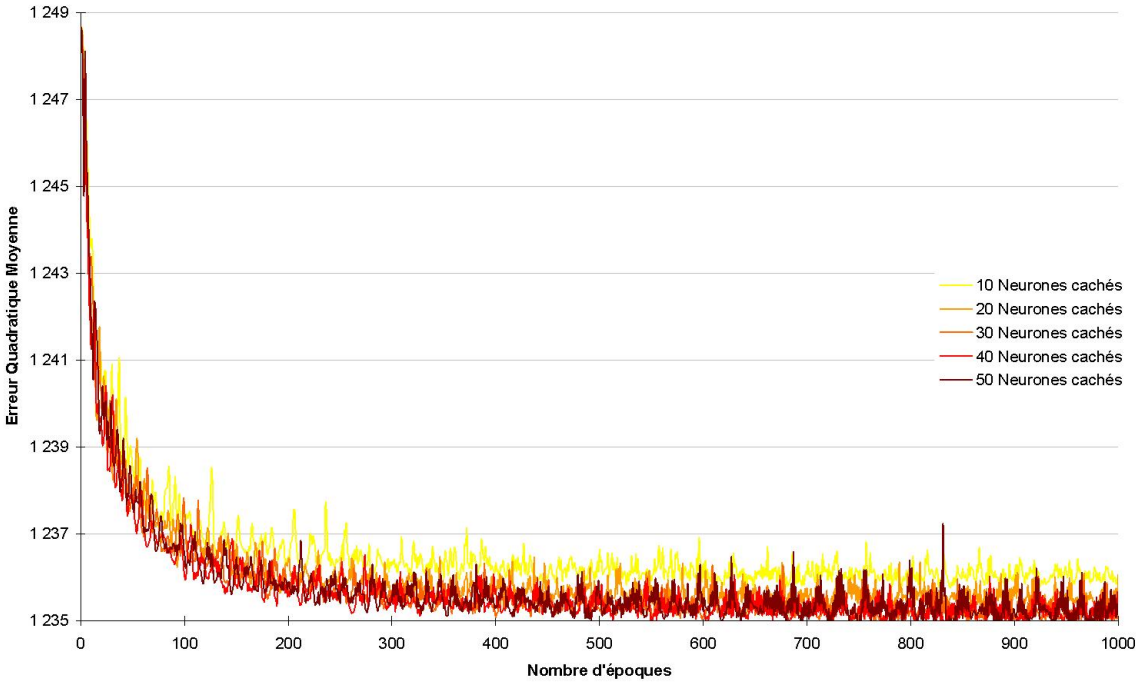


FIG. 15.2 – Validation (brute) - Nombre de neurone cachés

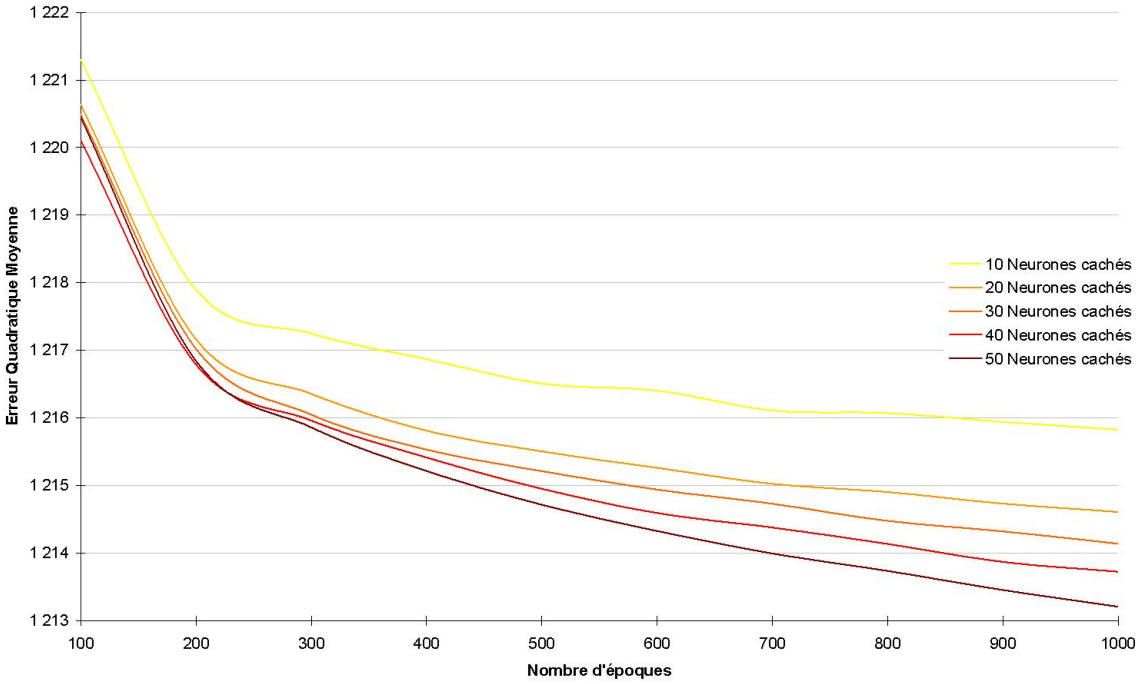


FIG. 15.3 – Entraînement (lissé) - Nombre de neurones cachés

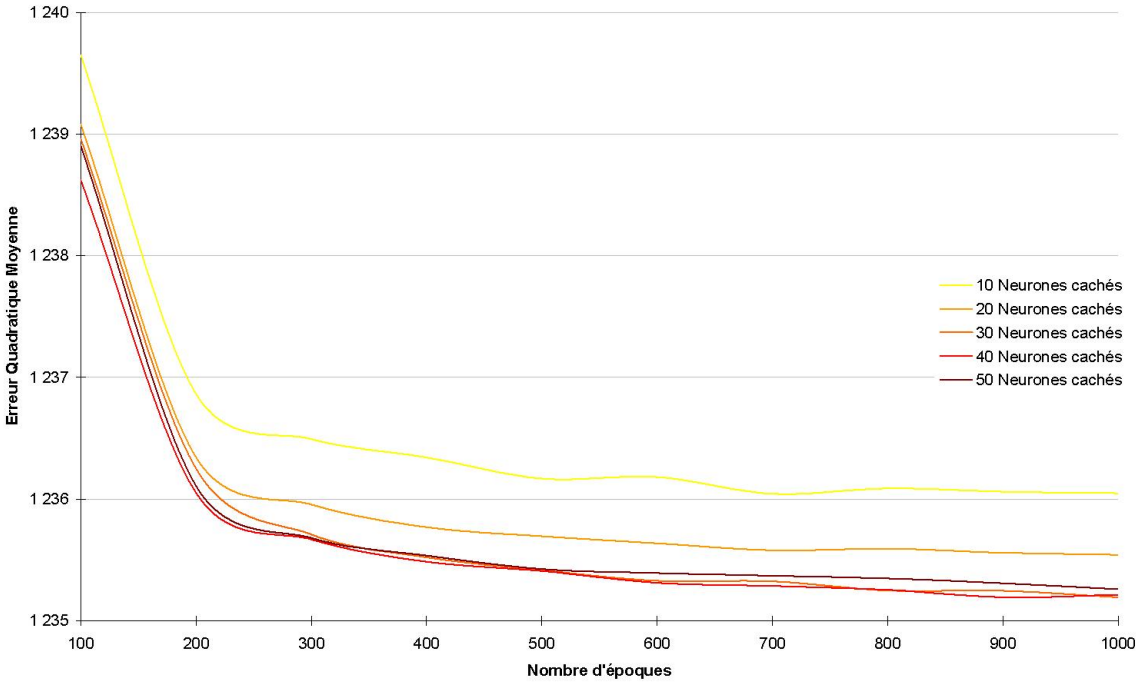


FIG. 15.4 – Validation (lissée) - Nombre de neurone cachés

15.2 Nombre de couches cachées

Comme nous l'avons vu précédemment, un réseau de neurones doté d'une seule couche cachée peut approximer n'importe quelle fonction continue bornée avec la précision voulue, sous réserve qu'il y ait suffisamment de neurones cachés. L'ajout d'une deuxième couche cachée permet de lever la contrainte de continuité de la fonction à approximer. Dans notre cas, la fonction remboursement complémentaire que nous cherchons à prédire est bornée et vraisemblablement continue. L'utilisation d'une seule couche devrait donc suffire dans notre modélisation. Nous allons cependant vérifier ce point en ajoutant une seconde couche identique à la première (40 neurones cachés).

Le test est effectué sur 1 000 époques avec un pas d'apprentissage fixé à 0,001. Afin d'accroître la vitesse de convergence de l'algorithme, nous avons eu recours à une technique dite de "moment d'inertie" avec un momentum valant 0,8. Cette technique, qui sera détaillée ultérieurement, permet d'accélérer la convergence.

Le comparatif se fait ici entre les deux réseaux suivants :

- une simulation avec une seule couche cachée de 40 neurones
- une simulation avec deux couches cachées de 40 neurones chacune

Les résultats obtenus sont les suivants :

	Nombre de couches cachées	
	1	2
EQM Min (validation)	1 234,54	1 235,18
Epoque	367ème	193ème
EQM (test)	1 285,69	1 286,10

L'ajout d'une seconde couche complexifie l'architecture du réseau qui passe de 3 900 à 4 760 poids. La phase d'apprentissage sur une époque est donc plus longue sur le réseau à deux couches cachées (plus de poids à ajuster). Les nombres d'époques nécessaires à la convergence ne sont alors pas comparables.

La remontée de l'erreur sur la base de validation est observée pour nos deux simulations.

L'erreur décroît plus rapidement en phase d'apprentissage sur le réseau avec une couche supplémentaire. Elle passe en dessous de celle du réseau à une seule couche cachée à partir de la 150ème époque. Par contre, la capacité de généralisation est moins performante sur le réseau avec deux couches cachées puisque sur la base de validation, l'erreur commise est systématiquement supérieure à celle obtenue avec une seule couche cachée.

L'apport d'une seconde couche cachée ne présente donc aucun d'intérêt, ce qui vient confirmer notre intuition. Pour cela, l'idée de rajouter une seconde couche cachée ne sera pas retenue pour la suite.

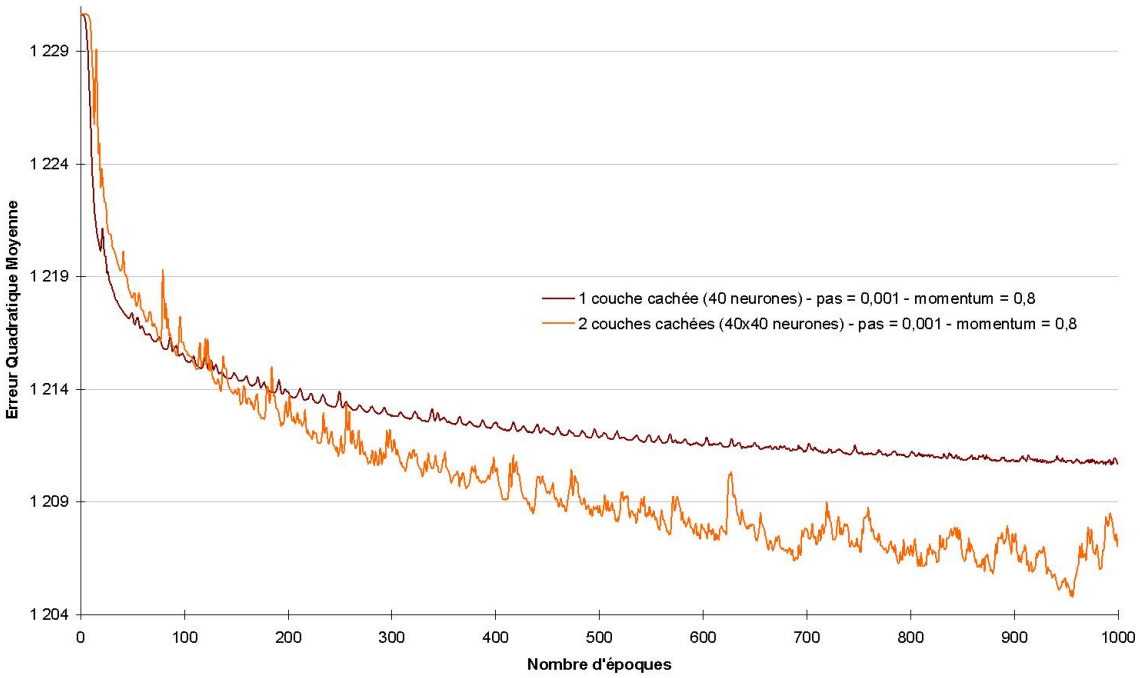


FIG. 15.5 – Entraînement (brut) - Nombre de couches cachées

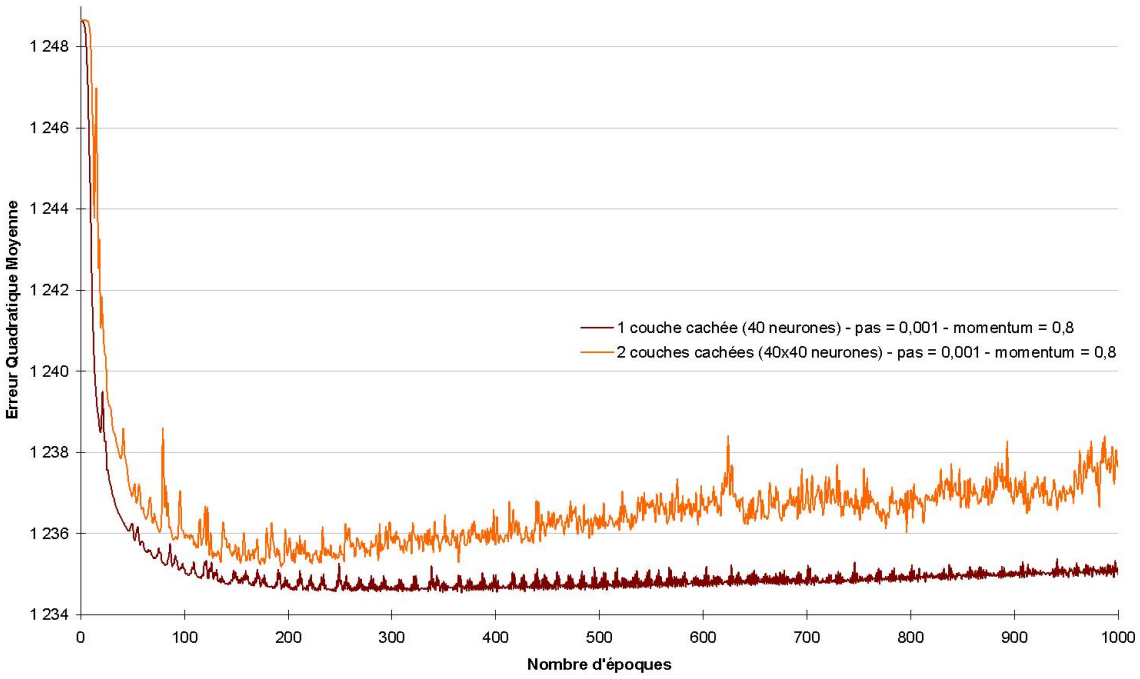


FIG. 15.6 – Validation (brute) - Nombre de couches cachées

15.3 Technique d'apprentissage du premier ordre

15.3.1 Rétropropagation de l'erreur (Backprop version batch)

Pas d'apprentissage

En conservant le modèle de base défini précédemment, en augmentant uniquement le nombre de neurones dans sa couche cachée de 20 à 40, cinq scénarii de pas d'apprentissage vont être simulés :

- une simulation avec un pas de 0,1
- une simulation avec un pas de 0,05
- une simulation avec un pas de 0,01
- une simulation avec un pas de 0,001
- une simulation avec un pas de 0,0001

Le scénario de pas à 0,1 est rapidement écarté car, au bout de quelques itérations seulement, l'erreur croît exponentiellement avec plus de 700 000 000 à la $17^{\text{ème}}$ itération.

Sur la base d'apprentissage, l'amplitude des oscillations est plus importante sur les simulations réalisées avec les valeurs de pas les plus élevées. Sur les 4 scénarii restants, la tendance générale de l'erreur est baissière, cette diminution semble s'atténuer au fil des itérations. Conformément à la théorie, plus le pas de l'erreur est faible plus le modèle est précis et plus la convergence est lente. Ceci se remarque notamment avec la simulation ayant la plus faible valeur de pas à 0,0001, qui se détache légèrement des autres en convergeant moins rapidement.

Sur la base de validation, le même constat est observé à la différence qu'avec un pas de 0,001 la descente de la pente est plus marquée que sur les trois autres. A partir de la $600^{\text{ème}}$ itération, la courbe avec un pas de 0,001, passe en dessous de celle à 0,01. Pour cette raison et en prenant également en considération le temps de convergence, la valeur de pas à 0,001 est retenue.

	Valeur du pas d'apprentissage			
	0,05	0,01	0,001	0,0001
EQM Min (validation)	1 234,98	1 234,88	1 234,86	1 237,00
Epoque	$733^{\text{ème}}$	$910^{\text{ème}}$	$1\,000^{\text{ème}}$	$1\,000^{\text{ème}}$
EQM (test)	1 285,86	1 285,75	1 285,68	1 287,74

Les simulations avec des pas de 0,001 et 0,0001 laissent présumer que l'erreur peut encore continuer à diminuer à condition de rajouter des simulations supplémentaires. En effet, le minimum n'est atteint qu'à la dernière itération et la pente de l'erreur, orientée à la baisse, est encore présente. Ceci vient confirmer le fait qu'après avoir déterminer les autres paramètres, le modèle retenu sera testé sur une durée plus longue ou accéléré par l'ajout d'un momentum. De nouvelles simulations avec la plus petite valeur de pas à 0,0001 ne sont pas non plus définitivement écartées.

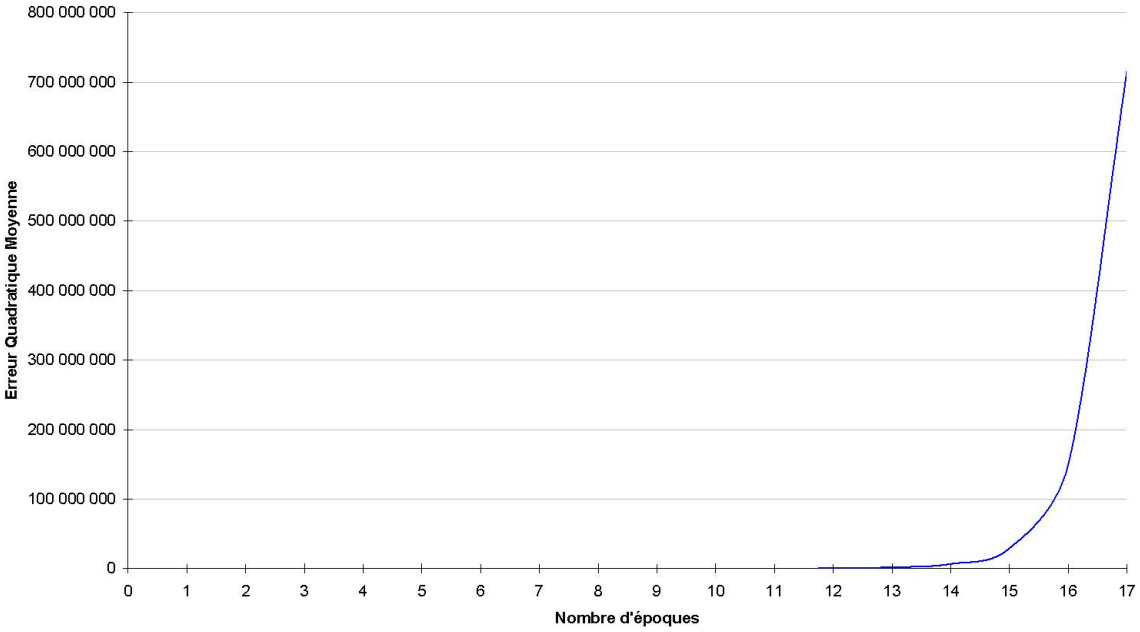


FIG. 15.7 – Entraînement (lissé) avec un pas de 0,1

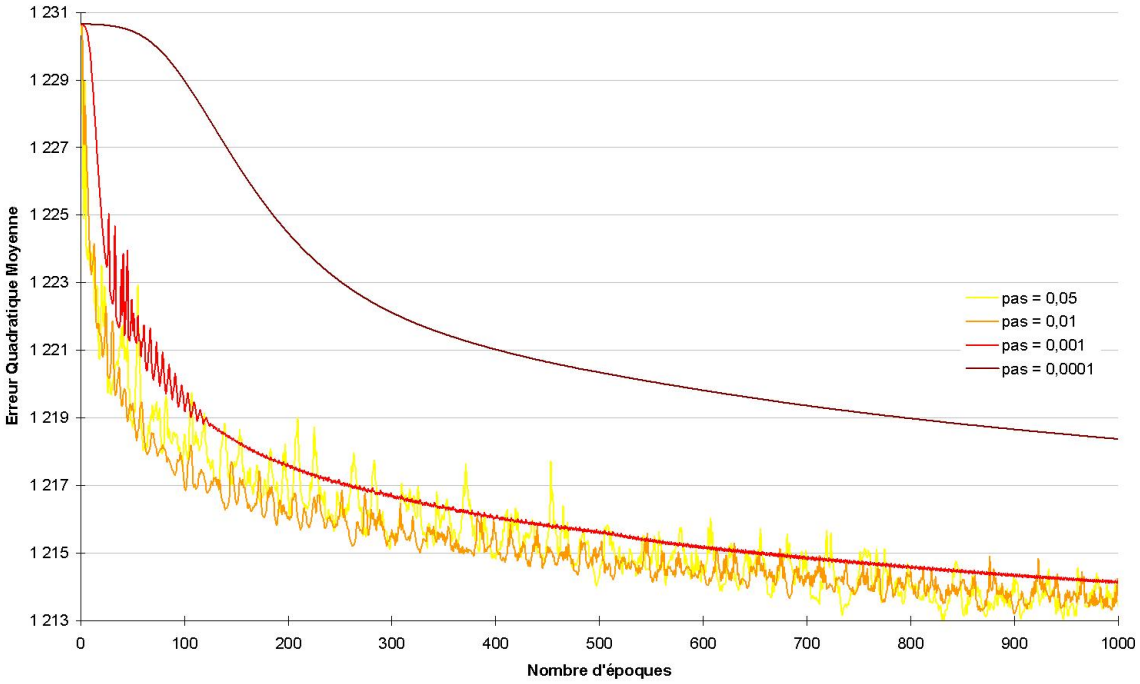


FIG. 15.8 – Entraînement (brut) - Pas d'apprentissage

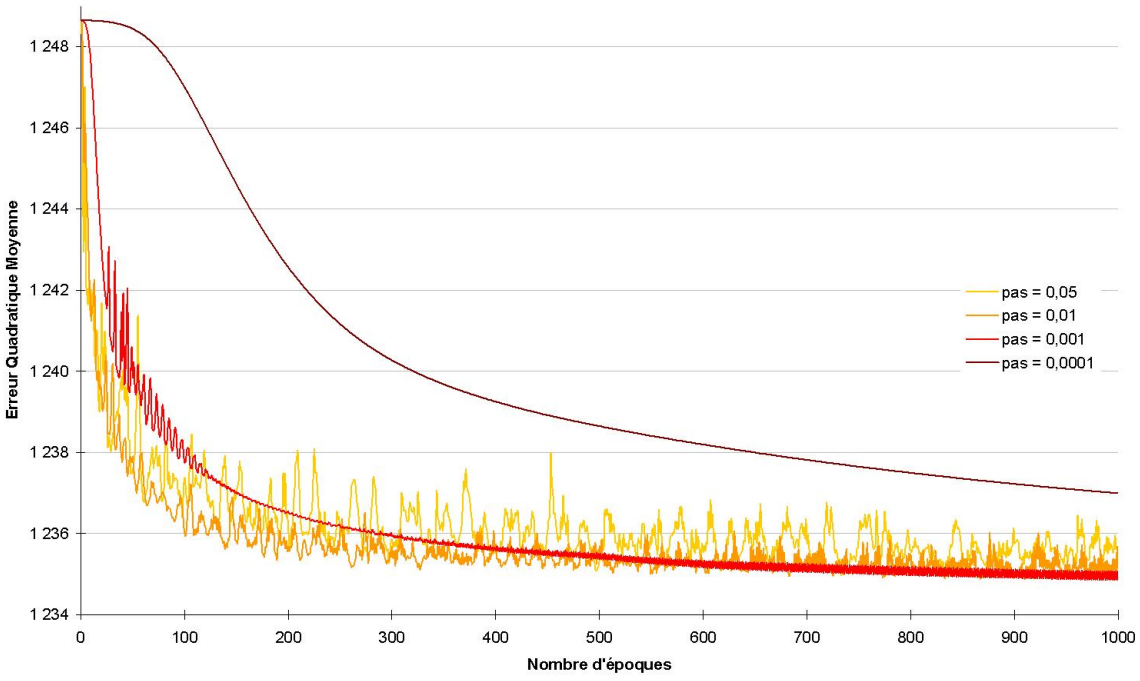


FIG. 15.9 – Validation (brute) - Pas d'apprentissage

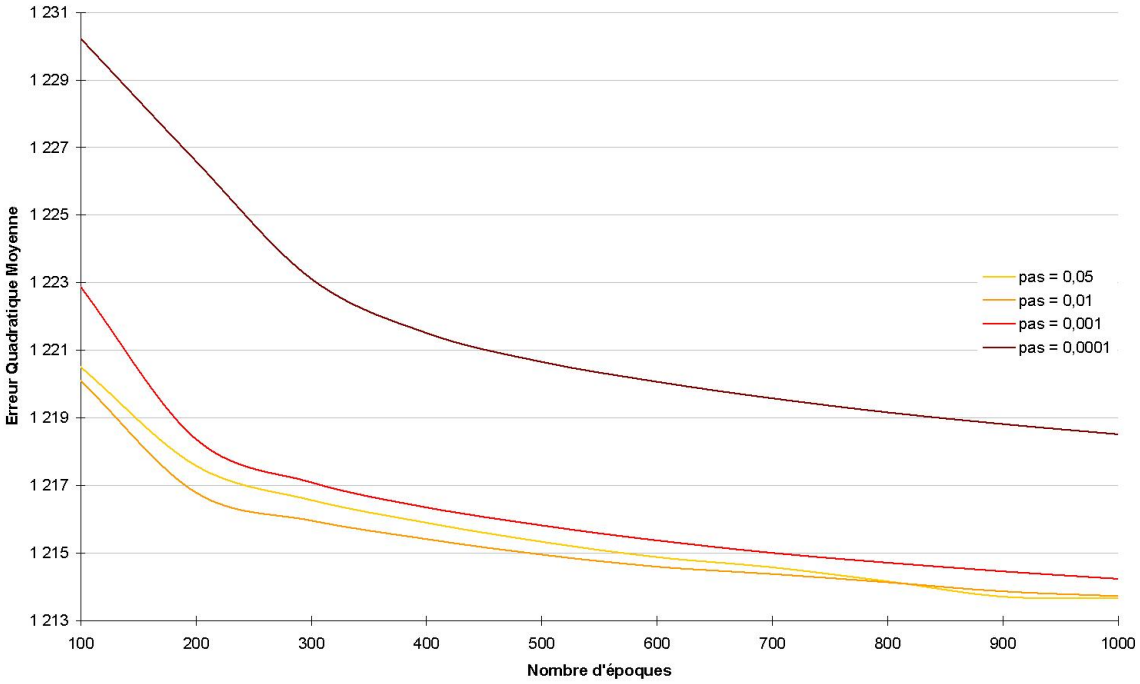


FIG. 15.10 – Entraînement (lissé) - Pas d'apprentissage

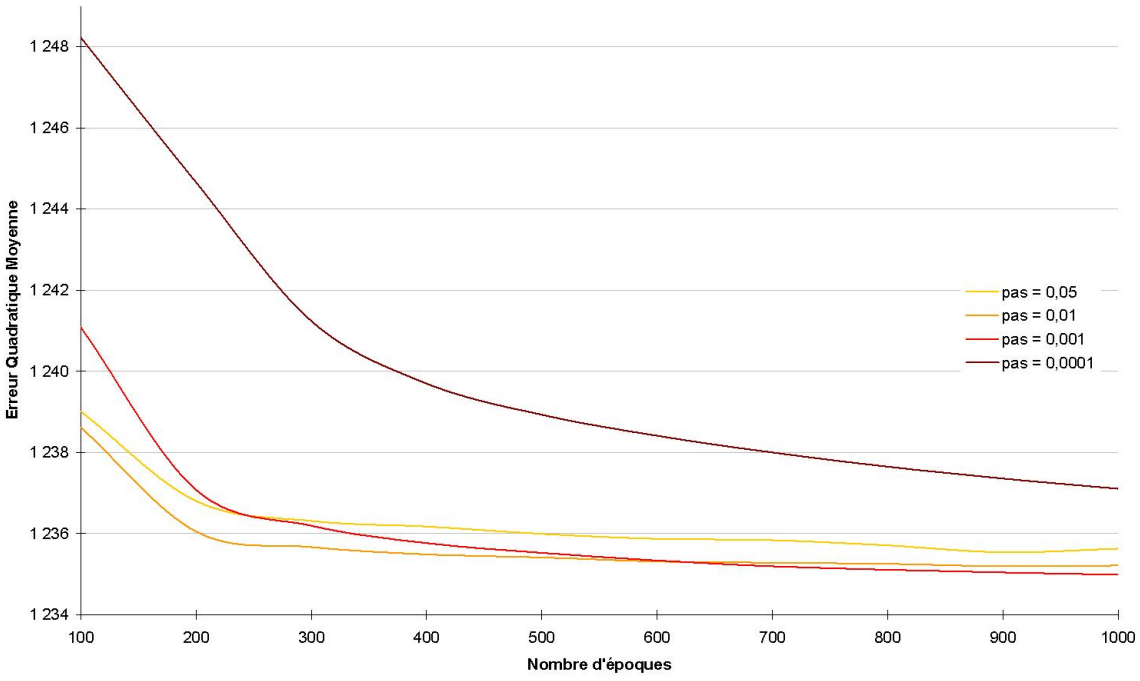


FIG. 15.11 – Validation (llissé) - Pas d'apprentissage

15.3.2 Accélération avec un moment d'inertie

Pour accélérer l'optimisation du réseau, l'ajout d'un terme d'inertie (momentum) permet de conserver les informations des précédents changements de poids et d'éviter les oscillations. De plus, cette technique du premier ordre inventée par Rumelhart en 1986 [RUM86] permet d'éviter que l'algorithme converge vers un minimum local.

