Conservatoire National des Arts et Métiers
Certificat de spécialisation analyste de données massives

STA211 - Entreposage et fouille de données
Responsable : Ndeye NIANG KEITA

NOTE DE SYNTHÈSE
ET
PROFILAGE DE CATÉGORIES SOCIO-PROFESSIONNELLES

Sébastien Gardoll 28 janvier 2017

Table des matières

Table des figures

Liste des tableaux

NOTE DE SYNTHÈSE

1 Introduction data mining

Le data mining ou fouille de données, est un processus d'extraction de connaissances issu de la statistique et de l'intelligence artificielle (apprentissage automatique), à l'usage de prise de décisions. Il doit son essor essentiellement au coût devenu bas de l'acquisition des données (souvent de seconde main), du matériel informatique dont le stockage et la puissance de calcul, et au développement mathématiques et technologiques (NoSQL, MapReduce, etc.). Son but est de découvrir des connaissances profitables à partir d'un grand volume d'informations hétérogènes. Le data mining ne cherche pas nécessairement à expliquer le phénomène étudié mais cherche au moins à prédire une réponse du phénomène.

L'origine des données est diverse : elles peuvent être collectées par sondage, centralisées dans les bases ou les data warehouses des entreprises, provenant d'Internet et son déluge de données, etc. Les data warehouses sont des rationalisations de l'information disséminée dans les différentes bases de données métiers de l'entreprise. L'information dans une data warehouse est classée par thème transversal aux différents métiers de l'entreprise sous forme de bases de données multidimensionnelles. Pour des besoins de performance, les métadonnées sont généralement séparées et ces entrepôts sont subdivisés par thèmes, en de plus petits ensembles : les data marts. Les data warehouses sont accompagnées par un ensemble d'outils d'analyses (OLAP) à l'intention des stratèges de l'entreprise. Enfin, Internet met à portée des données qualifiées de big data, notamment les réseaux sociaux, le web (texte, vidéos, images et son) ou le comportement des utilisateurs du web (ex : clics sur les pages d'un web marchand).

1.1 Processus

Le data mining est une étude qui suit un processus dont les phases sont les suivantes : l'analyse orientée métier (business understanding) qui a pour but de fixer les objectifs que le processus devra atteindre ; l'exploration des données qui cherche à comprendre les données, à évaluer leur fiabilité, les décrire et découvrir leurs relations et leur appartenance à des groupes (modalités ; voir 3) ; la modélisation (1.2) qui cherche à produire le modèle offrant le meilleur compromis entre la qualité de la modélisation, le temps passé à l'optimisation du modèle et le temps de calcul des prédictions du modèle ; le déploiement qui consiste à mettre en production les modèles entraînés en les déployant sur le système hôte et à mettre en place des mécanismes de surveillance.

1.2 Familles de méthodes

Méthodes supervisées
Les méthodes supervisées ont pour but la recherche d'une relation entre des données décrites par un ensemble de valeurs prises par de variables aléatoires et la réponse d'un phénomène renvoyée à partir de ces valeurs. Les méthodes supervisées se divisent en deux sous groupes : les méthodes explicatives (boîtes blanches) qui modélisent la relation par une fonction algébrique ou un ensemble de règles logiques et les méthodes non explicatives (boîtes plus ou moins grises) qui produisent un modèle complexe non linéaire (ex : réseaux de neurones). La figure 1 donne un aperçu de la modélisation supervisée.

Méthodes non supervisées
Les méthodes non supervisées cherchent à mettre en évidence des sous groupes cohérents (exclusifs ou non) dans un ensemble de données, sans que l'appartenance des données au(x) sous groupe(s) ne soit connue. La figure 2 illustre la modélisation non supervisée.

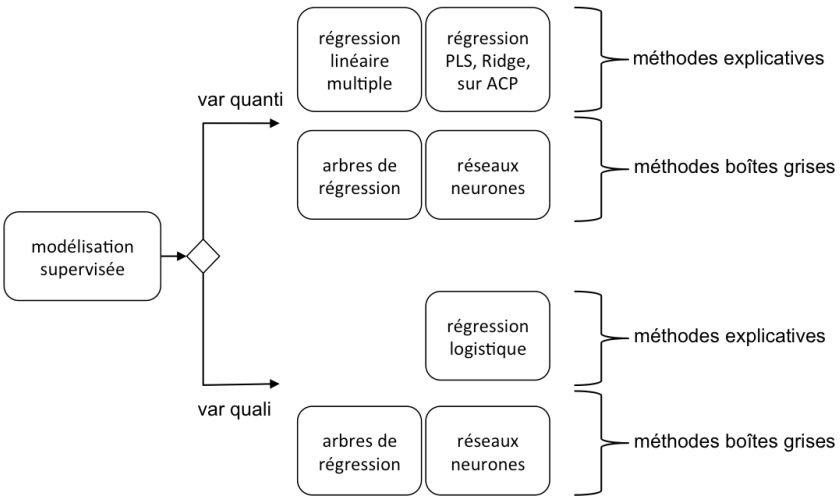

Figure 1 : Modélisation supervisée

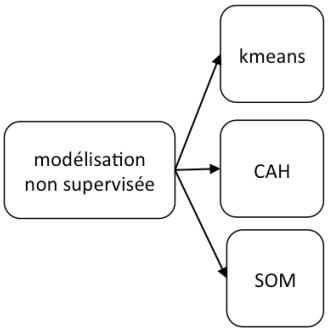

Figure 2 : Modélisation non supervisée

Méthodes mixtes
Ces méthodes ne sont pas exclusives, elles peuvent s'agréger soit elles-mêmes, méthodes d'ensemble-agrégatives (6), avec comme exemples la méthode random forest et gradient boosting machine, soit entre elles (méthodes de stacking) afin de produire un modèle agrégatif plus performant que ces composants pris séparément. La méthode de partitionnement non supervisée mixte (4) est un autre exemple de complémentarité de méthodes.

2 Codage optimal

Le codage optimal (optimal scaling) a pour but de transformer les données brutes d'une étude afin de réaliser des analyses qui ne pourraient pas être effectuées autrement. Cet ensemble de techniques cherchent une nouvelle échelle de mesure des données, respectant les propriétés de l'échelle originelle, tout en optimisant les critères de performance des modèles entraînés sur les données mesurées sur la nouvelle échelle.

Formellement une échelle mesure toute la gamme de valeurs, ordonnées ou non, que prend une variable aléatoire. Les échelles de mesures peuvent prendre une ou l'ensemble des propriétés suivantes :

Égalité Deux variables sont égales si elles ont la même valeur.

Ordre Comparaison entre les valeurs de l'échelle.

Égalité des intervalles ou Égalité des distances relatives des valeurs.

Zéro absolu Référence signifiant l'absence de mesure.

Deux cas se présentent : les données quantifiables et les données qualitatives. Le cas des données quantifiables est simple, il s'agit de les transformer en appliquant une fonction mathématique ou d'effectuer un découpage en intervalles de valeurs correspondant à autant de classes pour une transformation en valeurs qualitatives. Les données qualitatives nécessitent une transformation afin de les rendre numériques. Cette transformation peut être la construction d'un tableau disjonctif complet ou d'un tableau de contingence à partir des valeurs des données qualitatives. La transformation des données nominales consiste à conserver la propriété d'égalité : les deux variables égales dans l'échelle d'origine, le sont également dans la nouvelle échelle. La transformation de données ordinales doivent en plus respecter la propriété de l'ordre, soit basée sur une quantification strictement monotone, soit basée sur une quantification faiblement monotone (des valeurs à l'origine inégales peuvent devenir égales sur la nouvelle échelle).

Une des manières d'obtenir un codage optimal est l'entraînement statistique. Il existe au moins quatre algorithmes pour les données qualitatives : ALSOS, HOMALS, PRINCALS et AFDM (en utilisant le coefficient de corrélation). ALSOS est basé sur l'optimisation d'une fonction de coût par moindre carrés alternés (ALS). Il s'agit d'optimiser les paramètres du codage avec les paramètres du modèle contraint et vice versa. HOMALS est une application de l'ACM qui converge vers l'optimal global et dont les solutions sont emboîtées les unes dans les autres. Cependant, l'ordre de variables ordinales dans la nouvelle échelle est perdu. PRINCALS est une amélioration de HOMALS qui permet le mélange de données mesurées sur des échelles ayant des propriétés différentes et garde l'ordinalité des données. Cependant, les solutions obtenues par PRINCALS ne sont que des optima locaux.

3 Exploration des données

L'exploration des données est illustrée par la figure 3.

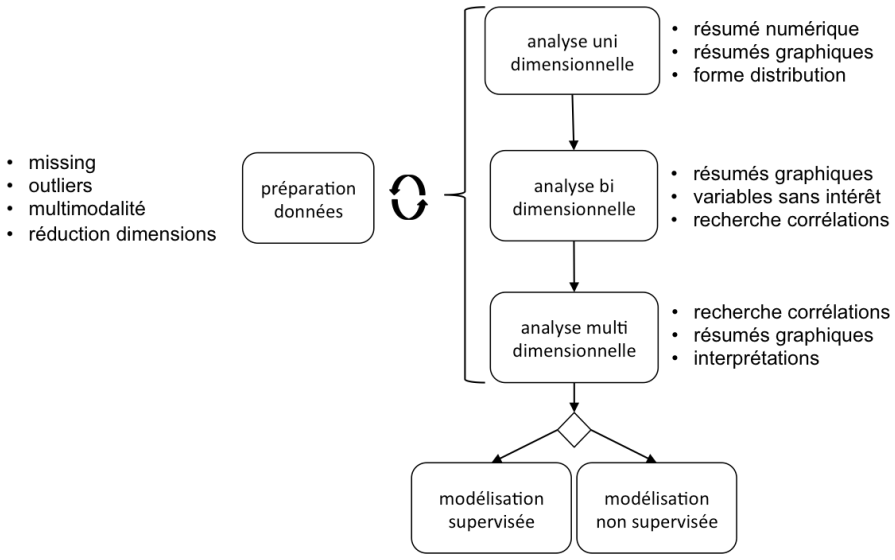

Figure 3 : Exploration des données

Analyse unidimensionnelle
L'étude unidimensionnelle des données permet de se familiariser avec la nature des variables explicatives et les données étudiées par les résumés numériques, les résumés graphiques et la forme de distribution.

Analyse bidimensionnelle
L'analyse bidimensionnelle cherche essentiellement des relations entre les variables explicatives prise

deux à deux, et, dans l'objectif d'une modélisation supervisée, la variable à expliquer et chacune des variables explicatives. L'étude de la corrélation étant sensible aux outliers, les représentations graphiques comme le nuage de points et les parallel box plots servent à les détecter visuellement. Ces représentations peuvent également servir à découvrir des sous groupes de données ou distribution multimodale. Cette dernière présente plusieurs maxima locaux qui pourraient correspondre à des sous groupes distincts. Si le cas est avéré, chaque groupe pourrait être modélisé de façon indépendante. Enfin dans le cadre d'une modélisation supervisée, l'absence de corrélation entre la variable à expliquer et une variable explicative démontre l'inutilité de cette dernière.

Outliers et missings
Les outliers sont des valeurs s'éloignant singulièrement des autres valeurs. Il n'existe pas de formalisation mathématique et en cette absence, il n'existe pas de critère de rejet de ces valeurs. Cependant, le rejet est facilité si la loi de distribution des données est connue ou si l'on sait que le phénomène mesuré est sujet à l'erreur. L'élimination d'observations comportant des outliers et des missings est envisageable lorsque le nombre de données est suffisamment grand sinon l'application de méthodes d'imputation est inévitable. Elles sont souvent basées sur la prédiction de la valeur, soit de façon simple avec la valeur moyenne ou médiane de la série, soit par la prédiction d'un modèle plus ou moins local (k plus proches voisins, etc.). A noter que des méthodes de modélisation comme CART (7) sont robustes à ces deux problèmes.

Analyse multidimensionnelle
L'ACP est une analyse multidimensionnelle sur variables quantitatives uniquement. Son principe consiste à projeter les données d'entrée dans un espace de plus petite dimension, formé par des axes, deux à deux orthogonaux, calculés à partir de combinaisons linéaires des variables explicatives, en maximisant l'inertie des données (variance des variables). Ces combinaisons ou composantes principales sont autant de nouvelles variables décorrélées les unes des autres. La qualité des composantes est proportionnelle à la contribution de celle-ci à l'explication de l'inertie des données, c'est à dire à l'information contenue dans les données. Une autre mesure est la qualité de représentation de l'observation. Il existe plusieurs méthodes pour le choix des composantes à retenir (elbow, seuil, etc.). Comme l'ACP se base sur la notion de distance, elle est sensible à la méthode de calcul de celle-ci (euclidienne, Manhattan, cosinus, etc.) et les données sont généralement centrées et réduites. L'ACP ne donne pas de bon résultat (faible contribution des composantes à l'inertie) si les variables ne sont pas corrélées ou non linéairement corrélées entre elles. Elle permet :

- La visualisation des observations projetées sur un hyperplan (plan factoriel) formé par les axes et l'estimation d'éventuels groupes (voir aussi cartes de Kohonen 9.4).
- La détection d'outliers par l'étude des dernières composantes et la contribution des observations à l'inertie expliquée par les composantes.
- La décorrélation des variables afin d'utiliser des méthodes de modélisation sensibles à la multicolinéarité.
- La visualisation, au moyen du cercle de corrélation, de la corrélation entre les variables et les composantes (proximité du cercle) et les variables entre elles (regroupement).

La projection des données sur le plan factoriel et le cercle de corrélation sont interprétables car les axes sont des combinaisons linéaires des variables explicatives. La recherche de groupes d'observations projetées dans l'espace des axes est possible par interprétation externe et par partitionnement (4). L'interprétation externe consiste à introduire une variable qualitative pertinente servant à distinguer les groupes en étiquetant les observations par les différentes modalités de la variable ajoutée. La caractérisation des groupes est traitée section 4.

Les méthodes suivantes sont comme l'ACP, des méthodes de réduction de dimension. Elles se distinguent par la nature et le nombre de variables des données traitées : AFC (données à deux variables

qualitatives, transformées en tableau de contingence), ACM (données à plusieurs variables qualitatives, transformées en tableaux disjonctif complet) et AFDM (données à mélange de variables quantitatives et qualitatives).

4 Partitionnements

Partitionnement hiérarchique (CAH)
Le CAH cherche une suite de groupes emboîtés les uns dans les autres. Plus le groupe est profond, plus les éléments de ce groupe sont similaires. Le CAH est traditionnellement représenté par un dendrogramme. Le CAH ne fait pas l'hypothèse du nombre de groupes. Le CAH est agglomératif lorsqu'il s'agit de rassembler les éléments entre eux pour former des groupes puis itérer sur ces derniers et il est divisif lorsqu'il s'agit de diviser l'ensemble des éléments en deux groupes puis itérer sur ces derniers.

CAH agglomératif
Le critère d'agglomération des groupes est basé sur un critère de similarité comme la minimisation de la distance entre les éléments les plus proches ou les plus éloignés de chaque groupe pris deux à deux, la minimisation de la distance entre le centre de gravité de chaque groupe pris deux à deux ou la minimisation de la perte d'inertie intragroupe (équation 4) après agglomération des groupes pris deux à deux (critère de Ward). Bien sûr, ces méthodes sont très sensibles au choix du calcul de la distance.

CAH divisif
Le critère de division est basé sur la corrélation des éléments aux deux premières composantes principales issues de l'ACP sur les éléments de l'ensemble courant, chacune des composantes représentant un groupe. Ce principe est répété sur les deux groupes d'éléments ainsi créés. L'arrêt de la division intervient lorsque que la corrélation est plus petite qu'un certain seuil (ex : 1).

Partitionnement non hiérarchique
Pour le partitionnement direct, le nombre de groupes est fixé d'avance. Comme il n'est souvent pas possible de calculer la distance de toutes les combinaisons des éléments en temps raisonnable, ce partitionnement a recourt à des heuristiques comme les centres mobiles où les solutions obtenues ne sont que des optima locaux. Ces heuristiques suivent un algorithme à peu près similaire. Au départ, il s'agit de choisir aléatoirement les centres provisoires des groupes puis de regrouper les éléments selon leur proximité (voir les critères de similarité CAH) à ces centres, et enfin, de recalculer le nouveau centre des groupes et de recommencer itérativement. On arrête l'itération lorsque les éléments ne changent plus de groupe. La variante kmeans recalcule le centre d'un groupe à chaque affectation d'un élément à ce dernier.

Partitionnement mixte
Il s'agit d'utiliser séquentiellement les deux types de partitionnement afin de tirer parti de leurs avantages. Appliquer un partitionnement direct avec un grand nombre de groupes pour obtenir des groupes fortement homogènes. Puis un partitionnement hiérarchique sur le résultat précédent afin de réunir les groupes inutilement séparés et d'évaluer un nombre optimal de groupes. Enfin consolider par partitionnement direct avec le nombre de groupes, précédemment obtenu.

Caractérisation
Généralement le partitionnement a pour but de résumer les éléments des groupes. Si les éléments sont des observations, le groupe peut représenter un profil d'observations. Le partitionnement de variables explicatives, effectué sur la matrice de dissimilarité entre les variables, permet de déduire la multicolinéarité des variables entre elles. Elle sert également de réduction de dimensions, en résumant des groupes de variables explicatives. Par ailleurs, le centre de gravité d'un groupe est un résumé possible cependant, seul, il ne tient pas compte de la dispersion des éléments. La valeur test le complète en

mesurant la représentativité du centre pour un élément donné.

5 Performance

L'estimation des performances d'un modèle supervisé repose sur le calcul de l'erreur du modèle, basé sur l'écart de prédiction du modèle pour un jeu de données étiquetées. Sachant que la mesure de l'erreur est biaisée si on évalue celle-ci à partir du jeu de données utilisé pour l'entraînement du modèle, la validation consiste à créer trois jeux de de données suffisamment grands : un jeu d'entraînement, un jeu de validation d'hyper-paramètres et un jeu de test d'erreur. Le jeu d'entraînement sert à produire des modèles d'une même méthode ayant des hyper-paramètres différents. En calculant l'erreur des précédents modèles sur le jeu de validation, on parvient à choisir les hyper-paramètres optimaux. Enfin, le modèle est entraîné sur l'union des jeux d'entraînement et de validation et son erreur est calculé à partir du jeu de test. Il est alors possible de comparer les méthodes entre-elles.

Validation croisée
La validation croisée consiste à générer des échantillons à partir de l'échantillon de données étiquetées disponible moins un jeu de test. La variante LKOV retire k éléments de l'échantillon de départ qui serviront de données de validation. Les autres éléments servent à l'entraînement du modèle. On procède ainsi pour toutes les combinaisons possibles de k éléments. La variante LOOV est le cas particulier où k est égale à 1. K-fold cross validation est une variante non exhaustive qui consiste à diviser aléatoirement en k sous ensembles équivalents l'échantillon de départ. L'un des sous ensembles sert de jeu de données pour la validation du modèle entraîné sur les données des k-1 ensembles restants. L'opération est renouvelée k fois au total, ce qui diminue sensiblement les temps de calculs.

Bootstrap
Cette méthode génère des échantillons par tirage aléatoire avec remise d'éléments de l'ensemble de données de départ. On simule ainsi des échantillons plus ou moins indépendants (répétitions) dont la taille est la même que l'échantillon de départ.

6 Agrégation de modèles

L'erreur de modélisation est liée à la variance et le biais du modèle (équation 3), l'effort de modélisation porte donc sur leur minimisation (compromis biais-variance). Or pour des variables explicatives dont l'indépendance n'est pas connue, la variance du modèle peut s'exprimer sous la forme :

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

Où σ est la variance , ρ le coefficient de corrélation des paires de variables, B le nombre de prédicteurs. L'idée de l'agrégation de modèle est de diminuer la variance du modèle en entraînant un grand nombre de modèles, augmentation de B, qui se « ressemblent » le moins possible, diminution de ρ. L'agrégation de modèle ne cherche donc pas l'entraînement du meilleur modèle possible au risque de sur-apprentissage mais aboutit à la création d'un méta-modèle agrégeant la prédiction de modèles de la même méthode (à la différence du stacking), de performance généralement moyenne, évitant ainsi le sur-apprentissage. Cependant, l'éventuel aspect explicatif de la méthode agrégée est perdu car le méta-modèle, vu le nombre de ses composants, est considéré comme trop complexe à interpréter. La prédiction du méta-modèle est respectivement la moyenne et le vote, pondérée ou non, des prédictions quantitatives et qualitatives des modèles.

