Interactive Clustering Tree: Une méthode de classification descendante adaptée aux grands ensembles de données

Ricco Rakotomalala*,Tanguy Le Nouvel**

* Laboratoire ERIC - Université Lyon 2
5 av. Mendès France - 69500 Bron.
ricco.rakotomalala@univ-lyon2.fr .

** SPAD, 63 Boulevard de Ménilmontant - 75011 Paris.
tlenouvel@spad.eu

Résumé. Nous présentons une nouvelle méthode de segmentation non supervisée, particulièrement bien adaptée aux gros volumes de données et aux besoins opérationnels du secteur Banque Assurance. Cette méthode de classification descendante hiérarchique présente la segmentation finale sous la forme d'un arbre de décision dont l'appartenance aux classes (ou segments) dépend de règles logiques faisant intervenir les variables de l'analyse. De ce fait, la méthode hérite des propriétés inhérentes aux arbres de décision (interactivité, choix des variables de coupure, élagage/développement de l'arbre). Il en résulte une segmentation très simple d'interprétation et directement opérationnelle pour l'affectation d'un nouvel individu à l'une des classes. Enfin, elle intègre la possibilité de construire l'arbre avec d'autres variables que celles dont on mesure l'inertie. En ce sens, nous pouvons la considérer comme une généralisation au cas multicibles, plusieurs variables à prédire simultanément, des méthodes supervisées. Cet article présente les fondements théoriques de la méthode et s'appuie sur un exemple pratique pour illustrer les résultats obtenus et les comparer aux méthodes usuelles.

1 Introduction

La typologie, appelée également classification ou apprentissage non-supervisé, vise à segmenter une population en groupes homogènes d'individus du point de vue d'un ensemble de variables, de telle sorte que : deux individus d'un même groupe se ressemblent le plus possible ; deux individus de groupes distincts diffèrent le plus possible. La typologie synthétise et résume de manière concise une réalité complexe matérialisée par un ensemble de variables. Cette méthode exploratoire est un formidable outil pour exploiter les masses d'informations gigantesques stockées par les compagnies d'assurance et les organismes bancaires ainsi que les sociétés de vente par correspondance. La multiplicité des thématiques abordées : Risque, Tarification, Marketing opérationnel, Fidélisation, étude des points de vente, sont autant de terrains d'application de ce type de méthode.

Plus généralement, cette méthode constitue une alternative intéressante à tout projet de connaissance clients, qu'il s'agisse de décrire et comprendre des comportements anciens et/ou d'anticiper et prédire des comportements futurs. En effet, bien que la typologie soit souvent opposée aux méthodes de prédiction (scoring) dans lesquelles nous devons expliquer une variable d'intérêt, la pratique montre que la possibilité de mettre en évidence des classes homogènes d'observations, que nous pouvons caractériser, permet d'assurer une meilleure compréhension des phénomènes de causalité et une plus grande robustesse des modèles de prédiction. Il existe un grand nombre de méthodes de classification. Certaines d'entre elles sont bien adaptées au traitement des grands ensemble de données, nous pouvons citer notamment les méthodes des nuées dynamiques (plus généralement les centres mobiles), les cartes de Kohonen, les méthodes mixtes, etc., Nakache et Confais (2005) en recensent un très grand nombre. Malgré leurs qualités respectives, ces méthodes bien souvent présentent des inconvénients qui rendent leur utilisation malaisée : la difficulté à déterminer le nombre de classes ; la difficulté à interpréter les classes obtenues ; une industrialisation complexe dans la mesure où les règles d'affectation d'un nouvel individu requièrent de nombreux calculs ; l'impossibilité pour le praticien d'interagir avec le processus d'induction de manière à faire intervenir son expertise dans la recherche des solutions.