En notant M le momentum, la modification de la variation de poids à la k$^{\text{ème}}$ itération s'opère comme suit :

$$\boxed{\Delta w_{ij}(k) = \varepsilon \delta_i y_{ij} + M \Delta w_{ij}(k-1)}$$

La valeur du momentum est généralement comprise entre 0 et 1.

En conservant notre modèle de base définit précédemment avec toujours 40 neurones dans sa couche cachée et un pas d'apprentissage de 0,001, cinq scénarii de pas d'apprentissage vont être simulés :

- une simulation sans momentum
- une simulation avec un momentum de 0,5
- une simulation avec un momentum de 0,7
- une simulation avec un momentum de 0,8
- une simulation avec un momentum de 0,9

L'ajout d'un momentum atteint bien l'effet escompté en accélérant nettement la vitesse de convergence. Toujours avec 1 000 époques sur la base de validation, la remontée de l'Erreur Quadratique Moyenne apparaît plus distinctement. Le minimum d'erreur est obtenu sur la simulation avec un Momentum valant 0,8. Cette valeur offrant les meilleurs résultats est donc celle qui sera retenue par la suite.

	Momentum				
	0	0,5	0,7	0,8	0,9
EQM Min (validation)	1 234,86	1 234,67	1 234,62	1 234,54	1 234,88
Epoque	1 000$^{\text{ème}}$	857$^{\text{ème}}$	403$^{\text{ème}}$	367$^{\text{ème}}$	231$^{\text{ème}}$
EQM (test)	1 285,68	1 285,53	1 285,71	1 285,69	1 285,74

Sur la base de validation, l'Erreur Quadratique Moyenne (E.Q.M.) atteint son minimum avec une valeur de : 1 234,54 à la 367$^{\text{ème}}$ itération alors qu'en l'absence de Momentum, l'algorithme n'avait pas encore atteint son minimum à la 1 000$^{\text{ème}}$ itération.

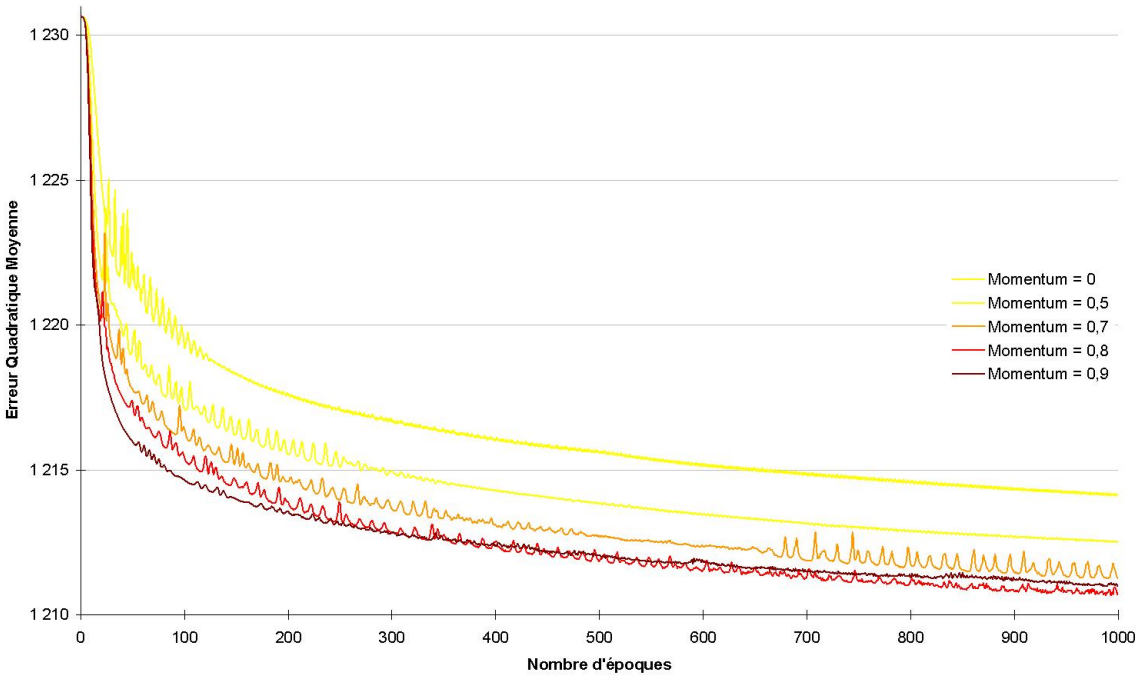


FIG. 15.12 – Entraînement (brut) - Momentum

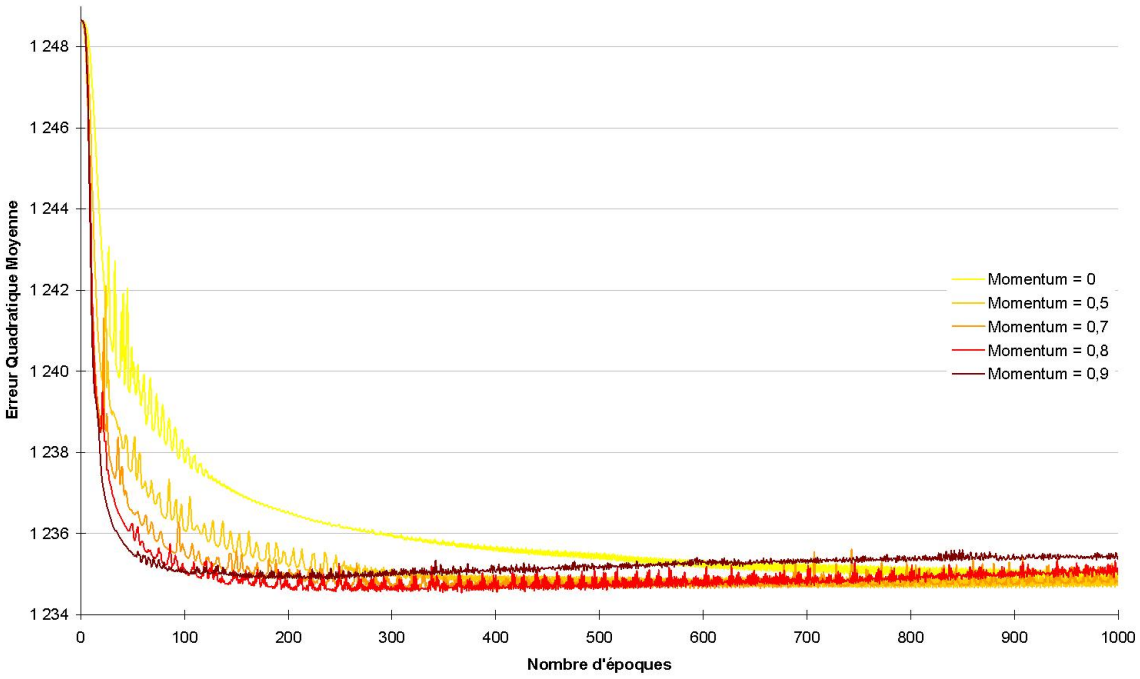


FIG. 15.13 – Validation (brute) - Momentum

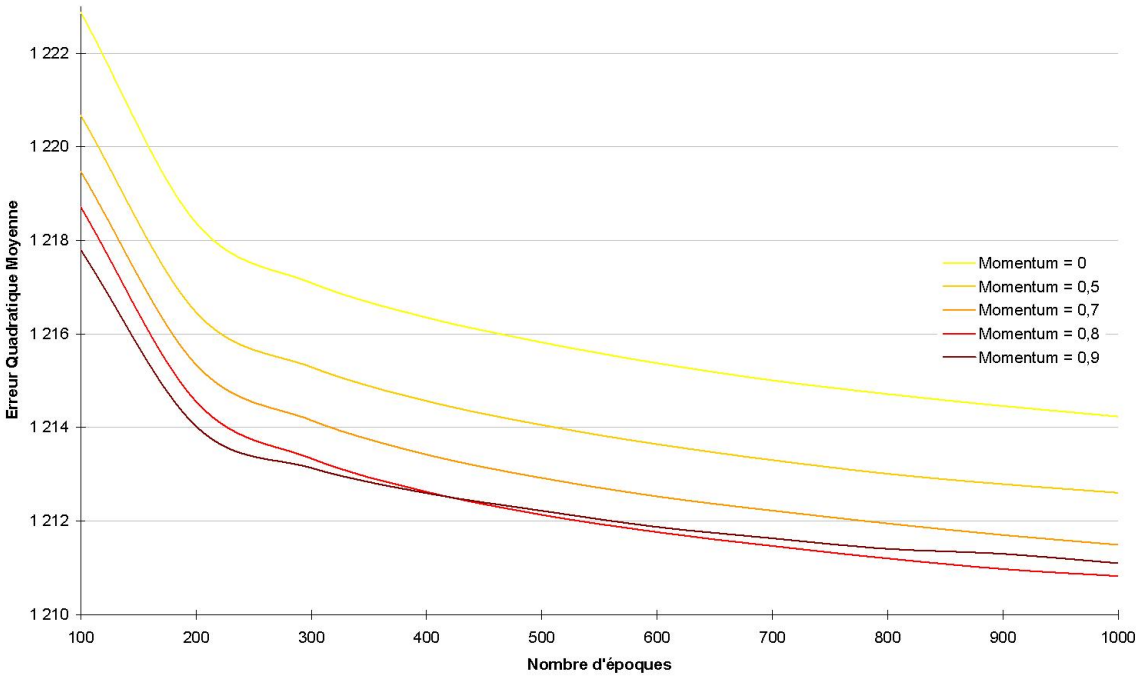


FIG. 15.14 – Entraînement (lissé) - Momentum

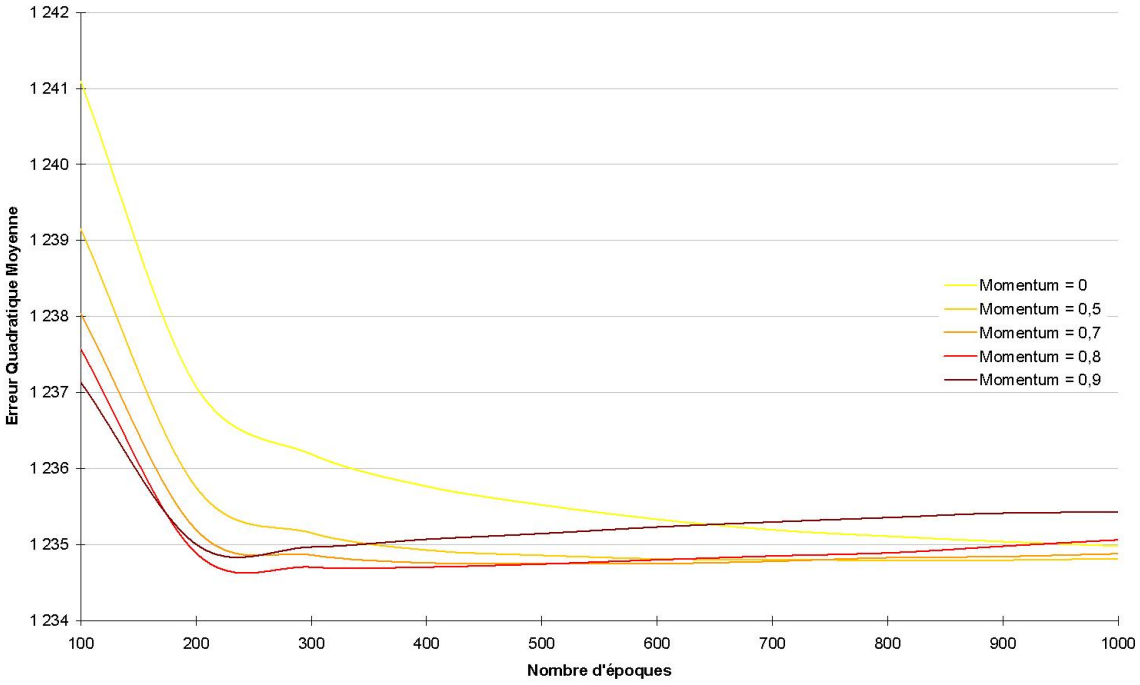


FIG. 15.15 – Validation (lissée) - Momentum

Lorsque les différents pas ont été testés, la simulation avec le pas le plus faible (0,0001) n'avait pas atteint un minimum sur la base de validation, l'erreur continuait encore à décroître à la 1 000$^{\text{ème}}$ itération. Une simulation complémentaire est relancée avec ce même pas, mais cette fois-ci en l'accélérant avec un momentum 0,8 sur une plus longue période de 3 000 itérations.

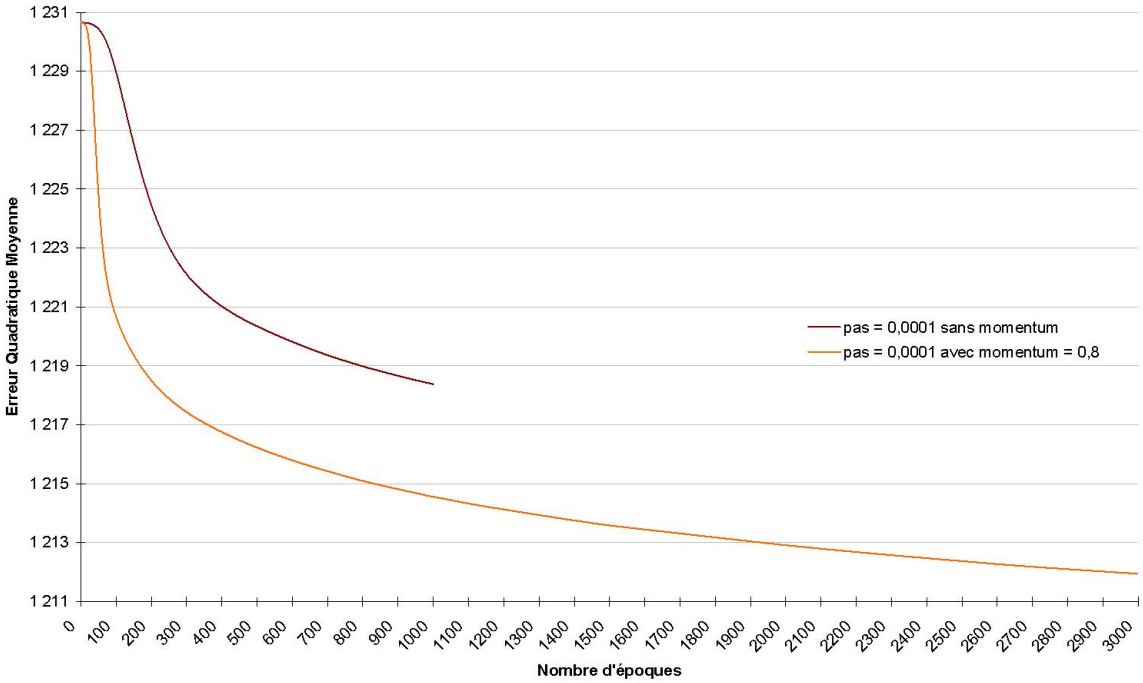


FIG. 15.16 – Entraînement (lissé) - Momentum

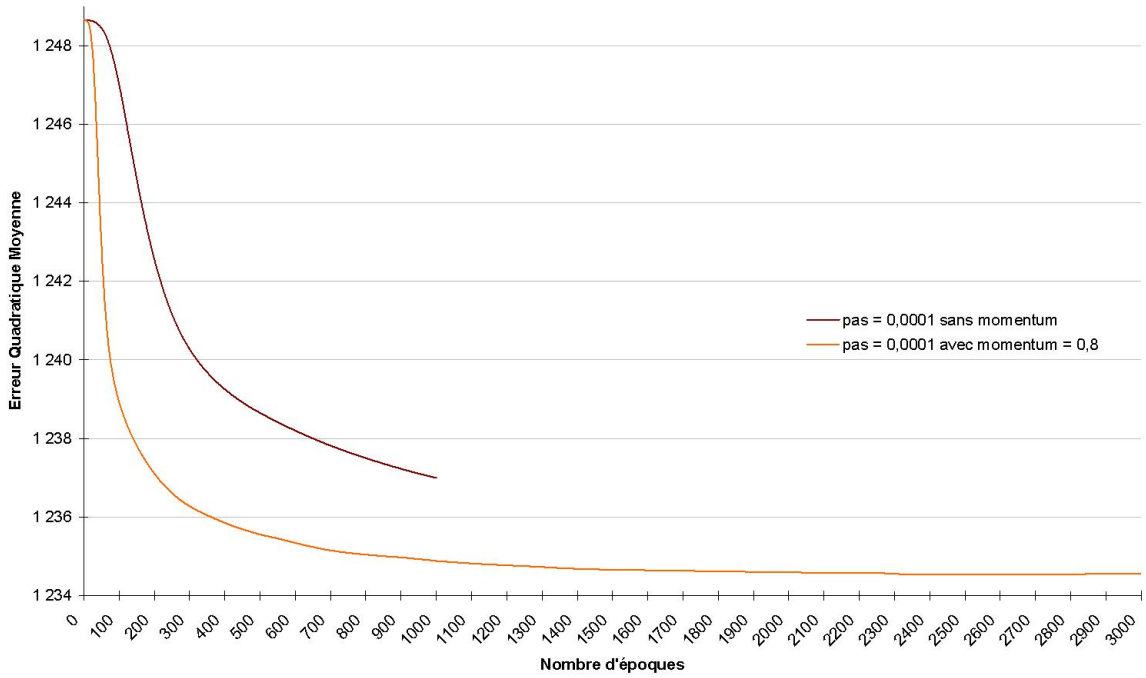


FIG. 15.17 – Validation (lissée) - Momentum

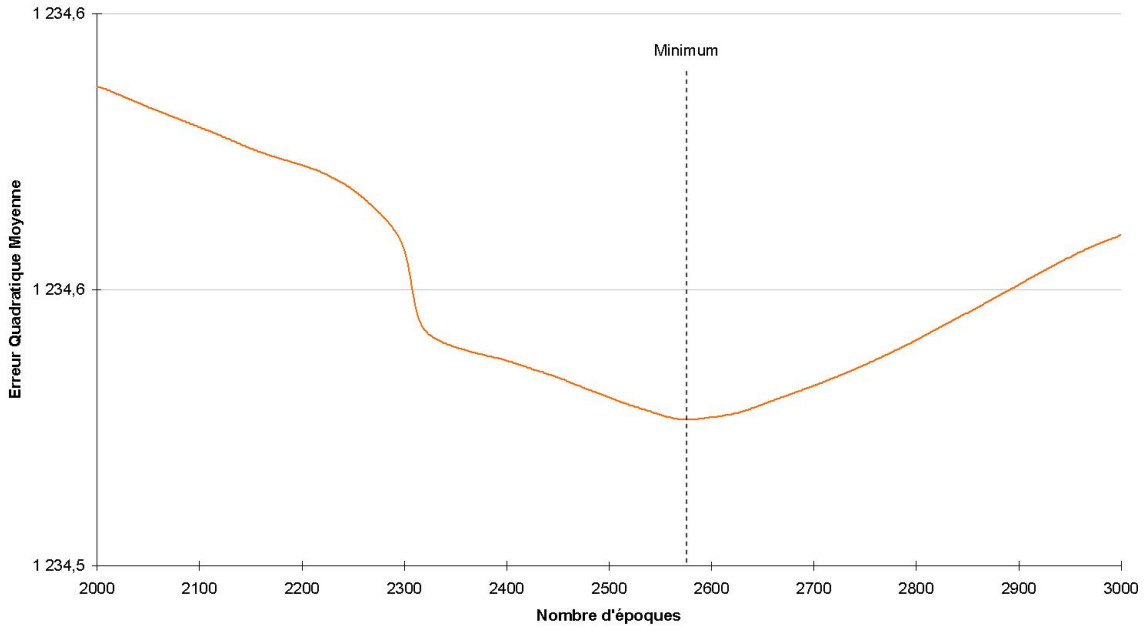


FIG. 15.18 – Simulation avec momentum : Focus validation itérations de 2 000 à 3 000

Sur la base de validation, l'Erreur Quadratique Moyenne (E.Q.M.) atteint son minimum avec une valeur de : 1 234,53 à la 2 573ème itération avant de remonter. L'erreur associée sur la base de test est de 1 285,50. La valeur de l'erreur minimum observée sur la base de validation est quasiment identique à celle observée avec un pas plus élevé à 0,001 : 1 234,53 contre 1 234,54. Ces résultats viennent confirmer notre choix d'un pas à 0,001, un pas plus

fin n'apportant pas de gain significatif.

Comme indiqué précédemment, la simulation ayant le plus grand nombre de neurones cachés (50) n'avait pas clairement convergé vers un minimum. Le lancement d'une nouvelle simulation complémentaire accélérée avec un pas de 0,001 et un momentum de 0,8 est alors jugé utile.

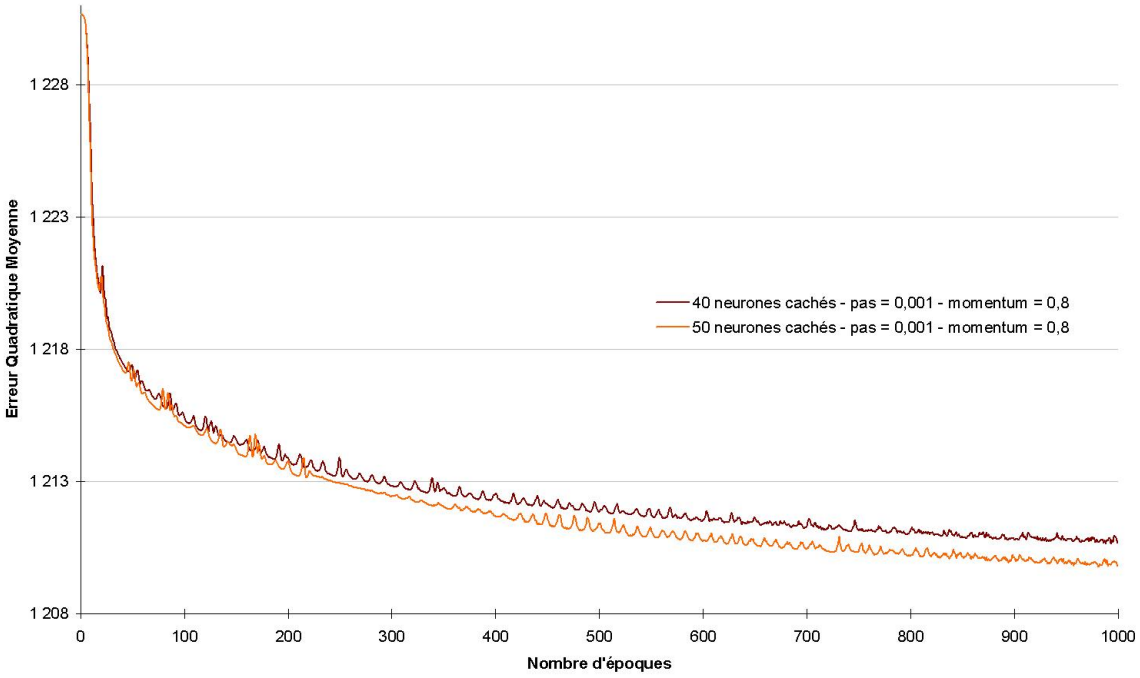


FIG. 15.19 – Entraînement (brut) - Nombre de neurones cachés

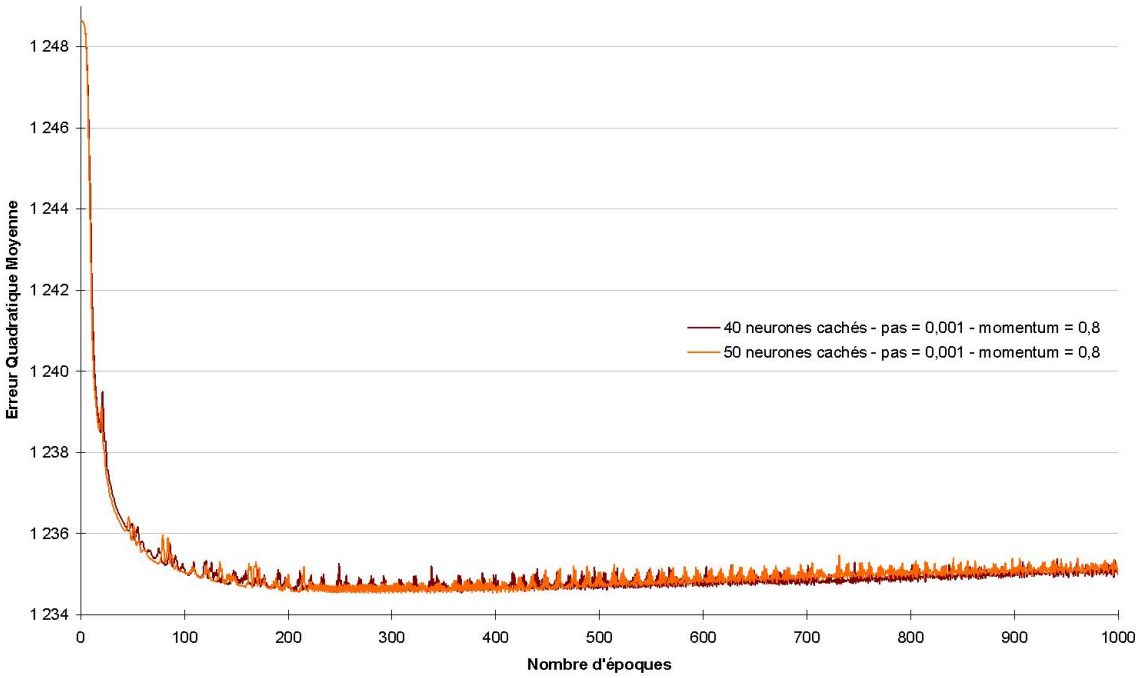


FIG. 15.20 – Validation (brute) - Nombre de neurones cachés

Avec l'ajout d'un moment d'inertie, la simulation atteint bien un minimum sur la base de validation à la 297ème itération avec une valeur de 1 234,53 et 1 285,38 sur la base de test. Même constat qu'avec la précédente simulation, l'ajout de 10 neurones supplémentaires dans la couche cachée n'apporte pas un gain de précision supplémentaire.

15.3.3 RProp

La technique RProp est une technique du premier ordre inventée par Riedmiller et Braun en 1993 [RIE93] qui présente l'avantage d'accélérer la convergence de l'algorithme. Dans l'algorithme classique de rétropropagation de l'erreur, les variations de poids sont proportionnelles à $\frac{\delta E}{\delta w_{ij}}$ or la grandeur de cette différentielle n'est pas la plus représentative de la modification à effectuer. Cette variante ne prend plus en compte la valeur de cette différentielle mais seulement les changements de signe.

Le principe de la règle de mise à jour des poids est la suivante :

$$\boxed{\Delta w_{ij} = -sign(\frac{\delta E}{\delta w_{ij}}) \times \varepsilon_{ij}}$$

où ε_{ij} (update value) représente le pas d'apprentissage qui est adapté selon la direction du gradient par rapport au gradient précédent : le pas augmente (up) lorsqu'il ne change pas de signe et diminue (down) s'il y a changement de signe.

A la k$^{\text{ème}}$ itération, le pas d'apprentissage $\varepsilon_{ij}(t)$ est borné par ε_{min} et ε_{max} de la façon suivante :

$$\varepsilon_{ij}(k+1) = \begin{cases} min(\varepsilon_{ij}(k)u, \varepsilon_{max}) & \text{si} \quad \frac{\delta E}{\delta w_{ij}}(k) \times \frac{\delta E}{\delta w_{ij}}(k-1) > 0 \\ max(\varepsilon_{ij}(k)d, \varepsilon_{min}) & \text{si} \quad \frac{\delta E}{\delta w_{ij}}(k) \times \frac{\delta E}{\delta w_{ij}}(k-1) < 0 \\ \varepsilon_{ij}(k) & \text{sinon} \end{cases}$$

où $0 < d < 1 < u$

La modification des poids ne s'opère que si l'algorithme se dirige "dans le bon sens" :

$$\Delta w_{ij}(k+1) = \begin{cases} -\varepsilon_{ij}(k) sign(\frac{\delta E}{\delta w_{ij}}) & \text{si} \quad \frac{\delta E}{\delta w_{ij}}(k) \times \frac{\delta E}{\delta w_{ij}}(k-1) \geq 0 \\ 0 & \text{sinon} \end{cases}$$

La mise à jour est effectuée au terme de chaque époque.

Cette technique est testée avec les paramètres suivants :

- up = 1,2
- down = 0,5
- min = 0,00001
- max = 50

La méthode Rprop est comparée avec la méthode backprop avec moment d'inertie sélectionnée précédemment.

Les résultats obtenus sont les suivants :

	Méthode	
	Backprop	Rprop
EQM Min (validation)	1 234,54	1 234,31
Epoque	367ème	253ème
EQM (test)	1 285,69	1 285,49

Lors de la phase d'entraînement et après une centaine d'époques, la méthode Rprop s'accélère pour enfin converger plus rapidement et dépasser la Backprop.