Bagging
Cette méthode prescrit l'entraînement des modèles en parallèle sur des ensembles de données différents, généré par boostrap (5), afin d'éviter la corrélation entre les modèles. L'utilisation conjointe du feature sampling, la sélection aléatoire des variables pour chaque échantillon, renforce cette décorrélation. Ces techniques sont mises en œuvre dans la méthode random forest (7).

Boosting
Cette méthode prescrit l'entraînement en série de modèles. Chaque modèle modélise l'information mal modélisée (éléments mal classés, etc.) par le modèle précédent. Cette technique est mise en œuvre dans la méthode gradient boosting machine (7).

7 Segmentation

La segmentation est une famille de méthodes supervisées non linéaires, basée sur la division successive, selon les variables explicatives, de l'échantillon de départ en sous groupes ou nœuds les plus homogènes vis à vis de la variable à expliquer. Le résultat obtenu est un arbre dont les nœuds sont affectés à une valeur de la variable à expliquer et un ensemble de règles qui expliquent cette répartition. Les règles sont facilement interprétables et la segmentation prend en charge les corrélations et la sélection des variables dont la nature est quelconque. Cependant, elle est très sensible au sur-apprentissage, aux choix des divisions (variables explicatives) et aboutit pour les variables quantitatives à une segmentation dyadique potentiellement problématique pour la modélisation de phénomène quantifiable continue. Déroulement de l'algorithme :

1. A chaque nœud, choisir la meilleure division selon un critère donné.
2. Déclarer selon une règle si le nœud est terminal ou intermédiaire.
3. Affecter aux nœuds une valeur de la variable à expliquer.
4. Estimer l'erreur du nœud (coût de la division).

Le critère de division est une mesure de la pureté des nœuds créés vis à vis de la variable à expliquer. Pour une variable quantitative, la pureté est définie par sa variance. Pour une variable qualitative, on prend généralement l'indice de Gini qui minimise les mal-classés ou le critère d'entropie qui pénalise l'équiproportion entre les modalités de la variable. La règle de l'indivision d'un nœud, autre que la nullité du cardinal des nœuds produits, est variable selon les méthodes (seuils). L'affectation est la moyenne pour une variable quantitative et la modalité la plus représentée ou celle qui provoque les moins de mal classés pour une variable qualitative. Enfin, l'erreur d'un arbre est la somme des erreurs des nœuds de l'arbre.

CART
Le problème du sur-apprentissage provient de l'impossibilité de calculer la meilleure séquence de règles du fait de la croissance exponentielle de l'arbre. CART est une méthode produisant un arbre binaire à l'aide du critère de Gini en élaguant l'arbre pour lequel le nombre de divisions est maximal, $A_{\max}$, de ses branches n'apportant pas une amélioration particulière de la pureté des nœuds. CART génère une séquence d'arbres produits par emboîtement des sous arbres de $A_{\max}$, les uns dans les autres. Calculant pour chaque arbre une erreur pénalisant l'impureté des nœuds et leur nombre (par échantillon test ou validation croisée), on peut choisir l'arbre ayant la plus petite erreur ou ceux ayant une erreur comprise dans l'intervalle de l'erreur de $A_{\max}$ et de l'écart type. Cependant, la solution obtenue est un optimum local.

Random forest
Plutôt que d'élaguer l'arbre et obtenir un seul prédicteur, random forest est une méthode agrégative (6)

aboutissant à un méta modèle composé d'un grand nombre d'arbres non élagués. Par conséquence, on obtient un gain de temps d'entraînement des arbres, également raccourci par leur construction en parallèle. Malgré l'aspect boîte grise, il est possible de mesurer l'importance de chaque variable explicative. La fréquence d'apparition de chaque variable dans les arbres est un critère basique, le MDA basé sur l'effet d'une permutation aléatoire des valeurs des variables et le MDG basé sur la décroissance de l'hétérogénéité selon le critère de Gini, sont des mesures plus représentatives.

Gradient boosting machine
Cette méthode reprend les principes du boosting des méthodes agrégatives (6) et généralise par la méthode de résolution descente de gradient, la recherche d'une solution d'une fonction de coût quelconque (libre choix).

8 Règles d'association

Définition
Soit un ensemble d'items I et un ensemble de transactions T où chaque transaction est un sous ensemble de I. Une règle d'association décrit la co-occurence de A (antécédent) et C (conséquent), deux sous ensembles disjoints de I de tel sorte que $A \Rightarrow C$. Une règle d'association est caractérisée par son support, sa confiance et des indices de pertinence. Le support est le nombre de transactions contenant A et C. Il reflète la fiabilité de la règle. La confiance mesure la proportion de transactions comportant C parmi les transactions contenant A. Elle reflète la précision de la règle. Cependant, le support et la confiance ne mesurent pas complètement l'intérêt des règles. Les indices de pertinence (lift, jaccard, confiance centrée, lœvinger, multiplicateur de côtes, etc.) viennent les compléter chacun dans un contexte donné. Ainsi l'indice de lift favorise les règles impliquant des événements rares, tandis que l'indice de jaccard favorise les règles où les probabilités de A, C et $A \cap C$ sont proches.

Génération des règles
Il existe plusieurs algorithmes (Apriori, DIC, etc.) décrivant la construction de règles pertinentes de façon optimisée. La combinaison des items des transactions aboutissant à un très grand nombre de règles, ces algorithmes décrivent un certain nombre d'étapes afin de construire les règles et spécifient généralement des seuils pour le support et la confiance afin d'éliminer à chaque étape les règles inintéressantes. Or, l'établissement de ces seuils est insuffisant. Par exemple, les règles impliquant un item dont la probabilité d'apparition parmi les transactions est forte (support élevé mais confiance faible), ne sont pas d'un grand intérêt. A contrario, les événements rares sont des règles très intéressantes qui concernent des items rarement impliqués dans les transactions, leur support est faible et décrivent une co-occurence exclusive, leur confiance est proche de 1. Les indices de pertinence complètent les critères de support et de confiance choisis moyennement ou peu discriminants, en aidant l'élimination des règles peu intéressantes. Concernant l'optimisation, on note que si $C' \subset C$, $A \Rightarrow C$ est valide si toutes les règles de la forme $A \Rightarrow C'$ le sont aussi.

9 Réseaux de neurones

Les réseaux de neurones artificiels, d'inspiration biologique, sont utilisés pour modéliser des problèmes supervisés, généralement des classements (traités dans cette section), et des problèmes non supervisés (9.4). Dans leur emploi supervisé, ils prennent comme entrée des vecteurs de données ou d'entrée et donne généralement une modalité d'une variable nominale en sortie.

9.1 Perceptron

Un neurone formel est une composition, donnée par l'équation 2, d'une fonction linéaire dite d'agrégation et d'une fonction non linaire dite d'activation, notée σ. La fonction d'agrégation est une combinaison de poids notés w_i et des n composantes d'un vecteur d'entrée notées x_i, seuillées par un scalaire que l'on appelle biais, noté b.

$$f(x) = \sigma(b + \underbrace{\sum_{i=1}^{n} w_i x_i}_{\text{agrégation}}) \tag{2}$$

Le perceptron est un classifieur binaire, composé d'un neurone dont la fonction d'activation est une fonction sigmoïde dont le résultat peut être interprété comme une probabilité ou potentiel d'activation du neurone lorsque cette probabilité tend vers 1. Afin de modéliser des problèmes multi-classes, on empile des neurones sous forme de couche où chaque neurone est associé à une classe du problème et est connecté aux composantes du vecteur d'entrée. On parle de réseau entièrement connecté. La méthode d'apprentissage des poids et du biais (paramètres) de chaque neurone est conduite par la minimisation d'une fonction de coût, optimisée par descente de gradient et validation (5). Cependant, le perceptron est limité aux séparations linéaires par hyperplan. Le perceptron multicouches (MLP) apporte une solution avec l'empilement de couches de neurones toutes interconnectées. Le MLP permet d'approximer n'importe quelle fonction avec un sur-apprentissage contenu (relativement peu de paramètres). Chaque couche peut être vue comme une représentation abstraite différente de la donnée : de la plus proche de la donnée pour le première couche, à la plus proche des classes à modéliser pour la dernière couche. L'apprentissage des paramètres du réseau est optimisé par l'algorithme de rétro-propagation du gradient. Cet algorithme repose sur deux principes, la descente de gradient, processus itératif, qui calcule pour une fonction de coût d'un modèle, les ajustements à apporter aux paramètres du modèle afin de l'optimiser et la rétro-propagation de l'erreur dont l'idée est de calculer l'erreur des neurones d'une couche en comparant leur résultat avec ce qu'il aurait valu pour activer certains neurones de la couche suivante. Cet algorithme initialise aléatoirement la valeur des paramètres et commence par la couche la plus basse pour laquelle la fonction de coût est calculable par validation, puis continue de proche en proche vers la première couche. Les fonctions de coût des réseaux de neurones n'étant généralement pas convexe, la solution trouvée est un optimum local toutefois proche de l'optimum global (effet « vallée plate »). Enfin pour des raisons de temps de calcul, la version stochastique de la descente de gradient est généralement préférée (approximation du gradient). Le sur-apprentissage est contrôlé par validation (5) et par l'utilisation de méthodes de régulation (désactivation aléatoire de neurones, etc.). Cependant, les MLP possédant un nombre important de couches (profonds), ont certaines limitations. Notamment, le gradient de leur fonction de coût peut prendre des valeurs extrêmement grandes ou à l'inverse des valeurs tendant vers zéro (évanouissement), lors du déroulement de l'algorithme, entraînant une perte d'information. D'autre part, leur nombre de paramètres croît exponentiellement car leurs neurones sont toutes interconnectées. L'entraînement du réseau fait donc face aux problèmes liés à la grande dimension (malédiction) et à une consommation importante de ressources informatiques. De plus, MLP n'est pas robuste aux modifications non sémantiques de l'information (ex : transformations spatiales) car il est insensible à sa structure et ne détecte pas ses propriétés invariantes. Ainsi, MLP n'est pas une bonne méthode pour les données de type signal (image, son, etc.).

9.2 Réseaux neuronaux convolutifs (CNN)

Les CNN résolvent les limitations des MLP profonds en traitant les données, avant des couches entièrement connectées MLP, afin d'en générer des représentations internes morcelées, robustes aux

modifications non sémantiques et plus proches de la sémantique des classes à modéliser tout en étant économe en nombre de paramètres. Géométriquement, les CNN démêlent et redressent les données dans un espace où leur classification est facilitée. Ce traitement, intégré dans l'architecture du réseau, consiste en l'alternance (fig 4) de couches convolutives, de couches de pooling et de couches d'activation effectuant une transformation non linéaire (RELU remplaçant les sigmoïdes pour un gain de performance et comme solution au problème d'évanouissement de gradient).

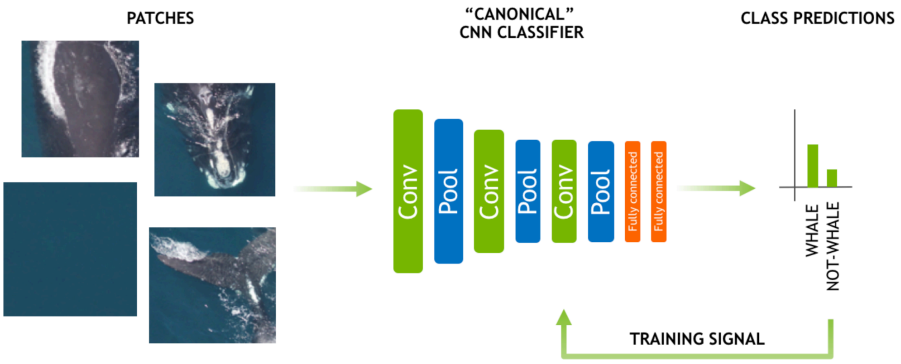

Figure 4 : Schéma MLP (Nvidia deep learning laboratories)

Couche convolutive
Une couche convolutive est un empilement de couches de neurones qui agissent comme des filtres traitant des morceaux d'une matrice de données. En effet, chacun des neurones est connecté à une subdivision de la matrice de données (ex : patch pour une image), d'où la notion de convolution. Il s'agit d'affecter à un petit nombre de paramètres des neurones, une valeur non nulle. De plus, ces valeurs sont partagées par l'ensemble des neurones d'une couche, ce qui réduit drastiquement le nombre de paramètres à entraîner. Ces caractéristiques assurent l'indépendance de localisation spatiale des motifs détectés d'où une robustesse aux translations spatiales. Enfin, chaque couche de neurones est comme un filtre (FIR), appris automatiquement, sensible à un aspect différent des classes à discriminer. Une couche de convolution traite une matrice en entrée et restitue un tenseur en sortie qui s'apparente à une pile d'abstractions plus proches des classes.

Couche de pooling
La couche de pooling consiste à résumer (max, moyenne, etc.) indépendamment (agrégation spatiale) des subdivisions d'une matrice de données, une forme de sous échantillonnage. Elle participe à la réduction du nombre de paramètres (donc au sur-apprentissage), de l'empreinte mémoire et participe à l'apprentissage de l'invariance par translation des données.

9.3 Deep learning

Il correspond à l'évolution actuelle des réseaux de neurones, de plus en plus profonds. Cette évolution est dynamisée ces dernières années par l'amélioration des ressources de calcul, notamment le calcul sur GPU qui profite aux couches convolutives bien parallélisables. Elle est également favorisée par l'émergence de frameworks de développement, d'une augmentation du nombre de données étiquetées (par transformations ou non) et quelques améliorations techniques comme le chevauchement et une meilleure gestion des marges des subdivisions des couches convolutives (padding), la diminution du phénomène de co-adaptation (sur-apprentissage) des neurones par une forme de moyenne ou vote de ceux-ci (dropout) et l'apparition de nouvelles techniques comme le deep feature (DF) et le Generative Adversarial Network (GAN). DF consiste à extraire les premières couches d'un réseau entraîné afin de les brancher à un classifieur modélisant un problème différent (MLP, SVM, etc.). L'extrait agit comme une méthode de vectorisation des données, intéressant palliatif aux problèmes d'accès aux ressources d'entraînement (nombre de données insuffisant, ressources de calcul, etc.), donnant de bons résultats.

Le GAN est une architecture cherchant à transformer un problème de modélisation non supervisé en un problème supervisé selon la théorie des jeux. Il se base sur la compétition entre un réseau cherchant à générer des données les plus réalistes possible vis à vis de l'étude, et un réseau qui les discernent à l'aide d'une base de données étiquetées.

9.4 Cartes de Kohonen (SOM)

SOM est une méthode de modélisation non supervisée et de représentation graphique de données, basée sur un réseau à une couche de neurones entièrement connectés. Les neurones forment une grille de topologie variable (carrée, linéaire, etc.). Les neurones étant munis d'indice, on définit une distance relative entre ceux-ci (graphe). Chaque neurone est associé à un vecteur, dit de référence, dans l'espace des données à modéliser et dont les composantes sont les poids du neurone. Cette grille revient à diviser l'espace d'entrée en zones, appelées cartes, où la contiguïté des cartes est fidèlement retranscrite sur la grille de neurones. L'entraînement des neurones démarre par l'initialisation aléatoire des poids des neurones, la solution obtenue est donc un optimum local. Il est donc nécessaire d'exécuter plusieurs fois SOM. Puis pour chaque vecteur d'entrée (version stochastique), le neurone dont le vecteur de référence est le plus proche, est déclaré gagnant. Les vecteurs de référence de ce neurone ainsi que ceux de ses voisins sont déplacés. L'amplitude du déplacement est fonction d'un pas d'apprentissage, de la différence entre les vecteurs et d'une fonction, dite température, définissant l'appartenance au voisinage (plusieurs noyaux possibles). Cette dernière assure que les déplacements des vecteurs de référence des neurones éloignés du neurone gagnant sont bien moins importants que ceux associés aux neurones proches du neurone gagnant. Les vecteurs de référence sont ainsi les centres de gravité des cartes, calculés en prenant en compte toutes les données selon leur distance relative (au contraire de kmeans). De proche en proche, les cartes sont déformées afin de rendre compte de la répartition des données successivement présentées (ou par lots pour la version batch). La grille donne une représentation en basse dimension de l'espace d'entrée. A noter que SOM est moins sensible aux outliers que kmeans. Le résultat de la fonction température évolue jusqu'à être constant vers les dernières itérations de l'algorithme. A la fin, le comportement de la méthode tend vers celui de kmeans. Si le nombre de cartes (groupes de données) est trop grand, SOM est souvent associée à une CAH afin de réduire leur nombre. L'interprétation des caractéristiques des groupes est classique (4). Enfin, au sujet de la différence entre ACP et SOM, l'ACP est une projection linéaire avec perte d'information alors que SOM est une projection non linéaire sans perte. SOM donne la proximité des données sans les transformer au contraire de l'ACP, cependant les indices des neurones n'ont pas de signification.

10 Données multivues non supervisées

Les données multivues sont des tableaux de données qui peuvent évoluer en nombre d'observations (n_t) et de variables explicatives (p_t). Si l'étude des tableaux séparés ne rendent pas compte de cette évolution, plusieurs méthodes existent pour l'analyser suivant les cas de figures (n_t constant ou p_t ou les deux). Ces méthodes non supervisées suivent les étapes suivantes :

1. Interstructure. A partir des tableaux pris comme objet d'étude (au lieu des observations), l'interstructure étudie globalement l'évolution entre les tableaux.
2. Compromis. C'est le résumé des tableaux selon des critères donnés.
3. Intrastructure. C'est la recherche d'un espace commun aux tableaux afin d'étudier les fines différences entre les observations des tableaux (l'étude monodimensionnelle étant hors sujet).
4. Trajectoires. C'est l'étude de l'évolution des variables ou des observations suivant les tableaux.

Double ACP (DACP)
DACP est une méthode qui ne s'applique que sur les tableaux où n_t et p_t sont constants. L'interstructure est obtenue selon les objectifs de l'analyse. Par exemple, à partir de l'ACP sur le tableau des moyennes des valeurs des variables explicatives de tous les tableaux. Une ACP sur ce résumé de tableaux donne l'évolution globale des tableaux sur le plan factoriel de l'ACP (interprétation classique). L'intrastructure est obtenue en appliquant une ACP sur les données des tableaux concaténés verticalement, correspondant au compromis. La projection des données sur le plan de cet ACP donne la trajectoire : l'évolution des observations (sélectionnées pour plus de clarté).

STATIS
Soit N tableaux-matrices de données X_t avec un nombre constant $n_t = n$ d'observations et p_t non constant, avec $t = 1, \dots, N$. Puisque les matrices de données ne sont pas de même taille, il n'est donc plus question d'appliquer une ACP sans transformer les données. Statis se base sur le coefficient RV, généralisation du coefficient de corrélation de Pearson qui s'applique sur les matrices (similitude). Cette méthode commence par calculer les matrices W_t composées des produits scalaires deux à deux des vecteurs observations provenant des tableaux de données. Les W_t sont donc des matrices carrées de taille $n_t \times n_t$, chacune exprimant les liens inter-individus des X_t. Puis Statis construit la matrice S carrée composée des coefficients RV calculés à partir des matrices W_t prises deux à deux. Ainsi S exprime l'évolution du lien inter-individus entre les tableaux de données. L'interstructure est obtenue en appliquant une ACP sur S. Le calcul de cette ACP est facilité par la nature du coefficient RV : sachant qu'il est calculé à partir des produits scalaires entre W_t, le calcul des composantes principales de la matrice S est obtenu en diagonalisant S puis en calculant les valeurs et vecteurs propres du résultat (substitution de la matrice notée W par S dans l'équation 5). Cependant, l'interprétation sur le plan factoriel de l'évolution des liens inter-individus est limitée car les axes ne sont pas interprétables d'où la recherche de l'intrastructure. La recherche de l'intrastructure consiste à la construction d'un espace commun aux données (indice « co »), formé par la matrice W_{co} qui est une combinaison linéaire des W_t. Les poids de cette combinaison sont les coordonnées des W_t sur le premier axe principal (par définition le plus corrélé avec les W_t) de l'interstructure. W_{co} est le compromis, reflet de la ressemblance des W_t dont on peut avoir une idée avec le cercle de corrélation provenant de l'interstructure : plus les W_t apparaissent groupés, plus W_{co} est représentatif de tous les W_t. L'ACP sur W_{co} donne l'intrastructure et la projection des données de départ sur le plan factoriel de cette ACP permet de visualiser la trajectoire des données et le cercle de corrélation de cette ACP donne une interprétation des axes factoriels selon les variables explicatives. Le partitionnement des trajectoires est parfois nécessaire pour en interpréter un grand nombre.