Interactive Clustering Tree (ICT), en français " Arbre de Classification Interactive ", est une méthode de classification descendante hiérarchique qui présente plusieurs caractéristiques intéressantes dès lors que la compréhension des résultats tient une place cruciale dans l'analyse. Elle intègre la description des groupes dans son processus de regroupement, sous la forme d'un arbre de décision que l'on connaît très classiquement en apprentissage supervisé (Zighed et Rakotomalala, 2000). L'interprétation est immédiate, il en est de même lorsque nous voulons affecter un nouvel individu à un groupe, les procédures de classement se présentent sous la forme de règles logiques intégrables très facilement dans un système informatique.Au delà de l'interprétation, la compréhension du processus d'appartenance aux groupes est un élément clé pour la validation des résultats, elle donne la possibilité à l'expert de comprendre les phénomènes de causalité, de la reconnaître, de la confirmer ou de l'invalider par rapport à ses connaissances. Cette caractéristique est cruciale. Dans de nombreux domaines, le praticien, les décideurs, ont besoin de comprendre un résultat pour se l'approprier. Une règle d'affectation opaque, aussi précise soit-elle, emporte difficilement l'adhésion. A l'instar de fonctionnalités disponibles dans les logiciels de fouilles de données implémentant les arbres de décision, il est possible avec cette méthode de piloter le processus d'induction. Ici également, l'expert peut intervenir, peaufiner la connaissance produite, orienter l'exploration vers des solutions qui ne correspondent peut être pas à des optimums, bien souvent trompeurs, mais en phase avec les contraintes que lui connaît et qui sont difficilement traduisibles en termes de variables ou de critères numériques. Cela peut se traduire par exemple, à pouvoir explicatif égal, à préférer des variables plus fiables ou moins coûteuses à recueillir lors de l'analyse.

Enfin dernière caractéristique notable, les arbres de classification peuvent appréhender, avec un temps de calcul raisonnable, des bases de données de grande taille - plusieurs centaines de milliers d'observations - sur des simples ordinateurs de bureau, voire des ordinateurs portables. A priori peu importante à partir du moment où les calculs peuvent être réalisés, la rapidité de traitement prend toute sa valeur dans la phase où le praticien cherche des combinaisons, construit

des variables intermédiaires, teste des idées, bref, tout le travail exploratoire dans lequel il évalue des solutions pour mettre en évidence des régularités significatives dans les données. S'il fallait patienter des heures pour vérifier la validité de chaque hypothèse, le travail d'exploration interactive des données dans laquelle l'expertise joue un rôle considérable sera très rapidement réduit à sa plus simple expression. Cet article est organisé de la manière suivante : dans un premier temps, nous présentons succinctement les éléments théoriques qui fondent la méthode, l'analogie avec les arbres de décision bien connus dans la littérature nous permettra d'aller à l'essentiel, la transposition de l'approche à la classification ; ensuite, nous détaillerons à l'aide du logiciel SPAD le traitement d'un exemple issu du domaine des assurances, nous mettrons l'accent sur l'interaction et les multiples possibilités d'interprétation des résultats, nous comparerons les résultats obtenus avec quelques méthodes classiques de typologie, l'idée est de vérifier si la contrainte introduite par la représentation par arbre dégrade significativement la qualité du partitionnement ; nous discuterons alors des pistes d'améliorations possibles de la méthode dans la quatrième section ; nous conclurons dans la cinquième et dernière section.

2 La méthode ICT- Fondements théoriques

Deux repères permettent de bien comprendre la méthode ICT. Le premier est de faire l'analogie avec les arbres de décision, bien connus en fouille de données. Il est possible de voir les arbres de classification comme une extension de cette approche où, au critère de pureté et de variance, est substitué un critère d'inertie calculé sur un ensemble de variables. Les feuilles de l'arbre représentent les classes produites par la typologie, l'objectif de l'apprentissage est de produire un arbre de classification où l'on minimiserait l'inertie intra-classes. Nous retrouvons dès lors les mêmes questions que pour l'induction des arbres de décision : comment caractériser l'homogénéité d'un sommet de l'arbre ; comment choisir la variable de segmentation lors du partitionnement d'un sommet ; comment regrouper les modalités lorsque la variable de segmentation comporte plus de deux modalités ; enfin, quelle règle adopter pour définir la bonne taille de l'arbre, et par conséquent le nombre de classes de la typologie. La méthode ICT peut être vue comme une extension multivariée de la segmentation binaire généralisée présentée dans l'ouvrage de Zighed et Rakotomalala (2000, chapitre 8). La partition est hiérarchique, la typologie en $(K+1)$ classes est le fruit de la scission d'un des groupes de la partition en K classes.