Pendant celle de validation, l'erreur minimum est obtenue plus rapidement avec la méthode Rprop dont la valeur est également la plus faible.

Dans le cas présent, la méthode Rprop est donc la plus performante des deux.

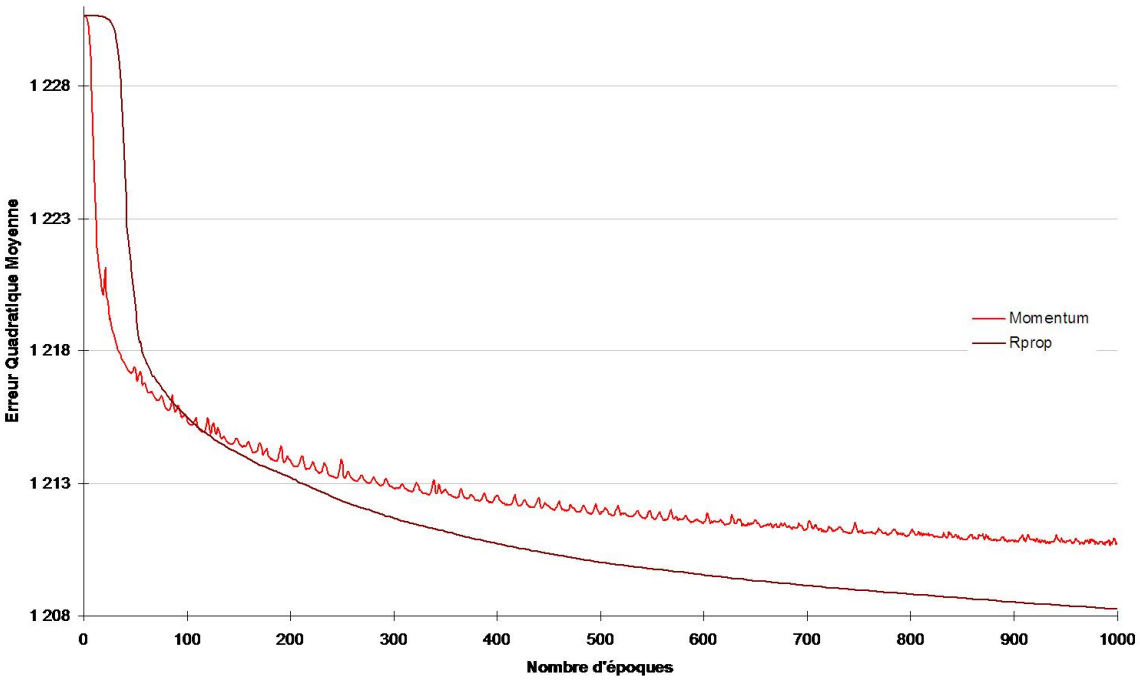


FIG. 15.21 – Entraînement (brut) - RPROP vs Momentum

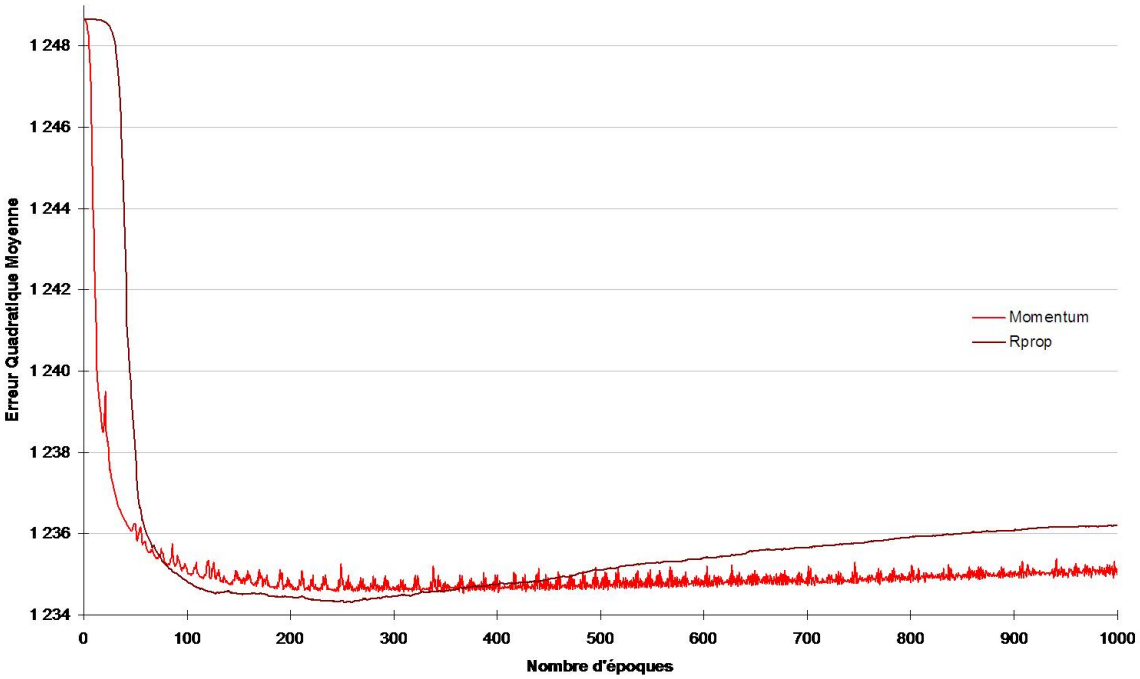


FIG. 15.22 – Validation (brute) - RPROP vs Momentum

15.4 Technique d'apprentissage du second ordre

15.4.1 Principe d'approximation du second ordre

Les techniques d'apprentissage du second ordre sont des algorithmes itératifs de descente du gradient où la fonction erreur est remplacée par son approximation quadratique au voisinage du point courant.

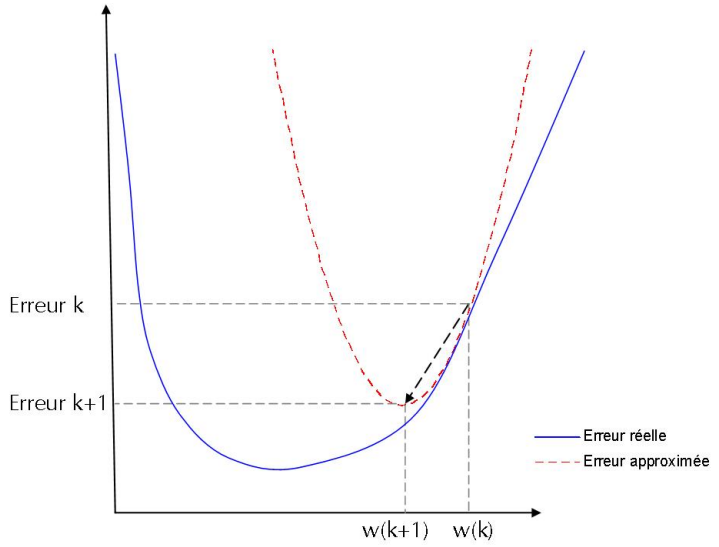


FIG. 15.23 – Hypothèse d'un comportement quadratique de l'erreur de Newton

Le développement de Taylor de la fonction coût au voisinage de w permet de faire l'approximation suivante :

$$E(w+h) \approx E(w) + \left(\frac{\partial E}{\partial w}\right)' h + \frac{1}{2} h' H(w) h$$

avec H représentant la matrice Hessienne du coût :

$$H = \begin{pmatrix} \frac{\partial^2 E}{\partial w_1^2} & \frac{\partial^2 E}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_n} \\ \frac{\partial^2 E}{\partial w_2 \partial w_1} & \frac{\partial^2 E}{\partial w_2^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_n \partial w_1} & \cdots & \cdots & \frac{\partial^2 E}{\partial w_n^n} \end{pmatrix}$$

En dérivant, le gradient s'écrit alors :

$$\frac{\partial E(w+h)}{\partial w} \approx \frac{\partial E(w)}{\partial w} + h' H(w)$$

L'objectif est de déterminer la quantité h de telle sorte à ce que le gradient s'annule (Méthode de Newton) :

$$\boxed{h = \Delta w = -H^{-1}(w) \frac{\partial E(w)}{\partial w}}$$

Généralement, le calcul de H^{-1} reste complexe. Pour cela, des méthodes permettant d'approximer le Hessien sont utilisées. Nous en développerons quelques unes.

15.4.2 Quickprop

La technique Quickprop est une méthode de second ordre inventée par Fahlman en 1988 [FAH88] qui repose sur deux hypothèses : l'erreur en fonction de chaque poids est une parabole convexe et les changements de pente ne sont pas affectés par les variations des autres poids.

Cet algorithme utilise un gradient modifié en ajoutant un terme supplémentaire :

$$\frac{\partial E^*(k)}{\partial w_{ij}} = \frac{\partial E^(n)}{\partial w_{ij}} + decay \times w_{ij}(k)$$

A la k ème itération, les variations de poids sont données comme suit :

$$\boxed{\Delta w_{ij}(k) = \alpha_{ij}(k)\Delta w_{ij}(k-1) - \varepsilon(k)\frac{\partial E^*(k)}{\partial w_{ij}}}$$

L'ajout du terme de descente de gradient supplémentaire $-\varepsilon(k)\frac{\partial E^*(k)}{\partial w_{ij}}$ permet de corriger les variations dues à l'influence des autres poids.

A la première itération, $\Delta_{ij}(0)$ est initialisé à 0, le pas suivant donne : $\Delta_{ij}(1) = -\varepsilon(1)\frac{\partial E^*(1)}{\partial w_{ij}}$. Pour les itérations suivantes, $\varepsilon(k)$ et $\alpha_{ij}(k)$ sont calculés comme suit :

$$\varepsilon_{ij}(k) = \begin{cases} \varepsilon_0 & \text{si} \quad \frac{\delta E^*(k)}{\delta w_{ij}} \times \Delta w_{ij}(k-1) > 0 \\ \varepsilon_0 & \text{si} \quad \Delta w_{ij}(k-1) = 0 \\ 0 & \text{sinon} \end{cases}$$

Le Hessien H est approximé comme suit :

$$\frac{\partial^2 E^*}{\partial w_{ij}^2}(k) \approx \frac{\frac{\partial E^*}{\partial w_{ij}}(k) - \frac{\partial E^*}{\partial w_{ij}}(k-1)}{\Delta w_{ij}(k-1)}$$

Le coefficient $\alpha_{ij}(k)$ est alors calculé de la manière suivante :

$$\alpha_{ij}(k) = \frac{\frac{\partial E^*}{\partial w_{ij}}(k)}{\frac{\partial E^*}{\partial w_{ij}}(k-1) - \frac{\partial E^*}{\partial w_{ij}}(k)}$$

Pour éviter que la taille du pas n'augmente trop brutalement et pour s'assurer également que la descente suive bien la bonne direction, la contrainte suivante est ajoutée au coefficient $\alpha_{ij}(k)$:

$$\alpha(k) = \begin{cases} \alpha_{max} & \text{si} \quad |\alpha_{ij}(k)| > \alpha_{max} \\ \alpha_{max} & \text{si} \quad \left(\frac{\partial E^*}{\partial w_{ij}}(k-1) - \frac{\partial E^*}{\partial w_{ij}}(k)\right) \Delta w_{ij}(k-1) > 0 \\ \alpha_{ij} & \text{sinon} \end{cases}$$

Les paramètres standard que nous retenons pour la simulation sont les suivants :

- $\varepsilon_0 = 0,1$
- $\alpha_{max} = 1,75$
- $decay = 0,0001$

En comparant la méthode Backprop accélérée avec un moment d'inertie avec les méthodes Rprop et Quickprop, les résultats obtenus sont les suivants :

	Méthode		
	Backprop	Rprop	Quickprop
EQM Min (validation)	1 234,54	1 234,31	1 234,79
Epoque	$367^{\text{ème}}$	$253^{\text{ème}}$	$979^{\text{ème}}$
EQM (test)	1 285,69	1 285,49	1 285,63

La méthode Rprop reste toujours la plus performante avec une convergence plus rapide et une erreur observée sur la base de validation plus faible.

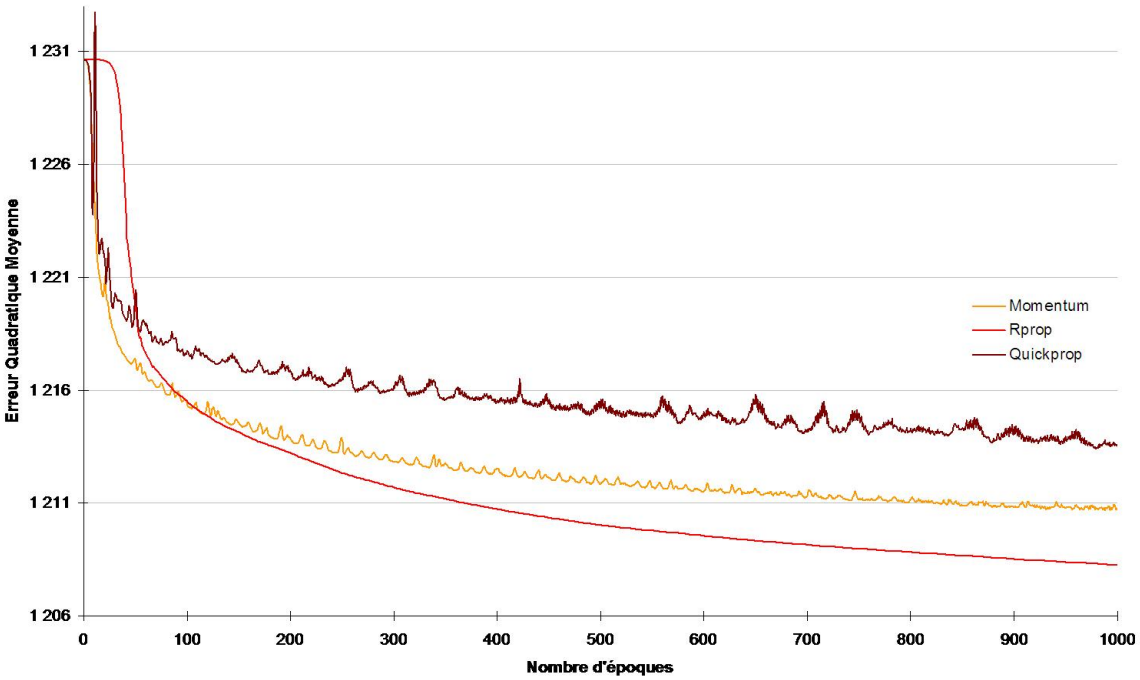


FIG. 15.24 – Entraînement (brut) - RPROP vs Momentum vs Quickprop

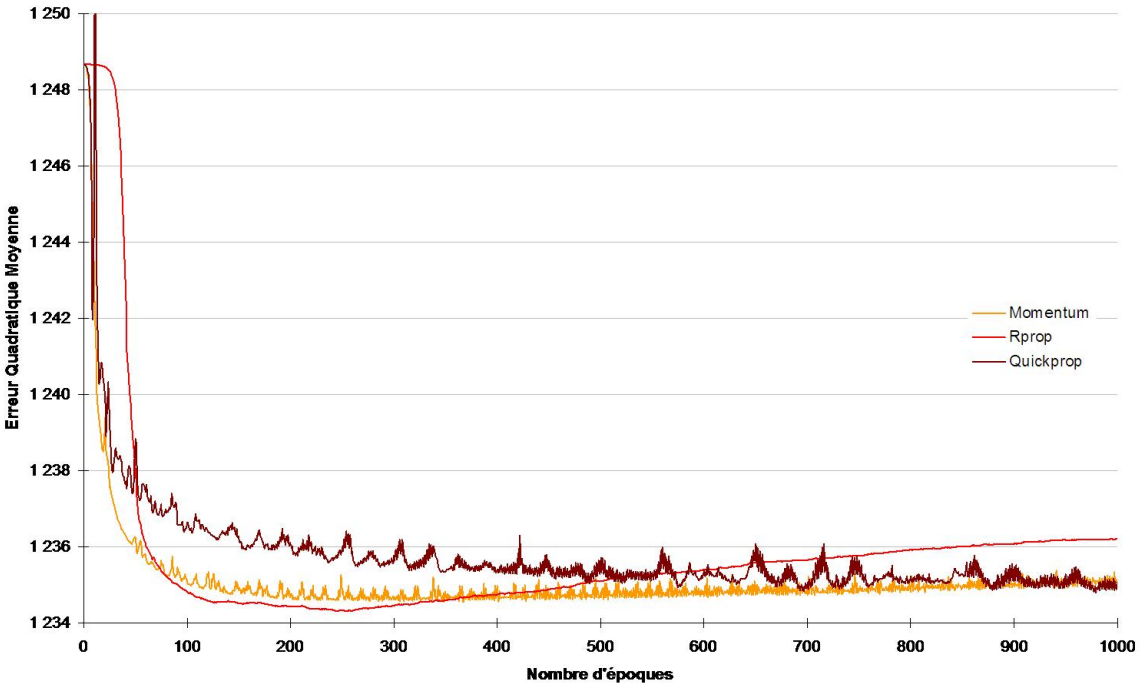


FIG. 15.25 – Validation (brute) - RPROP vs Momentum vs Quickprop

15.4.3 Gradients Conjugués

La méthode des gradient conjugués [TRI88] diffère des méthodes newtoniennes car sa complexité de calcul est plus faible (inférieure à $O(w^2)$). Cette spécificité la rend attractive pour les problématiques de grande dimension. Le constat de départ est que dans l'algorithme de descente du gradient, sous hypothèse d'une erreur quadratique, la même direction est reprise inutilement un grand nombre de fois. L'idée est de réduire le nombre d'itérations en recherchant dans une direction de descente de la meilleure solution. Chaque descente est réitérée de telle sorte à être conjuguée à la précédente.

Voici en exemple de la démarche avec 2 poids :
- Trouver un minimum dans la direction du gradient (inverse) : $\Delta w(1)$
- Trouver un minimum dans la direction de son conjugué $\Delta w(2)$, c'est à dire en conservant la direction du gradient pour conserver son apport tout en minimisant l'erreur : $\Delta w(1) H \Delta w(2) = 0$

Si l'hypothèse d'une erreur quadratique était vérifiée, l'algorithme convergerait avec un nombre d'itérations égal au nombre de poids.

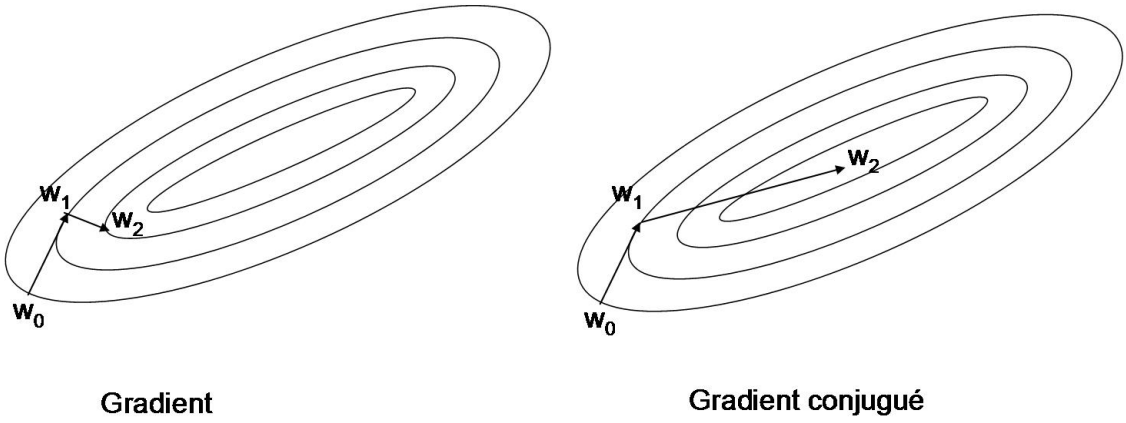


FIG. 15.26 – Principe du gradient conjugué

Comme tout algorithme de descente, chaque itération k est définie par un pas ε_k et une direction $d_k \in \mathbb{R}^p$ de telle sorte que : $\Delta_k = \varepsilon_k d_k$.

Cette technique, plus rapide que la gradient simple reste complexe à programmer car elle combine plusieurs algorithmes permettant d'obtenir le pas et la direction de descente.

Notons g_k le Gradient et H_k le Hessien à la k$^{\text{ème}}$ itération.

La plus grande pente corrigée est obtenue à partir d'un modèle quadratique par itération grâce à la formule [MOL90] :

$$\boxed{d_k = -g_{k-1} + \beta_k d_{k-1}}$$

où le scalaire β_k est ajusté de telle sorte que les directions de recherche d_k et d_{k-1} soient mutuellement conjuguées par rapport à H, c'est à dire qu'elles vérifient le système suivant :

$$d_k^T H d_{k-1} = 0, \forall d_k \neq d_{k-1}$$

Pour une fonction quadratique, la valeur de β_k est la suivante :

$$\beta_k = \frac{g_k^T H_{k-1} d_{k-1}}{d_{k-1}^T H_{k-1} d_{k-1}}$$

L'expression de β_k est approchée par l'une des deux formules suivantes :

- Formule de Polak-Ribière [POL69] : $\beta_k = \frac{g_k^T (g_k - g_{k-1})}{g_{k-1}^2}$
- Formule de Hestenes-Stiefel [HES52] : $\beta_k = \frac{g_k^T (g_k - g_{k-1})}{d_{k-1}^T (g_k - g_{k-1})}$

Quant au choix du pas de descente ε_k , il doit être une bonne estimation d'un minimiseur de $\varepsilon \mapsto E(w_{k-1} + \varepsilon d_k)$ selon différentes techniques (méthodes de dichotomies, de la section dorée, itératives, de Newton,...).

Cette méthode est comparée aux autres variantes d'accélération vues précédemment :

	Méthode			
	Backprop	Rprop	Quickprop	Gradients Conjugués
EQM Min (validation)	1 234,54	1 234,31	1 234,79	1 234,90
Epoque	367$^{\text{ème}}$	253$^{\text{ème}}$	979$^{\text{ème}}$	106$^{\text{ème}}$
EQM (test)	1 285,69	1 285,49	1 285,63	1 285,59

La méthode du Gradient Conjugué est celle qui converge le plus rapidement de toutes celles testées. En revanche, c'est toujours la technique RPROP qui bénéficie de la plus faible erreur.

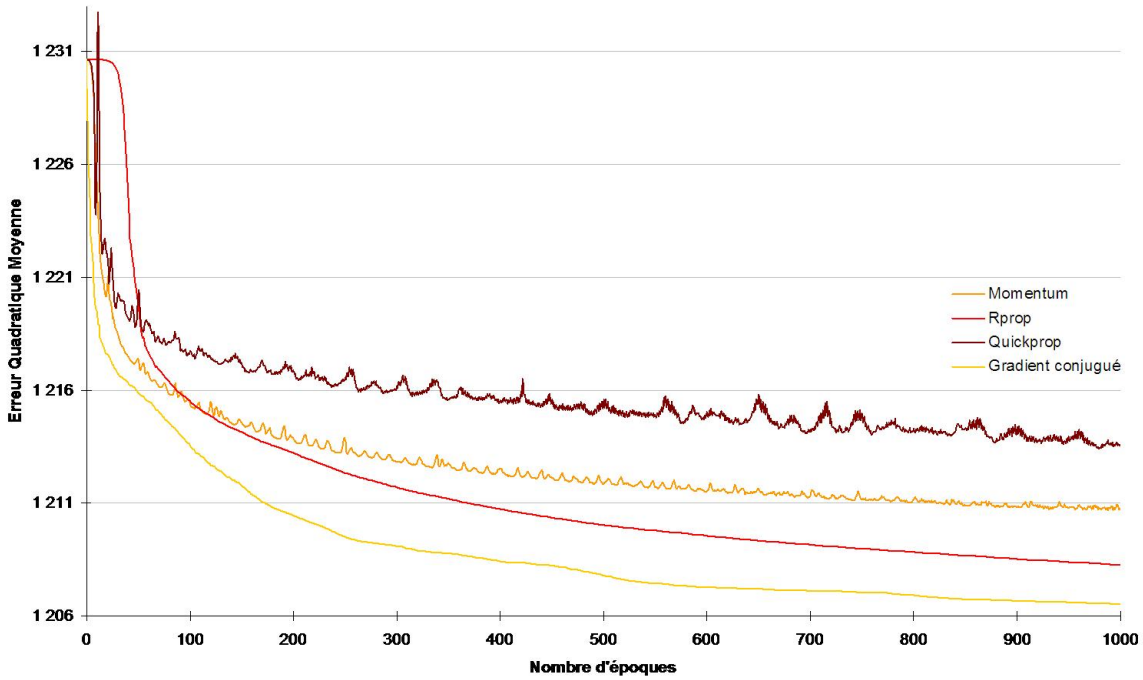


FIG. 15.27 – Entraînement (brut) - RPROP vs Momentum vs Quickprop vs Gradient Conj.

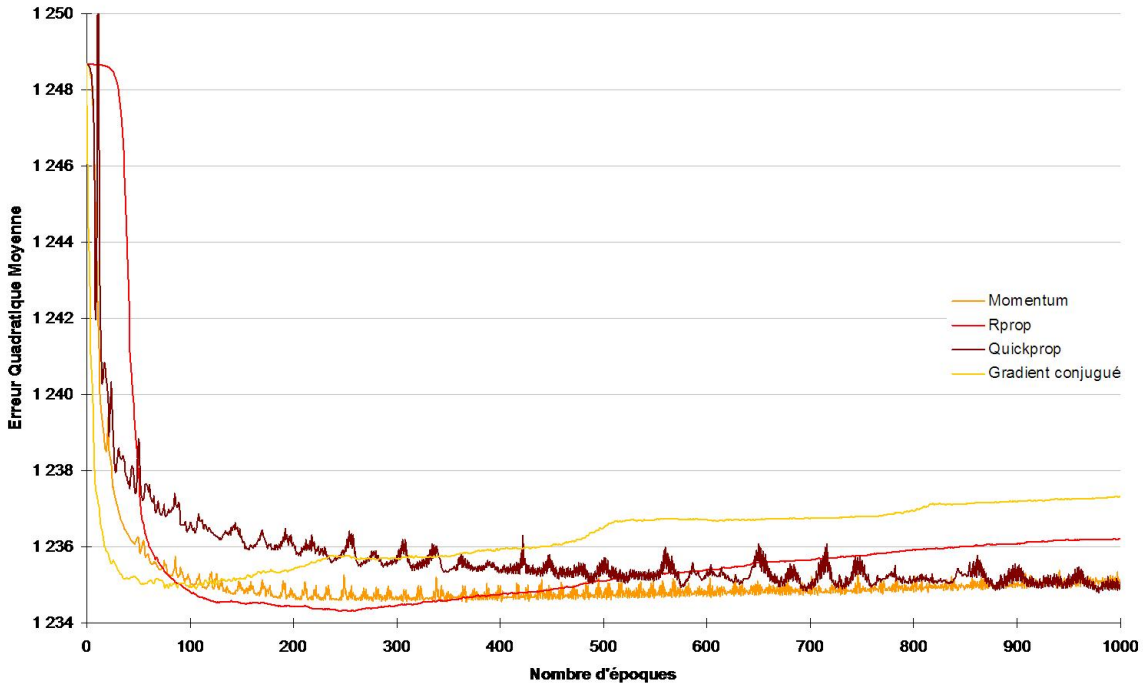


FIG. 15.28 – Validation (brute) - RPROP vs Momentum vs Quickprop vs Gradient Conj.

15.4.4 Levenberg-Marquardt

La méthode de Levenberg-Marquardt [MAR63] est un compromis entre la méthode du gradient et celle de Newton dans le sens où elle s'adapte selon la forme de la fonction erreur.

En notant I la matrice identité, g_k le gradient et H_k la Hessienne, les paramètres sont modifiés comme suit :

$$\boxed{\Delta_k = -\left[H_{k-1} + \lambda_{k-1} I\right]^{-1} g_{k-1}}$$

Plus la valeur de λ_{k-1} est faible plus la méthode se rapproche de celle du gradient, à l'inverse plus elle est forte plus elle s'apparente de celle de Newton.

La valeur de λ_{k-1} est modifiée à chaque itération afin d'être incrémentée si l'erreur croît et décrémentée si elle décroît. Dans un cadre général, le début de la phase d'optimisation est le plus souvent proche de la méthode du gradient et tend vers celle de Newton au voisinage de la solution.

Un des avantages majeur de cette technique réside dans sa vitesse de convergence et la précision de sa solution. La contrepartie est qu'elle nécessite une grande capacité de mémoire informatique proportionnelle au carré du nombre de poids pour le calcul du Hessien, ce qui restreint son utilisation à de petits réseaux (d'usage, moins de quelques centaines de poids). Or, dans le cas présent, le réseau étudié comporte près de 4 000 poids, ce qui ralenti considérablement l'exécution de l'algorithme : une itération dure en moyenne cent fois plus longtemps qu'avec les autres méthodes. Faute de mémoire, nous avons été contraints de stopper celui-ci prématurément à la $18^{\text{ème}}$ itération.

Toujours en comparaison avec les précédentes méthodes, sachant que les temps de traitement d'une itération peuvent différer fortement, voici les résultats obtenus :

	Méthode				
	Backprop	Rprop	Quickprop	Gradients Conjugués	Levenberg-Marquardt
EQM Min (validation)	1 234,54	1 234,31	1 234,79	1 234,90	1 234,93
Epoque	$367^{\text{ème}}$	$253^{\text{ème}}$	$979^{\text{ème}}$	$106^{\text{ème}}$	$18^{\text{ème}}$
EQM (test)	1 285,69	1 285,49	1 285,63	1 285,59	1 285,66

Compte tenu des contraintes fortes en mémoire, cet algorithme n'a servi que pour les phases préliminaires.