Autres
Statis Dual reprend la même démarche que Statis avec n_t non constant, p_t constant et V_t la matrice de covariance des variables au lieu de W_t. Statis Dual se concentre donc sur l'évolution des variables plutôt que des observations comme Statis. L'AFM qui est un méthode factorielle applicable à un mélange de variables quantitatives et qualitatives, structurées en groupes, est également une méthode d'analyse de données multivues, basée comme une extension de l'AFDM.

PROFILAGE DE CATÉGORIES SOCIO-PROFESSIONNELLES

11 Projet

L'Observatoire de la Qualité de l'Air Intérieur (OQAI) a mené en 2007 une étude nationale [1] concernant la pollution de l'air de 24 millions de résidences en France. L'OQAI a livré un jeu de données de 528 observations décrites par un peu plus de 30 variables rassemblées en trois ensembles : les variables concernant les caractéristiques des logements, les variables concernant les habitudes des personnes qui y résident et les variables décrivant les caractéristiques des ménages. Bien que les variables soient orientées pollution de l'air, ce projet a pour but de révéler des profils socio-professionnels basés sur ces données.

11.1 Données

Les données proviennent d'une base validée par l'OQAI et les données manquantes ont été complétées par l'AFSSET. J'ai sélectionné, parmi la trentaine de variables disponibles, celles qui me paraissent liées à la profession et aux habitudes ménagères. La sélection regroupe 16 variables quantitatives et 13 qualitatives, listées à la section B, résumées à la section C et décrites dans les documents suivants :

- Variables_habitudes.pdf
- Variables_logement.pdf
- Variables_menage.pdf

En bref, il s'agit d'indicateurs concernant les habitudes d'hygiène, l'occupation professionnelle, le niveau d'instruction, les revenus, le type de logement et sa surface, etc.

11.2 Problématiques et méthodes

Le premier objectif de cette étude vise la découverte de profils socio-professionnels. Ces profils se traduisent généralement par des sous ensembles d'observations distincts. Le deuxième objectif consiste à modéliser la Catégorie Socio-Professionnelle (CSP) des observations (variable PROFES1) et dans une certaine mesure à l'expliquer à l'aide des variables sélectionnées. Concernant les profils (14), j'ai choisi d'appliquer un partitionnement non supervisé à l'aide de la méthode SOM (14.2). Comme cette méthode n'accepte que des vecteurs, un codage optimal par la méthode AFDM (14.1) est appliqué aux données en préliminaire. Les groupes d'observations sont ensuite caractérisés (14.3). Concernant la modélisation de la CSP (15), j'ai choisi une modélisation explicative par la méthode CART appliquée à l'ensemble des observations (15). Cette étude suit le processus évoqué à la section 3 et s'ouvre donc sur l'étude uni (12) et bidimensionnelle (13) des données de l'OQAI.

12 Étude unidimensionnelle

12.1 Variables quantitatives

Les résumés des variables quantitatives donnés en annexe C montrent que la majeure partie des variables souffrent d'outliers (présence d'observations au-delà du troisième quartile des box plots), particulièrement les variables HSRF (surface de logement), NPVe1 (nombre de pièces de vie), REV3

1. http ://www.oqai.fr/userdata/documents/293_Rapport___Typologie_des_logements.pdf

(revenus) et ICOS1 à ICOS4 (indices d'hygiène). Le test de Shapiro-Wilk à 0,5% montre que aucune des variables ne suit de loi normale.

12.2 Variables qualitatives

Les résumés des variables qualitatives donnés en annexe D ne montrent pas de sous représentation de modalités mise à part les classes 1, 2, et 7 de la variable PROFES1 (respectivement agriculteur, artisans et autre profession) et les classes 2, 3, 5 et 6 de la variable NOCCUA (respectivement chômeur, étudiant, au foyer et autre inactif).

13 Étude bidimensionnelle

13.1 Corrélation entre variables quantitatives

Sachant que aucune variable quantitative ne suit de loi normale, la corrélation entre ces variables est évaluée à l'aide du test de corrélation de rang de Spearman. La figure 5 montre la matrice des coefficients en valeur absolue. Une corrélation de zéro signifie que le l'hypothèse nulle est rejetée au risque de 0,5%. On peut en déduire que les variables quantitatives (centrées et réduites) sont très faiblement ou pas corrélées entre elles.

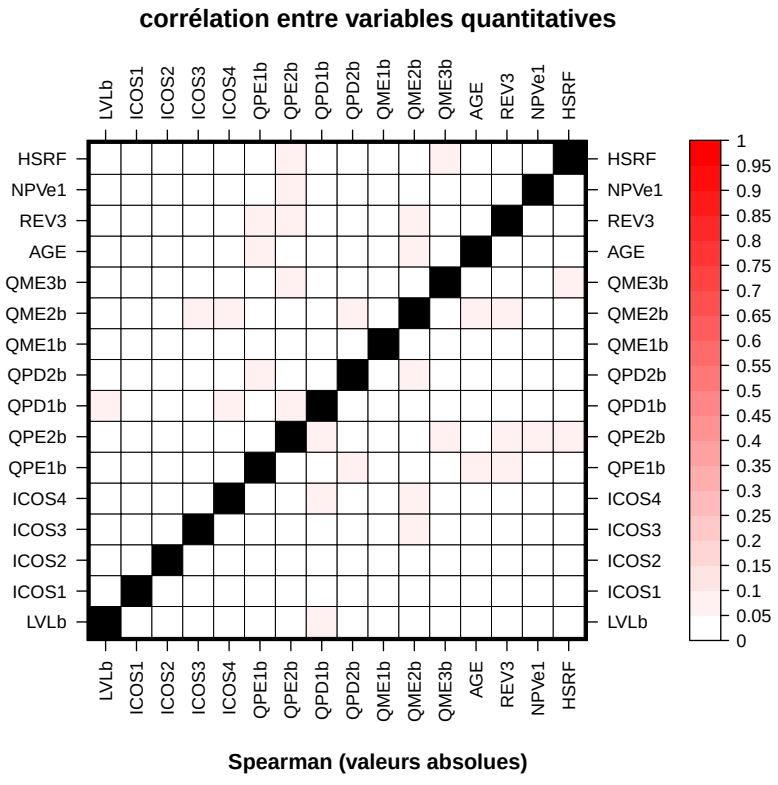

Figure 5 : Corrélation entre variables quantitatives

13.2 Corrélation entre variables qualitatives

La matrice de corrélation entre les variables qualitatives, montrée à la figure 6, donne l'intensité de la corrélation selon le v de Cramér si le test du χ^2 (approximation) montre une corrélation entre

les variables prises deux à deux au risque de 0,5%. Dans le cas contraire, l'intensité est forcée à zéro. Dans l'ensemble, les variables qualitatives sont peu ou pas corrélées entre elles. Hormis, la corrélation triviale entre la possession de chiens (ANI21n) avec celle de chats (ANI22n) et des traitements antiparasites pour eux (ANTCPn), cette étude montre une corrélation intéressante entre le type de logement (FC3) et le statut des habitants (FC8 ; propriétaires, locataires, etc.). La variable PROFES1 que l'on souhaite prédire à partir des autres variables, est étudiée en section 13.4.

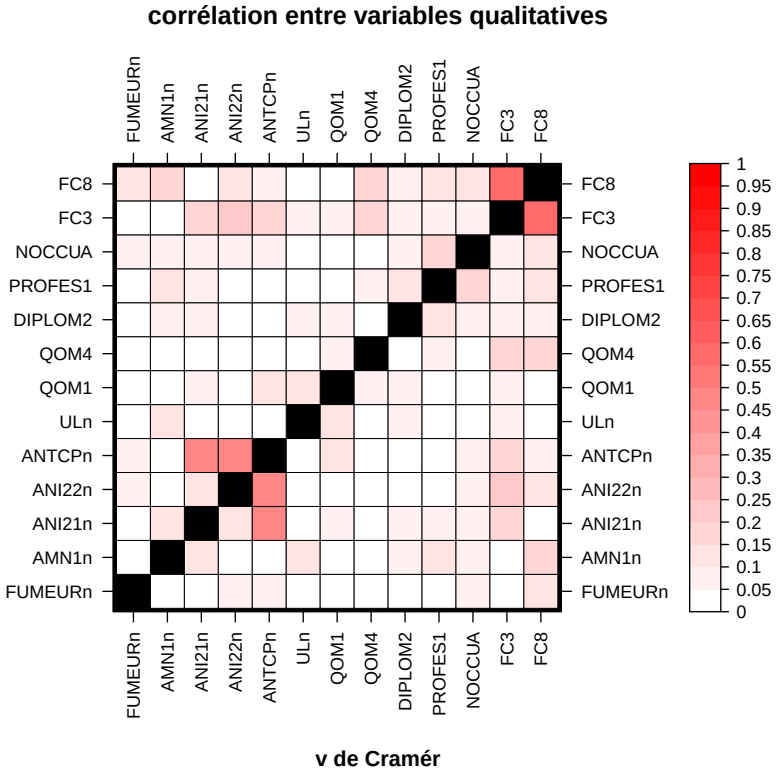

Figure 6 : Corrélation entre variables qualitatives

13.3 Corrélation entre variables qualitatives et quantitatives

La matrice de la figure 7, montre si l'une des modalités des variables qualitatives a une influence (rouge) ou non (blanc) sur les variables quantitatives. Autrement dit, elle montre si la moyenne des variables quantitatives pour un groupe d'observations présentant l'une des modalités des variables qualitatives, est différente pour les observations présentant les autres modalités. Cette matrice est calculée à partir du test de Kruslal-Wallis (au lieu de l'ANOVA) au risque de 0,5%, sachant la non normalité des variables quantitatives et le problème d'outliers. On note ici une bien meilleure interaction entre les variables, hormis LVLb (nombre de lessives), ICOS4 (produits de soin visage et corps), REV3 (revenus), NPVe1 (nombre de pièces de vie), HSRF (surface de logement) qui ne sont influencées que par très peu de variables qualitatives (< 4) ; NOCCUA (occupation) qui n'influence que très peu de variables quantitatives (< 4).

13.4 Corrélation entre CSP et les autres variables

La variable PROFES1 représente les Classes Socio-Professionnelles de l'étude de l'OQAI. Cette section montre ses corrélations avec les autres variables que l'on souhaite utiliser pour la prédire. Les figures 8 et 9 ne sont que des extraits des matrices précédemment présentées. La figure 9 montre que

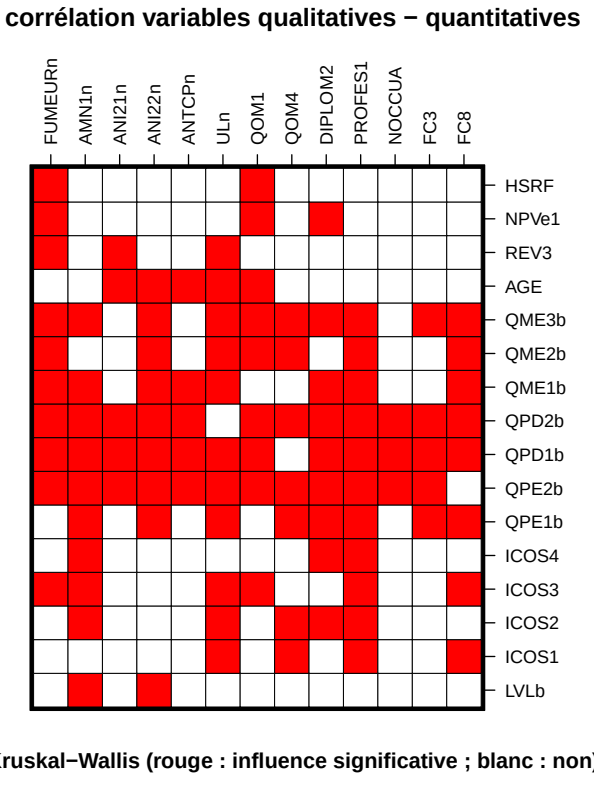

Figure 7 : Corrélation entre variables quantitatives et qualitatives

la variable des CSP influence un bon nombre de variables quantitatives sauf REV3 (ce qui est étonnant), HSRF, NPVe1, AGE et LVLb. La figure 8 montre une très faible corrélation entre PROFES1 et les variables qualitatives.

13.5 Conclusion de l'étude bidimensionnelle

Cette étude bidimensionnelle montre que les variables sont faiblement corrélées, ce qui laisse présager une certaine difficulté pour l'émergence de profils socio-professionnels, les observations ne se regroupant pas spécialement bien ensemble au sein de leur espace. Concernant l'entraînement d'un prédicteur pour la variables CSP, l'étude montre que les variables quantitatives n'expliquent que très faiblement la variable PROFES1. Cependant, le prédicteur peut s'appuyer sur un bon nombre de variables qualitatives mais leur contribution n'est pas quantifiée.

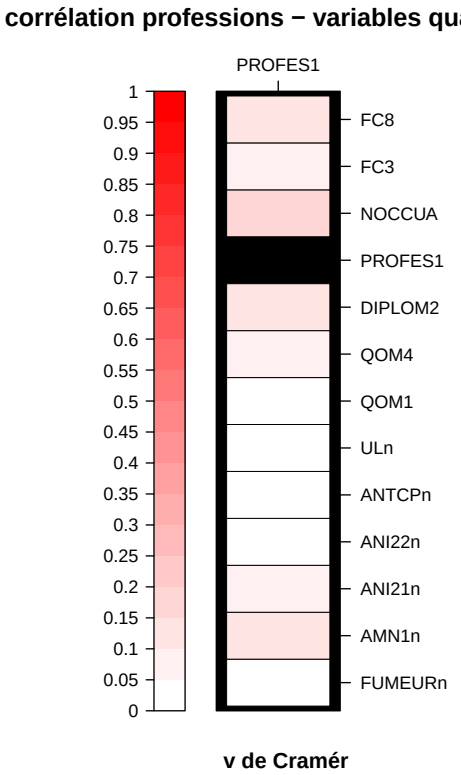

Figure 8 : Corrélation entre la variable PROFES1 et les variables qualitatives

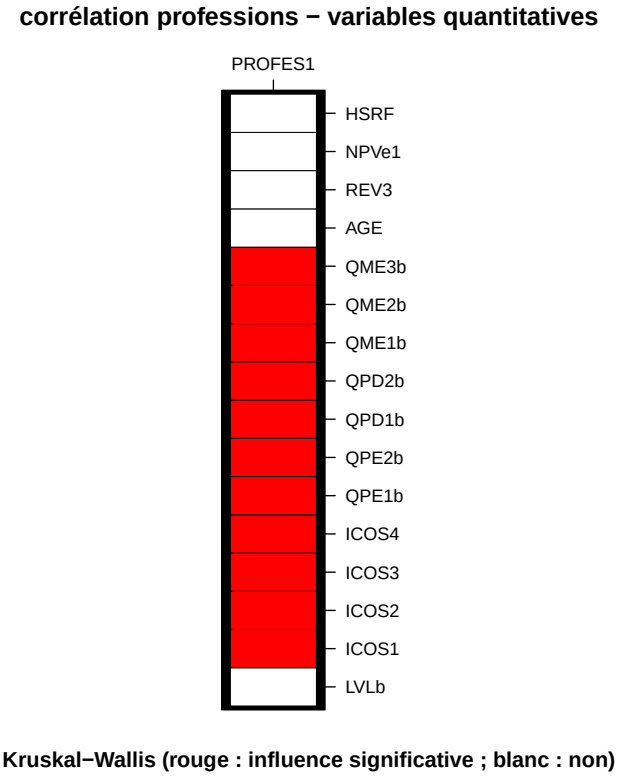

Figure 9 : Corrélation entre la variable PROFES1 et les variables quantitatives

14 Profils socio-professionnels

Le but de cette partie de l'étude est de révéler des profils socio-professionnels. Cependant, comme un certain nombre de variables quantitatives ont des outliers, les cartes de Kohonen (SOM) me semblent appropriées car elles leurs sont relativement tolérantes. SOM est une méthode (voir section 9.4) qui ne prend en entrée que des données quantitatives. Puisque les observations de l'étude sont décrites par un mélange de variables quantitatives et qualitatives, l'application d'une méthode de codage optimale est obligatoire (voir section 2). Cette section décrit l'utilisation de l'AFDM comme méthode de vectorisation des observations puis l'application de la méthode SOM et la description des profils trouvés.

14.1 Vectorisation

Selon Ricco Rakotomalala [2], le package FactoMineR contient une implémentation R de la méthode AFDM. Pour rappel, AFDM est une méthode factorielle qui généralise les méthodes de l'ACP et de l'ACM, prises à poids égaux.

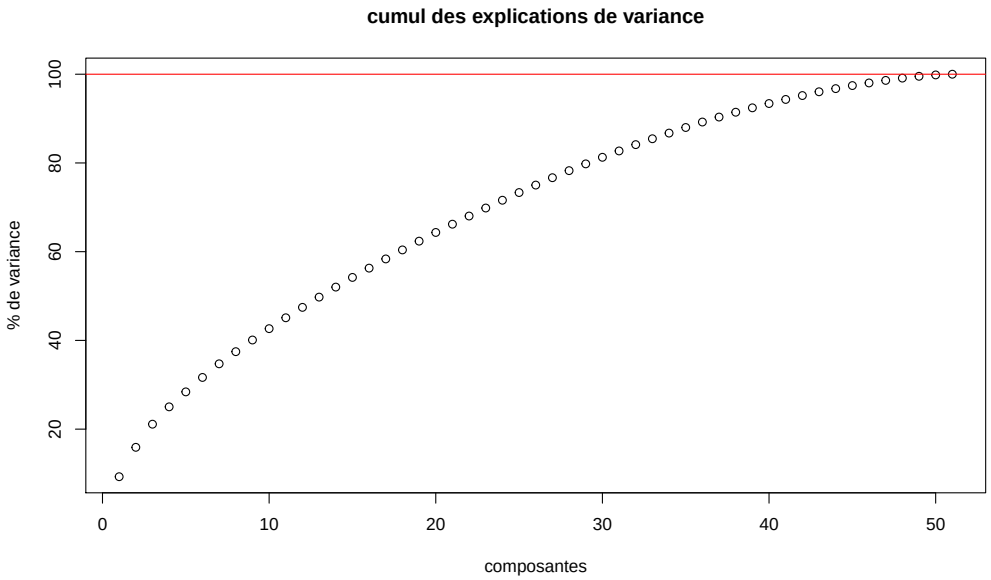

Figure 10 : Variance cumulée par nombre de composantes AFDM

Comme illustré à la figure 10, individuellement les composantes de l'AFDM représentent très mal les observations (9,3% et 6,6% d'explication de la variance pour les deux premières composantes). Les remarques soulevées lors de l'étude bidimensionnelle concernant la faiblesse de la corrélation entre les variables est encore une fois montrée. Le cercle de corrélation et la projection des observations sur le plan factoriel des deux premières composantes sont en annexe E. La projection des observations sur le plan factoriel composé par les deux premiers axes (fig 22) indique une difficulté certaine pour un partitionnement. Le résultat de l'AFDM est la vectorisation des observations en une matrice de 528 lignes et 51 colonnes (composantes).

2. https ://eric.univ-lyon2.fr/ ricco/tanagra/fichiers/fr_Tanagra_Clustering_Mixed_Data.pdf

14.2 Partitionnement

Le partitionnement des observations vectorisées par AFDM est effectuée par la méthode SOM dont le package kohonen en donne une implémentation R.

Le choix de la taille et la forme de la grille de cartes est conduit par des heuristiques. Selon Ricco Rakotomalala [3], la densité en observations distribuées sur chacun des neurones, doit être le plus homogène possible. D'autre part, les distances entre les neurones et les observations, données par la carte de proximité, indiquent l'éventuelle présence de frontières entre groupes de données. En effet, une surface où la proximité est forte indique un groupe d'observations similaires et une surface où la proximité est faible, indique des observations moins similaires et une frontière si cette surface sépare des surfaces à forte proximité.

Après plusieurs essais, la carte 5x5 hexagonale (voisinage circulaire) donne l'un des meilleurs résultats vis à vis des recommandations citées précédemment. La courbe d'apprentissage, donnant le nombre d'itérations nécessaires à l'apprentissage du réseau de neurones, est donnée en annexe F. La figure 11 qui donne la densité de répartition des observations, montre une relative faible homogénéité. Par exemple, 2 neurones sont presque vides d'observations. Ce résultat est à contraster avec celui de la carte 7x7 hexagonale qui montre plus de neurones vides (grille trop grande). En terme d'effectifs, on discerne 3 gros groupes (neurones) et quelques petits groupes.