Les travaux sur les méthodes de classification " divisives monothétiques " est sans conteste notre second repère (Chavent, 1998 ; Chavent et al., 1999). La segmentation a bien été évoquée ici ou là auparavant (Volle, 1976), ces travaux ont mis en place un cadre rigoureux. Deux termes résument très bien l'approche : il s'agit d'une méthode " divisive ", le point de départ à la partition grossière rassemblant toutes les observations, la démarche consiste à fractionner les individus de manière à constituer des groupes homogènes ; il s'agit d'une méthode " monothétique ", la subdivision est réalisée à partir des valeurs d'une variable, et non pas de toutes les variables simultanément. Les méthodes DIVAF et DIVOP (Chavent et al., 1999) nous ont inspiré pour mettre au point la méthode ICT que l'on peut considérer comme une variante où les contraintes de calcul, la rapidité essentiellement, tiennent une place importante. Un arbre de classification est donc un arbre de décision dont les critères usuels sont calculés non plus sur une variable unique, qui est la variable à prédire dans l'apprentissage supervisé, mais sur

un ensemble de variables qui servent à caractériser la proximité des individus dans les groupes. Trois informations permettent d'évaluer la pertinence du groupe représenté par un sommet de l'arbre (Figure 1) :

- la taille relative du groupe (sa fréquence dans l'ensemble de l'échantillon), qui permet de se donner une idée sur la réalité du phénomène : s'agit-il d'un groupe associé à quelques cas atypiques, ou bien s'agit-il d'un sous-ensemble de la population correspondant à un comportement ou à des caractéristiques distinctifs ;
- l'indice d'homogénéité (IH) représente l'homogénéité de la classe du point de vue des variables actives de l'analyse. Plus l'indice est élevé, plus la classe est homogène. Il est calculé comme suit :

$$IH = \left(1 - \frac{\text{Inertie Intra Classe}}{\text{Inertie Totale}}\right) * 100 \; ;$$

- le gain d'inertie totale expliquée si l'on segmente le groupe. Plus grande sera cette valeur, plus il sera intéressant de scinder le sommet pour discerner des classes avec des traits spécifiques, notre partition hiérarchique est donc indicée.