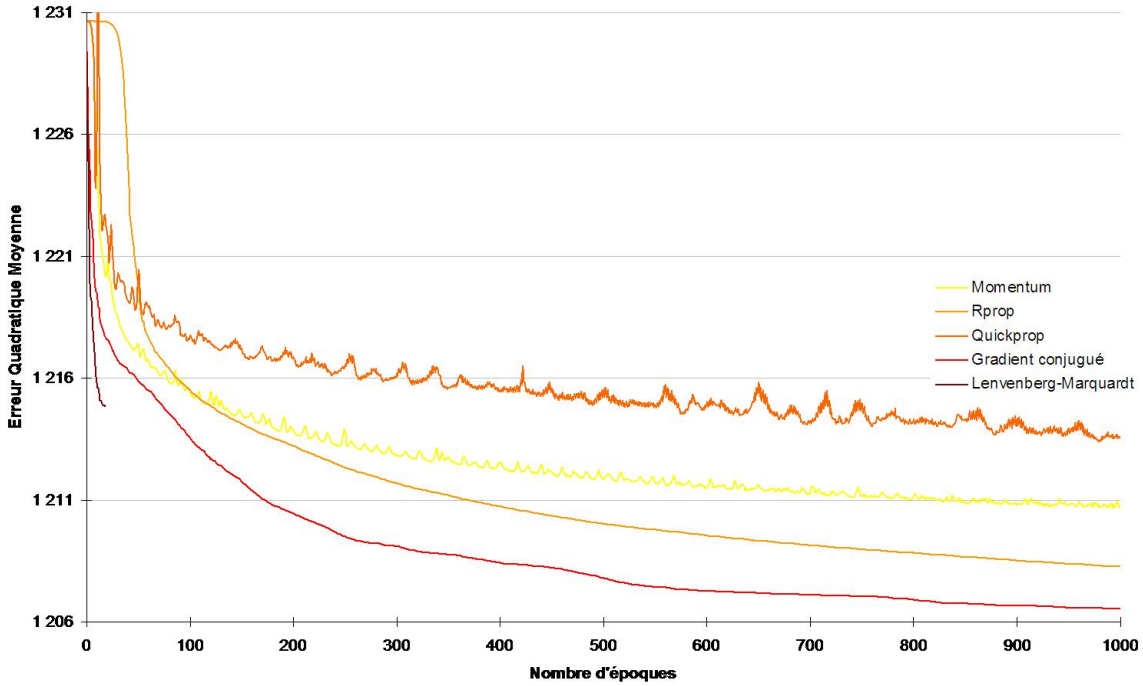


FIG. 15.29 – Ent. (brut) - RPROP vs Mom. vs Quickp. vs Grad. Conj. vs Lev.-M.

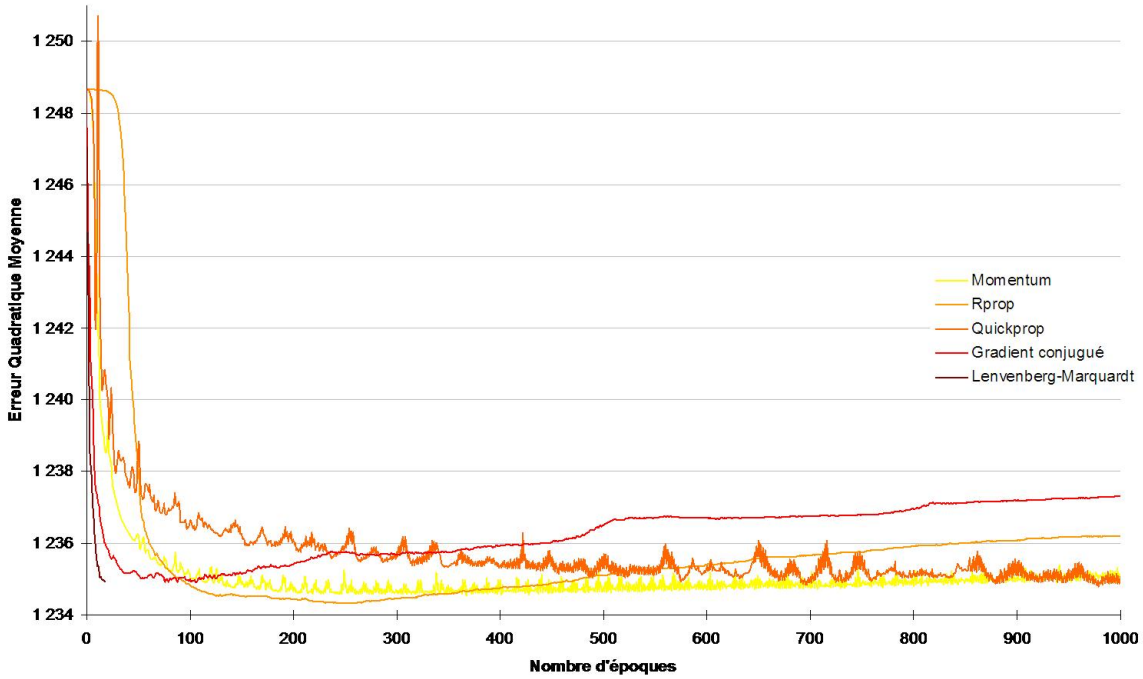


FIG. 15.30 – Valid. (brute) - RPROP vs Mom. vs Quickp. vs Grad. Conj. vs Lev.-M.

15.4.5 Quasi-Newton

Comme nous l'avons indiqué précédemment pour la méthode de Newton, l'inversion de la matrice Hessienne devient vite une opération complexe si la taille du réseau est importante.

Les méthodes dites de Quasi-Newton fournissent une solution dégradée en générant une séquence de matrices symétriques définies positives qui approximent itérativement la matrice Hessienne $M(w) \approx H(w)$ (ou bien son inverse).

En partant de la méthode de Newton :

$$h = \Delta w = -H^{-1}(w)\frac{\partial E(w)}{\partial w}$$

La solution approchée devient :

$$\boxed{h = \Delta w = -M^{-1}(w)\frac{\partial E(w)}{\partial w}}$$

La remise à jour de la matrice $M(w)$ entre deux itérations peut s'obtenir selon différentes méthodes, notamment la Méthode de Davidson-Fletcher-Powell (DFP) et celle de Broyden-Fletcher-Goldfard-Shanno (BFGS) inventée simultanément par Broyden [BRO70], Fletcher [FLE70], Goldfard [GOL70] et Shanno [SHA70] dans les années 1969-1970.

La mise à jour de la matrice M à partir de la méthode BFGS à la k$^{\text{ème}}$ itération s'obtient comme suit :

$$M_k = M_{k-1} + \frac{yy^T}{y^T s} - \frac{M_{k-1}ss^T M_{k-1}}{s^T M_{k-1} s}$$

où $s = w_k - w_{k-1}$ et $y = g_k - g_{k-1}$

Le coût est $O(n^2)$ en mémoire et $O(n^3)$ en calcul.

La méthode BFGS est celle retenue pour les mises à jour dans les simulations avec la méthode de Quasi-Newton.

La synthèse des résultats est la suivante :

	Méthode					
	Backprop	Rprop	Quickprop	Gradients Conjugués	Levenberg-Marquardt	Quasi - Newton
EQM Min (validation)	1 234,54	1 234,31	1 234,79	1 234,90	1 234,93	1 235,51
Epoque	367$^{\text{ème}}$	253$^{\text{ème}}$	979$^{\text{ème}}$	106$^{\text{ème}}$	18$^{\text{ème}}$	31$^{\text{ème}}$
EQM (test)	1 285,69	1 285,49	1 285,63	1 285,59	1 285,66	1 286,42

Le minimum d'erreur est atteint assez rapidement sur la base d'entrainement, cependant cette méthode offre les moins bons résultats en terme d'erreur.

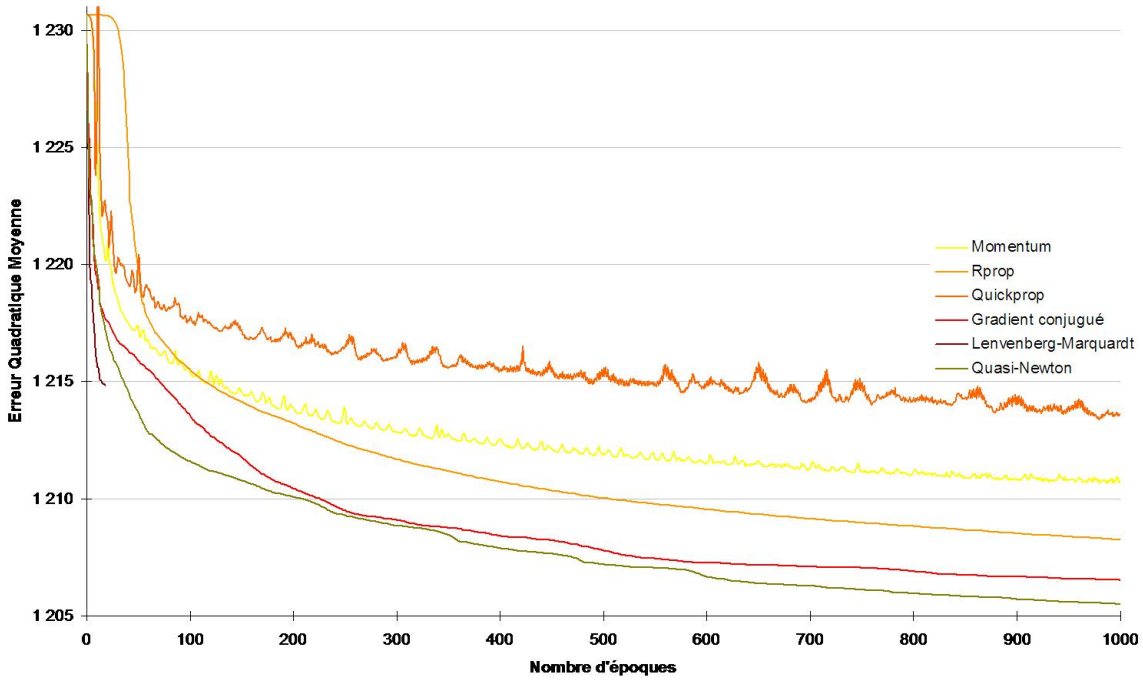


FIG. 15.31 – Ent. (brut) - RPROP vs Mom. vs Quickp. vs Grad. C. vs Lev.-M. vs Q.-New.

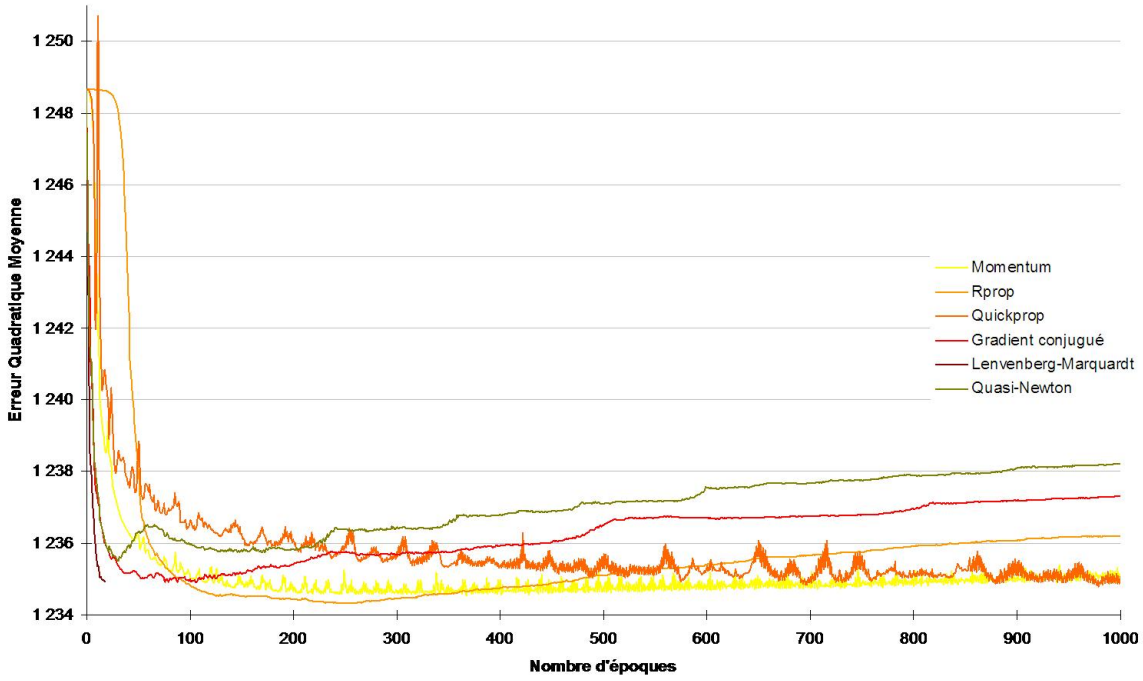


FIG. 15.32 – Valid. (brute) - RPROP vs Mom. vs Quickp. vs Grad. C. vs Lev.-M. vs Q.-New.

15.5 Pénalisation de la fonction coût : Weigth Decay

Cette technique de régularisation sert à pénaliser les grandes valeurs de poids qui dans certains cas peuvent être responsables de surapprentissage [BIS95]. Le principe est de rajouter à la fonction erreur, une pénalité qui croît avec la magnitude des poids. Le réseaux devient alors moins flexible en évitant une spécialisation sur les données d'apprentissage. L'amélioration de la capacité de généralisation de cette méthode a été montrée empiriquement en 1991 par Weigend, Rumelhart et Huberman [WEI91].

La pénalité peut se matérialiser de la façon suivante :

$$P = \lambda \sum_{i}^{N} w_i^2$$

L'apprentissage se fait en minimisant la fonction erreur à laquelle s'ajoute la pénalité :

$$\boxed{E(w_k)' = E(w_k) + P}$$

Une valeur trop faible du taux de régularisation λ fait perdre l'intérêt de régularisation, une valeur trop importante fait tendre les poids vers zéro et inhibe l'apprentissage (les données ne sont plus prises en compte). Le choix de la valeur de λ est donc important pour calibrer correctement la régularisation, en pratique elle ne dépasse pas 0,0001.

La technique de Weight Decay est appliquée à notre modèle avec la méthode Rprop (celle offrant les meilleurs résultats) avec un taux de régularisation de 0,0001. Comme la méthode Rprop avait convergé précédemment à la $253^{\text{ème}}$ itération, nous nous contenterons de 500 époques pour cette nouvelle simulation.

Les résultats obtenus sont les suivants :

	Méthode	
	Rprop sans Weight Decay	Rprop avec Weight Decay
EQM Min (validation)	1 234,31	1 2 34,53
Epoque	$253^{\text{ème}}$	$194^{\text{ème}}$
EQM (test)	1 285,49	1 285,50

L'apport d'une méthode de régularisation a permis d'accélérer la convergence sur la base d'apprentissage les premiers temps mais à partir de la $50^{\text{ème}}$ époque, l'erreur de la simulation sans Weight Decay repasse en dessous.

Sur la base de validation c'est le même constat, l'erreur minimum est atteinte plus rapidement avec la régularisation mais sa valeur reste supérieure. La pente de remontée de l'erreur est moins marquée avec la régularisation sur les dernières itérations. Ceci est conforme avec l'effet recherché, à savoir éviter le sur-apprentissage.

Sur la base de test, les valeurs des erreurs obtenues avec et sans régularisation sont quasi identiques.

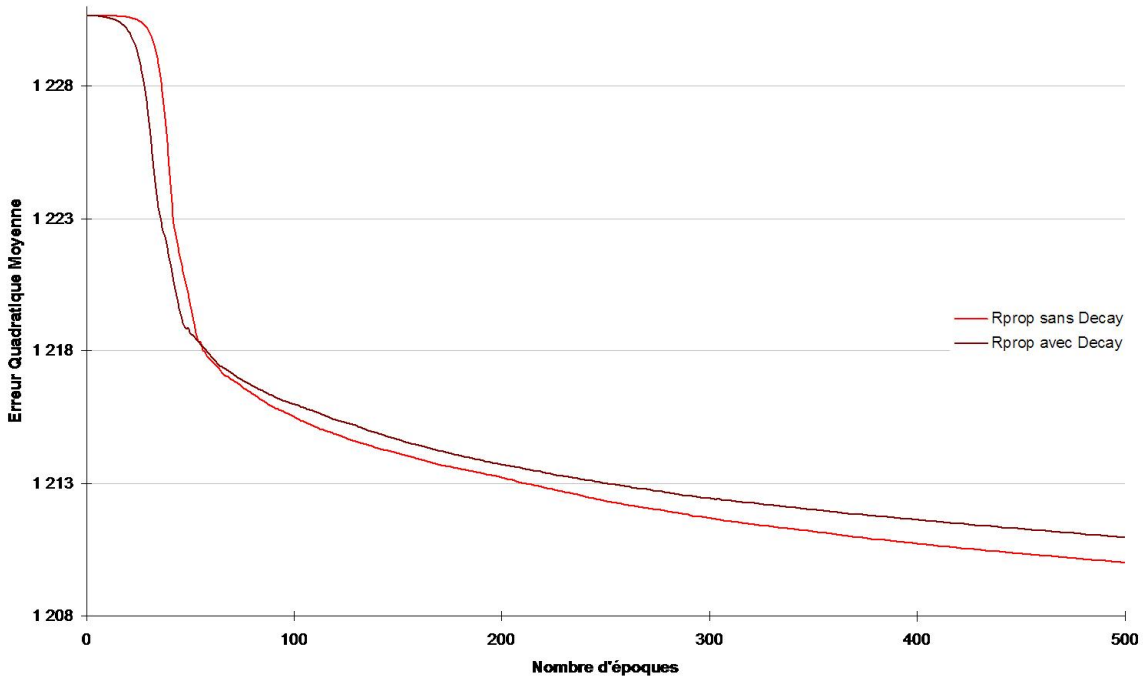


FIG. 15.33 – Entraînement (brut) - RPROP sans Decay vs RPROP avec Decay

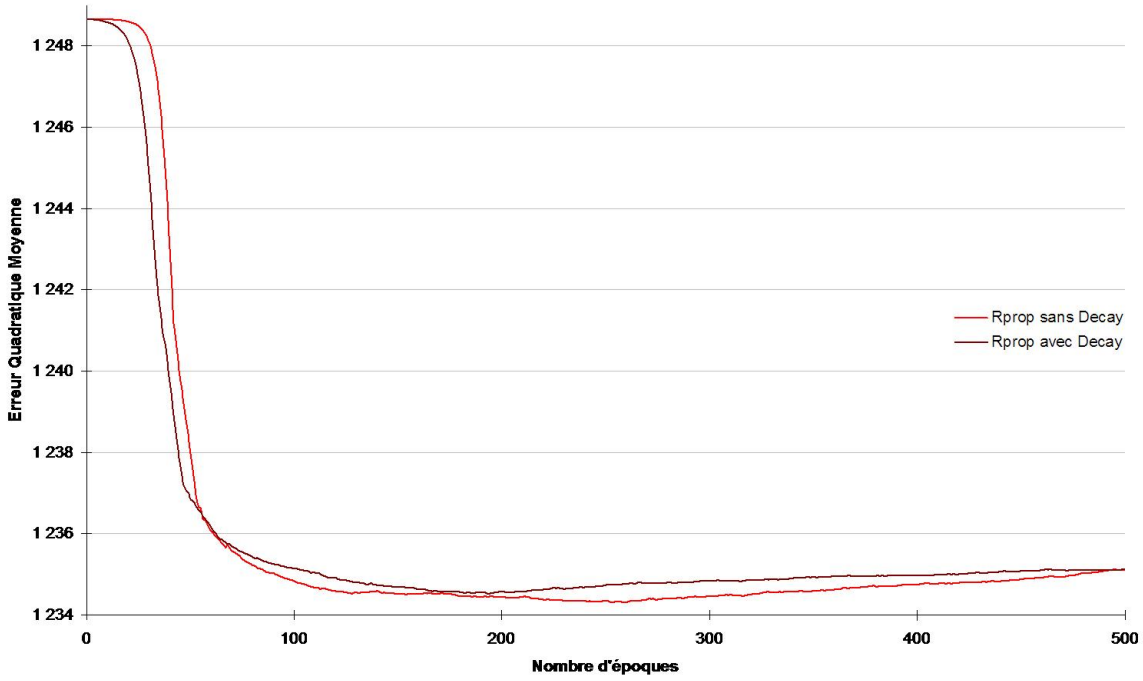


FIG. 15.34 – Validation (brute) - RPROP sans Decay vs RPROP avec Decay

15.6 Fonctions d'activation

En partant toujours du modèle offrant les meilleurs résultats, à savoir celui entraîné avec la méthod RPROP, 3 nouveaux scénarii de simulation sont envisagés avec des fonctions d'activation différentes :

- la fonction tangente hyperbolique
- la loi normale
- la fonction arc tangente
- la fonction logistique

Les résultats ainsi obtenus sont les suivants :

	Fonctions d'activation			
	Tangente Hyperbolique	Loi Normale	Arc Tangente	Fonction Logistique
EQM Min (validation)	1 234,31	1 234,70	1 234,77	1 234,17
Epoque	$253^{\text{ème}}$	$140^{\text{ème}}$	$255^{\text{ème}}$	$294^{\text{ème}}$
EQM (test)	1 285,49	1 285,32	1 285,31	1 285,01

La loi Normale s'adapte plus vite que les autres aux données de la base d'apprentissage, mais généralise moins bien sur la base de validation.

La fonction logistique est celle qui réduit le plus l'erreur sur la base d'apprentissage.

Le modèle le plus performant devient celui entrainé avec la méthode RPROP et une fonction d'activation logistique.

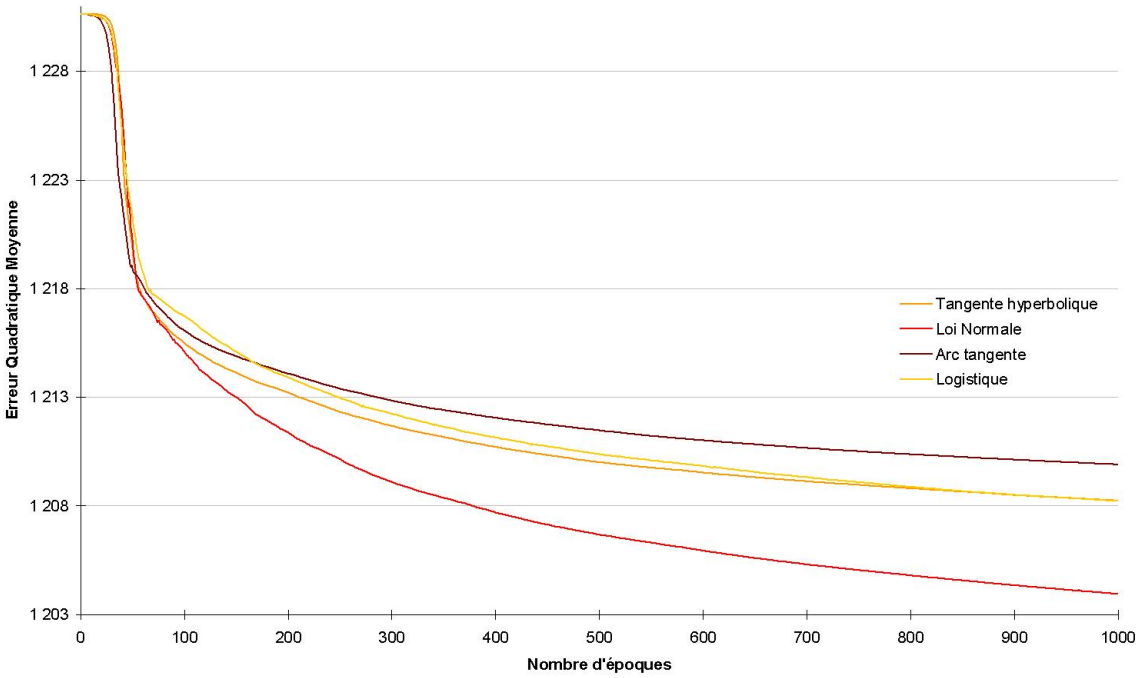


FIG. 15.35 – Entraînement (brut) - Fonctions d'activation

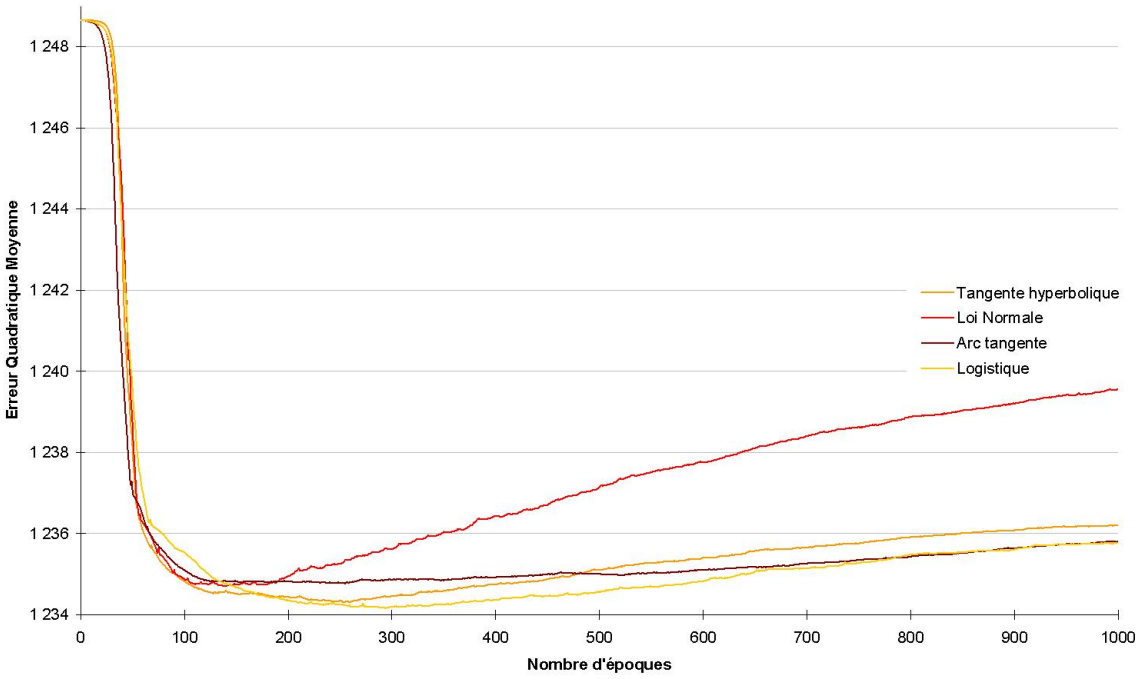


FIG. 15.36 – Validation (brute) - Fonctions d'activation

15.7 Fonction d'erreur à minimiser

La fonction d'erreur quadratique moyenne, utilisée jusqu'à présent, pénalise fortement les écarts importants entre les valeur prédites et celles attendues à cause de l'élévation au carré. La fonction de Huber (parfois appelée parabole tronquée) permet d'atténuer ce phénomène en réduisant la magnitude des valeurs extrêmes. La fonction erreur $\rho(z)$ est quadratique pour les faibles résidus et similaire à la $|y - \mu|$ pour les erreurs importantes. L'impact des valeurs aberrantes ou extrêmes sur le modèle est alors moindre, le modèle gagne ainsi en robustesse.

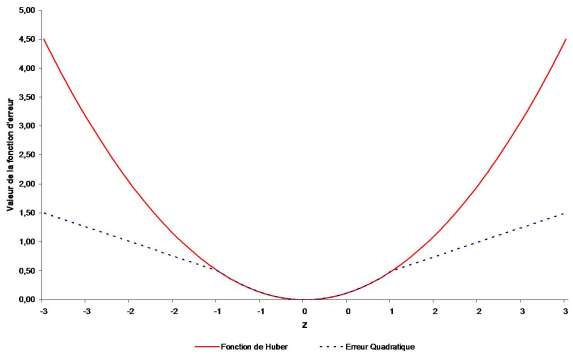


FIG. 15.37 – Fonction de Huber

La fonction erreur individuelle $\rho(z)$ est égale à :

$$\rho(z) = \begin{cases} \frac{z^2}{2} & \text{si} \quad |z| < 1 \\ -|z| - \frac{1}{2} & \text{si} \quad |z| >= 1 \end{cases}$$

où y est valeur cible, μ est la valeur prédite, σ est la dispersion, M est la constante de réglage de la M-estimation et $z = \frac{y-\mu}{\sigma M}$.

Jusqu'à présent, les meilleurs résultats ont été obtenus avec la technique d'apprentissage Rprop et une fonction d'activation logistique. Toujours en suivant la même démarche que précédemment, en partant de ce modèle, deux scnénarii de fonctions d'erreur sont comparés : la fonction d'erreur quadratique vs la fonction de Huber.

En paramétrant la fonction de Huber de façon standard avec les valeurs par défaut : $\sigma = 1$ et $M = 1,5$, les résultats ainsi obtenus sont les suivants :

	Fonctions d'erreur	
	Erreur Quadratique	Fonction de Huber
EQM Min (validation)	1 234,17	1 241,40
Epoque	294$^{\text{ème}}$	51$^{\text{ème}}$
EQM (test)	1 285,01	1 291,47

Sur la base d'apprentissage, comme sur la base de test, l'erreur commise est plus importante avec la fonction de Huber. L'usage de la fonction de Huber n'améliore pas la précision des résultats, celle-ci ne sera pas retenue. Le précédent modèle reste encore le plus performant.