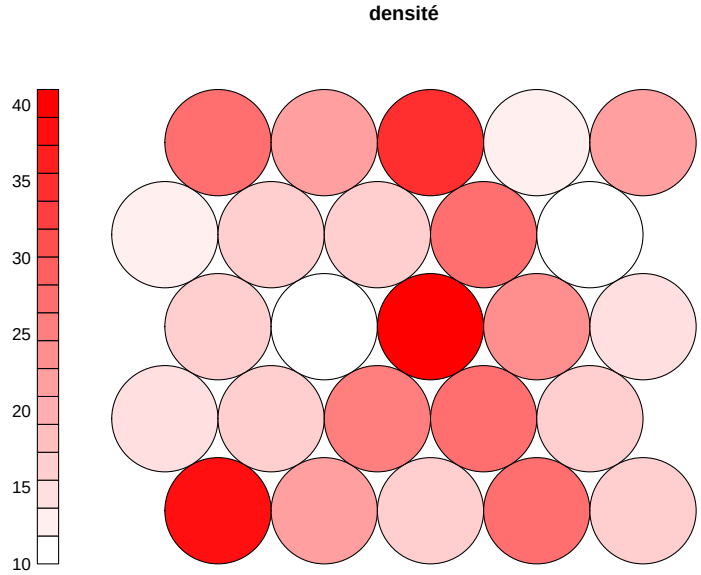

Figure 11 : Carte de densité

La figure 12 qui donne la proximité (rouge : proche, blanc : éloigné), suggère une séparation entre deux des trois gros groupes mais là, globalement, la séparation des groupes n'est franchement pas marquée.

Une CAH, donnée à la figure 13, est appliquée pour réduire le nombre de groupes. Le critère de similarité choisi est le critère de Ward, pondéré par le nombre d'observations contenues dans les neurones. Ce critère, adapté à la nature numérique des observations vectorisées de l'étude, favorise des groupes sphériques compacts. Le résultat de la CAH doit également satisfaire la contrainte de continuité des cartes sur la grille : les frontières des groupes formés par la CAH ne peuvent pas être discontinues. Sachant qu'il y a 7 classes dans la variable représentant les CSP (PROFES1) et considérant la forme du dendrogramme (coupure au niveau d'un saut important), j'ai fixé le nombre de groupes à 8.

3. https ://eric.univ-lyon2.fr/ ricco/tanagra/fichiers/fr_Tanagra_Kohonen_SOM_R.pdf

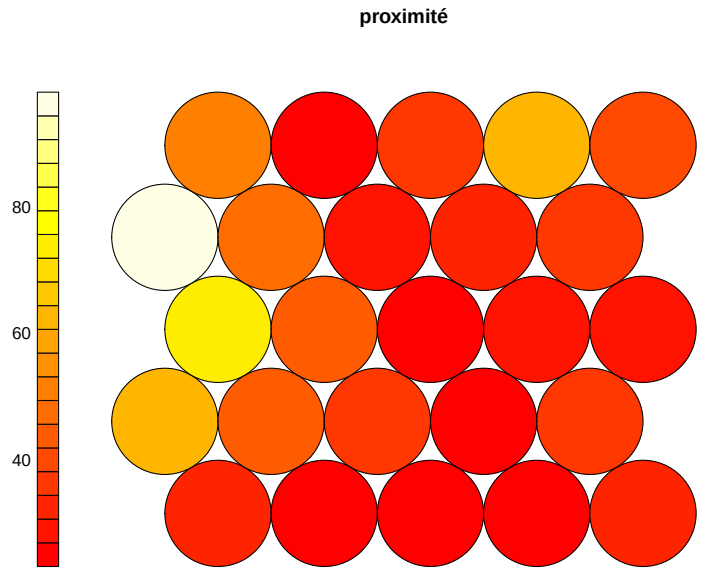

Figure 12 : Carte de proximité

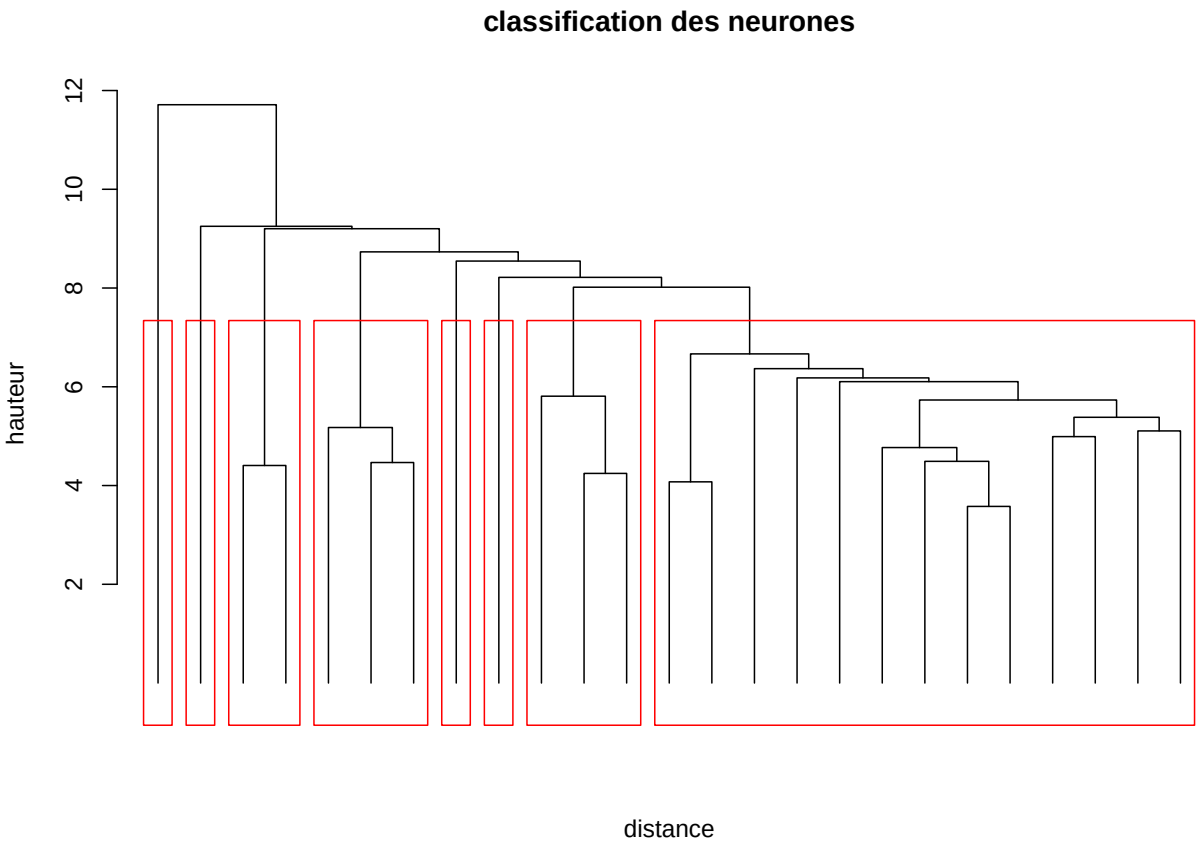

Figure 13 : Regroupement des neurones par CAH

La projection des frontières (traits noirs) des groupes trouvés par la CAH sur la grille de neurones est illustrée par la figure 14 où chacune des couleurs représente un groupe et les petits ronds dans les neurones donnent une idée du nombre d'observations.

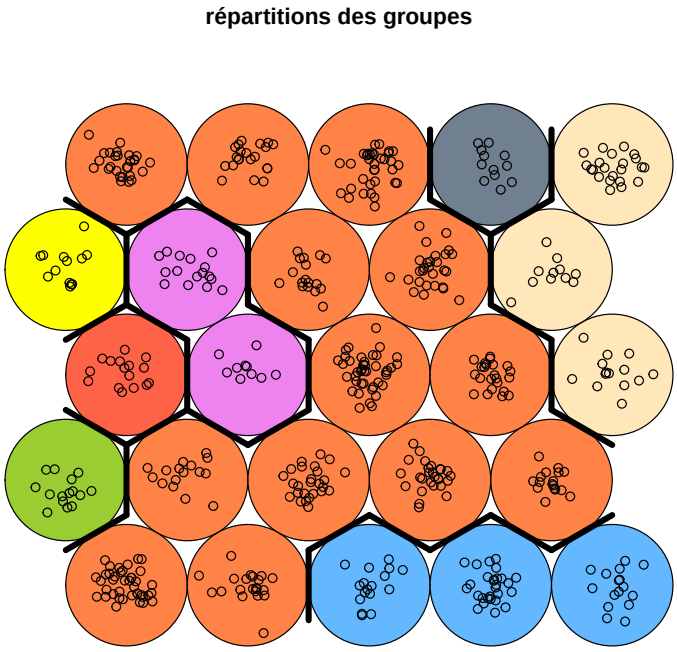

Figure 14 : Répartition des groupes CAH sur la grille de neurones

SOM et CAH aboutissent à la formation de 8 groupes ou profils socio-professionnels : 1 gros groupe rassemblant 338 observations, 2 petits groupes d'une cinquantaine d'observations et 5 très petits groupes d'une quinzaine d'observations.

14.3 Caractérisation des profils

Ces profils dont la répartition en observation est très déséquilibrée, montrent néanmoins des profils socio-professionnels plus ou moins bien marqués. Une ACP sur les moyennes des variables quantitatives par groupes est donnée en annexe G. Cependant, elle ne m'a pas permis de bien caractériser les profils. Cependant, l'étude des valeurs numériques de ces variables s'est avérée plus facile à interpréter.

Les tableaux 1 et 2 donnent la moyenne des variables quantitatives par profil. Les moyennes des groupes qui ne sont pas significativement différentes de celles de l'échantillon global (valeur test [4]), sont supprimées du tableau.

Les tableaux 3 et 4 affichent le rapport entre la moyenne des variables par groupe et la moyenne des variables de l'échantillon global. Ce rapport est exprimé en pourcentage lorsque le rapport est inférieur à 1 sinon il reste inchangé (symbole x).

Les ACM sur les données des variables qualitatives par groupes ne donnent pas de résultats satisfaisants : les composantes des ACM expliquent très faiblement la variance (au mieux à 25%, la plupart à 15% pour les deux premières composantes). Je me suis donc basé sur les bar plots des variables qualitatives par groupes, reproduits en annexe H. La description des variables quantitatives PROFES1,

4. http ://eric.univ-lyon2.fr/ ricco/tanagra/fichiers/fr_Tanagra_Comprendre_La_Valeur_Test.pdf

	effectifs	LVLb	ICOS1	ICOS2	ICOS3	ICOS4	QPE1b	QPE2b	QPD1b
prof1	338	3.3	9.2	9.4	3.1	7.7	1.9	-	-
prof2	61	-	-	-	-	-	-	-	-
prof3	15	-	-	-	-	-	-	2.2	-
prof4	16	-	-	-	-	-	3.5	-	-
prof5	26	-	-	-	-	-	-	-	-
prof6	47	6.1	17.2	17.9	8.1	14.4	3.4	-	-
prof7	13	1.4	-	-	-	4.7	1	-	-
prof8	12	5	14.9	-	-	-	-	-	-
all	528	3.6	10	10.1	3.6	8.4	2	1	1.9

Table 1 : Moyennes des variables quantitatives par profils 1/2

	QPD2b	QME1b	QME2b	QME3b	AGE	REV3	NPVe1	HSRF
prof1	-	-	2.8	0.9	50.7	2409.9	-	104
prof2	-	-	2.4	-	53.1	4092	6	145.3
prof3	-	-	-	-	61	1556.2	-	156
prof4	7.5	-	-	-	-	1379.2	-	-
prof5	-	-	-	-	38.6	1804.5	-	80.6
prof6	-	4.1	4.5	2.9	42.2	-	-	-
prof7	-	1.1	1.5	-	24.2	1451	2.1	34.8
prof8	-	-	-	-	39	-	-	-
all	4.3	2.4	3	1.1	49	2510.2	5.1	109.9

Table 2 : Moyennes des variables quantitatives par profils 2/2

	effectifs	LVLb	ICOS1	ICOS2	ICOS3	ICOS4	QPE1b	QPE2b	QPD1b
prof1	338	92%	92%	93%	86%	91%	93%	-	-
prof2	61	-	-	-	-	-	-	-	-
prof3	15	-	-	-	-	-	-	2.14x	-
prof4	16	-	-	-	-	-	1.75x	-	-
prof5	26	-	-	-	-	-	-	-	-
prof6	47	1.69x	1.72x	1.76x	2.26x	1.71x	1.69x	-	-
prof7	13	39%	-	-	-	56%	48%	-	-
prof8	12	1.37x	1.49x	-	-	-	-	-	-

Table 3 : Rapport des variables quantitatives par profils 1/2

	QPD2b	QME1b	QME2b	QME3b	AGE	REV3	NPVe1	HSRF
prof1	-	-	94%	83%	1.04x	96%	-	95%
prof2	-	-	80%	-	1.08x	1.63x	1.17x	1.32x
prof3	-	-	-	-	1.25x	62%	-	1.42x
prof4	1.73x	-	-	-	-	55%	-	-
prof5	-	-	-	-	79%	72%	-	73%
prof6	-	1.72x	1.50x	2.55x	86%	-	-	-
prof7	-	47%	50%	-	49%	58%	41%	32%
prof8	-	-	-	-	80%	-	-	-

Table 4 : Rapport des variables quantitatives par profils 2/2

DIPLOM2 et NOCCUA sont en annexe B.3. Chaque figure donne le bar plot d'une variable pour chaque groupe. Les barres représentent la proportion des classes en pourcentage. Elles sont en rouge lorsque la proportion est significativement différente de celle de l'échantillon global (valeur test [5]). Le bar plot de l'échantillon global, représenté en bleu, est donné comme référence.

Pour les descriptions succinctes des variables quantitatives, voir annexe B.1. Remarque : certaines des classes des variables qualitatives PROFES1 et NOCCUA sont nettement sous représentées, voir 12.2. Il faut également tenir compte de la grande disparité des revenus (REV3) avec une médiane à 2350€.

les quantités mentionnées dans la liste suivante se réfèrent aux moyennes des profils

Profil 1
Le profil rassemblant le plus d'observations (338) et par conséquent celui qui est le plus proche de l'échantillon total. Il s'agit de personnes de 50 ans (AGE) percevant les deuxièmes plus forts revenus (REV3) avec 2410€. Le profil est un mélange de professions (PROFES1 ; voir fig 36) sauf agriculteur et « autre profession », un mélange de diplômés (DIPLOM2 ; voir fig 35), avec uniquement des actifs (majoritaires) et des retraités (NOCCUA ; voir fig 37).

Profil 2
Deuxième profil par la taille en observations, il s'agit de tous les diplômés bac +5 et plus. Les personnes du profil ont 53 ans et perçoivent les plus forts revenus (4092€). Ce sont uniquement des cadres supérieurs, actifs (majoritaire) ou retraités. Ils possèdent le plus grand nombre de pièces de vie (NPVe1) avec la deuxième place pour la surface totale de logement (HSRF). Ce profil se distingue des autres par une plus grande proportion à faire appel à une aide ménagère (AMN1n ; voir fig 28) et donc à légèrement moins utiliser balais et serpillières la semaine (QME2b).

Profil 3
Profil rassemblant tous les agriculteurs, uniquement des retraités et des actifs à parts égales. C'est le profil le plus vieux (61 ans). En majorité le niveau est les études primaires ou un enseignement technique court. Les personnes de ce profil perçoivent le troisième plus bas revenus. Elles possèdent la plus grande surface totale de logement et sont les seules à utiliser deux fois plus de produits ménagers, autres que ceux utilisés pour nettoyer les surfaces (QPE2b).

Profil 4
Profil qui rassemble toutes les personnes dont l'occupation est « autre inactif ». Ces personnes sont majoritairement sans diplôme, dont la profession est mélangée, agriculteurs et cadres supérieurs exclus. Ces personnes ont les plus bas revenus. Ce profil se distingue des autres par l'utilisation de

5. http ://eric.univ-lyon2.fr/ ricco/tanagra/fichiers/fr_Tanagra_Comprendre_La_Valeur_Test.pdf

nettoyants de surfaces par semaine (QPE1b) significativement plus grande et par l'utilisation de parfums d'ambiance sous d'autres formes que les aérosols (QPD2b) nettement supérieure à la moyenne des observations totales.

Profil 5
Ce profil correspond à tous les chômeurs. Il est composé par des personnes de 38 ans, dont la profession est un mélange d'ouvrier (majoritairement), suivi de professions intermédiaires et de cadres supérieurs. Ce profil se distingue des autres car il rassemble des personnes qui ne font appel à aucune aide ménagère et ne possèdent presque pas de chien (ANI21n ; voir fig 29).

Profil 6
Troisième profil par la taille en observation, il ne rassemble que des actifs, majoritairement des ouvriers suivi de professions intermédiaires et d'employés. Leur niveau est majoritairement celui d'un enseignement technique court. Ce profil qui est dans la moyenne d'âge et de revenus, se distingue par des personnes qui ne font appel à aucune aide ménagère et dont les indicateurs de lessives, de produits ménagers et d'hygiène corporelle (LVlb, ICOS1 à ICOS4 et QPE1b) ainsi la fréquence du ménage du logement (QME1b à QME3b) sont très significativement supérieurs aux autres profils.

Profil 7
Il rassemble tous les étudiants. Ce profil est le plus jeune (24 ans), avec des personnes majoritairement bac +2 suivi de bac +3/4, occupant une activité très majoritairement « autre profession »(petits boulots ?). Elles ont le deuxième plus bas revenus, n'ont pas de chien et très peu de chats (ANI22n, voir fig 30) et vivent dans des logements qui ont la plus petite surface et le moins grand nombre de pièces. Ce profil se distingue également par des indicateurs de lessives, de produits ménagers, d'hygiène corporelle (respectivement LVlb, QPE1b et ICOS4) ainsi que de fréquence de ménage du logement (QME1b et QME2b), très significativement inférieurs aux autres profils.

Profil 8
Ce profil rassemble toutes les « personnes au foyer ». Ces sujets sont très majoritairement d'anciens employés avec à parts égales un niveau d'enseignement technique court ou sans diplôme. Ils ont le même âge (38 ans) que le profil des chômeurs mais ont des revenus dans la moyenne de l'échantillon contrairement à ces derniers. Ce profil se distingue par la possession de très peu de chats et des indicateurs de lessives et d'utilisation de déodorant significativement supérieurs à la moyenne de l'échantillon.

15 Modélisation CSP

Le but de cette partie de l'étude est de modéliser et d'expliquer les CSP, soit la variable PROFES1, par les autres variables. A cette fin, j'ai choisi la méthode CART (voir section 7). Cet algorithme permet de segmenter les observations sans recodage et tolère, dans une certaine mesure, les outliers (voir section 12). Selon Ricco Rakotomalala [6], le package Rpart donne une implémentation de CART en R. Cette section décrit l'application de CART sur les données de l'étude de l'OQAI.

Rpart utilise en interne la validation croisée pour calculer l'erreur associée aux arbres générés. Pour l'évaluation des performances de l'arbre optimal, le calcul de la matrice de confusion est effectué sur un jeu de données test n'ayant pas été utilisé pour l'entraînement des arbres. A ce propos, le package caret dispose d'outils d'échantillonnage afin de créer le jeu de test.

La figure 15 montre l'évolution de l'erreur calculée par validation croisée (X-val Relative Error) en fonction de la complexité (cp) des arbres trouvés lors de l'entraînement de CART.