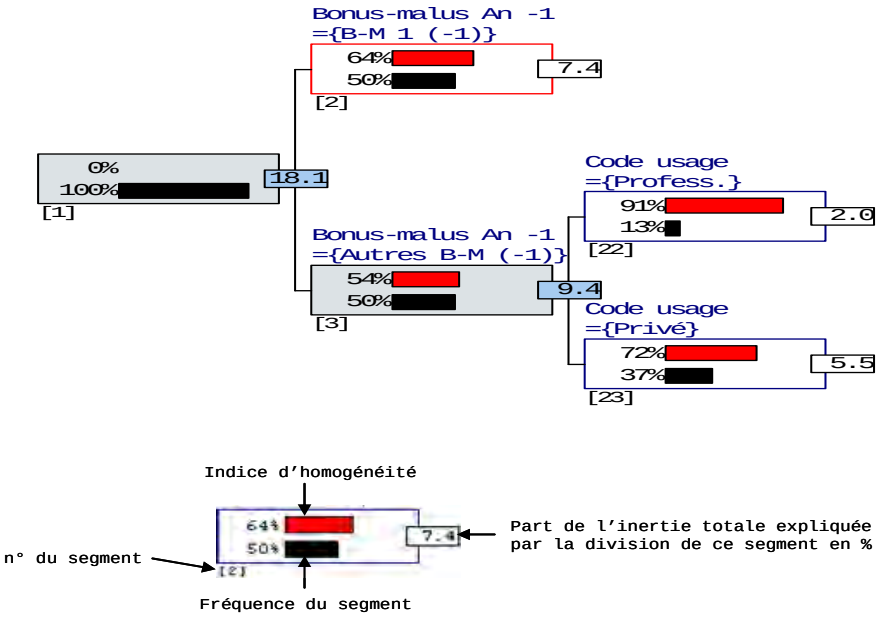


FIG. 1 – *Arbre de classification ICT*

2.1 Caractériser l'homogénéité d'un groupe d'observations

Nous utilisons l'inertie intra-classes pour mesurer l'homogénéité de la partition. Calculer une inertie ne pose pas de difficultés, s'agissant de variables continues. Il faut faire attention aux problèmes d'échelle dans le cas où les variables seraient exprimées dans des unités différentes, le mieux est alors de normaliser les données afin de pouvoir utiliser une distance simple telle que la distance euclidienne. Dans ce cas bien sûr nous ne tenons pas compte des interactions entre les variables. La situation est un peu plus difficile en ce qui concerne les données catégorielles. Nous devrions alors passer par d'autres distances telles que la distance du KHI-2. Pour ne pas multiplier la gestion de cas particuliers, toujours préjudiciable lorsque nous voulons traiter une large palette de problèmes, nous nous sommes inspirés de la classification sur facteurs (Lebart et al., 1995). Nous faisons précéder les calculs d'une analyse factorielle : une analyse en composantes principales lorsque les variables sont numériques, une analyse des correspondances multiples lorsque les variables sont catégorielles. Nous construisons alors la classification à partir des axes factoriels. Lorsque les variables sont mixtes, une possibilité serait de discrétiser, découper en intervalles, les variables continues avant de procéder à l'analyse des correspondances multiples. Outre le fait qu'elle nous permet de traiter dans un cadre unifié les différentes configurations, la classification sur facteurs comporte un certain nombre d'avantages qui la démarque de la classification à partir des variables originelles. Les variables, les facteurs, sont cette fois-ci réellement deux à deux indépendants, l'utilisation d'une distance euclidienne simple est totalement justifiée. En ne conservant que les premiers axes principaux, nous n'utilisons que l'information " utile " pour construire les groupes, ce lissage des données permet d'évacuer le bruit qui est concentré sur les derniers axes, générateurs d'instabilité.

2.2 Sélectionner la variable de segmentation

Evaluer une segmentation permet de juger de son opportunité ; elle permet également de choisir parmi les variables de segmentation candidates, celle qui est la plus intéressante ; elle permet enfin de choisir parmi les feuilles de l'arbre, celle qui induira la prochaine subdivision la plus pertinente au sens de l'inertie expliquée. Parmi les différents critères à notre disposition en classification (Nakache et Confais, 2005), nous nous sommes tournés vers le critère de WARD qui quantifie le gain d'inertie. Pour la segmentation d'un sommet S (de taille n) en deux feuilles S_A et S_B (d'effectifs n_A et n_B) à l'aide de la variable X_j, le critère de WARD s'écrit :

$$\Delta(X_j) = \frac{n_A \times n_B}{n_A + n_B} d^2(g_A, g_B).$$

$d^2(g_A, g_B)$ représente le carré de la distance euclidienne entre les deux centres de gravité des feuilles S_A et S_B. Le processus de sélection de la variable de segmentation sur un sommet, dans le cas où elles seraient toutes binaires, revient à choisir celle qui induit le gain d'inertie maximal.

$$X_{j^*} = arg \max_j \Delta(X_j).$$

Lorsque la variable X_j est continue, nous opérons en deux temps : tout d'abord nous trions les données selon les valeurs de la variable de segmentation, puis nous testons toutes les coupures candidates de manière à optimiser le critère d'inertie. Il existe des algorithmes de tris efficaces de type QUICKSORT ou HEAPSORT, l'évaluation de chaque coupure peut être réalisée en

temps linéaire. Dans la pratique, le traitement des variables continues est très rapide et ne pose pas de problèmes particuliers. Bien entendu, tout comme dans les arbres de décision, comme le découpage est binaire à chaque noeud, la même variable continue peut être réintroduite à différents niveaux de l'arbre.