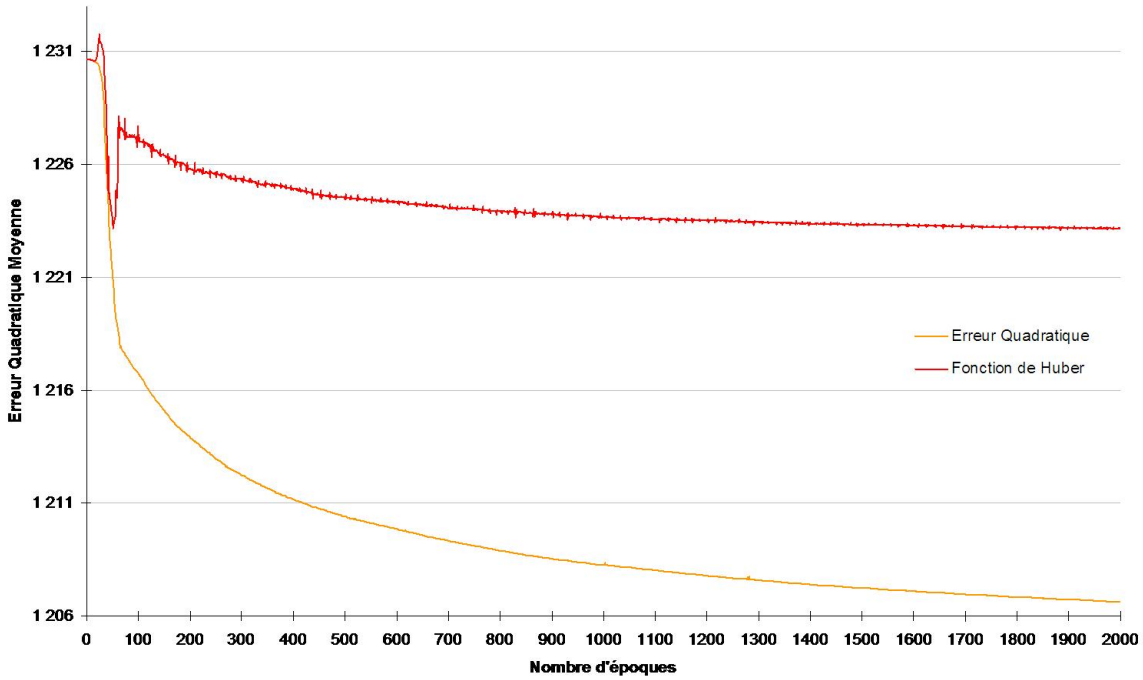


FIG. 15.38 – Entraînement (brut) - Fonctions d'erreur

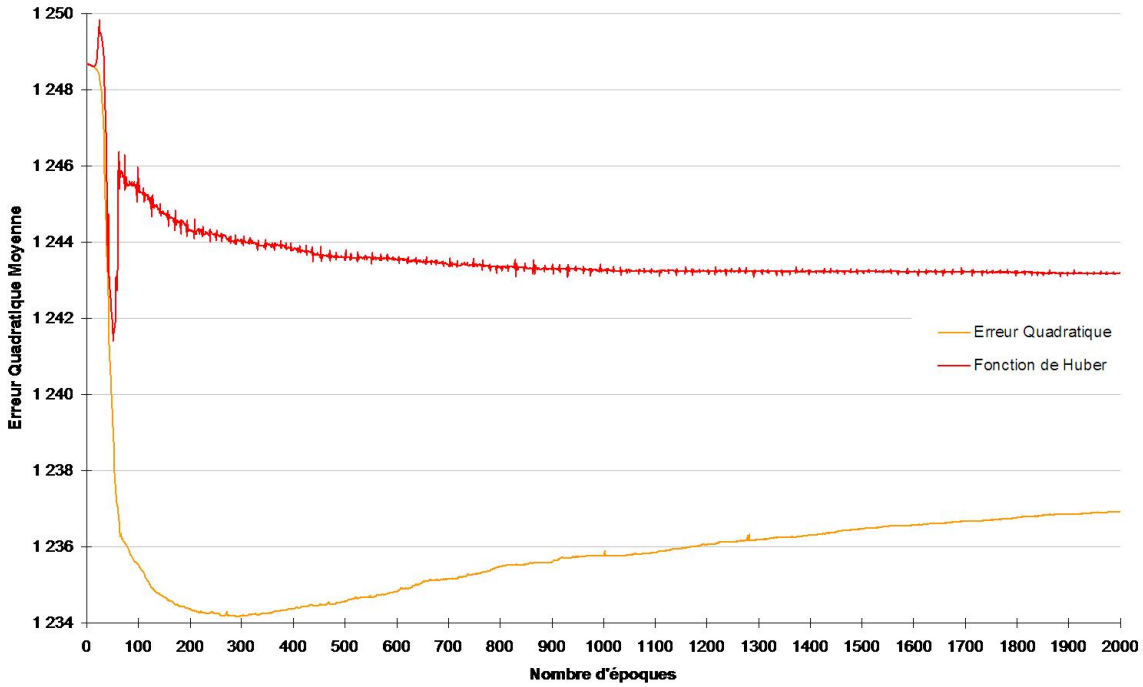


FIG. 15.39 – Validation (brute) - Fonctions d'erreur

15.8 Liaisons directes

Précédemment, l'information entre la couche d'entrée et celle de sortie transitait systématiquement par la couche cachée. Une variante consiste à rajouter des liaisons directes entre les couches d'entrée et de sortie. Cette pratique permet d'ajouter à la prédiction une partie linéaire. Une régression linéaire est alors intégrée au modèle en plus des vertus non linéaires que lui confère les réseaux de neurones. Ce compromis peut parfois faire gagner en performance. La contrepartie réside dans l'augmentation du nombre de poids au détriment de la robustesse.

Cette alternative est testée ici avec toujours la fonction d'activation logistique et un apprentissage RPROP.

Le comparatif avec et sans liaisons directes donne les résultats suivants :

	Liaisons directes	
	Sans	Avec
EQM Min (validation)	1 234,17	1 234,36
Epoque	$294^{\text{ème}}$	$312^{\text{ème}}$
EQM (test)	1 285,01	1 285,27

La version avec liaisons directes converge plus rapidement dans les premiers instants, mais au final la version sans liaison directe atteint un niveau d'erreur inférieur sur les bases de validation et de test.

L'ajout de liaisons directes ne procure donc pas de meilleurs résultats, cette variante n'est alors pas retenue.

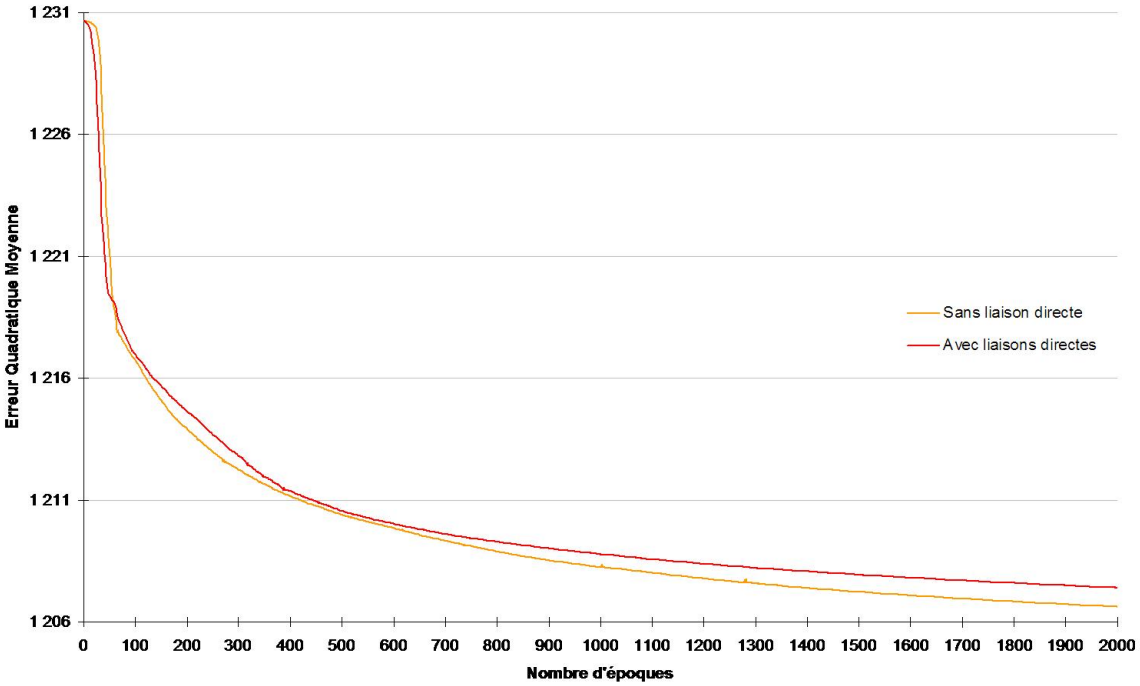


FIG. 15.40 – Entraînement (brut) - Liaisons directes

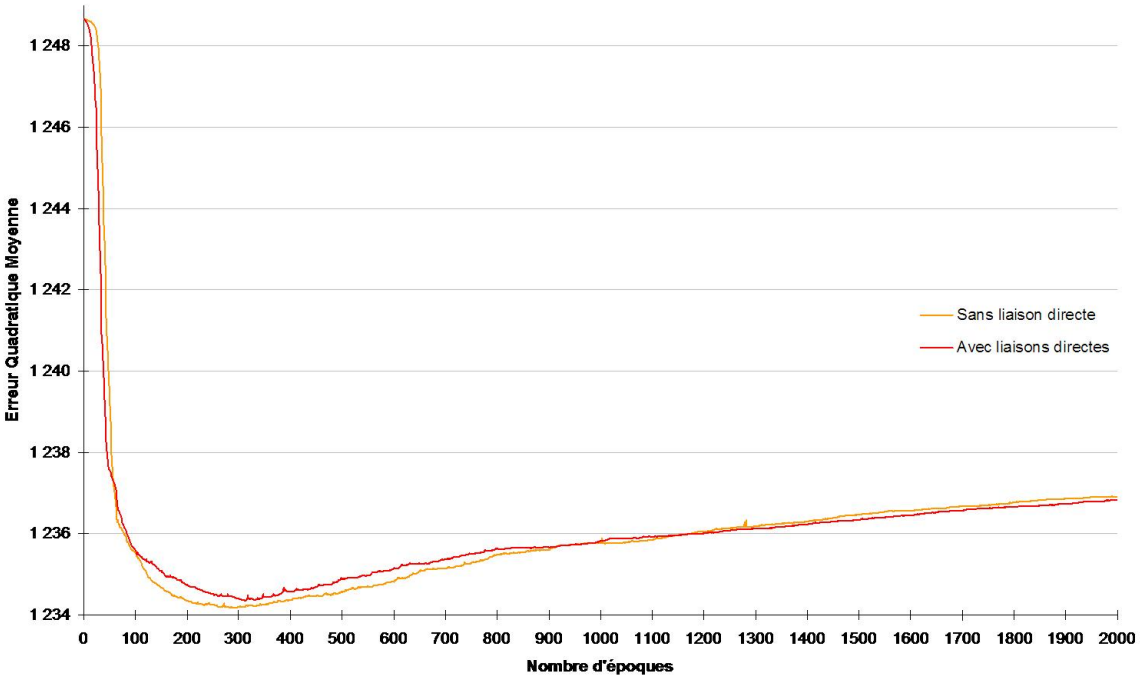


FIG. 15.41 – Validation (brute) - Liaisons directes

15.9 Elagage

La technique dite d'élagage ou de *pruning* en anglais a pour objectif de limiter le sur-apprentissage en réduisant la complexité du modèle. Le principe est de supprimer, une fois l'apprentissage terminé, les connexions du modèle ayant la plus faible influence sur l'erreur de sortie du réseau.

Nous nous sommes contentés de sélectionner les variables d'entrée qui détenaient les valeurs de poids les plus faibles pour les ôter du modèle. Pour chacune d'entre elle, les poids de leurs liaisons avec toutes les cellules de la couche cachée ont été sommés. Arbitrairement, celles qui totalisaient une valeur inférieure à un, ont été écartées du réseau.

Variables	Liaisons							
	H11	H110	H111	H112	...	H18	H19	Total
Régime=Alsace	0,0258	3,9588	0,1783	0,0607	...	0,0182	0,005	0,2281
Région=Ch.-Ardenne	0,1828	0,0184	0,1543	0,0289	...	0,2906	0,0553	0,2619
VP3	0,0335	0,0004	0,0207	0,3271	...	0,1187	0,0012	0,3509
Choix=Base	0,0342	0,0118	1,028	0,1828	...	0,1558	0,054	0,4553
VP4	0,1645	0,0221	0,7295	0,0652	...	0,0103	0,268	0,4835
VS3	0,1024	3,5795	0,3897	0,0117	...	0,0247	0,07	0,5233
VS2	7,8623	0,037	0,9163	0,0879	...	0,0148	0,2884	0,6612
VP2	1,6382	0,1882	0,4035	0,0313	...	1,6758	4,7902	0,7628
VS1	0,118	0,7539	0,0579	0,8255	...	0,2296	0,4142	0,7665
Région=Fr.-Comté	0,0924	0,1215	0,0241	0,2916	...	1,6309	0,1238	0,8276

Onze variables totalisent des poids inférieurs à un. Deux d'entre elles, qui indiquent des régions, sont issues d'une seule et même variable catégorielle comportant 21 modalités, la variable région. A l'exception de ces deux variables, les neuf restantes sont supprimées du modèle.

Avec toujours la fonction d'activation logistique et un apprentissage RPROP, le comparatif est réalisé entre les simulations avec et sans élagage.

Les résultats obtenus sont les suivants :

	Elagage	
	Sans	Avec
EQM Min (validation)	1 234,17	1 234,47
Epoque	$294^{\text{ème}}$	$243^{\text{ème}}$
EQM (test)	1 285,01	1 285,34

Cette variante n'apporte pas de gain d'efficacité et ne sera pas retenue.

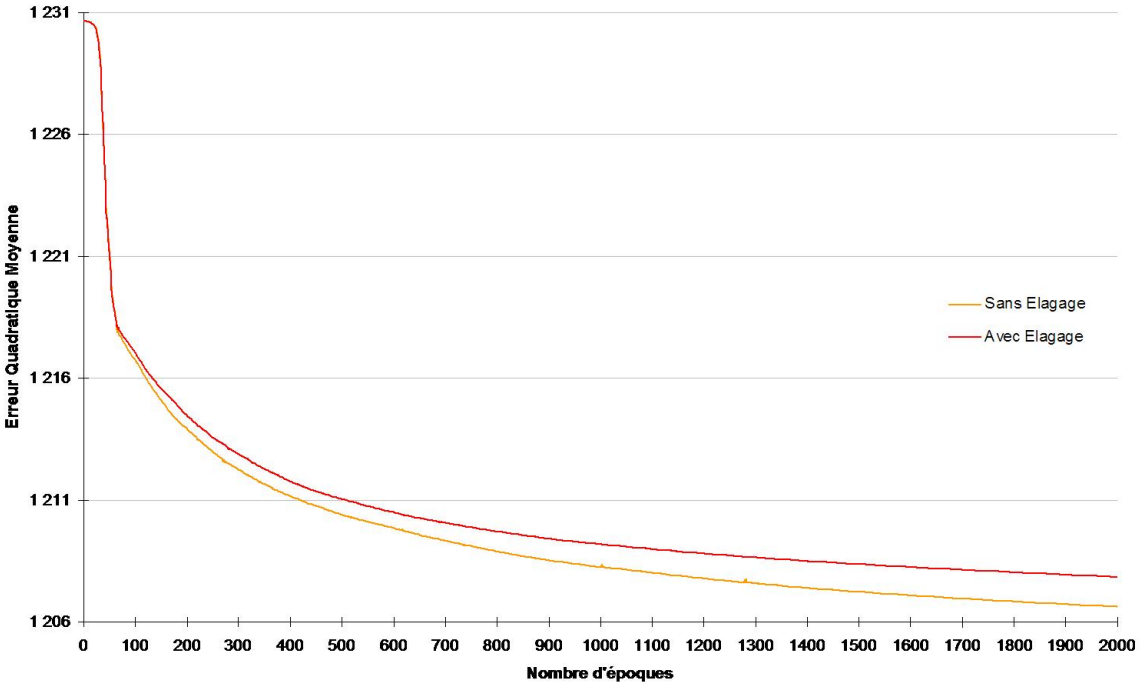


FIG. 15.42 – Entraînement (brut) - Elagage

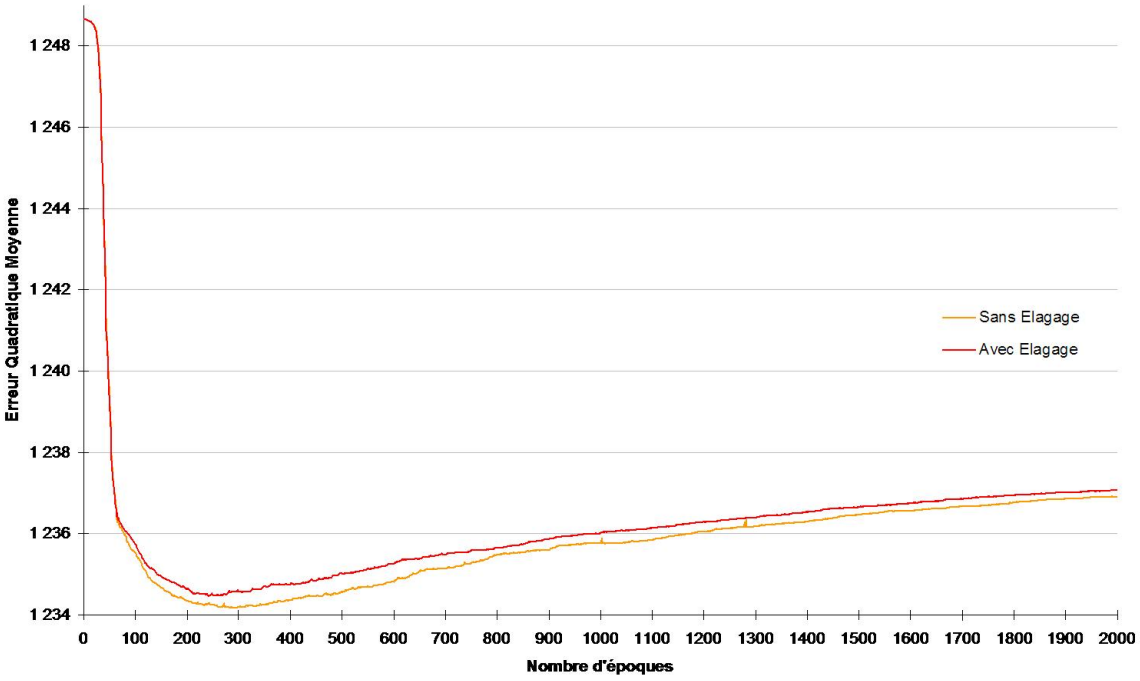


FIG. 15.43 – Elagage (brute) - Liaisons directes

Chapitre 16

Comparatif et validation du modèle

16.1 Comparaison avec un Modèle Linéaire Généralisé (G.L.M.)

16.1.1 Présentation générale

Le terme de "modèles linéaires généralisés" fut introduit par Nelder et Wedderburn en 1972 [NEL72], cette famille de modèle fut ensuite exposée plus précisément en 1989 par Mc Cullagh et Nelder [CUL89].

Ces modèles servent à étudier la liaison entre une variable à expliquer Y et un ensemble de variables explicatives X_i.

Ils recouvrent plusieurs sous-catégories :

- le modèle linéaire général (régression multiple et analyses de la variance et covariance)
- le modèle log-linéaire
- la régression logistique
- la régression de Poisson

Ces modèles présentent trois aspects :

- la composante aléatoire, définie par une loi de probabilité appartenant à la famille des lois exponentielles, qui caractérise la variable à expliquer :

$$f(y_i, \theta_i, \phi, \omega_i) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\omega_i + c(y_i, \phi, \omega_i)\right)$$

  Les fonctions a, b et c dépendent du type de loi utilisée. θ_i est appelé le paramètre canonique, il est inconnu et fonction de l'espérance. ω_i est un poids. La famille exponentielle comprend notamment les lois de probabilités Normale, Binomiale, Poisson, Gamma et Gamma Inverse. Pour la loi Normale, $\theta = \mu$, $b(\theta) = \theta^2/2$, $\phi = \sigma^2$ et $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)$

- la composante déterministe, s'exprime sous forme de combinaison linéaire :

$$\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_p X_p$$

Les interactions entre les variables peuvent être prises en compte en effectuant un changement de variable du type : $X_3 = X_1 X_2$.
Sur le même principe, des effets non linéaires d'ordre supérieurs peuvent être représentés en posant par exemple : $X_4 = X_1^2$.

- le lien $g(\mu)$ entre l'espérance de la variable à expliquer notée μ et la composante déterministe :

$$g(\mu) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_p X_p$$

Dans la plupart des cas, les équations maximisant la vraisemblance pour déterminer les paramètres ne sont pas linéaires et ne peuvent pas être reformulées de façon à simplifier le problème. En pratique, les estimations sont obtenues grâce à des logiciels utilisant des algorithmes itératifs.

16.1.2 Régression linéaire multiple

Le modèle de régression linéaire multiple généralise le cas simple avec p variables explicatives et n observations. Il se présente de la manière suivante $\forall i = 1, ..., n$:

$$Y_i = \alpha_0 + \alpha_1 X_{i,1} + \alpha_2 X_{i,2} + ... + \alpha_p X_{i,p} + \varepsilon_i$$

où Y est la variable à expliquer, $X_{i,j}$ les variables explicatives, α_j les coefficients de régression et ε_i l'erreur du modèle.

Matriciellement, l'écriture se condense en $Y = X\alpha + \varepsilon$:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

16.1.3 Méthode d'estimation des Moindres Carrés Ordinaires (MCO)

En partant du modèle complet :

$$y_i = \alpha_0 + \alpha_1 x_{i,1} + ... + \alpha_p x_{i,p} + \varepsilon_i$$

Le but est de trouver une estimation des paramètres $\widehat{a}_i$ de telle sorte que :

$$\widehat{y_i} = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_{i,1} + ... + \widehat{\alpha}_p x_{i,p}$$

Le principe repose sur la minimisation de la somme des erreurs quadratiques entre l'observation et l'estimation $\varepsilon_i = y_i - \widehat{y}i$:

$$\min_{\widehat{\alpha}_0,..,\widehat{\alpha}_p} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\widehat{\alpha}_0,..,\widehat{\alpha}_p} \sum_{i=1}^{n} (y_i - \widehat{\alpha}_0 - \widehat{\alpha}_1 x_{i,1} - ... - \widehat{\alpha}_p x_{i,p})^2$$

Ce qui se traduit par la recherche des solutions du système suivant :

$$\frac{\partial \left(\sum \widehat{\varepsilon}_i^2\right)}{\partial \widehat{\alpha}_j} = 0$$

Le système de $p+1$ équations normales ainsi obtenu se résout en calculant $\widehat{\alpha}$ comme suit :

$$Y = X\widehat{\alpha}$$

$$X^T Y = X^T X \widehat{\alpha}$$

$$\widehat{\alpha} = (X^T X)^{-1} X^T Y$$

$\widehat{\alpha}$ est un estimateur sans biais et de variance minimale de α

16.1.4 Application

Dans notre problématique, la fonction à prédire étant continue, nous avons choisi d'effectuer une régression linéaire avec une fonction lien identité. Les variables catégorielles sont codées selon la méthode GLM. C'est-à-dire que chaque valeur est traduite par une nouvelle variable qui vaut "1" ou "0" selon que la variable corresponde ou non à cette valeur.

Par exemple, si le type de bénéficiaire C_BENEF est "$adhrent$", il sera codé par : "$C\_BENEF = adhrent$" $= 1$, "$C\_BENEF = conjoint$" $= 0$ et "$C\_BENEF = Enfant$" $= 0$.

Valeurs de	Nouvelles variables		
C_BENEF	Adhérent	Conjoint	Enfant
"Adhérent"	1	0	0
"Conjoint"	0	1	0
"Enfant"	0	0	1

Compte tenu du grand nombre de variables de notre modèle, une infinité de modèles linéaires peuvent être envisagés. Les différentes interactions entre les variables ainsi que les effets polynomiaux engendrent une multitude de combinaisons possibles. Pour pallier à ce problème, trois méthodes de sélection automatique de variables existent :

- la méthode d'élimination, (Backward) : Tous les effets candidats sont présents initialement. Les effets jugés non significatifs sont retirés au fur et à mesure.

- la méthode de sélection (Forward) : Aucun effet n'est présent initialement. A chaque itération, les variables pertinentes sont ajoutées une à une.

- la méthode mixte (Stepwise) : C'est un mix des deux où alternent successivement un pas de sélection avec un pas d'élimination.

Ces méthodes continuent jusqu'à ce qu'aucune variable ne puisse entrer ou sortir du modèle ou bien qu'un critère d'arrêt soit satisfait.

La contribution des variables au modèle est déterminée à l'aide de la statistique F de Fisher :

$$F = \frac{\frac{RSS_1 - RSS_2}{p_1 - p_2}}{\frac{RSS_2}{n - p_2}}$$

où RSS_1 et RSS_2 représentent les sommes des carrés des résidus des modèles 1 et 2, p_1 et p_2 leur nombre de paramètres tels que $p_1 < p_2$ et n le nombre d'observations.

Cette statistique indique dans quelle mesure le modèle 2 apporte un intérêt significatif par rapport au modèle 1. La $p\text{-}value$ de la statistique F : $p\text{-}value_F$ définit ainsi un seuil de contribution des variables.

Ne pouvant tester l'intégralité des modèles, nous avons dû procéder à une préselection de variables candidates.

Les modèles que nous avons choisi de tester sont les suivants :

- tous les facteurs simples
- tous les facteurs simples et toutes les interactions possibles d'ordre 2
- la méthode Backward avec un seuil de sortie de $p\text{-}value_F > 0,05$, systématiquement tous les facteurs simples, toutes les interactions entre $SEXE$, AGE, C_BENEF et PDA, toutes les interactions entre les garanties, AGE^2 et $AGE \times C\_BENEF \times SEXE$
- la méthode Forward avec un seuil d'entrée de $p\text{-}value_F < 0,05$ et les mêmes variables que précédemment
- la méthode Stepwise avec un seuil d'entrée de $p\text{-}value_F < 0,05$, un seuil de sortie de $p\text{-}value_F > 0,05$ et les mêmes variables que précédemment

Le choix des variables candidates dans les méthodes automatiques correspond à celles qui ont, a priori, le plus d'effet sur la consommation médicale. Par exemple, les niveau de garantie des prothèses dentaires PDA est réputé être le plus représentatif de la qualité d'un régime frais de santé. Il a donc été retenu dans le choix des interactions. De même pour l'âge, le sexe et le type de bénéficiaire qui sont des facteurs importants.

Afin de respecter une cohérence avec le modèle de réseau de neurones, le critère de sélection du modèle sera celui dont l'Erreur Quadratique Moyenne sur la base de validation sera la plus faible. Pour les méthodes automatiques, le dernier modèle de l'algorithme ne sera donc pas forcément celui retenu.

Les résultats ainsi obtenus sont les suivants :

	Type de modèle				
	Facteurs simples	+ Interac. d'ordre 2	Méthode Backward	Méthode Forward	Méthode Stepwise
EQM Min (validation)	1 237,18	1 237,33	1 235,91	1 235,88	1 235,88
EQM (test)	1 288,09	1 288,30	1 286,64	1 286,82	1 286,82

Dans le cas présent, les méthodes de sélection automatique apportent un intéret par rapport aux méthodes simples en procurant des résultats de meilleure qualité.

Les méthodes Forward et Stepwise aboutissent au même modèle qui comprend tous les effets simples et les interactions suivantes :

- $AGE \times C\_BENEF$
- $SEXE \times C\_BENEF$
- $AGE \times PDA$
- $AGE \times SEXE \times C\_BENEF$

La méthode Backward est celle qui donne les meilleurs résultats. Elle comporte tous les effets simples ainsi que les interactions suivantes :

- $AGE \times C\_BENEF$
- $SEXE \times C\_BENEF$
- $SEXE \times PDA$
- $PDA \times C\_BENEF$
- $PDA \times AGE$
- $AGE \times AGE$
- $PDA \times MON$
- $VER \times MON$
- $VP \times MON$
- $VP1 \times MON$
- $VP2 \times MON$
- $VP4 \times MON$
- $VS1 \times MON$
- $VS3 \times MON$
- $VS4 \times MON$
- $AGE \times SEXE \times C\_BENEF$

Les statistiques correspondantes sont les suivantes :

Statistiques	Entrainement	Validation	Test
Akaike's Information Criterion	4079155.5333	.	.
Average Squared Error	1482123.5628	1527473.0647	1655430.5614
Average Error Function	1482123.5628	1527473.0647	1655430.5614
Degrees of Freedom for Error	286971	.	.
Model Degrees of Freedom	98	.	.
Total Degrees of Freedom	287069	.	.
Divisor for ASE	287069	215302	215302
Error Function	425471729058	328868005773	356417510739
Final Prediction Error	1483135.8471	.	.
Maximum Absolute Error	92963.413716	89226.994339	194872.49372
Mean Square Error	1482629.7049	1527473.0647	1655430.5614
Sum of Frequencies	287069	215302	215302
Number of Estimate Weights	98	.	.
Root Average Sum of Squares	1217.4249722	1235.9098125	1286.6353646
Root Final Prediction Error	1217.8406493	.	.
Root Mean Squared Error	1217.6328285	1235.9098125	1286.6353646
Schwarz's Bayesian Criterion	4080191.1461	.	.
Sum of Squared Errors	425471729058	328868005773	356417510739
Sum of Case Weights Times Freq	287069	215302	215302

Les résultats de ce modèle n'ont pas permis de battre les performances de la modélisation neuronale. Sur la base de test, la valeur de l'Erreur Quadratique Moyenne est de 1 286,64 sur le modèle G.L.M. contre 1 285,01 sur le réseau de neurones, ce qui conforte notre choix.