6. https ://eric.univ-lyon2.fr/ ricco/tanagra/fichiers/fr_Tanagra_R_CART_algorithm.pdf

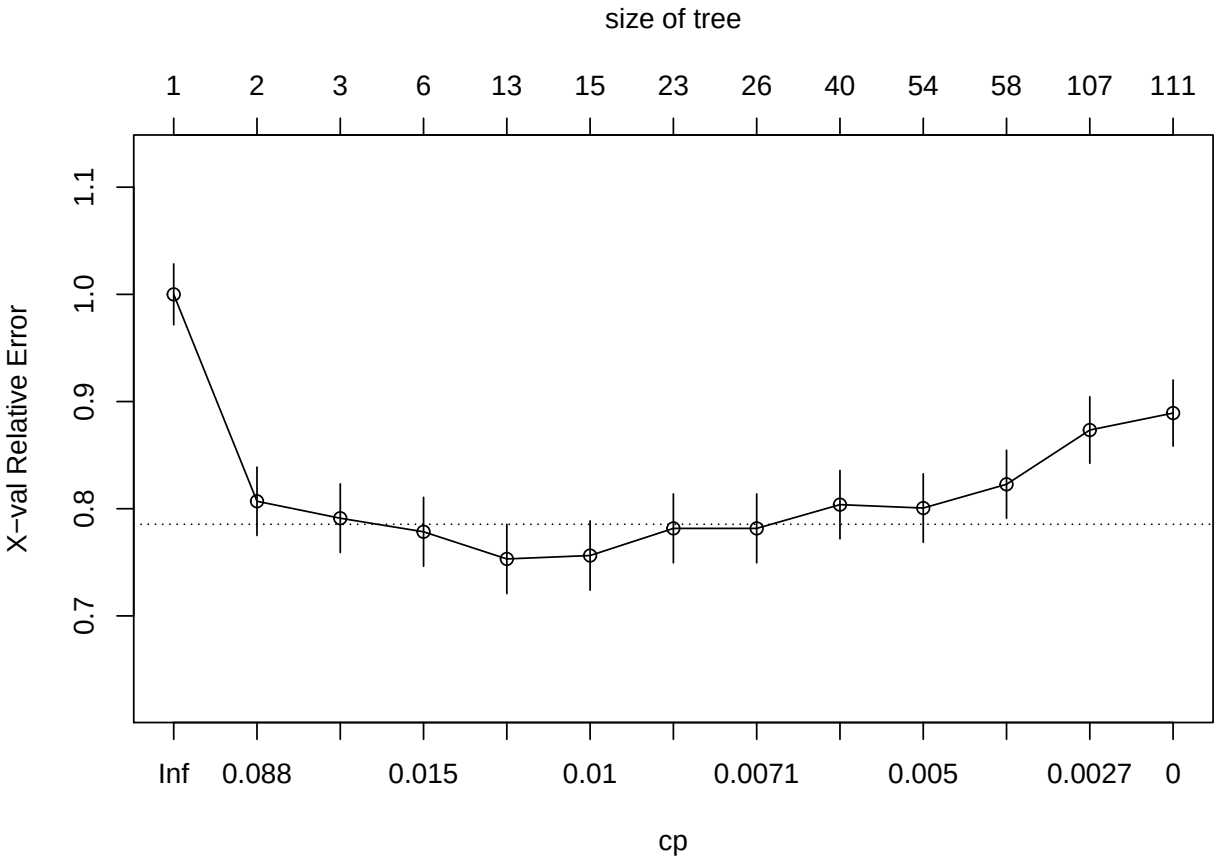

Figure 15 : Évolution de l'erreur par validation croisée selon la complexité des arbres

Lors de cette étude, j'ai choisi automatiquement l'arbre ayant la moins grande erreur de validation. Les arbres sous la ligne en pointillé à la figure 15, sont les arbres sélectionnés par la méthode de l'écart type.

La figure 16 montre l'arbre sélectionné (après élagage de l'arbre maximal). Les descriptions des variables PROFES1, DIPLOM2 et NOCCUA sont en annexe B.3. Pour une description exhaustive des classes des variables qualitatives, voir les fichiers pdf accompagnant ce rapport (listés en section 11.1).

Le calcul de la matrice de confusion 5 par la package caret donne une justesse (accuracy [7]) ou taux de vrai positifs et négatifs parmi toutes les données de 42%. Ce résultat très médiocre provient des très faibles performances en sensibilité du prédicteur (taux de récupération des vrais positifs). La précision (precision) et la spécificité (specificity) montre que le prédicteur a tendance à faire une segmentation très grossière. Ce résultat s'explique également par la sous représentation des classes 1, 2 et 7 de la variables PROFES1. Concernant les règles de segmentation, on retrouve quelques points en commun avec les profils découverts précédemment. Par exemple, les diplômés bac +5 et plus se retrouvent dans le même segment comme le profil 2 (oui pour DIPLOM2 = 9), la discrimination par les revenus (REV3 >= 2624) ou encore les étudiants qui sont tous dans le même segment (non pour NOCCUA = 1, 2, 4, 5, 6) comme le profil 7.

7. https ://en.wikipedia.org/wiki/Sensitivity_and_specificity

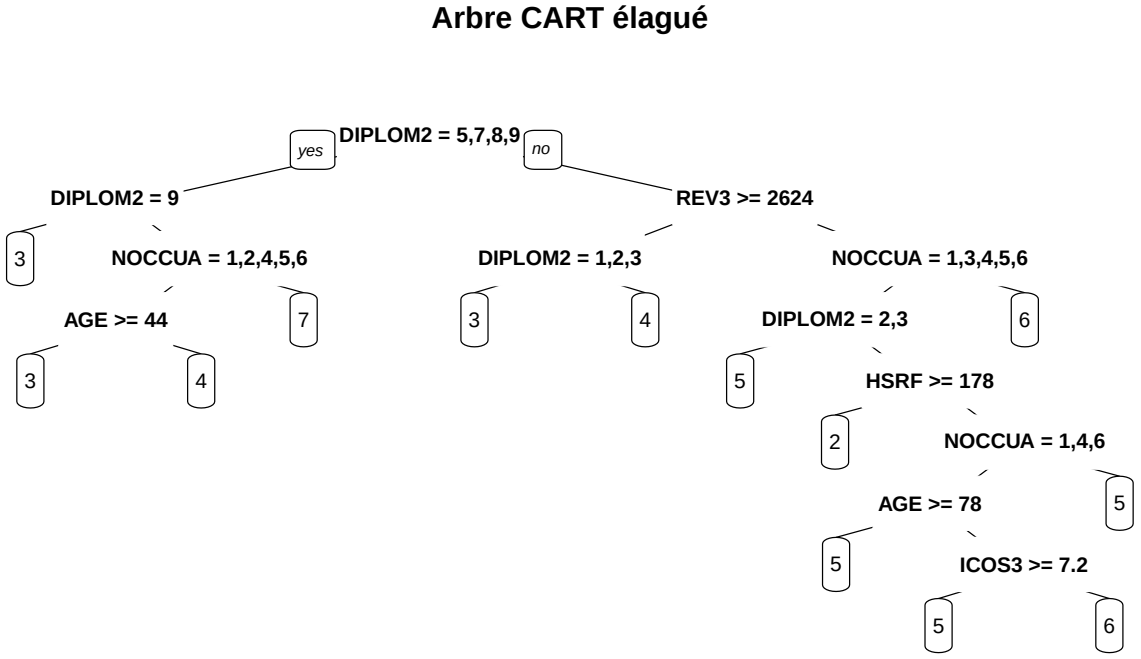

Figure 16 : Arbre optimal pour toutes les observations

	Class : 1	Class : 2	Class : 3	Class : 4	Class : 5	Class : 6	Class : 7
Sensitivity	0.00	0.17	0.63	0.17	0.45	0.59	0.00
Specificity	1.00	0.97	0.79	0.90	0.85	0.76	1.00
Pos Pred Value	*NaN*	0.25	0.52	0.33	0.41	0.39	*NaN*
Neg Pred Value	0.97	0.95	0.86	0.78	0.87	0.87	0.98
Precision	*NA*	0.25	0.52	0.33	0.41	0.39	*NA*
Recall	0.00	0.17	0.63	0.17	0.45	0.59	0.00
F1	*NA*	0.20	0.57	0.22	0.43	0.47	*NA*
Prevalence	0.03	0.06	0.26	0.23	0.19	0.21	0.02
Detection Rate	0.00	0.01	0.16	0.04	0.09	0.12	0.00
Detection Prevalence	0.00	0.04	0.32	0.12	0.21	0.32	0.00
Balanced Accuracy	0.50	0.57	0.71	0.53	0.65	0.67	0.50

Table 5 : Matrice de confusion pour CART sur toutes les observations

16 Conclusion

Ce travail, basé sur une sélection des données de l'étude de l'OQAI qui avait pour origine le but d'expliquer les sources de pollution de l'air des logements en France, a permis de révéler quelques profils socio-professionnels plus ou moins bien délimités et l'entraînement d'un prédicteur explicatif des CSP aux performances médiocres. La méthode SOM appliquée après une AFDM des données, mélange de variables quantitatives et qualitatives, a réussi notamment à isoler les profils des étudiants, des chômeurs et des agriculteurs, modalités pourtant sous représentées. Cependant, la plus grande fraction des observations restent agglomérées dans un profil fourre-tout qui ne se distingue pas significativement de la moyenne de l'échantillon global. Le prédicteur des CSP entraîné par la méthode CART, a une justesse très médiocre. Cependant, les règles de segmentation restent cohérentes avec les profils trouvés par SOM. Ces résultats sont à mettre en perceptive avec la faible corrélation des variables entre elles. Enfin, dans la perceptive d'une amélioration des contours des profils, les méthodes cluster wise sont une piste intéressante.

ANNEXES

A Équations

Erreur de modélisation

$$E(Y - \hat{Y})^2 = \underbrace{Var(\hat{F}(X)) + \left[Biais(\hat{F}(X))\right]^2}_{\text{réductible}} + \underbrace{Var(\varepsilon)}_{\text{irréductible}} \tag{3}$$

Avec Y la réponse réelle, $\hat{Y}$ la réponse calculée par le modèle $\hat{F}$, X le vecteur de données et ε une erreur systématique aléatoire de moyenne nulle et de distribution normale.

Inertie

$$Inertie = \underbrace{Inertie_{\text{intra}} + Inertie_{\text{inter}}}_{\text{constant}} \tag{4}$$

ACP

$$WDc = \lambda c \tag{5}$$

$$c = Xu \tag{6}$$

Avec W le matrice des produits scalaires des observations prises deux à deux, D matrice diagonale définissant le poids des variables explicatives, λ les valeurs propres des composantes principales c, u les vecteurs propres de X la matrice des données d'entrée.

B Variables

B.1 Quantitatives

LVLb Nombre de lessives par mois

ICOS1 Utilisation déodorant

ICOS2 Utilisation d'eau de toilette

ICOS3 Utilisation de produits de soin cheveux

ICOS4 Utilisation de produits de soin visage et corps

QPE1b Utilisation produits nettoyant de surfaces par semaine

QPE2b Utilisation d'autres produits ménagés par semaine

QPD1b Utilisation parfum ambiance sous forme d'aérosol par semaine

QPD2b Utilisation parfum ambiance sous d'autres formes par semaine

QME1b Utilisation aspirateur par semaine

QME2b Utilisation balai/serpillière par semaine

QME3b Autre nettoyage par semaine

AGE Âge en années

REV3 Revenus par tranches

NPVe1 Nombre de pièce de vie

HSRF Surface total du logement

B.2 Qualitatives

FUMEURn Fumeurs

AMN1n Aide ménagère

ANI21n Chiens

ANI22n Chats

ANTCPn Traitement parasites chiens et chats

ULn Traitement parasites autres

QOM1 Fréquence de sortie des ordures

QOM4 Type de stockage des ordures

DIPLOM2 Diplôme

PROFES1 Profession

NOCCUA Occupation

FC3 Type de logement

FC8 Statut des habitants

B.3 Classes

PROFES1

- 1 : agriculteur
- 2 : artisans
- 3 : cadre supérieur
- 4 : profession intermédiaire
- 5 : employé
- 6 : ouvrier
- 7 : autre

DIPLOM2

- 1 : sans diplôme
- 2 : fin du premier cycle enseignement général
- 3 : fin du second cycle enseignement général
- 4 : enseignement technique court
- 5 : enseignement technique long
- 6 : fin études primaire
- 7 : enseignement supérieur bac +2

- 8 : enseignement supérieur bac +3/4
- 9 : enseignement supérieur bac +5 et plus

NOCCUA

- 1 : exerce une profession
- 2 : chômeur
- 3 : étudiant
- 4 : retraite ou pré-retraite
- 5 : au foyer
- 6 : autre inactif

C Résumés des variables quantitatives

Table 6 : Résumés des variables quantitatives 1/4

LVLb	ICOS1	ICOS2	ICOS3
Min. :0.000	Min. : 0.000	Min. : 0.00	Min. : 0.0
1st Qu. :1.500	1st Qu. : 3.500	1st Qu. : 4.50	1st Qu. : 0.0
Median :3.500	Median : 8.000	Median :10.50	Median : 1.0
Mean :3.624	Mean : 9.986	Mean :10.14	Mean : 3.6
3rd Qu. :7.000	3rd Qu. :14.000	3rd Qu. :14.00	3rd Qu. : 7.0
Max. :7.000	Max. :35.000	Max. :35.00	Max. :28.0

Table 7 : Résumés des variables quantitatives 2/4

ICOS4	QPE1b	QPE2b	QPD1b
Min. : 0.000	Min. :0.00	Min. :0.000	Min. : 0.000
1st Qu. : 3.500	1st Qu. :1.00	1st Qu. :0.000	1st Qu. : 0.000
Median : 7.000	Median :1.00	Median :0.000	Median : 0.000
Mean : 8.438	Mean :2.02	Mean :1.013	Mean : 1.932
3rd Qu. :11.125	3rd Qu. :3.50	3rd Qu. :1.000	3rd Qu. : 3.500
Max. :35.000	Max. :7.00	Max. :7.000	Max. :14.000

Table 8 : Résumés des variables quantitatives 3/4

QPD2b	QME1b	QME2b	QME3b
Min. : 0.000	Min. :0.000	Min. :0.000	Min. :0.000
1st Qu. : 0.000	1st Qu. :1.000	1st Qu. :1.000	1st Qu. :0.000
Median : 1.000	Median :1.000	Median :3.500	Median :1.000
Mean : 4.324	Mean :2.363	Mean :2.994	Mean :1.119
3rd Qu. : 7.000	3rd Qu. :3.500	3rd Qu. :3.500	3rd Qu. :1.000
Max. :14.000	Max. :7.000	Max. :7.000	Max. :7.000

Table 9 : Résumés des variables quantitatives 4/4

AGE	REV3	NPVe1	HSRF
Min. :18.00	Min. : 535	Min. : 1.000	Min. : 7.0
1st Qu. :38.00	1st Qu. :1299	1st Qu. : 4.000	1st Qu. : 70.0
Median :48.00	Median :2349	Median : 5.000	Median : 97.0
Mean :48.99	Mean :2510	Mean : 5.085	Mean :109.9
3rd Qu. :60.00	3rd Qu. :2899	3rd Qu. : 6.000	3rd Qu. :130.0
Max. :89.00	Max. :7600	Max. :14.000	Max. :700.0