Un effet "T-scores" se définit comme étant égal à l'estimation du paramètre divisé par son erreur standard. Les scores sont ensuite classés de façon décroissante selon leur valeur absolue. Ici, les trente premiers d'entre eux sont représentés. Les histogrammes teintés en orange sont signés positivement et ceux en bleu négativement.

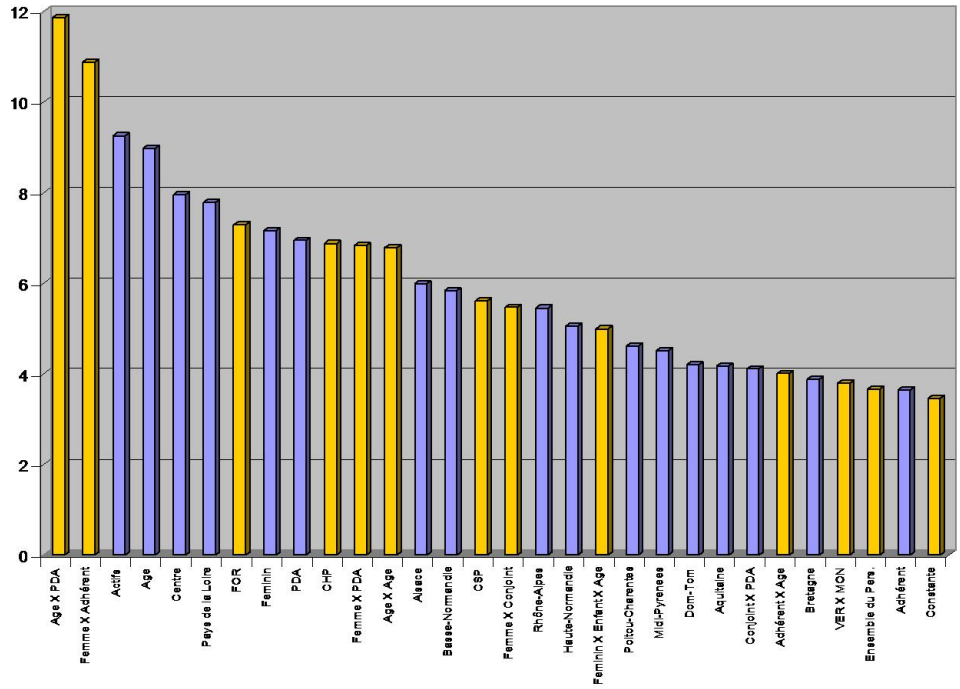


FIG. 16.1 – Effets T-Scores

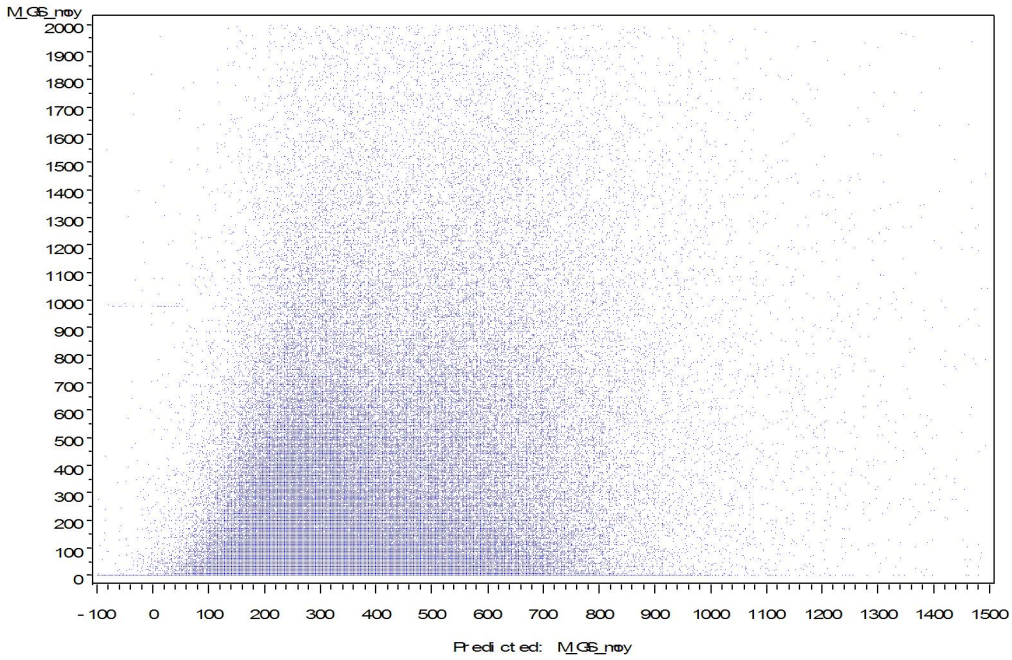


FIG. 16.2 – Représentation de la valeur prédite / valeur attendue

16.2 Comparaison avec un Modèle Memory-Based Reasoning

16.2.1 Principe

Cette méthode est appelée aussi k plus proches voisins (en anglais, K.N.N. K-Nearest Neighbor) inventée en 1986 par Stanfill et Waltz [STA86] puis reprise par Weiss et Kulikowski en 1990 [WEI90].

L'idée de base est d'utiliser les exemples déjà connus pour effectuer une prédiction de manière simple sur des nouveaux exemples. Cette technique est adaptée pour des problèmes de classification ou de régression d'une fonction inconnue.

La spécificité de cette méthode est qu'aucun modèle n'est induit à partir des données. L'intégralité des exemples est stockée en mémoire en l'état.

Pour prédire à partir d'un nouvel exemple, l'algorithme sélectionne ses k plus proches voisins dans la base mémorisée, puis la réponse affecte la réponse la plus fréquente dans le cadre d'une classification ou la valeur moyenne de la fonction à prédire.

Les deux seuls paramètres sont le nombre des voisins les plus proches : k et la fonction de similarité qui permet de définir un critère de proximité entre les exemples :

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

16.2.2 Application

Cette technique est testée avec la distance euclidienne et k=16 (nombre de voisins les plus proches). Une des méthodes de recherche des plus proches voisins les plus rapide intitulée Kd-Tree (K Dimentionality Tree) fut proposée en 1975 par J.L. Bentley. Cet algorithme génére un arbre binaire qui partitionne les données. La technique qui sera utilisée est une modification de celle-ci, la méthode Rd-Tree (Reduced Dimentionality Tree), elle permet d'accroître ses performances.

Les Erreurs Quadratiques Moyennes obtenues sur les bases d'entraînement, de validation et de test ont respectivement les valeurs suivantes : 1 263,41, 1 319,98 et 1 363,90.

	Modélisation		
	Réseau de neurones	G.L.M.	Memory Based Reasoning
EQM Min (validation)	1 234,17	1 235,91	1 319,98
EQM (test)	1 285,01	1 286,64	1 363,90

Le modèle de réseau de neurone reste celui qui offre les meilleures performances. Sur la base de test, la valeur de l'Erreur Quadratique Moyenne reste toujours plus faible sur le modèle de réseau de neurones avec $1285,01$. Le modèle de réseau de neurones n'est pas battu par les deux modèles G.L.M. et Memory Based Reasoning.

Le coefficient de corrélation entre les valeurs prédites x et celles attendues y, noté $\rho_{xy} = \frac{covxy}{\sigma x \sigma y}$ donne également une préférence pour le modèle de réseau de neurones.

Il en est de même pour la moyenne des valeurs absolues de l'erreur, notée $\overline{E_{abs}}$.

	Modélisation		
	Réseau de neurones	G.L.M.	Memory Based Reasoning
Erreur absolue moyenne $\overline{E_{abs}}$	404,4243	405,3017	462,8759
Coefficient de corrélation ρ_{xy}	0,14266	0,13312	0,09930

L'exemple ci-dessous est un comparatif entre les prédictions des trois modèles dans le cadre d'un adhérent actif de sexe masculin vivant en Ile-de-France travaillant dans le secteur de services aux entreprises bénéficiant d'un contrat ensemble du personnel avec de bonnes garanties :

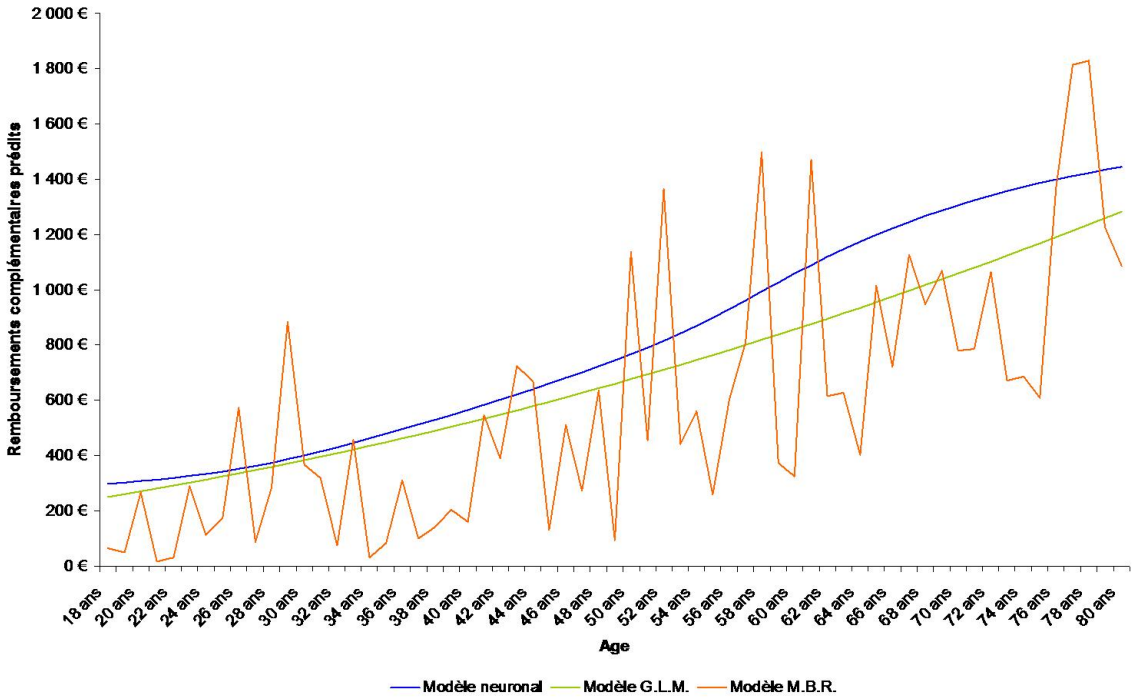


FIG. 16.3 – Prédictions selon les trois modèles en fonction de l'âge

16.2.3 Compléments : Analyse de la répartition et des résidus des prédictions

Les différentes répartitions des montants de remboursements annuels sur la base de test peuvent être observées entre d'une part la consommation 2007 réelle, et d'autre part les prédictions des trois modèles.

Sur le réel observé, le pic de la répartition est proche de zéro, cette configuration provient du fait qu'il y a 22% des bénéficiaires qui n'ont jamais eu recours au remboursement d'actes au cours de l'année. Cette surreprésentation du montant nul n'apparaît pas dans les prédictions.

Sur la méthode M.B.R., la forme de la répartition ressemble à celle du réel observé avec une concentration plus marquée autour de la moyenne. L'explication vient du principe même de la méthode qui effectue à chaque prédiction une moyenne sur k données. Mécaniquement, plus la valeur de k augmente, plus la variance diminue.

Les modèles de réseau de neurones et GL.M. ont une distribution assez similaire.

Les fonctions de répartition de l'écart entre les valeurs prédites par les trois modèles et le réel observé sont assez proches de celle de la loi normale avec une asymétrie plus ou moins prononcée.

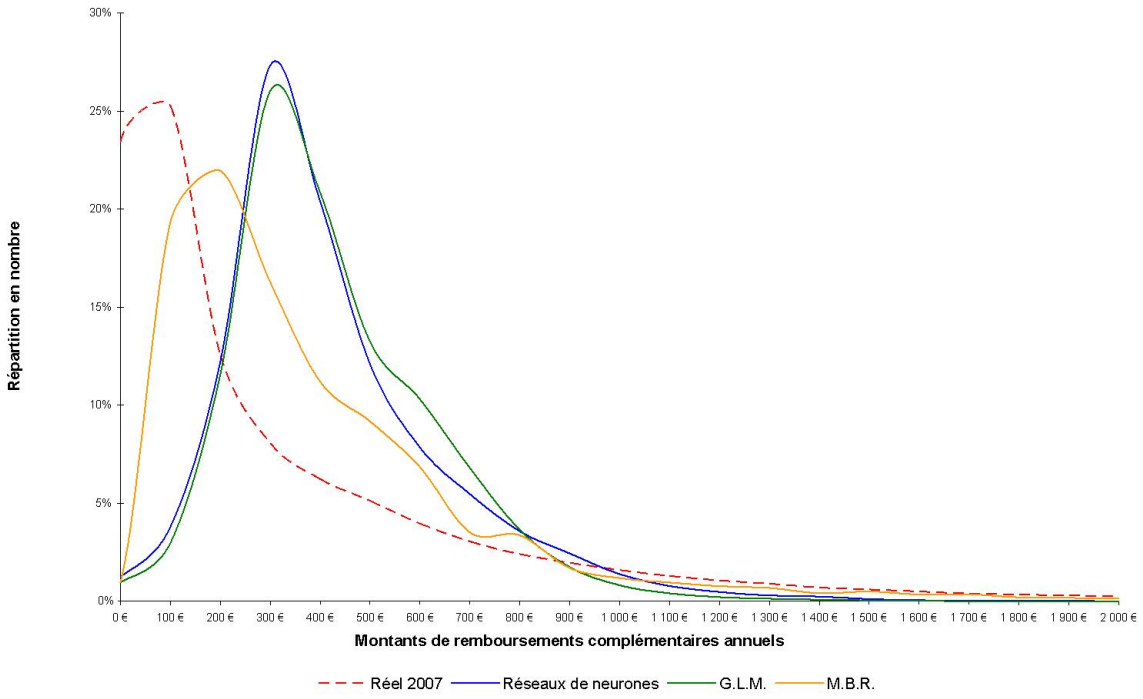


FIG. 16.4 – Répartition des montants de consommation médicale : prédictions et réel

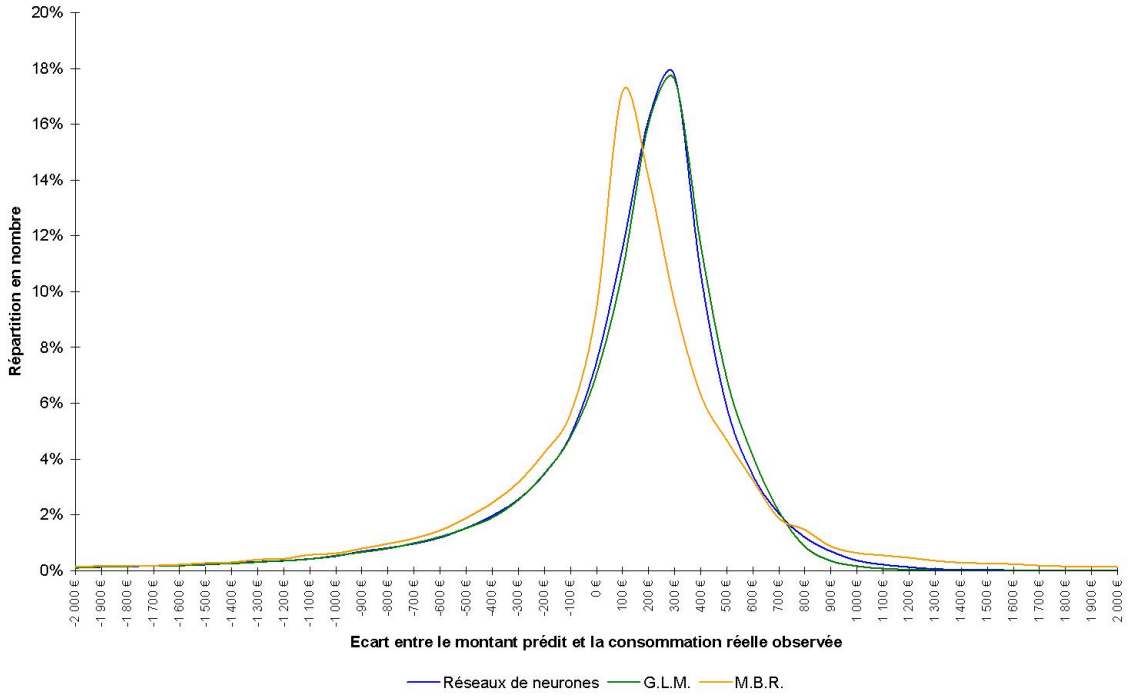


FIG. 16.5 – Répartition des erreurs entre les valeurs prédites et le réel observé

16.3 Comparaison avec la méthode classique de tarification : fréquence × coût moyen

16.3.1 Méthodologie

Jusqu'à présent, le modèle de réseau de neurones a été ajusté progressivement en faisant varier différents paramètres pour ne retenir à chaque fois que ceux donnant les meilleurs résultats. Une fois les paramètres de ce meilleur modèle déterminés, il a été comparé avec deux autres modélisations avancées de la consommation médicale. Ses prévisions se sont avérées être plus précises que sur les deux autres modèles. Ce premier constat a permis de justifier notre choix porté vers un modèle neuronal.

Afin de vérifier en pratique l'intérêt de notre modèle, il a semblé nécessaire de le tester également dans des conditions réelles d'utilisation en le comparant avec les pratiques tarifaires usuelles sur le marché.

Pour cela, nous nous plaçons dans un contexte normal de tarification où nous ne disposons pas encore de l'historique de consommation ligne à ligne. L'indicateur de niveau de garantie n'est alors plus estimé à partir de la consommation observée sur le groupe à tarifer mais à partir du tableau de garanties issu du contrat.

La prévision de la consommation médicale sera appliquée sur la population active des 13 comptes les plus importants rentrés en portefeuille à compter de 2008. Le fait de ne pas prendre des comptes en portefeuille avant 2008 a pour objectif de ne pas tarifer des comptes qui ont déjà servi à batir le modèle.

Notre démarche consistera à comparer les trois éléments suivants :

- le montant de prestations prédit par le modèle = tarif estimé ($\Pi_{prédit}$)
- le montant de la prime pure réellement encaissée = tarif appliqué ($\Pi_{appliqué}$)
- le montant réel des prestations versées = tarif d'équilibre ($\Pi_{équilibre}$)

Plus le tarif se rapproche du montant réel des prestations versées, plus la précision de son estimation est de bonne qualité.

Le passage de la prime commerciale à la prime pure réellement encaissée s'est fait en retirant la CMU en l'assimilant à une taxe (5,90% en 2009) et les chargements :

$$\Pi_{appliqué} = \Pi_{nette\ commerciale} \times \frac{1 - chargements}{1 + CMU}$$

Les prestations versées de survenance 2009 sont arrêtées à fin juin 2010. A cette date d'observation un taux de 1,70% de P.S.A.P. a été appliqué pour reconstituer un exercice complet en intégrant la queue de sinistre :

$$\Pi_{équilibre} = Prestations\ 2009_{30-06-10} \times (1 + PSAP_{30-06-10})$$

Le tarif prédit par le modèle a été majoré d'une hypothèse de dérive de la consommation médicale de 1,83% par an (soit 3,70% sur deux ans, ce qui équivaut à la moitié de la valeur de l'évolution de la C.M.T. entre 2007 et 2008) :

$$\Pi_{prédit} = \Pi_{modèle\ RN} \times (1 + C.M.T._{2007-2008})$$

Cette hypothèse repose sur l'observation de l'historique de l'évolution des remboursements complémentaires sur notre portefeuille qui progresse deux fois moins vite que la C.M.T. (comme le montre l'évolution des remboursements Gras Savoye entre 2006 et 2007 de 1,87%, cf. page 54).

De par notre rôle de courtier, les taux de cotisations des contrats de notre portefeuille sont systématiquement validés par notre service actuariat grâce à des outils de tarification internes. En cas d'écarts trop conséquents entre notre vision et celle des assureurs, nos résultats sont alors confrontés.

Dans cette démarche, le tarif appliqué est un moyen détourné d'observer la cotation moyenne obtenue à partir d'outils classiques.

16.3.2 Exemple

Détaillons l'exemple suivant que l'on nommera anonymement "contrat N° 18" dans tout ce qui suit. Il s'agit d'une entreprise de transports de 663 salariés basée essentiellement en Ile-de-France dont le régime obligatoire couvre l'ensemble du personnel sans distinction selon la C.S.P. Le régime ne permet pas la souscription d'une option facultative.

La première étape consiste à créer les indicateurs de garantie en partant des principales garanties issues du contrat :

GARANTIES FRAIS DE SANTE - CONTRAT N° 18
ENSEMBLE DU PERSONNEL

Poste	Garantie		Indicateur
	Soins courants		
Consultations-visites de généralistes	200% B.R.	→	99,3%
Consultations-visites de spécialistes	200% B.R.	→	97,0%
	Optique		
Verres unifocaux	6% P.M.S.S.	→	57,1%
Verres multifocaux	11% P.M.S.S.	→	59,1%
Monture	6% P.M.S.S.	→	87,7%
	Dentaire		
Prothèses dentaires remboursées	400% B.R.	→	91,1%
Soins dentaires	300% B.R.	→	96,7%
	Hospitalisation		
Honoriaires-Chirurgie	100% F.R.	→	100%
Chambre particulière	2,5% P.M.S.S. par jour	→	84,7%
	Divers		
Cure thermale	1,5% PMSS par jour (max. 18 jours)	→	75,0%
Forfait maternité	25% P.M.S.S.	→	55,9%

Une fois ce travail de codification réalisé, les données démographiques sont incorporées individuellement dans la base à tarifer en précisant tous les champs nécessaires au modèle (type de bénéficiaire, age, sexe, csp, activité, localisation géographique, garanties, ...).

Les renseignements concernant la démographie sont saisis pour chaque individu :

DEMOGRAPHIE									
NO PERS	SEXE	C BENEF	L REGION	CLAS INSEE	AGE	REGIME	COLLEGE	ACTIVITE	CHOIX
51733562	M	AD	Île-de-France	Transports	57	Général	Ens. Pers.	Actifs	Sans
51737078	F	AD	Île-de-France	Transports	55	Général	Ens. Pers.	Actifs	Sans
51766432	F	CJ	Île-de-France	Transports	31	Général	Ens. Pers.	Actifs	Sans
51773810	M	AD	Île-de-France	Transports	22	Général	Ens. Pers.	Actifs	Sans
51792989	F	CJ	Île-de-France	Transports	40	Général	Ens. Pers.	Actifs	Sans
51792994	M	AD	Île-de-France	Transports	40	Général	Ens. Pers.	Actifs	Sans
51792995	M	EF	Île-de-France	Transports	13	Général	Ens. Pers.	Actifs	Sans
51792997	F	EF	Île-de-France	Transports	11	Général	Ens. Pers.	Actifs	Sans
51793163	F	AD	Île-de-France	Transports	29	Général	Ens. Pers.	Actifs	Sans
51800629	F	AD	P.A.C.A.	Transports	36	Général	Ens. Pers.	Actifs	Sans
51800722	M	CJ	P.A.C.A.	Transports	39	Général	Ens. Pers.	Actifs	Sans
51801590	F	AD	Île-de-France	Transports	34	Général	Ens. Pers.	Actifs	Sans
51805019	M	CJ	Île-de-France	Transports	31	Général	Ens. Pers.	Actifs	Sans
51933838	M	EF	Île-de-France	Transports	21	Général	Ens. Pers.	Actifs	Sans
51955153	M	AD	Île-de-France	Transports	35	Général	Ens. Pers.	Actifs	Sans
51966062	M	AD	Île-de-France	Transports	32	Général	Ens. Pers.	Actifs	Sans
51970883	M	CJ	Île-de-France	Transports	32	Général	Ens. Pers.	Actifs	Sans
52058282	F	EF	Île-de-France	Transports	8	Général	Ens. Pers.	Actifs	Sans
52062027	F	AD	Île-de-France	Transports	31	Général	Ens. Pers.	Actifs	Sans
52110585	M	EF	Rhône-Al.	Transports	13	Général	Ens. Pers.	Actifs	Sans

Puis ceux concernant les garanties (dans le cas présent, les garanties sont identiques pour chaque individu) :

GARANTIES														
NO PERS	CGE	CSP	VER	VS	VP	VS1	VS2	...	PDA	SCV	CHP	HOH	CTH	MAT
51733562	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51737078	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51766432	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51773810	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51792989	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51792994	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51792995	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51792997	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51793163	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51800629	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51800722	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51801590	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51805019	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51933838	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51955153	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51966062	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
51970883	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
52058282	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
52062027	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%
52110585	99,3%	97%	0%	57,1%	59,1%	0%	0%	...	91,1%	96,7%	84,7%	100%	75%	55,9%

Ensuite, il ne reste plus qu'à appliquer le modèle à l'ensemble de la population. Le tarif uniforme prédit est obtenu en rapportant la charge totale de sinistres prédite au nombre d'adhérents. Le montant total des prestations s'obtient en sommant les prestations prédites pour chaque bénéficiaire qui s'élève à 577 105 €. En rapportant ce montant au 633 adhérents, l'estimation du tarif uniforme sur 2007 est de 870 € par famille. En intégrant l'évolution de la C.M.T. entre 2007 et 2009, la projection du tarif sur 2009 est de : $\Pi_{prédit} = 902$ €.

Le montant des cotisations encaissées au titre de l'exercice 2009 s'élève à 695 439 €, soit en moyenne 958 € par famille. En retirant les chargements de 8,5% et la C.M.U. à 5,90%, le montant de la prime nette encaissée est de : $\Pi_{appliqué} = 825$ €.

Le montant des prestations versées au titre de 2009 jusqu'à fin juin 2010 s'élève à 771 470 € auquel s'ajoutent 1,70% de PSAP, soit un total de 784 585 €. Le montant de la prime d'équilibre par famille est de : $\Pi_{équilibre} = 1\ 063$ €.

Sur cet exemple, le tarif prédit se trouve être plus proche de la consommation réelle que le tarif appliqué.

16.3.3 Application

Les entreprises testées sont au nombre de 13 et représentent près de 24 000 adhérents et 59 000 bénéficiaires.