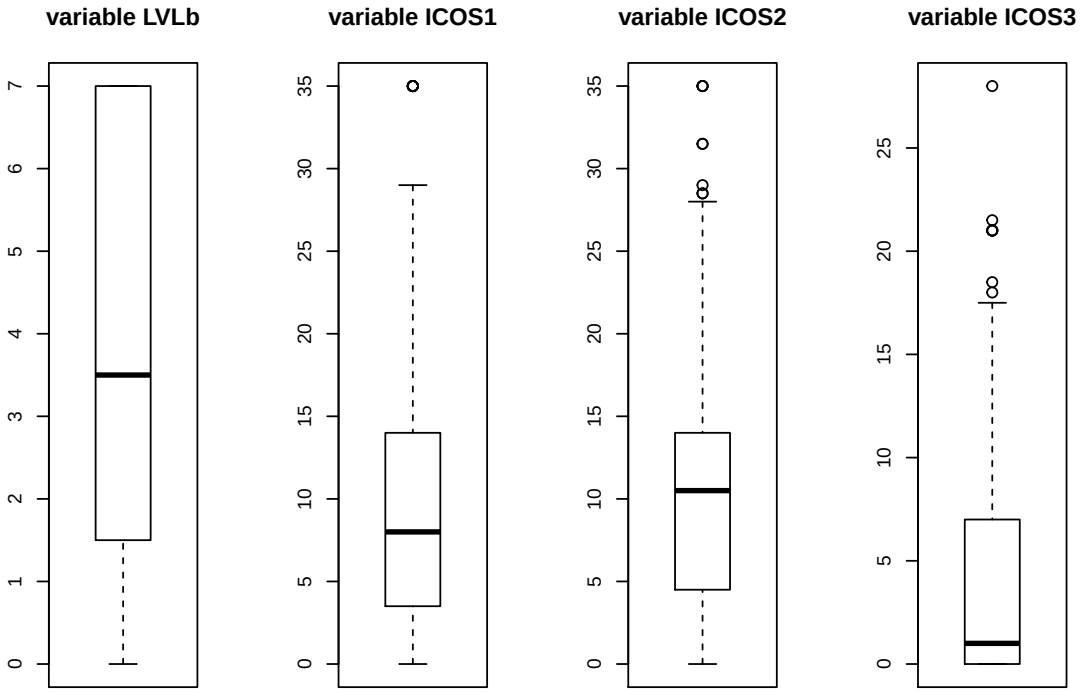

Figure 17 : Box plots LVLb, ICOS1, ICOS2, ICOS3

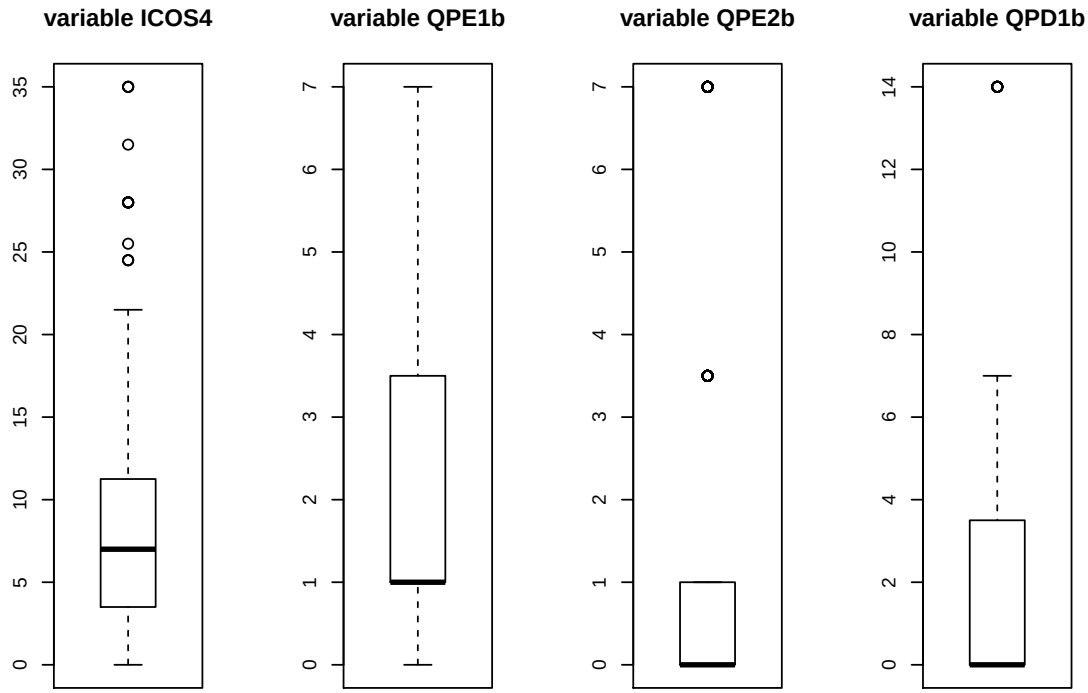

Figure 18 : Box plots ICOS4, QPE1b, QPE2b, QPD1b

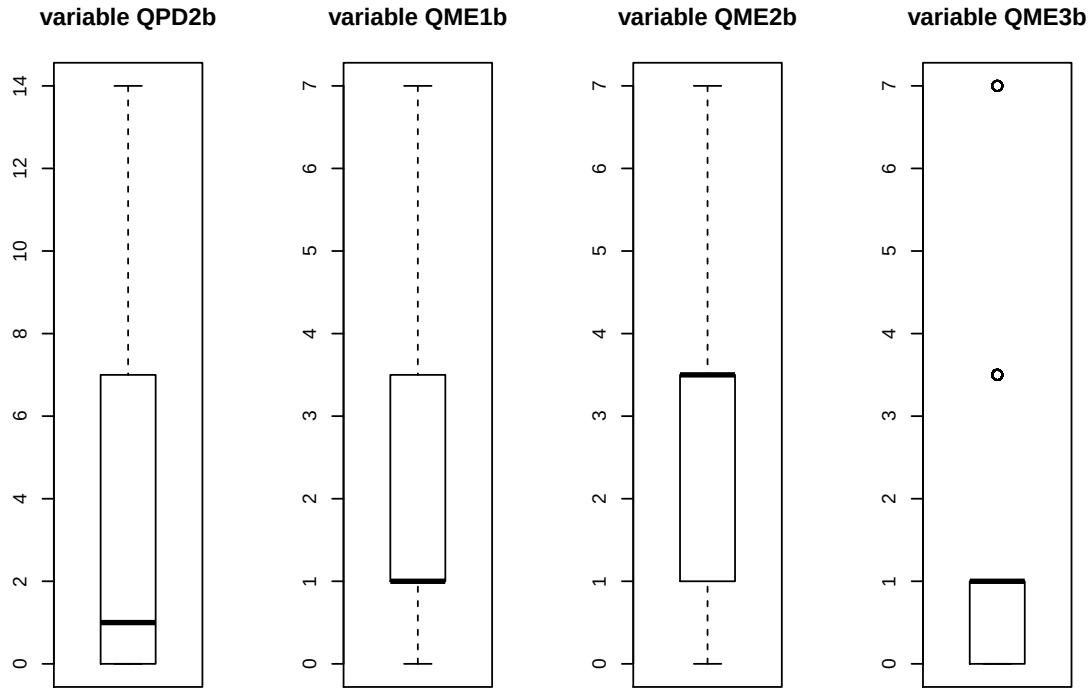

Figure 19 : Box plots QPD2b, QME1b, QME2b, QME3b

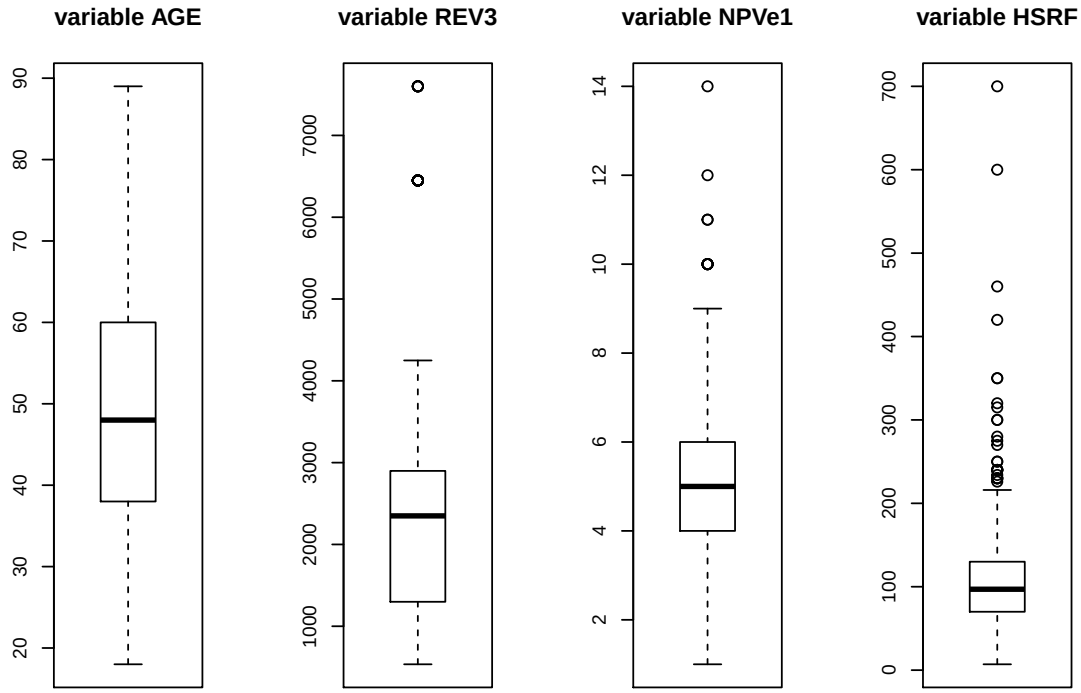

Figure 20 : Box plots AGE, REV3, NPVe1, HSRF

D Résumés des variables qualitatives

Table 10 : Résumés des variables qualitatives 1/2

FUMEURn	AMN1n	ANI21n	ANI22n	ANTCPn	ULn	QOM1
1 :300	1 :436	1 :373	1 :385	1 : 79	1 :437	1 :246
2 : 55	2 : 92	2 :155	2 :143	2 :166	2 : 91	2 :246
3 :115				3 :283		3 : 36
4 : 58						

Table 11 : Résumés des variables qualitatives 2/2

QOM4	DIPLOM2	PROFES1	NOCCUA	FC3	FC8
1 :155	1 : 57	1 : 16	1 :311	1 :207	1 :195
2 : 84	2 : 30	2 : 31	2 : 27	2 :321	2 :333
3 :251	3 : 42	3 :135	3 : 14		
4 : 38	4 :123	4 :120	4 :148		
	5 : 45	5 :103	5 : 12		
	6 : 53	6 :110	6 : 16		
	7 : 55	7 : 13			
	8 : 64				
	9 : 59				