Leurs secteurs d'activités sont les suivants :

- Services aux particuliers (8 246 salariés)
- Services aux entreprises (3 690 salariés)
- Industries agricoles et alimentaires (3 257 salariés)
- Commerce de gros (2 371 salariés)
- Industrie des biens de consommation (2 176 salariés)
- Activités immobilières (1 391 salariés)
- Transports (1 229)
- Industries des biens intermédiaires (981 salariés)
- Industries des biens d'équipement (540 salariés)

Ces entreprises sont composées de 24 contrats :

ENTREPRISE	ACTIVITE	SALARIES	CHARGEMENTS	CONTRATS
Entreprise A	Services aux particuliers	7 748	8%	N°1 - Régime obligatoire
Entreprise B	Industries agricoles et alimentaires	3 257	7,74%	N°2 - Régime base N°3 - Régime option 1 N°4 - Régime option 2
Entreprise C	Services aux entreprises	2 963		N°5 - Régime Cadre base N°6 - Régime Cadre option N°7 - Régime ETAM N°8 - Régime Non cadre
Entreprise D	Industrie des biens de consommation	2 176	9%	N°9 - Régime base N°10 - Régime option 1 N°11 - Régime option 2
Entreprise E	Commerce de gros	1 852	7%	N°12 - Régime obligatoire
Entreprise F	Activités immobilières	1 391	15%	N°13 - Régime base N°14 - Régime option
Entreprise G	Industries des biens intermédiaires	981	13%	N°15 - Régime base N°16 - Régime option
Entreprise H	Services aux entreprises	727	10%	N°17 - Régime obligatoire
Entreprise I	Transports	726	8,5%	N°18 - Régime obligatoire
Entreprise J	Industries des biens d'équipement	540	12%	N°19 - Régime obligatoire
Entreprise K	Commerce de gros	519	11%	N°20 - Régime obligatoire
Entreprise L	Transports	503	11%	N°21 - Régime Cadre N°22 - Régime Non cadre
Entreprise M	Services aux particuliers	498	11%	N°23 - Régime Cadre N°24 - Régime Non cadre

Les montants totaux des primes sont les suivants :

MONTANTS TOTAUX DE PRIMES				
ENTREPRISE	$\Pi_{prédit}$	$\Pi_{appliqué}$	$\Pi_{équilibre}$	Effectifs
Entreprise A	9 309 383 €	7 474 523 €	8 845 661 €	7 748
Entreprise B	3 347 102 €	2 366 063 €	3 810 500 €	3 257
Entreprise C	2 539 634 €	2 231 838 €	2 563 629 €	2 963
Entreprise D	2 657 857 €	2 157 608 €	2 161 029 €	2 176
Entreprise E	2 363 759 €	2 207 763 €	2 388 484 €	1 852
Entreprise F	1 253 882 €	1 106 076 €	994 590 €	1 391
Entreprise G	1 094 303 €	651 272 €	883 419 €	981
Entreprise H	612 581 €	480 405 €	571 602 €	727
Entreprise I	656 352 €	598 981 €	784 584 €	726
Entreprise J	313 301 €	402 180 €	501 080 €	540
Entreprise K	576 854 €	292 823 €	328 788 €	519
Entreprise L	459 337 €	340 291 €	406 907 €	503
Entreprise M	353 409 €	326 029 €	320 747 €	498
Total	25 537 761 €	20 635 858 €	24 561 027 €	23 881

Les tarifs purs (sans chargements et sans C.M.U.) sont les suivants :

TARIFS ANNUELS PAR FAMILLE				
ENTREPRISE	$\Pi_{prédit}$	$\Pi_{appliqué}$	$\Pi_{équilibre}$	Effectifs
Entreprise A	1 202 €	965 €	1 142 €	7 748
Entreprise B	1 028 €	726 €	1 170 €	3 257
Entreprise C	857 €	753 €	865 €	2 963
Entreprise D	1 221 €	992 €	993 €	2 176
Entreprise E	1 276 €	1 192 €	1 290 €	1 852
Entreprise F	901 €	795 €	715 €	1391
Entreprise G	1 115 €	664 €	901 €	981
Entreprise H	843 €	661 €	786 €	727
Entreprise I	904 €	825 €	1 081 €	726
Entreprise J	580 €	745 €	928 €	540
Entreprise K	1 113 €	565 €	634 €	519
Entreprise L	913 €	677 €	809 €	503
Entreprise M	710 €	655 €	644 €	498
Moyenne	1 069 €	864 €	1 028 €	23 881

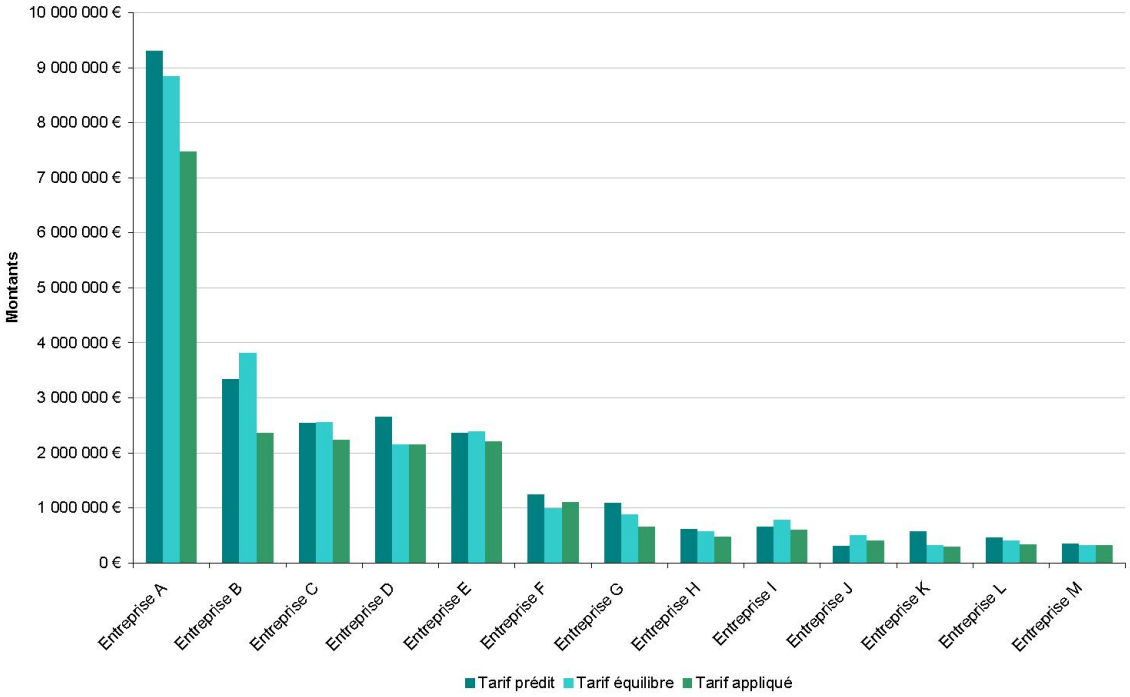


FIG. 16.6 – Montants totaux de primes

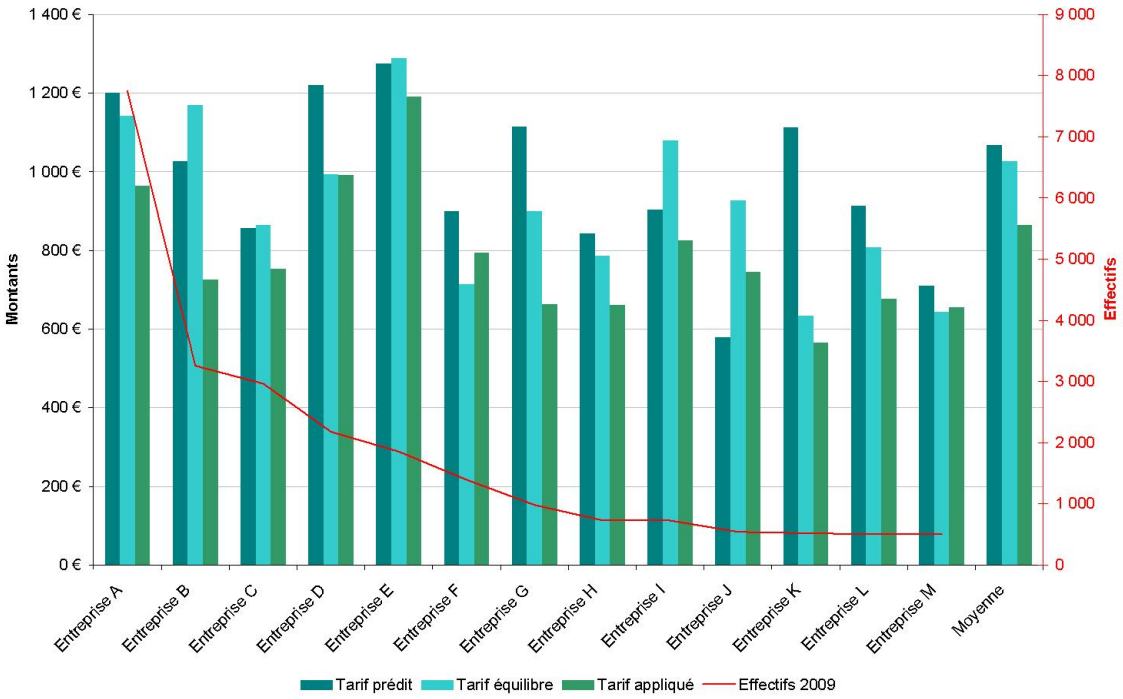


FIG. 16.7 – Tarifs annuels par famille

Sur l'ensemble des effectifs couverts, dans 79% des cas, le modèle de réseau de neurones fournit de meilleures estimations du risque. Au global, le tarif prédit est supérieur de 4% au tarif d'équilibre alors que les tarifs pratiqués sont inférieurs de 16%. L'application du modèle de tarification neuronale dans des conditions normales d'utilisation a été concluant en apportant des résultats plus précis que ceux appliqués.

16.4 Tests de cohérence

16.4.1 Test sur les effets âge, sexe et type de bénéficiaire

Ce test consiste à vérifier la cohérence des prédictions en faisant varier uniquement les trois critères principaux qui sont : le sexe, l'âge et le type de bénéficiaire. Pour éviter qu'il puisse y avoir des interférences avec les autres paramètres en entrée du modèle, le même exemple est repris pour tous les comparatifs : le référentiel est constitué de bénéficiaires vivant en Ile-de-France dont l'adhérent travaille dans le secteur de services aux entreprises bénéficiant d'un contrat ensemble du personnel avec de bonnes garanties.

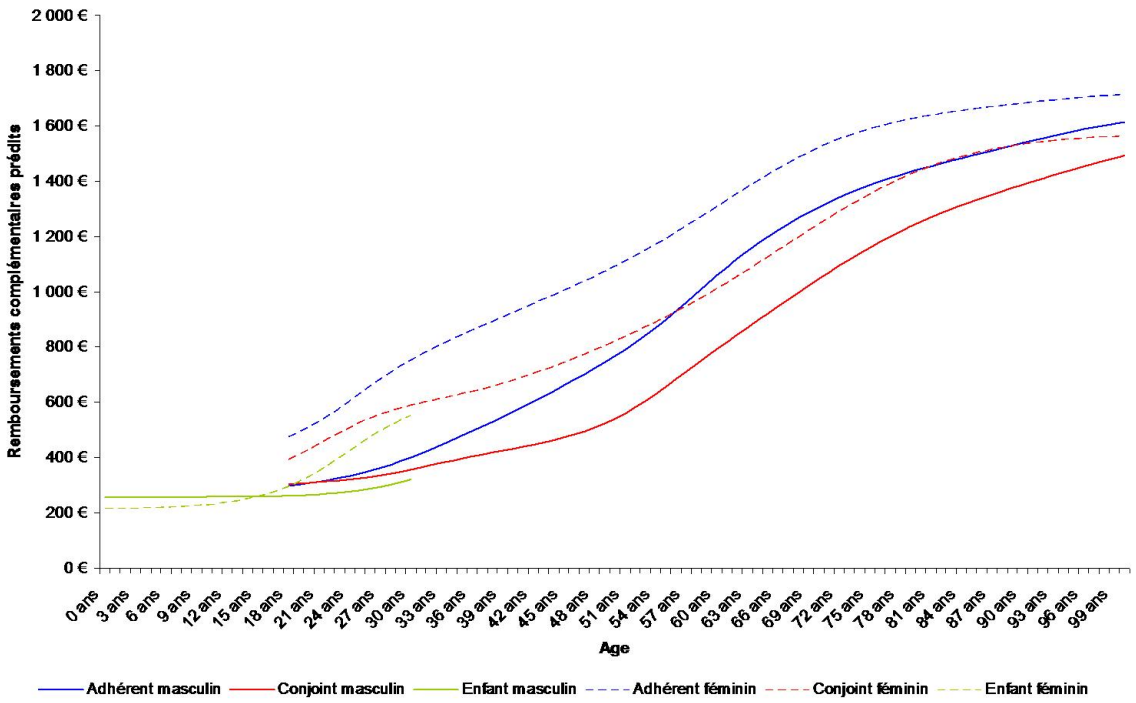


FIG. 16.8 – Effets âge, sexe et type de bénéficiaire

Ce premier test est concluant car le modèle reproduit bien les tendances attendues :

- la consommation croît avec l'âge,
- les femmes consomment davantage que les hommes,
- cet écart diminue pour les âges élevés,
- les adhérents consomment plus que les conjoints.

16.4.2 Test sur la C.S.P. et le niveau de garanties

Cette fois, l'exemple de référence est celui d'un adhérent de sexe masculin vivant en Ile-de-France travaillant dans le secteur de services aux entreprises. Les paramètres qui vont varier sont la C.S.P. et les niveaux de garanties, les exemples testés sont :

- un contrat cadre avec des garanties basses
- un contrat cadre avec des garanties hautes
- un contrat non cadre avec des garanties basses
- un contrat non cadre avec des garanties hautes

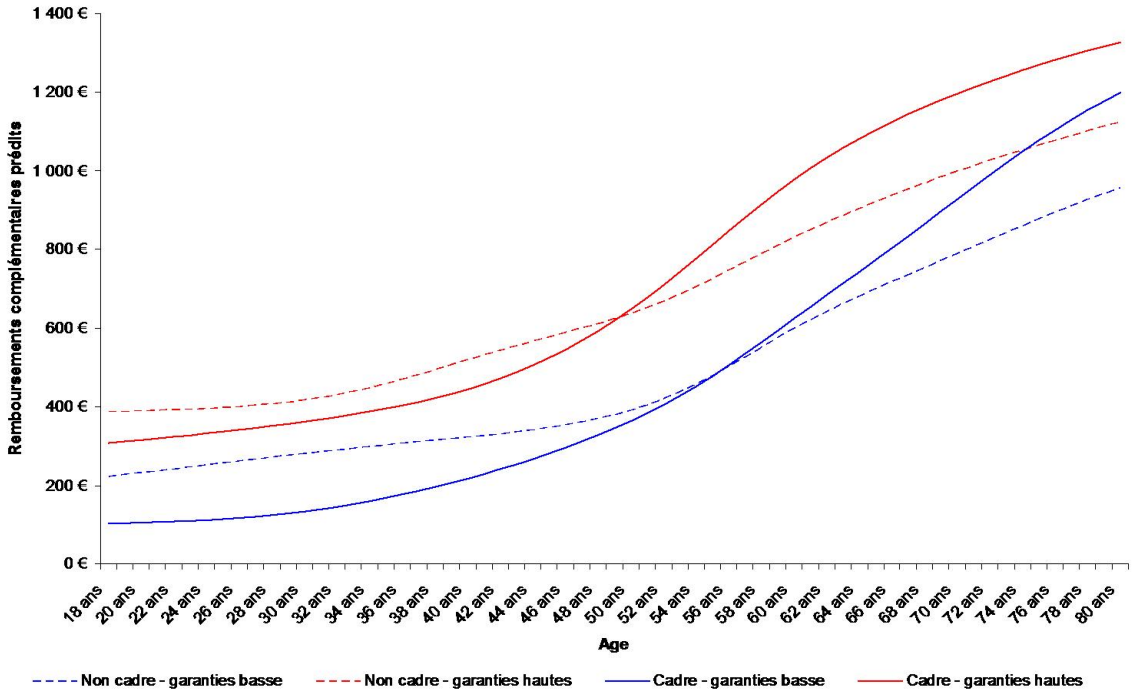


FIG. 16.9 – Effets C.S.P. et niveau de garanties

Ce deuxième test est également concluant :

- la consommation croît avec le niveau des garanties aussi bien pour les cadres que pour les non cadres
- à partir de la cinquantaine, les cadres consomment davantage que les non cadres.

En pratique, les contrats cadres offrent le plus souvent des garanties supérieures à celles des non cadres.

16.4.3 Test sur la région

En reprenant toujours le même exemple d'un adhérent de sexe masculin travaillant dans le secteur de services aux entreprises et bénéficiant d'un contrat ensemble du personnel, l'impact de six régions caractéristiques suivantes est testé :

- Ile-de-France
- Rhône-Alpes
- Provence Alpes Côte d'Azur
- Bretagne
- Auvergne
- Pays de la Loire

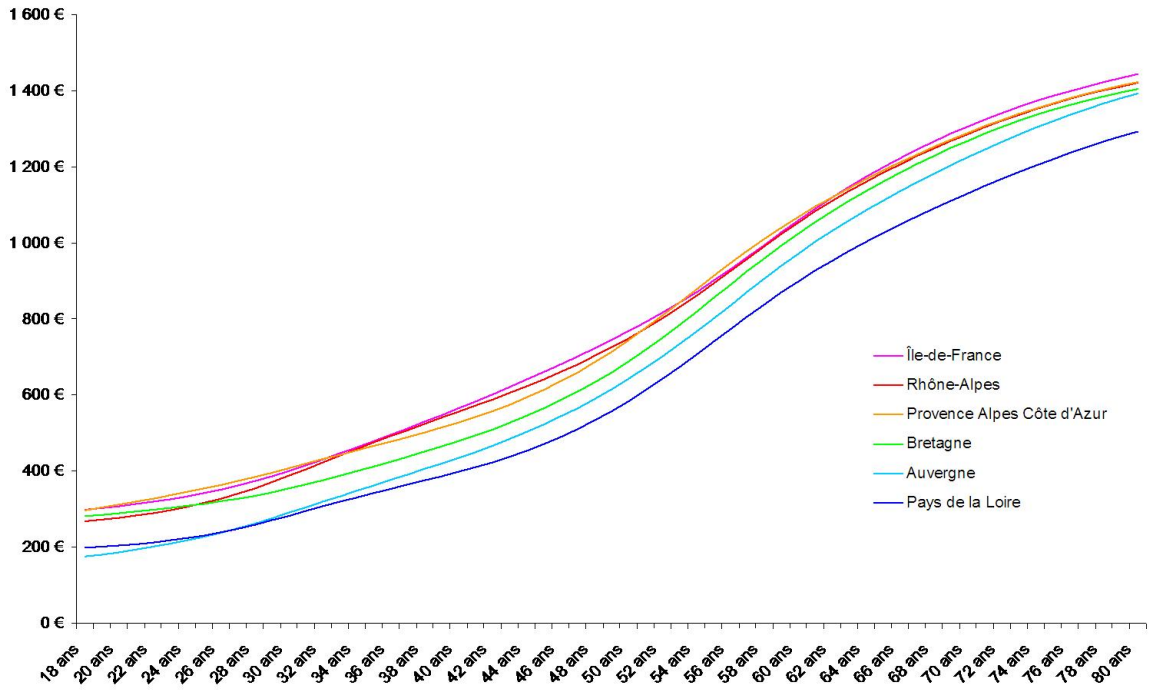


FIG. 16.10 – Effets Région

Ce troisième test est lui aussi concluant. Pour quasiment tous les âges, les niveaux de consommations les plus élevés sont observés en Ile-de-France, Provence Alpes Côte d'Azur et Rhône-Alpes. Cet axe est, en effet, réputé comme étant marqué par une consommation médicale élevée.

16.4.4 Test sur l'activité

En reprenant l'exemple d'un adhérent de sexe masculin vivant travaillant en Ile-de-France et bénéficiant d'un contrat ensemble du personnel, six principales activités sont testées :

- Industries agricoles et alimentaires
- Transports
- Construction
- Services aux entreprises
- Commerce de gros
- Industries des biens intermédiaires

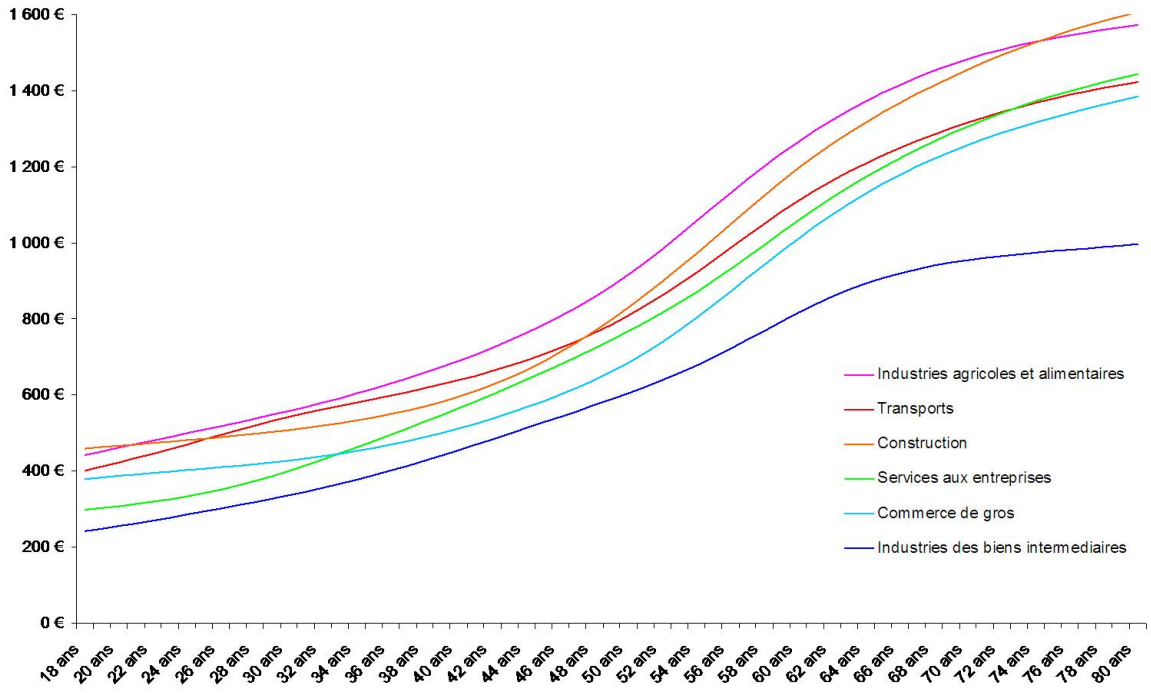


FIG. 16.11 – Effets Activité

Ce test, plus difficile à interpréter que les précédents car nous ne nous attendons pas à une hiérarchie précise des différentes activités en fonction de leur consommation. Pour autant, nous avons pu vérifier que pour chaque secteur, la courbe de la consommation était bien croissante avec l'âge et que les valeurs des prédictions restaient dans des fourchettes vraisemblables. Le bilan de ce test ressort également positif.

Toutefois, il n'est pas possible de faire un parallèle entre ces résultats et ceux indiqués en première partie dans l'analyse préliminaire des données car tous les effets n'avaient pas été neutralisés comme c'est le cas ici.

Chapitre 17

Conclusions

Au final, le modèle de réseau de neurones ayant obtenu les meilleurs performances est caractérisé de la la façon suivante :

Architecture

- 40 neurones cachés
- une seule couche cachée
- fonction d'activation logistique
- pas de liaison directe
- pas d'élagage
- toutes les cellules d'entrée sont reliées à toutes celles de le couche cachée

Apprentissage

- répartition de la base aléatoire : 40% entrainement, 30% validation et 30% test
- poids initiaux aléatoires (sauf pour la couche de sortie)
- une phase préliminaire (50 lancés sur 20 itérations avec la méthode de Levenberg-Marquardt)
- technique d'apprentissage : RPROP
- critère d'arrêt : Early Stopping avec au maximum 2000 itération
- fonction de coût à minimiser : erreur quadratique

Résultats

Avec ce modèle, sur la base de validation, la valeur de l'erreur quadratique moyenne la plus faible est obtenue à la $294^{\text{ème}}$ itération avec 1 234,17. En appliquant le modèle avec ces poids à la base de test, l'erreur quadratique moyenne vaut 1 285,01.

Un premier comparatif a été mené entre le modèle de réseau de neurones retenu et d'autre modèles avancés comme le modèle G.L.M. (Modèle Général Linéaire) et le M.B.R. (Memory Based Reasoning, méthode des k plus proches voisins). Dans un second temps, il a été confronté avec des méthodes classiques de tarification dans des conditions réelles d'application. Dans ces deux situations, ses performances se sont montrées concluantes. Enfin, une série de tests mise en œuvre pour s'assurer de la cohérence des résultats n'a relevé aucune anomalie.

Récapitulatif des simulations

Couches cachées	Neurones cachés	Spécificité	Fonction d'activation	Fonction coût	Technique	Pas	Momentum	EQM (test)
Modèle de base								
1	20		Tan h	E.Q.M.	Backprop	0,0100	0	1286,03
Sélection du nombre de neurones cachés								
1	10		Tan h	E.Q.M.	Backprop	0,0100	0	1286,72
1	20		Tan h	E.Q.M.	Backprop	0,0100	0	1286,03
1	30		Tan h	E.Q.M.	Backprop	0,0100	0	1286,10
1	40		Tan h	E.Q.M.	Backprop	0,0100	0	1285,75
1	50		Tan h	E.Q.M.	Backprop	0,0100	0	1285,83
Sélection du nombre de couches cachées								
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,8	1285,69
2	40		Tan h	E.Q.M.	Backprop	0,0010	0,8	1286,10
Sélection du pas d'apprentissage								
1	40		Tan h	E.Q.M.	Backprop	0,1000	0	$+\infty$
1	40		Tan h	E.Q.M.	Backprop	0,0500	0	1285,86
1	40		Tan h	E.Q.M.	Backprop	0,0100	0	1285,75
1	40		Tan h	E.Q.M.	Backprop	0,0010	0	1285,68
1	40		Tan h	E.Q.M.	Backprop	0,0001	0	1287,74
Sélection du momentum								
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,0	1285,68
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,5	1285,53
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,7	1285,71
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,8	1285,69
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,9	1285,74
Sélection de la technique d'apprentissage								
1	40		Tan h	E.Q.M.	Backprop	0,0010	0,8	1285,68
1	40		Tan h	E.Q.M.	Rprop			1285,49
1	40		Tan h	E.Q.M.	Quickprop			1285,63
1	40		Tan h	E.Q.M.	Grad. conj.			1285,59
1	40		Tan h	E.Q.M.	Lev.-Marq.			1285,66
1	40		Tan h	E.Q.M.	Quasi-Newt.			1286,42
Weight decay								
1	40		Tan h	E.Q.M.	Rprop			1285,49
1	40	W. decay	tanh	E.Q.M.	Rprop			1285,50
Sélection de la fonction d'activation								
1	40		Tan h	E.Q.M.	Rprop			1285,49
1	40		Normale	E.Q.M.	Rprop			1285,32
1	40		Arc Tan	E.Q.M.	Rprop			1285,31
1	40		Logistique	E.Q.M.	Rprop			1285,01
Sélection de la fonction de coût								
1	40		Logistique	E.Q.M.	Rprop			1285,01
1	40		Logistique	Huber	Rprop			1291,47
Liaisons directes								
1	40		Logistique	E.Q.M.	Rprop			1285,01
1	40	L. directes	Logistique	E.Q.M.	Rprop			1285,27
Elagage								
1	40		Logistique	E.Q.M.	Rprop			1285,01
1	40	Elagage	Logistique	E.Q.M.	Rprop			1285,34

CONCLUSION

L'étude descriptive a permis d'approfondir la connaissance de notre portefeuille et de mieux appréhender les mécanismes du comportement de consommation médicale. Les éléments apportés par cette première analyse ont pu être réutilisés au travers de différentes études internes, de communications clients et d'actions commerciales.

Un indicateur synthétique reflétant le niveau des garanties contractuelles a été défini. Il permet de comparer de manière homogène et de hiérarchiser des garanties exprimées différemment. Au-delà de son utilité pour la modélisation neuronale développée ici, il trouve également un intérêt beaucoup plus vaste dans de nombreuses applications telles que les benchmarks sectoriels, les comparatifs clients ou encore les études comportementales.

Une méthodologie complète utilisant la technique des réseaux de neurones a pu être expérimentée pour tarifer des contrats collectifs frais de santé. Cette nouvelle approche permet de tirer profit de la souplesse et de la précision que leur confère leur propriété fondamentale d'approximateurs universels parcimonieux.

Le confrontation de la démarche neuronale proposée avec d'autres outils mathématiques avancés s'est avérée être concluante. Les Modèles Linéaires Generalisés et Memory Based Reasoning ne sont pas parvenus à battre ses performances, ce qui a conforté le choix porté pour cette technique.

Elle a pu mettre en avant sa valeur ajoutée en étant appliquée dans des conditions réelles d'utilisation sur un panel des 13 contrats représentant près de 60 000 bénéficiaires. Dans 79% des cas, les résultats obtenus par le modèle de réseau de neurones ont été plus proches de la réalité que les tarifs appliqués provenant d'outils classiques de tarification. Le coût du risque global prédit pour l'ensemble du périmètre ressort supérieur de 4% au tarif d'équilibre alors que les tarifs pratiqués étaient inférieurs de 16%.

Les travaux de ce mémoire ont finalement débouchés sur plusieurs sujets pour aboutir à la finalité recherchée : batir un modèle de tarification performant. Cette nouvelle approche tarifaire a pu démontrer l'intérêt d'une modélisation neuronale car celle-ci offre un niveau précision supplémentaire comparativement aux méthodes traditionnellement utilisées.

La modélisation neuronale présente un avantage majeur dans le pilotage de régimes frais de santé. En offrant une vision plus précise du risque, elle contribue à pérenniser dans le temps sa gestion. Nous répondons ici à la nécessité pour l'entreprise de maîtriser ses budgets de protection sociale dans un environnement trés concurrentiel aggravé par la crise économique et un contexte social parfois difficile.