E AFDM

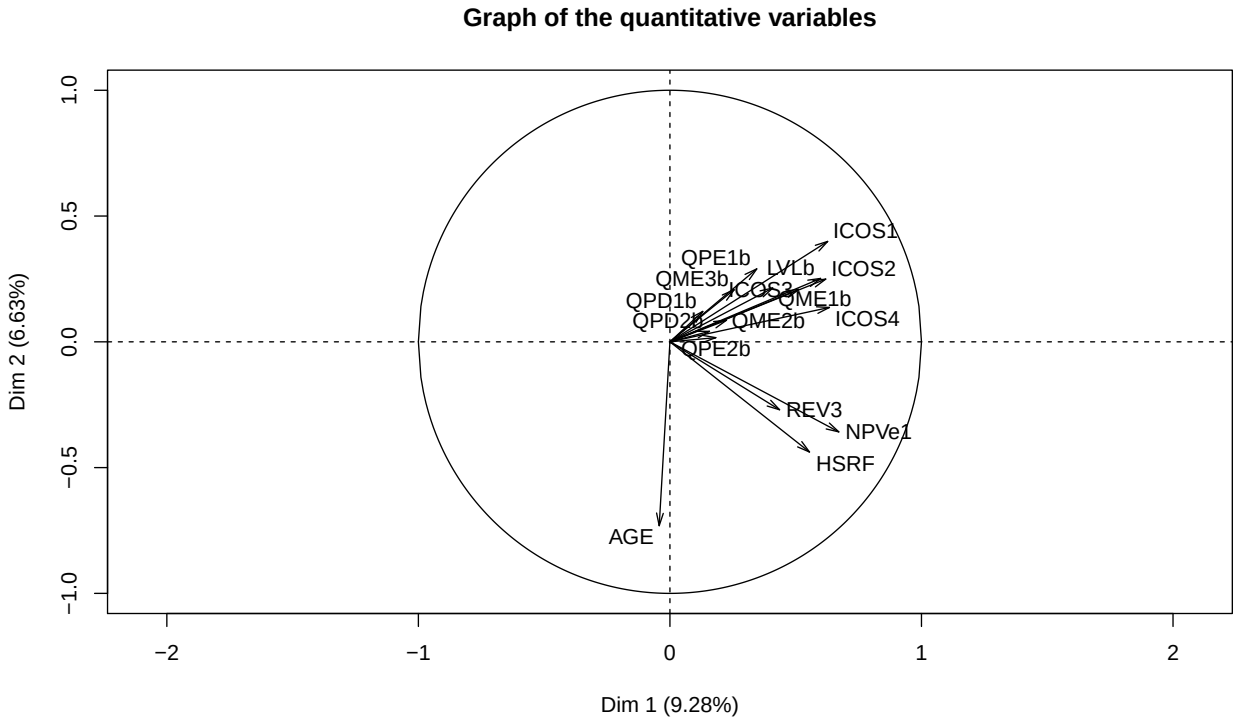

Figure 21 : Cercle de corrélation des variables quantitatives

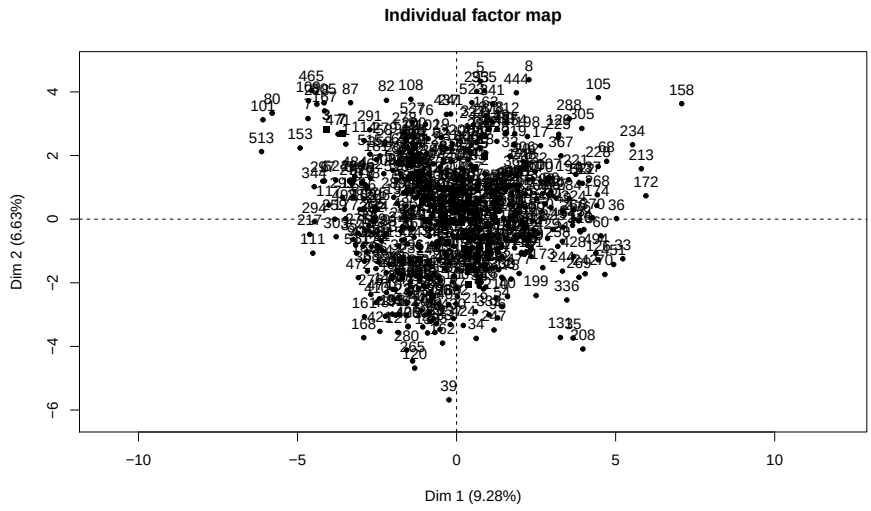

Figure 22 : Projection des observations sur la plan

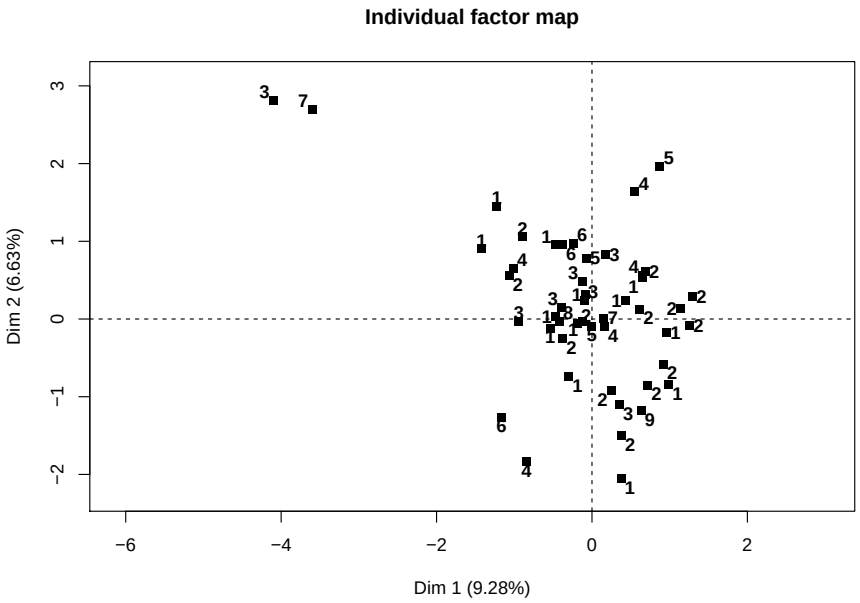

Figure 23 : Projection des modalités des variables qualitatives sur le plan

F Cartes de Kohonen

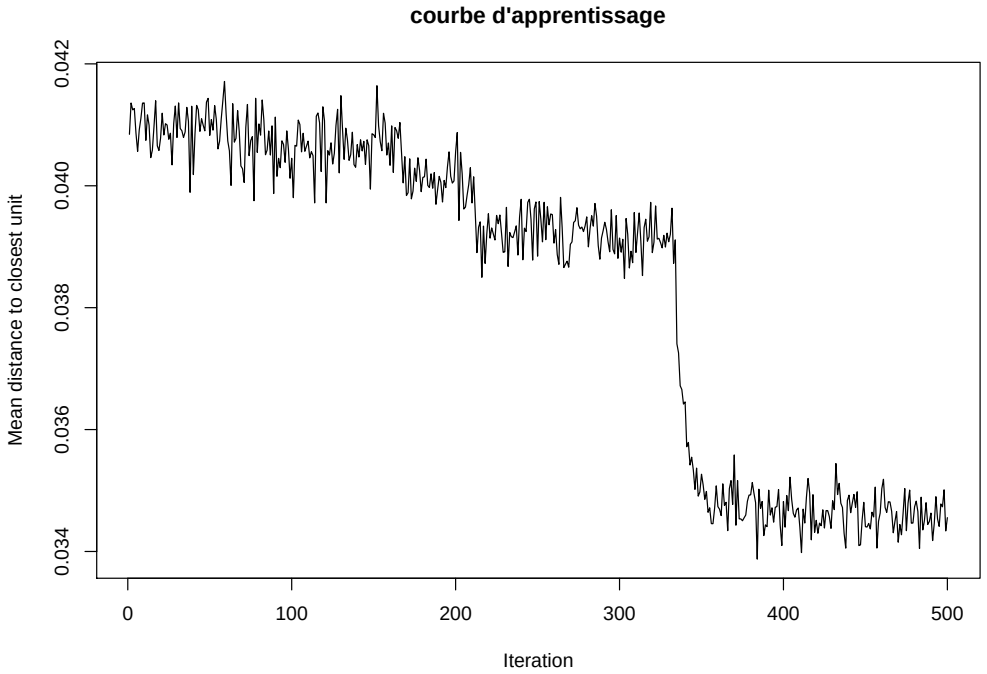

Figure 24 : Nombre d'itérations pour l'apprentissage SOM

G ACP sur les moyennes des profils

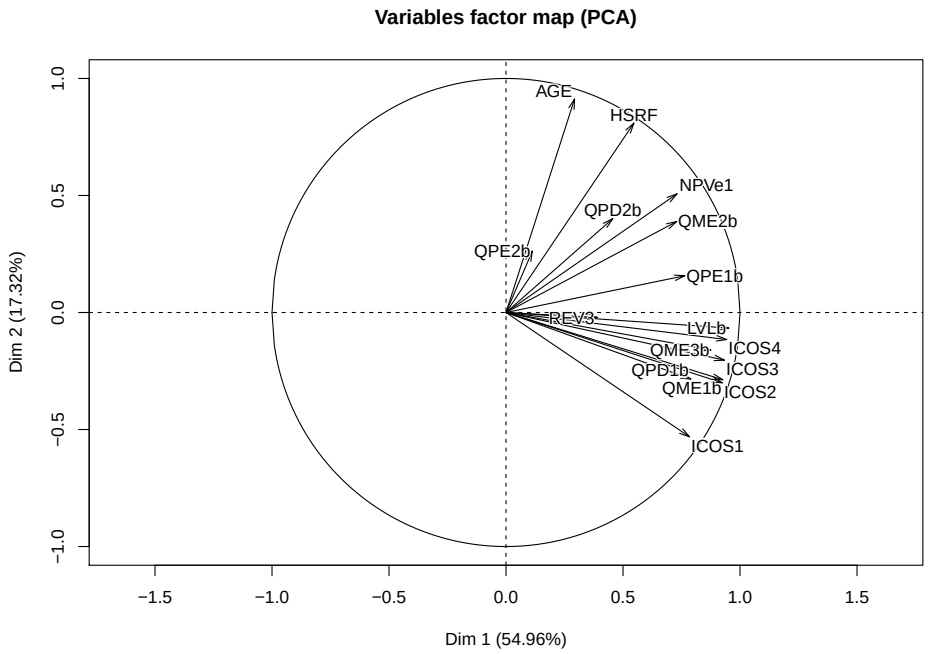

Figure 25 : Cercle de corrélation des variables quantitatives

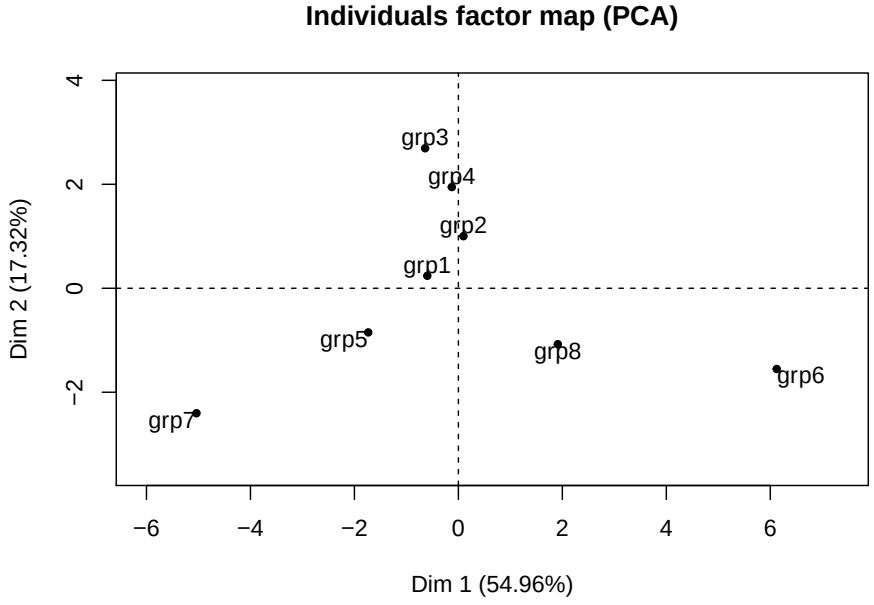

Figure 26 : Projection des Profils

H Variables qualitatives par profils

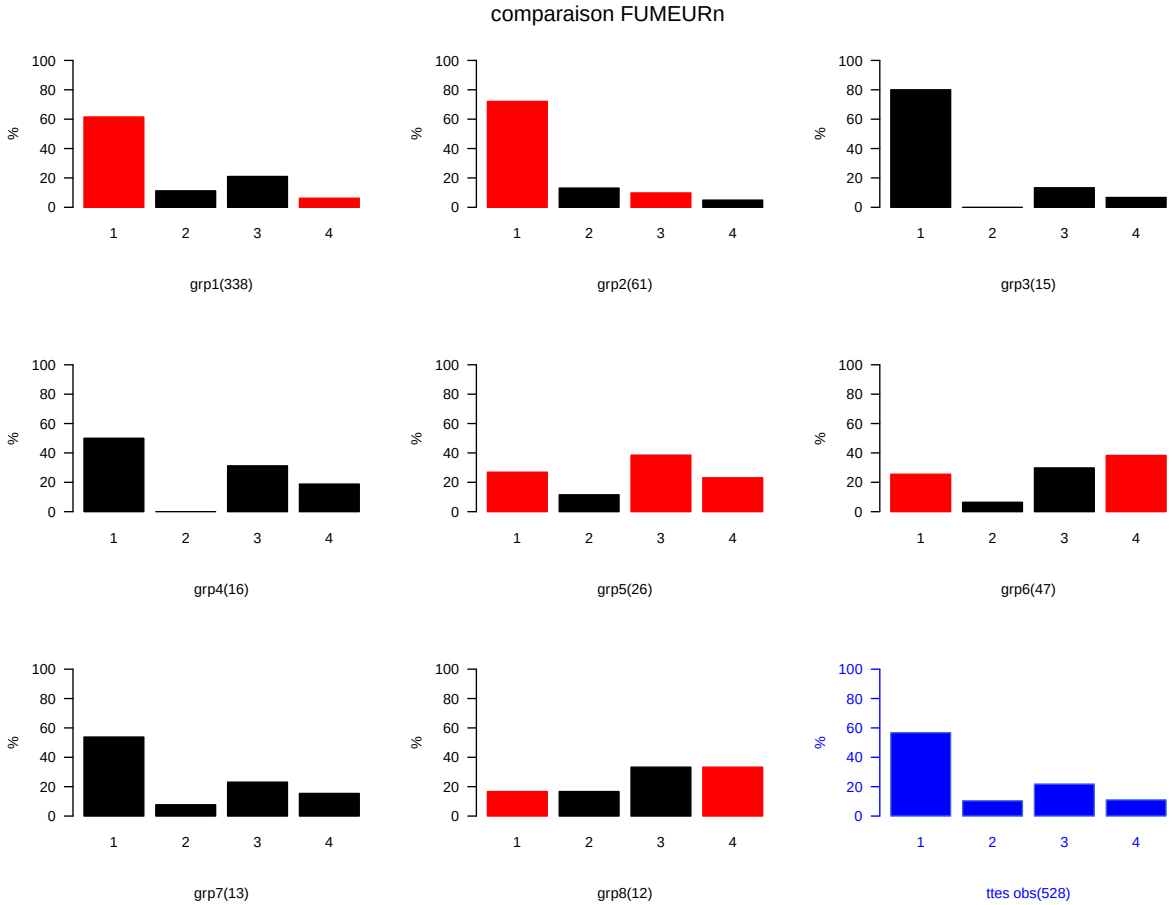

Figure 27 : Inter-comparaison FUMEURn

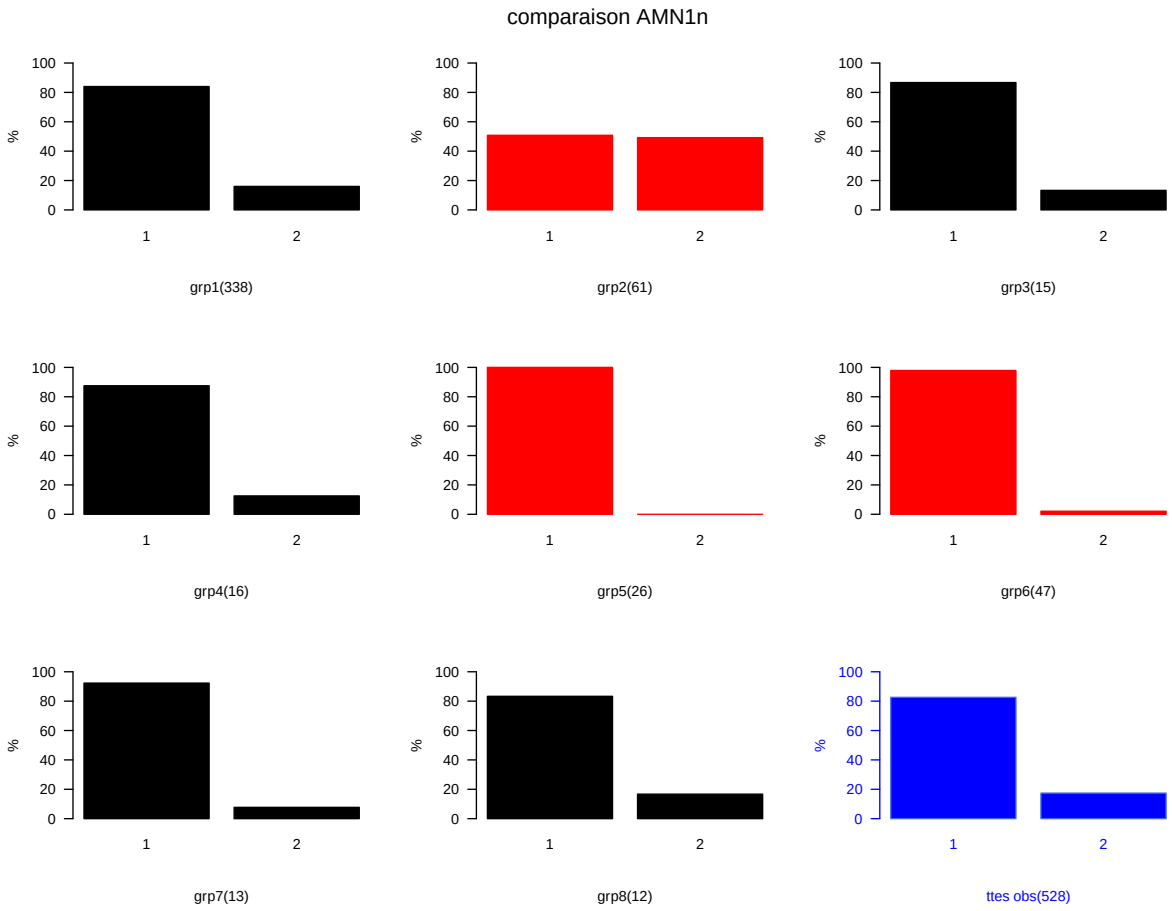

Figure 28 : Inter-comparaison AMN1n

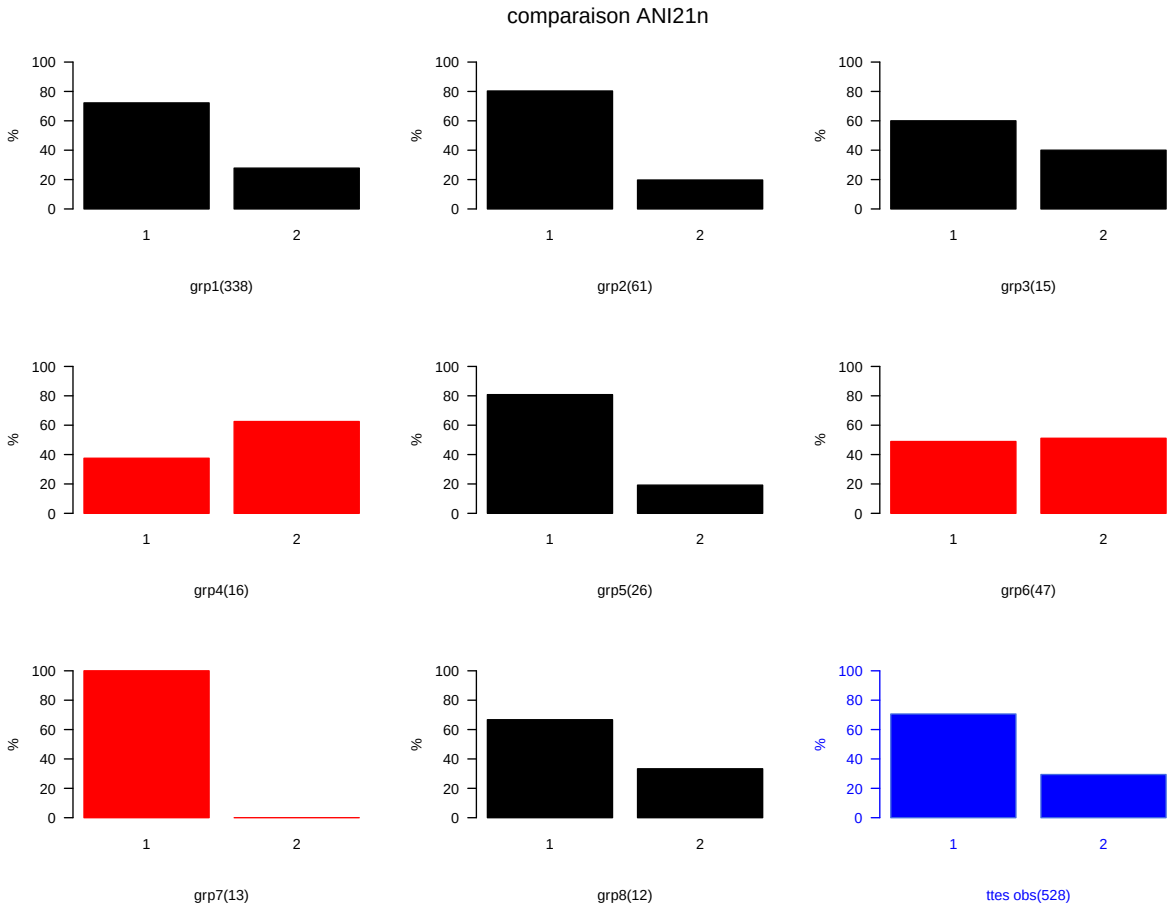

Figure 29 : Inter-comparaison ANI21n

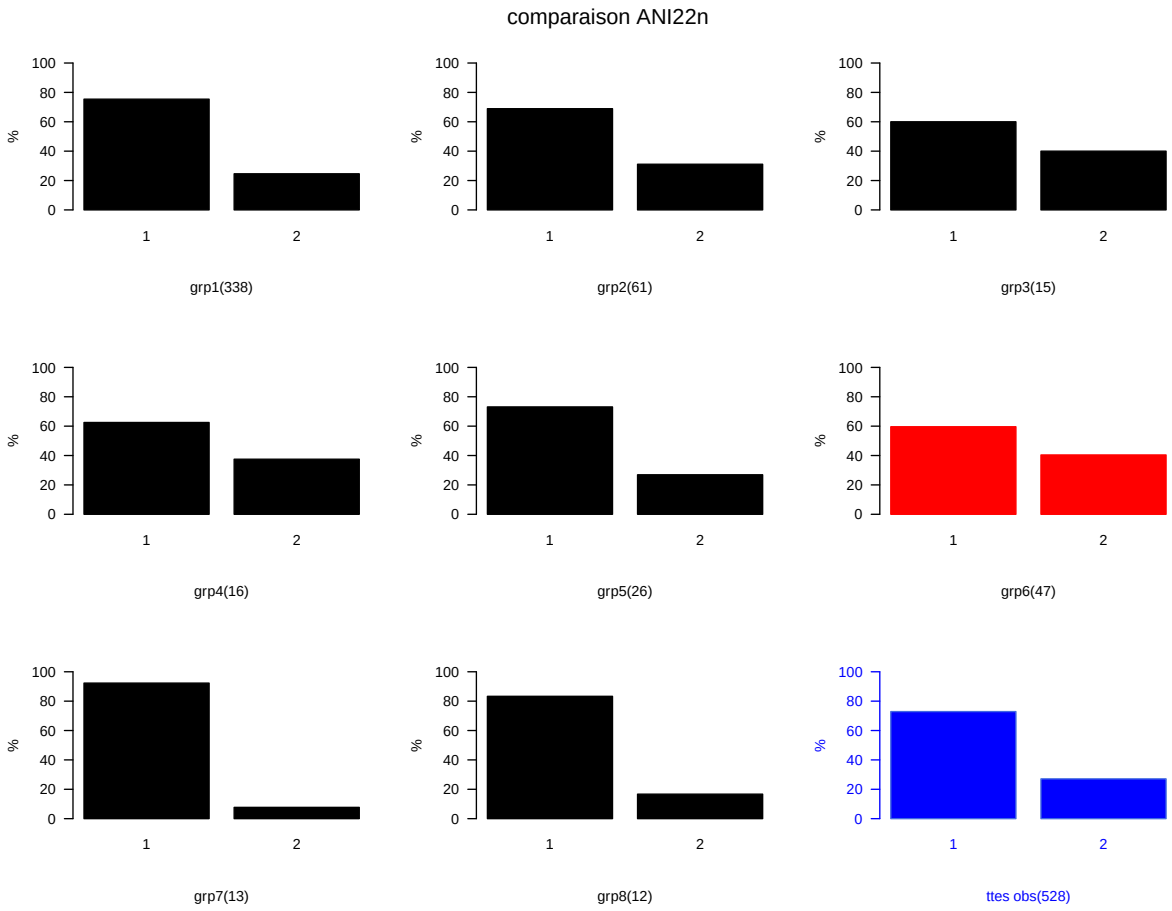

Figure 30 : Inter-comparaison ANI22n

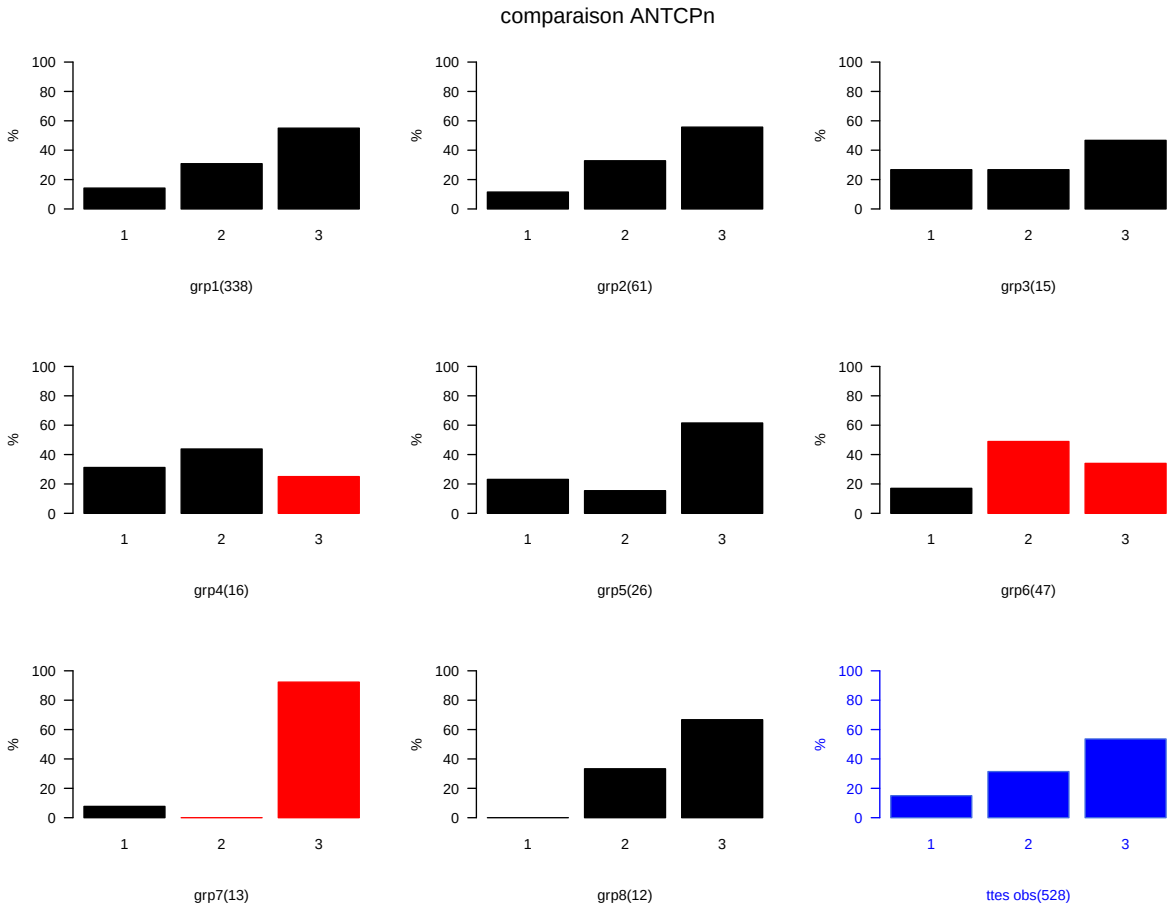

Figure 31 : Inter-comparaison ANTCPn

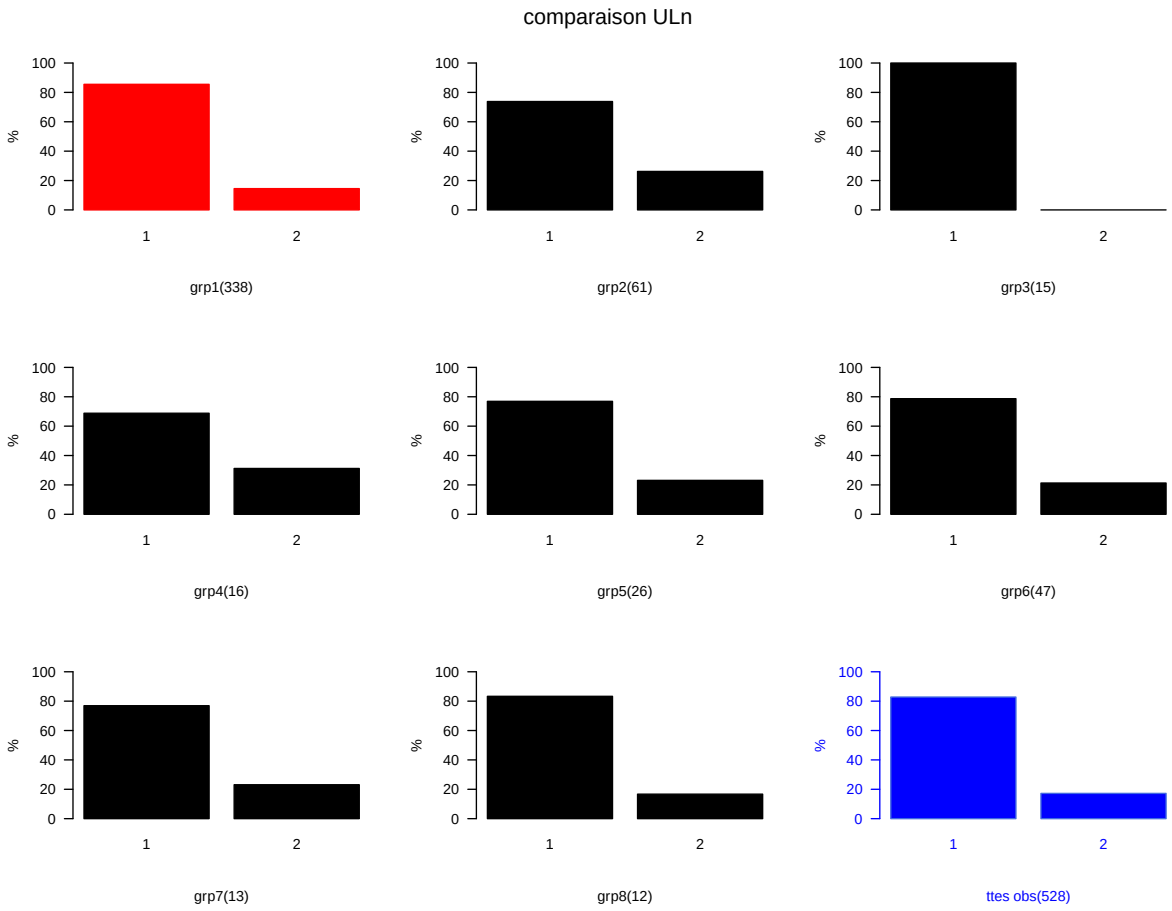

Figure 32 : Inter-comparaison ULn

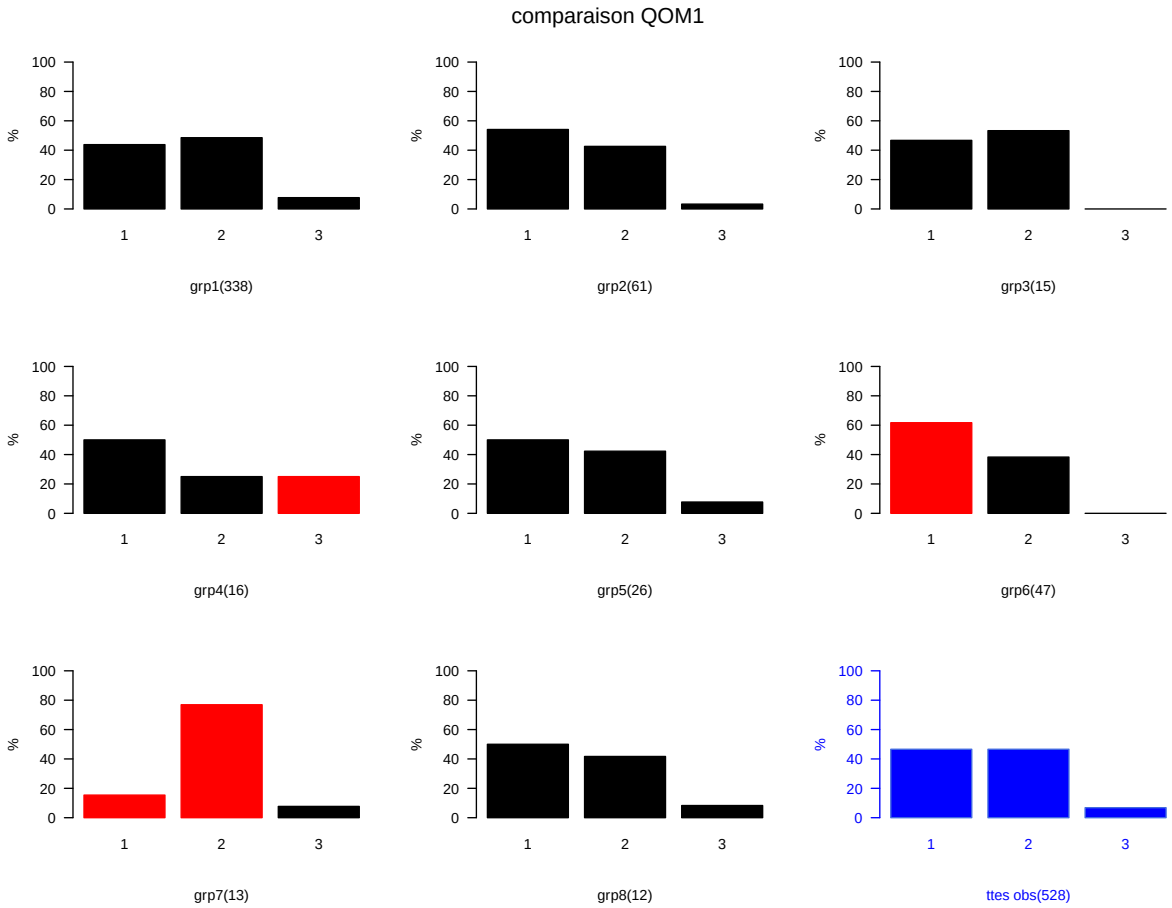

Figure 33 : Inter-comparaison QOM1

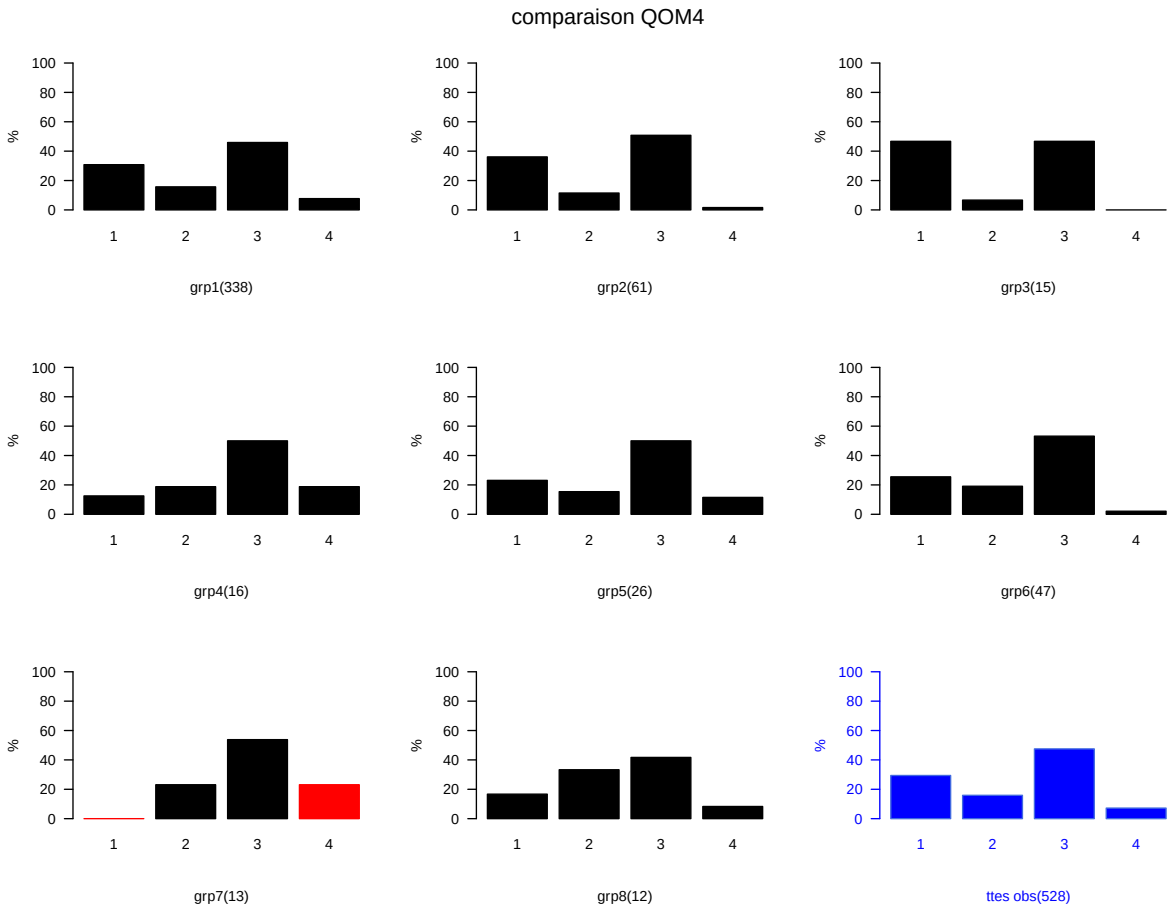

Figure 34 : Inter-comparaison QOM4

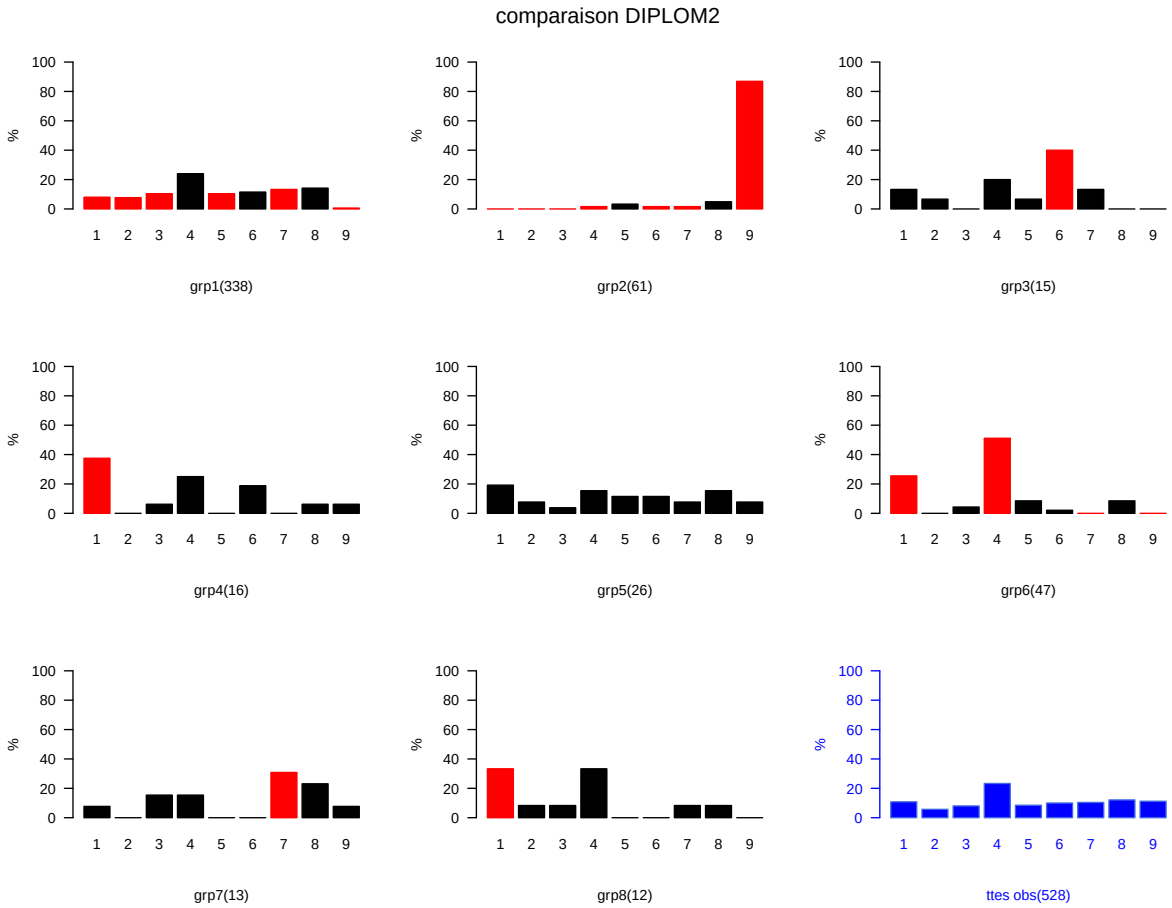

Figure 35 : Inter-comparaison DIPLOM2

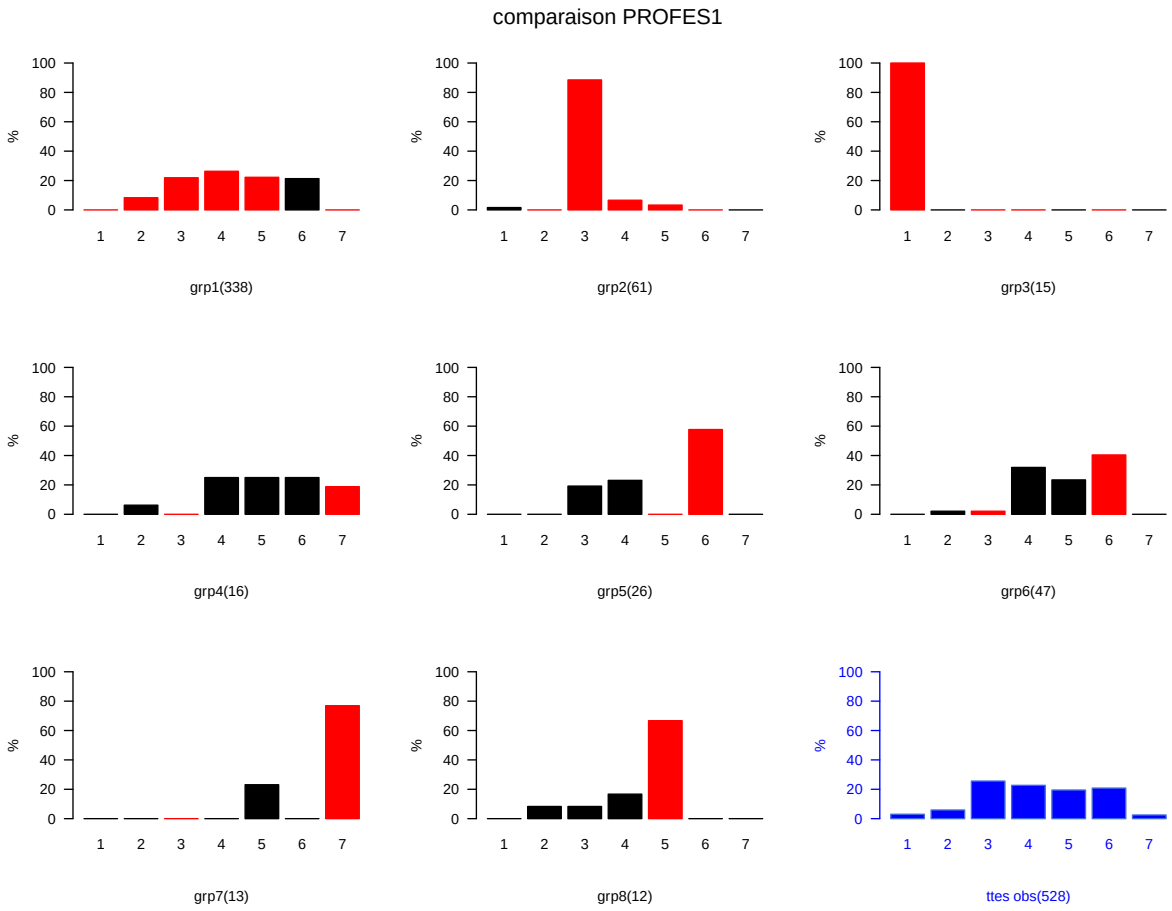

Figure 36 : Inter-comparaison PROFES1

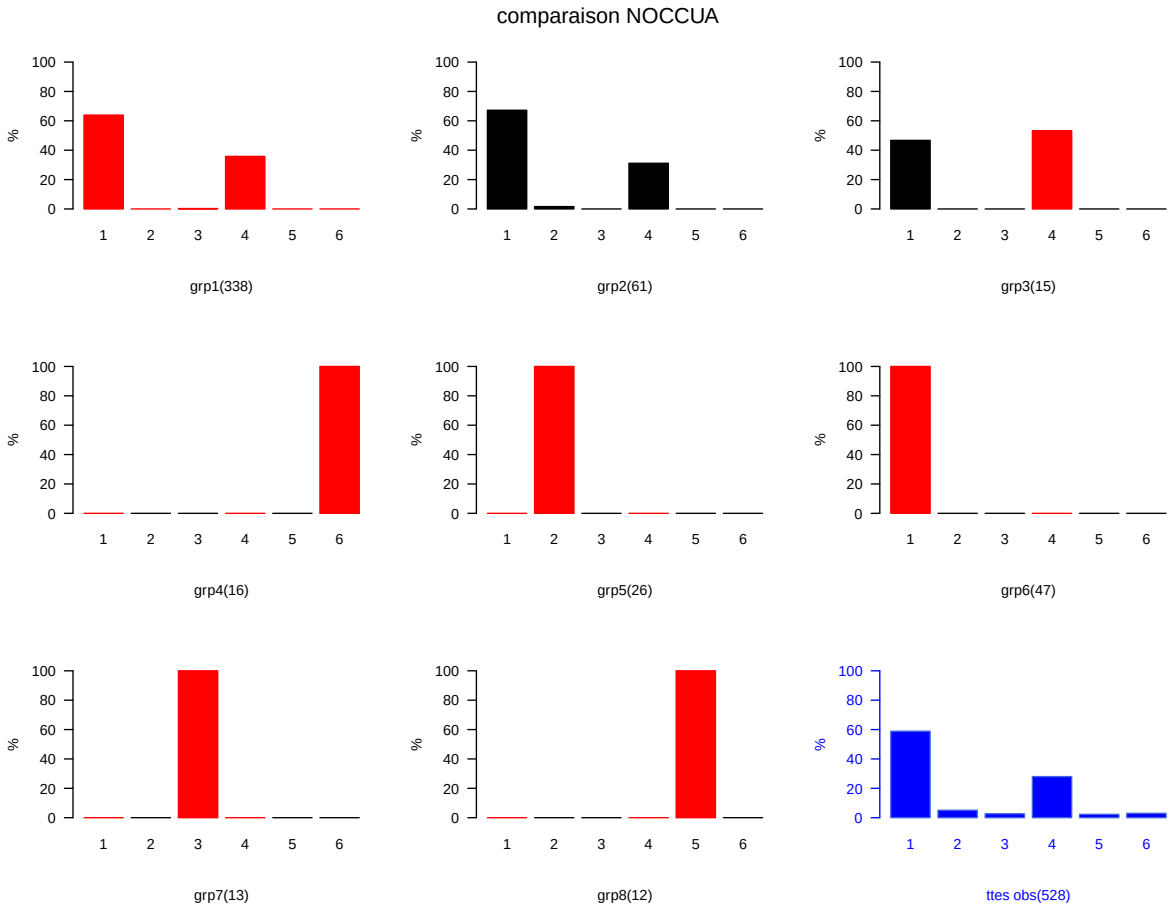

Figure 37 : Inter-comparaison NOCCUA

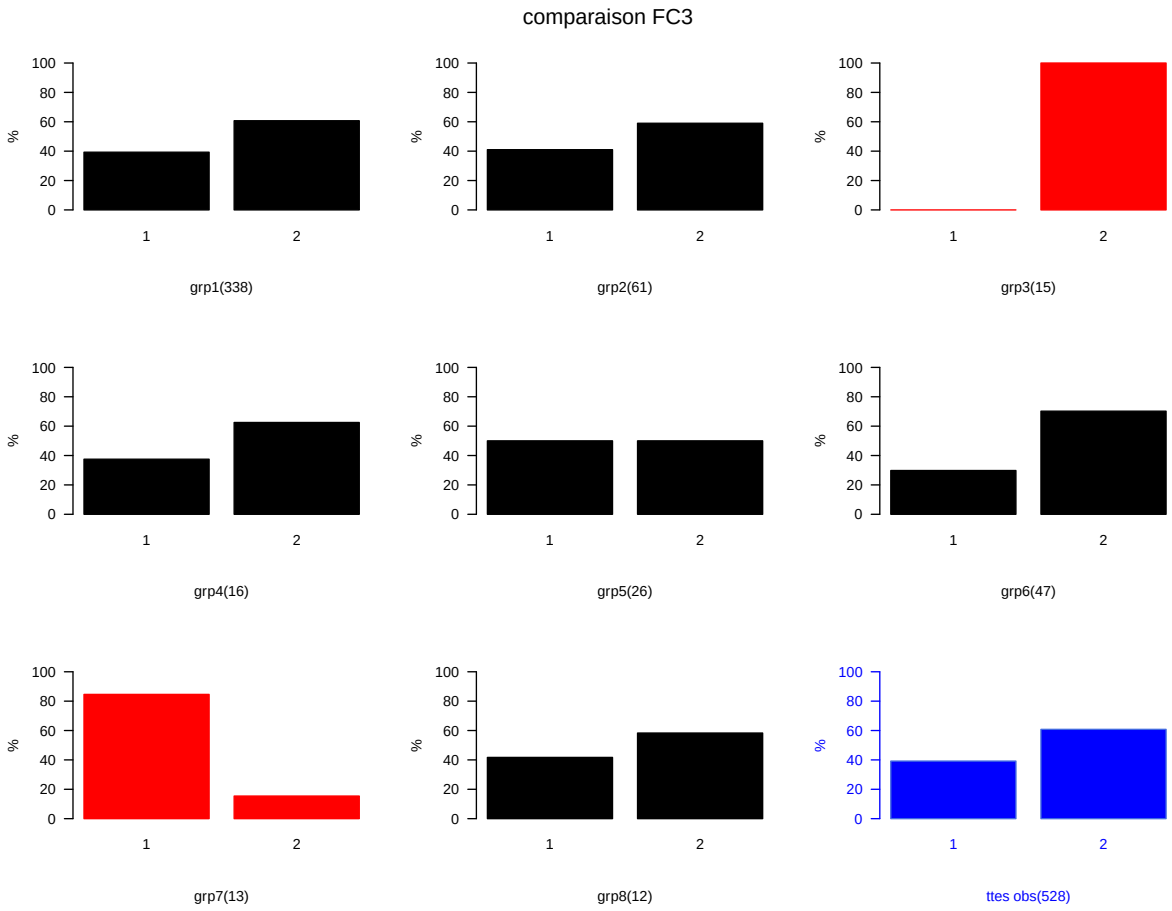

Figure 38 : Inter-comparaison FC3

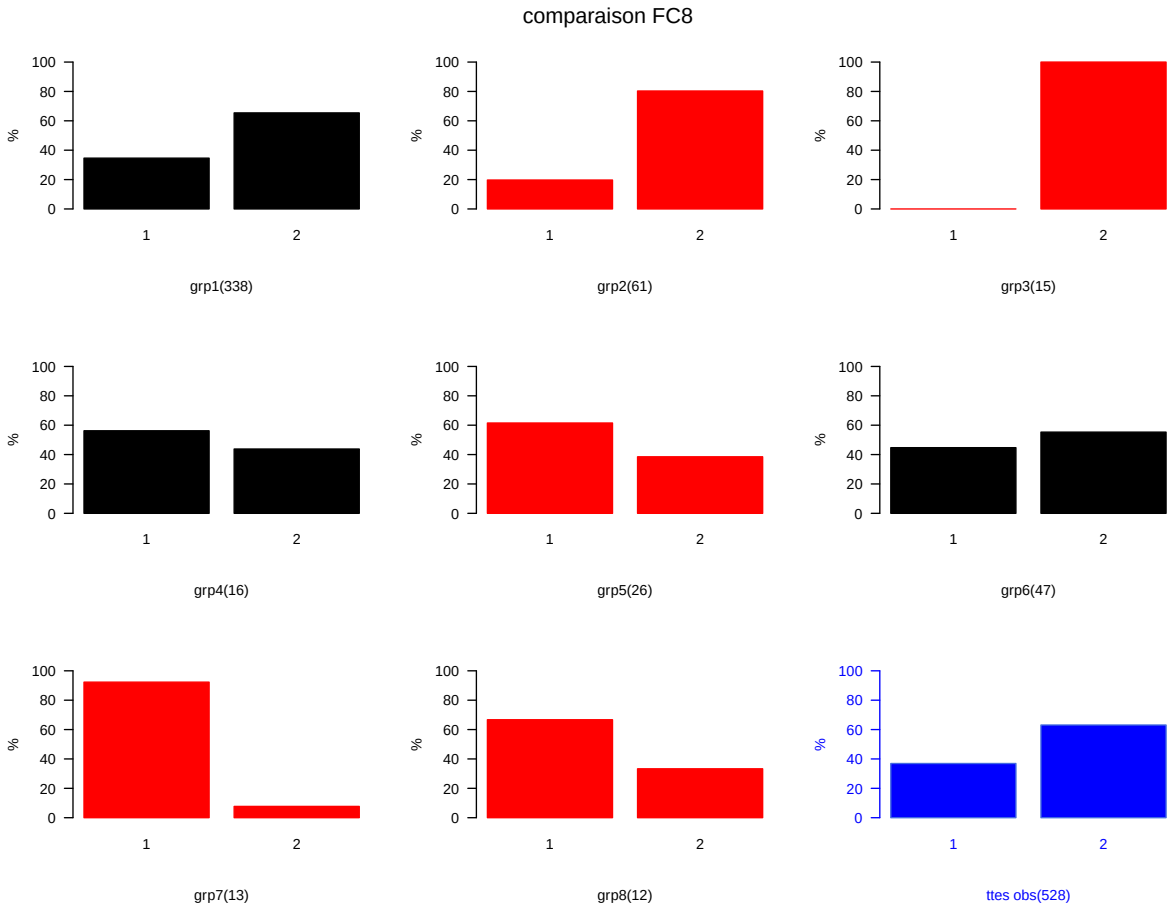

Figure 39 : Inter-comparaison FC8