En synthèse :

Caractéristiques	**Méthode proposée**	**Méthode classique**
Modélisation	Réseau de Neurones	Fréquence × Coût moyen
Approche	Tarif global	Tarif par poste
Influences des garanties sur les frais réels et les fréquences	Prises en compte	Non prises en compte
Liaisons entre les garanties	Prises en compte	Non prises en compte
Pouvoir explicatif des coefficients	Non	Oui
Précision des prédictions	Haute	Moyenne

Bibliographie

[RUB87] Rubin D.B. *: Multiple imputation for nonresponse in surveys, John Wiley & sons, 1987*

[SCH97] Schafer J.L. *: Analysis of incomplete multivariate data by simulation, New-York : Chapman and Hall, 1997*

[SCH98] Schafer J.L. et Olsen M.K. *: Multiple Imputation for Multivariate Missing Data Problems : a Data Analystś Perspective, 1998*

[ROB96] Robert C.P. *: Méthodes de Monte Carlo par Chaînes de Markov, Paris : Economica, 1996*

[VAT02] Vaton S. *: Notes de cours sur les méthodes de Monte Carlo par Chaînes de Markov, 2002*

[DEM77] Dempster A.P., Laird N.M., Rubin D.B. *: Maximum Likelihood from incomplete data via the EM Algorithm (with discussion), Journal of the Royal Statistical Society, vol. B 39, 1977*

[GEM84] Geman S. and Geman D. *: Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6, no. 6, 721-741., 1984*

[DRO02] Droesbeke J.J., Fine J. et Saporta G. *: Methodes bayésiennes en statistique, Editions TECHNIP, 2002*

[TAN87] Tanner M.A. and Wong W.H. *: The calculation of posterior distributions by data augmentation, Journal of American Statistical Association, 82, 528-540,1987*

[SCH02] Schafer J.L. and Graham J.W. *: Missing data : Our view of the state of the art, Psychological Methods, Vol. 7, No. 2, 147-177, 2002*

[SAP06] Saporta G. *: Probabilités analyse de données et statistiques, Editions TECHNIP, 2006*

[TAS04] Tassi P. *: Méthodes statistiques, Editions Economica, 2004*

[VAL42] Valéry P. *: Mauvaises pensées et autres, 1942*

[OUVOO] Ouvrard J.-L. *: Probabilités, volume 2, Cassini, Paris, 2000*

[BERO6] Bertrand F. et M. *: Mesures d'association, 2006/2007*

[ROY82] Royston P. *: An extension of Shapiro and Wilk's W Test for Normality to Large Samples, Applied Statistics, 31, 115-124, 1982*

[ROY95] Royston P. *: A remark on Algorithm AS 181 : The W Test for Normality. Applied Statistics, 44, 547-551, 1995*

[MAR70] Mardia K. V. *: Measues of multivariate skewness and kurtosis, Biometrika 57, 519-530, 1970*

[MAR80] Mardia K. V. *: Tests of univariate and multivariate normality, In Krishnaiah, P.R. (ed.), Handbook of statistics, vol. 1, ch.9, Amsterdam : North-Holland, 1980*

[WAR63] Ward J.H. *: Hierarchical grouping to optimize an objective function, J. A. S.A. 58, 236-244, 1963*

[CUL43] Mac Culloch W.S. et Pitts W.H. *: A logical calculus of the ideas immanent in nervous activity, Bulletin of mathematical biophysics, 5, p. 115-133, 1943*

[LET59] Lettvin J.Y., Maturana H.R., Mac Culloch W.S. et Pitts W.H. *: What's the frog's eye tells the frog's brain, Proces. Inst.Radio. Engeniors, 47 : 1940-1951, 1959*

[HEB49] Hebb D.O. *: The organisation of behavior : A neuropsychological theory, Wiley, New-York, 1949*

[ROS57] Rosenblatt F. *: The Perceptron : a perceiving and recognizing automaton, Project PARA, Cornell Aeronautical Lab. Report 85-460-1, 1957*

[ROS58] Rosenblatt F. *: The Perceptron : a probabilistic model for information storage and organization in the brain, Psychological review, vol. 65, p. 386-408, 1958*

[WID60] Widrow B. et Hoff M.E. *: Adaptative switching circuits, IRE Wescon Convention record : part 4, computers :Man-machine systems , p. 96-104, Los Angeles, CA 1960*

[MIN69] Minsky M. et Papert S. *:Perceptrons, Introduction, p.1-20 : "Perceptrons cannot learn arbitrary rules, Cambridge MA : MIT Press, 1969*

[HOP82] Hopfield J.J. *: Neural Networks and physical systems with emergent collective computational abilities, P.N.A.A. U.S.A, vol. 79, 1982*

[HIN83a] Hinton G.E. et Sejnowski T.J. *: Analyzing cooperative computation. In proceedings of the 5th annual congress of cognitive science society, Rochester, New-York, may 1983*

[HIN83b] Hinton G.E. et Sejnowski T.J. *: Optimal perceptual inference. In proceedings of the IEEE conference of Computer Vision and Pattern Recognition (CVPR), p. 448-453, IEE Computer Society, Washington DC, june 1983*

[WER74] Werbos P.J. *: Beyond regression : New tools for prediction and analysis in the behavorial sciences, Doctoral dissertation, Applied mathematics, Harvard university, Boston, MA, 1974*

[RUM86] Rumelhart D.E. et Mac Clelland J. *: Parallel distributed processing, Cambridge, Massachussets : MIT Press, 1986*

[FAU94] Fausett Laurene V. *: Fundamentals of neural networks : architectures, algorithmes and applications. Prentice Hall, 1994*

[HOR89] Hornik K., Stinchcombe M., White H. *: Multilayer feedforward networks are universal approximators, Neural Networks, vol 4 : 359-366, 1989*

[FUN89] Funahashi K. *: On the approximate realization of continuous mapping by neural networks, neural networks, N°2, 183-192, 1989*

[BOR07] Borne P., Benrejeb M. et Haggège J. *: Les réseaux de neurones, Technip, 137-142, 2007*

[COM91] Common P. *: Classification supervisée par réseaux multicouches, Traitement du signal, volume 8, N°6, 389-390, 1991*

[CYB89] Cybenko G. *: Approximation by superpositions of a sigmoidal function, Mathématics of control, Signals and systems, vol 2 : 303-314, 1989*

[KAR95] Karpinski M. and Macintyre A. *: Polynomial bounds for VC dimension of sigmoidal neural networks, Proceeding of the 27th ACM Symposium on theory of computing (STOC 1995), 200-208, ACM Press*

[SMI91] Smieja F.H. *: Hyperplane "Spline" Dynamics, Network Plasticity an Back-Propagation, Learning. GMD Report, GMD, St Augustin, Germany, 28/11/1991*

[SAP06] Saporta G. *: Probabilités Analyse de Données et Statistique, Technip, 2006*

[RIE93] Riedmiller and Braun *: A direct adaptive method for faster backpropagation learning : The RPROP algorithm, 586-591, San Francisco, CA, 1993*

[RUM86] Rumelhart D.E, Hinton G.E. and Williams R.J. *: Learning Internal Representations by Error Propagation, Parallel Distributed Processing : Explorations in the Microstructures of Cognition. - M.I.T. Press. - Vol. 1, p. 318-362., 1986*

[FAH88] Fahlman S.E. *: An Empirical Study of Learning Speed in Backpropagation Networks. Technical Report CMU-CS-88-162, CMU, September 1988*

[TRI88] Trigeassou J.C. *: Recherche de modèles expérimentaux assistée par ordinateur. Collection Informatique Prepas Université, 1988.*

[MOL90] Moller M.F. *: A Scalled Conjugate Gradient Algorithm for Fast Supervised Learning. Technical Report PB-339 -Compter Science Dept, University of Aarhus, Denmark, November 1990.*

[POL69] Polak E. et Ribière G. *: Note on the convergence of methods of conjugate directions. Revue Française d'Informatique et de Recherche Opérationnelle, 3(16) : 35-43, 1969.*

[HES52] Hestenes M.R. and Stiefel E. *: Methods of Conjugate Gradients for Solving Linear Systems, National Bureau of Standards Journal of research 49, p. 409-436, 1952.*

[MAR63] Marquardt D.W. *: An algorithm for least-squares estimation of nonlinear parameters, Journal of the Society for Industrial and Applied Mathematics, vol. 11, N° 2, p. 431-441, 1963*

[BRO70] Broyden C.G. *: The Convergence of a Class of Double-rank Minimization Algorithms, Journal of the Institute of Mathematics and Its Applications, 6, p. 76-90, 1970*

[FLE70] Fletcher R. *: A new Approach to Variable Metric Algorithms, Computer Journal, 13, p. 317-322, 1970*

[GOL70] Goldfarb D. *: A Family of Variable Metric Updates Derived by Variational Means, Mathematics of Computation, 24, p. 23-26, 1970*

[SHA70] Shanno D.F. *: Conditioning of Quasi-Newton Methods for Function Minimization, Mathematics of Computation, 24, p. 647-656, 1970*

[WEI91] Weigend A.S., Rumelhart D.E., Huberman B.A. *: Generalisation by weight-elimination with appication to forecasting. 1991*

[BIS93] Bishop C. *: Curvature-driven smoothing : a learning algorithm for feedforward networks, IEEE Trans. on Neural Networks, vol. 4, n°5, p. 882-884, 1993*

[BIS95] Bishop C. *: Neural Networks for Pattern Recognition, Oxford University Press, 1995*

[NEL72] Nelder J.A. et Wedderburn R.W.M.*Generalized Linear Models, J.R. Statis. Soc., A135, p. 370-384, 1972*

[CUL89] Mc Cullagh P. and Nelder J.A.*Generalized Linear Models, 2nd ed., Chapman and Hall, 1989*

[STA86] Stanfill C. and Waltz D.*Toward memory-based reasoning. Communications of the ACM, Vol. 29, p. 1213-1228, December 1986*

[WEI90] Weiss S.M. and Kulikowski C.*Computer Systems That Learn. Morgan Kaufmann, 1990*

Quatrième partie

ANNEXES

Annexe A

Algorithme Espérance - Maximisation (E.M.)

L'algorithme EM a été conçu par Dempster, Laird et Rubin en 1977 [DEM77], il permet de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. Il trouve son utilité essentiellement dans les domaines suivants : la classification de données, en apprentissage machine, ou en vision artificielle.

La vraisemblance d'un échantillon $X = (x_1, x_2, ..., x_N)$ est définie de la manière suivante :

$$L(\theta|X) = f(X|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

Cette fonction exprime la vraisemblance des paramètres selon les données. Le logarithme facilitant les calculs, le Log-vraisemblance est plus souvent utilisé : $\log L(\theta|X)$

L'expression de la vraisemblance des données pour déterminer un paramètre inconnu θ est souvent trop complexe à maximiser analytiquement, l'algorithme EM permet de contourner cette difficulté.

Le but de la méthode est de compléter les données en ajoutant un complément Y tel que :

$$(X, Y) \sim f(X, Y|\theta)$$

Y correspond au vecteur des données manquantes et le couple (X, Y) à celui des données complétées.

Le principe de l'algorithme EM est d'optimiser la log-vraisemblance des données complétées plutôt que celle des données observées.

La densité conditionnelle de Y sachant les données X est égale à :

$$k(Y|\theta, X) = \frac{f(X, Y|\theta)}{g(X|\theta)}$$

avec $g(X|\theta)$, la représentation marginale de la vraisemblance : $g(X|\theta) = \int_Z f(x, z|\theta)dz$.

La vraisemblance liée aux données complétées est notée :

$$L^c(\theta|X,Y) = f(X,Y|\theta)$$

Celle associée aux données observées :

$$L(\theta|X)$$

Ces deux vraisemblances sont liées par la relation suivante :

$$log\ L(\theta|X) = \mathbb{E}_Y[log\ L^c(\theta|X,Y)|\theta_0,X] - \mathbb{E}_Y[log\ k(Y|\theta,X)|\theta_0,X]$$

L'espérance est intégrée suivant la loi conditionnelle de Y sachant le données θ_0 et X : $k(Y|\theta,X)$, pour tout θ_0.

La maximisation de la vraisemblance observée $L(\theta|x)$ est obtenue en ne maximisant que le premier terme : $\mathbb{E}_Y[log\ L^c(\theta|X,Y)|\theta_0,X]$ puis en réitérant le processus afin de palier l'omission du second terme ainsi que la dépendance en θ.

En notant l'espérance de la log-vraisemblance issue des données complètes :

$$Q(\theta|\theta_0,X) = \mathbb{E}_Y[log\ L^c(\theta|X,Y)|\theta_0,X]$$

Par alternance d'étapes d'Espérance et Maximisation, l'algorithme EM construit itérativement une suite d'estimateurs $\widehat{\theta}_i$

$$Q(\widehat{\theta}_{i+1}|\widehat{\theta}_i,X) = \max_{\theta}\ Q(\theta|\widehat{\theta}_i,X)$$

Théorème A.0.1 (Algorithme Espérance-Maximisation)**.** *La technique utilisée est itérative où deux étapes sont appliquées alternativement :*

- *Etape "E" : Evaluation de l'espérance*
  A partir des dernières variables observées, l'espérance de la vraisemblance est calculée :

$$Q(\theta|\widehat{\theta}_i,X) = \mathbb{E}_Y[log\ L^c(\theta|X,Y)|\widehat{\theta}_i,X] = \int_{y\in Y} log\ L^c(\theta|X,y) \times k(y|\widehat{\theta}_i,X)dy$$

- *Etape "M" : Maximisation*
  Le maximum de vraisemblance des paramètres est ensuite estimé en maximisant la vraisemblance trouvée à l'étape précédente

$$\widehat{\theta}_{i+1} = arg\ \max_{\theta}\ Q(\theta|\widehat{\theta}_i,X)$$

La vraisemblance observée croît à chaque itération jusqu'à ce qu'un point fixe de Q soit obtenu.

Annexe B

Algorithme de Hastings-Metropolis

L'algorithme de Hastings-Metropolis a pour objectif de simuler une variable aléatoire X définie par sa densité de probabilité $f(x)$. La méthode génère une chaîne de Markov qui converge vers la loi objectif. Des variables aléatoires sont simulées selon une loi instrumentale pour ensuite appliquer un mécanisme d'acceptation-rejet.

L'algorithme de Hastings-metropolis simule une suite $x_1, x_2, ...$ en procédant comme suit :

Théorème B.0.2 (Algorithme de Hastings-Metropolis)**.**

- *1°) Générer* $Y_t \sim q(y, x_t)$.

  La loi q est dite loi instrumentale, la densité conditionnelle $q(y|x)$ *doit être simulable facilement et soit disponible analytiquement soit symétrique* ($q(y|x) = q(x|y)$. *La valeur ainsi obtenue est notée* y_t.

- *2°) Calculer le ratio de Hastings-Metropolis :*

$$\rho(x_t, y_t) = min\left(\frac{f(y_t)}{f(x_t)}\frac{q(x_t|y_t)}{q(y_t|x_t)}, 1\right)$$

- *3°) Accepter* X_{t+1} *avec la probabilité* $\rho(x_t, y_t)$, *sinon poser* $X_{t+1} = x_t$

Sous certaines conditions sur le noyau $q(x, y)$ *la variable aléatoire* X_t *converge en distribution vers la loi objectif* $f(x)$.

L'idée sous-jacente est de perturber la valeur courante de x_t selon un processus stochastique et d'accepter aléatoirement le résultat y_t de celle-ci. Le théorème suivant assure que la chaîne (x_t) admet bien f comme loi limite stationnaire.

Théorème B.0.3. *Pour toute loi conditionnelle q, la chaîne* (x_t) *produite par l'algorithme de Hastings-Metropolis admet* f *comme loi stationnaire.*

Démonstration. Le noyau de transition associé à l'algorithme de Hastins-Metropolis s'écrit :

$$K(x,y) = \int_{z\in\Omega} q(z|x)\left[\delta_z(y)\rho(x,z) + \delta_x(y)(1-\rho(x,z)\right] dz$$

avec $\delta_x()$ et $\delta_z()$ désignant respectivement la masse de Dirac en x et celle de z .

Schématiquement, cette expression exprime les faits suivants :

- si y est différent de x, le passage d'un état x à un état y ne peut se réaliser qu'après l'acceptation du tirage d'une transition de x vers y,
- si y est égal à x, la stagnation en x peut être produite soit en acceptant le tirage d'une transition de x vers lui-même, soit en rejetant le tirage d'une transition quelconque.

Démontrer que f est une mesure invariante de la chaîne de Markov issue de l'algorithme de Hastings-Metropolis revient à établir l'égalité suivante :

$$\iint_{x,y} \mathbb{I}_A(y) f(x) K(x,y) dxdy = \int_A f(x) dx$$

Quel que soit l'ensemble mesurable A, on a la relation suivante :

$$\iint_{x,y} \mathbb{I}_A(y) f(x) K(x,y) dxdy = \iiint_{x,y,z} \mathbb{I}_A(y) f(x) q(y|x)\left[\delta_z(y)\rho(x,z) + \delta_x(y)(1-\rho(x,z)\right] dxdydz$$

En posant $D = \{(x,y); f(y)q(x|y) > f(x)q(y|x)\}$, la relation devient :

$$\begin{aligned}
\iint f(x)K(x,y)\mathbb{I}_A(y)dxdx &= \iint_D \mathbb{I}_A(y)\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}q(y|x)f(x)dxdy \\
&+ \iint_{\overline{D}} \mathbb{I}_A(y)q(y|x)f(x)dxdy \\
&+ \iint_D \mathbb{I}_A(x)\left(1-\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}\right)q(y|x)f(x)dxdy \\
\iint f(x)K(x,y)\mathbb{I}_A(y)dxdx &= \iint_D \mathbb{I}_A(y)f(y)q(x|y)dxdy \\
&+ \iint_D \mathbb{I}_A(x)f(x)q(y|x)dxdy \\
&+ \iint_{\overline{D}} \mathbb{I}_A(y)f(x)q(y|x)dxdy \\
&- \iint_D \mathbb{I}_A(x)f(y)q(x|y)dxdy \\
&= \iint \mathbb{I}_A(y)f(y)q(x|y)dxdy \\
&= \int_A f(y)dy
\end{aligned}$$

En effectuant un changement de variable $(x,y) \to (y,x)$ sur les deux dernière intégrales, D se voit transformé en son complémentaire $\overline{D}$ ce qui les annulent. En opérant le même changement sur la première intégrale, le domaine d'intégration réunit D et $\overline{D}$ pour reconstituer

l'ensemble de l'espace Ω.

Le passage de la dernière égalité provient de la propriété des noyaux de transition et donc de $q(x|y) : \int q(x|y)dx = 1, \forall x$ □

La mise en évidence que f est une loi stationnaire par la chaîne de Markov issue de l'algorithme de Hastings-Metropolis n'est pas suffisante, cette propriété doit être complétée par le fait que cette loi peut être atteinte pour n'importe quels points de départ : condition dite d'"ergodicité" que nous ne développerons pas ici.

Annexe C

Echantillonnage de Gibbs

L'échantillonneur de Gibbs, proposé par Geman et Geman dans [GEM84] est une méthode *M.C.M.C.* qui permet de simuler une distribution $f(X)$ lorsque $X = (x_1, x_2, ..., x_N)$ avec $N \geq 2$ et que les lois conditionnelles $f_i(x_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_N)$ notées $f_i(x_i|x_{-i})$ peuvent être simulées aisément.

L'échantillonneur de Gibbs multidimensionnel par construction nécessite la connaissance des lois conditionnelles de f. Cette méthode est la composition de N algorithmes de Metropolis-Hastings avec une acceptation systématique où la loi de proposition est imposée par la méthode.

Le principe de base de cet algorithme est de réactualiser une à une chacune des composantes et de décomposer ainsi le problème en autant de sous-cas qu'il y a de dimensions.

La preuve est donnée par le théorème suivant :

Théorème C.0.4 (Théorème)**.**

L'algorithme d'échantillonnage de Gibbs correspond à la composition de N algorithmes de Hastings-Metropolis avec des probabilités d'acceptation uniformément égales à 1.

Démonstration. $\forall i \in [1, .., N]$, la loi instrumentale de l'étape i du théorème de Hastings-Metropolis [3.2.6.] est donnée par la relation suivante :

$$q_i(y|x) = \delta_{(x_1,...,x_{i-1},x_{i+1},...,x_N)}(y_1, ..., y_{i-1}, y_{i+1}, ..., y_N) \times f_i(y_i|x_1, ..., , x_{i-1}, x_{i+1}, ..., x_N)$$

Le rapport de probabilité $\rho(y, x)$ de l'algorithme de Hastings-Metropolis devient :

$$\begin{aligned}
\frac{f(y)q_i(x|y)}{f(x)q_i(y|x)} &= \frac{f(y)\delta_{(y_1,...,y_{i-1},y_{i+1},...,y_N)}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_N) \times f_i(x_i|y_1, ..., , y_{i-1}, y_{i+1}, ..., y_N)}{f(x)\delta_{(x_1,...,x_{i-1},x_{i+1},...,x_N)}(y_1, ..., y_{i-1}, y_{i+1}, ..., y_N) \times f_i(y_i|x_1, ..., , x_{i-1}, x_{i+1}, ..., x_N)} \\
&= \frac{f(y)\delta_{(y_1,...,y_{i-1},y_{i+1},...,y_N)}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_N) \times f_i(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)}{f(x)\delta_{(y_1,...,y_{i-1},y_{i+1},...,y_N)}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_N) \times f_i(y_i|x_1, ..., , x_{i-1}, x_{i+1}, ..., x_N)} \\
&= \frac{f(y)f_i(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)}{f(x)f_i(y_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)} \\
&= \frac{f_i(y_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)f_i(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)}{f_i(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)f_i(y_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)} \\
&= 1
\end{aligned}$$

□

Nous vous invitons à vous référer à l'ouvrage de Jean-Jacques Droesbeke, Jeanne Fine et Gilbert Saporta [DRO02] pour approfondir le sujet.

Le théorème de Hammersley Clifford constitue une première justification de l'algorithme de Gibbs en affirmant que la connaissance des lois conditionnelles $f_i(x_i|x_{i-1})$ permet de connaître également celle de la loi jointe $f(x)$.

Théorème C.0.5 (Théorème de Hammersley-Clifford)**.**

Sous condition de positivité (le support de f est le produit cartésien des supports des f_i), la probabilité $f(x)$ se factorise en un facteur de densités conditionnelles :

$$f(x_1, ..., x_N) \propto \prod_{i=1}^{N} \frac{f_{l_i}(x_{l_i}|x_{l_1}, ..., x_{i-1}, y_{l_{i+1}}, ..., y_{l_N})}{f_{l_i}(y_{l_i}|x_{l_1}, ..., x_{l_{i-1}}, y_{l_{i+1}}, .., y_{l_N})}$$

Quelle que soit la permutation l sur $\{1, 2, ..., N\}$ et $y \in Y$.

L'algorithme d'échantillonnage de Gibbs est structuré de la façon suivante :

Théorème C.0.6 (Algorithme d'échantillonnage de Gibbs)**.**

S'il existe un entier $N \geq 2$ tel que $x \in X$ se décompose en $(x_1, ..., x_N)$ et si les densités conditionnelles sont simulables, l'algorithme d'échantillonnage de Gibbs associé à cette décomposition a pour transition de x^i à x^{i+1} :

- *1°) Pas 0 - Initialisation de $x^{(0)} = (X_1^{(0)}, ..., X_N^{(0)})$*

- *2°) i^{eme} pas - Simulation de :*

$$x_1^{(i+1)} \sim f_1(x_1|x_2^{(i)}, ..., x_N^{(i)})$$

$$x_2^{(i+1)} \sim f_2(x_2|x_1^{(i+1)}, x_3^{(i)} ..., x_N^{(i)})$$

$$\vdots$$

$$x_k^{(i+1)} \sim f_k(x_k|x_1^{(i+1)}, ..., x_{k-1}^{(i+1)}, x_{k+1}^{(i)}, ..., x_N^{(i)})$$

$$\vdots$$

$$x_N^{(i+1)} \sim f_N(x_N|x_1^{(i+1)}, ..., x_{N-1}^{(i+1)})$$

- *3°) Incrémentation de i et retour à l'étape précédente tant que le critère d'arrêt n'est pas satisfait*

Sous certaines conditions, ce processus génère une chaîne de Markov $(x^t)_t$ dont les éléments convergent en distribution vers des tirages de $f(x)$, densité jointe a posteriori lorsque le nombre d'itérations k tend vers l'infini.

Annexe D

Echantillonnage de Gibbs avec complétion

Cette section présente une généralisation de cet algorithme qui permet parfois de simplifier la simulation des lois.

Définition : La densité g est une complétion de f si :

$$\int_Z g(x,z)dz = f(x)$$

Les lois conditionnelles de g étant parfois plus simple à simuler que celles de f, il est parfois préférable d'utiliser la forme plus générale de cet algorithme après complétion :

Soit $y = (x,z)$ une complétion des données x en z, de densité $g(y) = g(y_1,...,y_N)$ telle que ses densités conditionnelles : $g_1(y_1|y_2,y_3,..,y_N), g_2(y_2|y_1,y_3,...,y_N),...,g_N(y_N|y_1,y_2,...,y_{N-1})$ soient simulables.

L'algorithme devient :

1°) Pas 0 - Initialisation de $y^{(0)} = (Y_1^{(0)},...,Y_N^{(0)})$
2°) i^{eme} pas - Simulation de :

$$y_1^{(i+1)} \sim g_1(y_1|y_2^{(i)},...,y_N^{(i)})$$
$$y_2^{(i+1)} \sim g_2(y_2|y_1^{(i+1)},y_3^{(i)}...,y_N^{(i)})$$
$$\vdots$$
$$y_k^{(i+1)} \sim g_k(y_k|y_1^{(i+1)},...,y_{k-1}^{(i+1)},y_{k+1}^{(i)},...,y_N^{(i)})$$
$$\vdots$$
$$y_N^{(i+1)} \sim g_N(y_N|y_1^{(i+1)},...,y_{N-1}^{(i+1)})$$

3°) Incrémentation de i et retour à l'étape précédente tant que le critère d'arrêt n'est pas satisfait

Dans cette extension de l'algorithme de Gibbs, la chaîne de Markov converge en distribution vers des tirages de $g(y)$, densité jointe complétée.

Annexe E

Algorithme *Data Augmentation* (D.A.)

E.1 Imputation Step

Pour chaque itération, partant d'un vecteur moyenne μ et d'une matrice de covariance Σ donnés, les valeurs manquantes sont tirées selon une distribution conditionnelle aux données observées.

Supposons pour un motif de valeurs manquantes donné (par exemple, $[Y_{obs(1)}, Y_{obs(3)}, Y_{obs(4)}, Y_{mqt(2)}, Y_{mqt(5)}]$) :

$\mu = [\mu'_{obs}, \mu'_{mqt}]'$ le vecteur moyenne partitionné des deux jeux de variables Y_{obs} et Y_{mqt}, où μ_{obs} est le vecteur moyenne des variables Y_{obs} et μ_{mqt} celui des variables Y_{mqt}

et

$\Sigma = \left[\begin{array}{c|c} \Sigma_{obs,obs} & \Sigma_{obs,mqt} \\ \hline \Sigma'_{obs,mqt} & \Sigma_{mqt,mqt} \end{array}\right]$ la matrice de covariance partitionnée pour ces variables

où $\Sigma_{obs,obs}$ est la matrice de covariance pour les variables observées, $\Sigma_{mqt,mqt}$ celle pour les variables manquantes et $\Sigma_{obs,mqt}$ celle entre les variables observées et les manquantes.

Le calcul de la matrice de covariance conditionnelle aux données observées peut être réalisé en employant le *sweep operator* (Goodnight 1979) sur le pivot de la sous-matrice $\Sigma_{obs,obs}$:

$$\left[\begin{array}{c|c} \Sigma^{-1}_{obs,obs} & \Sigma^{-1}_{obs,obs}\Sigma_{obs,mqt} \\ \hline -\Sigma'_{obs,mqt}\Sigma^{-1}_{obs,obs} & \Sigma_{mqt|obs} \end{array}\right]$$

avec $\Sigma_{mqt|obs} = \Sigma_{mqt,mqt} - \Sigma'_{obs,mqt}\Sigma^{-1}_{obs,obs}\Sigma_{obs,mqt}$ pouvant être utilisé pour calculer la matrice de covariance conditionnelle de Y_{mqt} sous contrôle des Y_{obs}.

La distribution conditionnelle de $Y_{mqt}|Y_{obs} = y_{obs}$, pour une observation s'inscrivant dans le motif de données manquantes précédent, est une loi normale multivariée avec le vecteur moyenne :

$$\mu_{mqt|obs} = \mu_{mqt} + \Sigma'_{obs,mqt}\Sigma^{-1}_{obs,obs}(y_{obs} - \mu_{obs})$$

et la matrice de covariance :

$$\Sigma_{mqt|obs} = \Sigma_{mqt,mqt} - \Sigma'_{obs,mqt}\Sigma^{-1}_{obs,obs}\Sigma_{obs,mqt}$$

E.2 Estimation bayesienne du vecteur moyenne et de la matrice de covariance

Soit $Y = (y_1', y_2', ..., y_n')$ une matrice de taille $n \times p$ composée de n vecteurs indépendants y_i de dimension p suivant chacun une distribution multivariée normale de moyenne inconnue μ et de matrice de covariance est notée Λ.

La matrice de dispersion $A = (Y - \overline{Y})'(Y - \overline{Y}) = \sum_i (y_i - \overline{y})(y_i - \overline{y})' \sim W(n-1, \Lambda)$

$W(n, \Lambda)$ étant une distribution de Wishart de degré de liberté n. La distribution de Wishart est une version multidimensionnelle de la loi du χ^2 qui caractérise les matrices de covariance estimées de lois gaussiennes multivariées.

En notant $\Phi = \Lambda^{-1}$ la matrice de précision, si $A \sim W(n, \Lambda)$ alors $A^{-1} \sim W^{-1}(n, \Phi)$ ou W^{-1} désigne la distribution de Wishart Inverse.

Les distributions a priori de la matrice de covariance Σ et du vecteur moyenne μ sont données par les lois suivantes :

$$\Sigma \sim W^{-1}(m, \Phi)$$
$$\mu|\Sigma \sim N\left(\mu_0, \frac{1}{\tau}\Sigma\right)$$

ou τ est fixé positivement.

Leurs distributions a posteriori sont les suivantes :

$$\Sigma|Y \sim W^{-1}\left(n + m, (n-1)S + \Phi + \frac{n\tau}{n+\tau}(\overline{y} - \mu_0)(\overline{y} - \mu_0)'\right)$$

$$\mu|(\Sigma, Y) \sim N\left(\frac{1}{n+\tau}(n\overline{y} + \tau\mu_0), \frac{1}{n+\tau}\Sigma\right)$$

où $(n-1)S$ est la matrice de dispersion.

E.3 Posterior Step

Une fois les données complétées, la phase Distribution des paramètres a posteriori revient à déterminer la distribution de :

$$P(\mu, \Sigma / Y_{imput}) = \frac{P(Y/\mu, \Sigma)P(\mu, \Sigma)}{P(Y)}$$

Cette étape consiste à simuler les vecteurs paramètres en partant des données augmentées. A chaque itération, cette étape simule le vecteur moyenne a posteriori ainsi que la matrice de covariance a posteriori en multipliant les distributions a priori par la vraisemblance.

Pour beaucoup de modèles classiques, il existe des lois a priori conjuguées. Ces lois a priori assurent que la loi a posteriori soit de même forme paramétrique que la loi a priori. Pour le modèle gaussien multivarié notamment, les lois a priori conjuguées sont la loi a priori normale pour la moyenne μ et la loi de Wishart inverse pour la matrice de covariance Σ (cf section

précédente).

Plusieurs options peuvent être retenues selon le niveau d'information a priori dont on dispose sur les paramètres (aucune, sur μ et Σ ou sur Σ uniquement) .

En réitérant le processus suivant :

$$\theta^0 \rightarrow Y^0_{manq} \rightarrow \theta^1 \rightarrow Y^1_{manq} \rightarrow \theta^2 \rightarrow Y^2_{manq} \rightarrow ...$$

La chaîne de Markov (MCMC) ainsi obtenue converge en probabilité vers :

$$P(Y_{manq}, \theta / Y_{obs})$$

De proche en proche, les itérations de l'algorithme génèrent alternativement des paramètres et des données manquantes approximativement échantillonnés selon une distribution conditionnée par les données initialement observées.