2.3 Regrouper les modalités d'une variable catégorielle

Lorsque la variable X_j est catégorielle, comportant M modalités, nous avons la possibilité de produire une feuille par modalité de la variable de segmentation. Cette stratégie, très simple, se heurte néanmoins à différents écueils, le principal inconvénient est de fragmenter trop rapidement les données, certaines feuilles comportent trop peu d'observations, elles ne correspondent pas forcément à un comportement spécifique au regard des variables de l'étude. Le regroupement des modalités de manière à construire un arbre binaire n'est donc pas anodin, généralement il permet de produire des arbres plus concis, plus fiables et plus lisibles (Zighed et Rakotomalala, 2000). Nous distinguons deux cas. Si la variable est ordinale, il suffit d'ordonner les modalités et de tester les combinaisons binaires, il y a $(M-1)$ cas à tester, nous nous retrouvons dans une procédure analogue au traitement des variables continues. La situation est un peu plus compliquée lorsque la variable est nominale. Tester toutes les combinaisons possibles revient à tester $(2^{M-1}-1)$ cas, ce qui est impraticable dès que le nombre de modalités augmente. Pour donner un ordre d'idée, si $M = 20$, il y a 524.287 situations à évaluer. Il faut impérativement trouver une stratégie qui assure d'obtenir de bons résultats avec un temps de calcul raisonnable. Dans la méthode DIVOP par exemple (Chavent et al., 1999), les auteurs préconisent le calcul d'une nouvelle analyse factorielle localement sur chaque noeud de l'arbre, les modalités sont alors ordonnées sur le premier axe factoriel, ce qui permet de se ramener dans la situation des variables ordinales. Nous avons opté pour une solution plus simple en réalisant tout simplement une classification ascendante hiérarchique sur les modalités de la variable de segmentation. Il s'agit d'une approche pas à pas, le nombre de tests est connu à l'avance, la complexité de calcul est quadratique. Les inconvénients de ce type d'optimisation sont bien connus, nous sommes exposés à des solutions sous optimales. Mais on peut se demander finalement si, en lissant ainsi l'exploration de l'espace des solutions, nous ne nous prémunissons pas du sur-apprentissage. Les combinaisons optimales sont souvent spécifiques aux données d'apprentissage, elles ne sont pas toujours directement transposables à la population. Quoiqu'il en soit, l'implémentation logicielle étant interactive, le praticien a la possibilité de réviser les solutions proposées et d'introduire les combinaisons qui lui semblent plus en phase avec son expertise.

2.4 Définir le nombre de classes

Définir le nombre de classes à retenir est un problème complexe, il existe pléthore d'indicateurs qui permettent de s'en donner une idée (Nakache et Confais, 2005 ; chapitre 7). Plutôt que de définir une procédure automatique de détection du nombre " optimal " de classes, toujours hasardeuse, nous exploitons le fait que notre approche produise un dendogramme, une hiérarchie indicée de partitions, pour proposer des outils d'interprétation et de décision analogue à ceux de la classification ascendante. Le graphique représentant la réduction de l'inertie intra-classes associée à chaque segmentation est notamment d'une aide précieuse. Nous pouvons ainsi nous donner une idée de la bonne partition en observant le " coude " dans ce graphique,

correspondant à une réduction de l'inertie intra-classes peu informative au regard de l'augmentation de la complexité, l'augmentation du nombre de classes, de l'arbre de classification. Nul doute qu'à l'avenir, il faudra réfléchir à une méthode de détection du nombre de classes, ne serait-ce que pour donner au praticien une indication sur la " bonne " solution. Nous pensons faire le parallèle avec la méthode CART (Breiman et al., 1984). L'idée serait de fractionner les données d'apprentissage en données d'expansion (growing set) et de validation (pruning set). Le premier échantillon sert à construire l'arbre le plus spécialisé possible, puis dans une seconde phase, nous utilisons le second échantillon pour élaguer l'arbre selon le principe du coût complexité. Nous pouvons raisonnablement penser que si l'inertie intra-classes calculée sur le fichier d'expansion continue à décroître à mesure que nous augmentons la taille de l'arbre, il en est autrement sur le fichier de validation, l'inertie doit stagner, voire se dégrader, lorsque nous introduisons des segmentations qui ne sont plus pertinentes. Nous menons des expérimentations en ce sens à l'heure actuelle. La classification présente néanmoins une différence fondamentale par rapport au classement et à la régression, l'objectif n'est pas tant de construire l'arbre le plus précis possible, mais plutôt de trouver le meilleur arbitrage entre la complexité (le nombre de classes de la typologie) et la précision de l'arbre (l'inertie intra-classes).

3 Exemple et comparaison avec les méthodes usuelles

Pour illustrer la méthode ICT, nous disposons d'un échantillon constitué de 1106 assurés belges observés en 1992. Il s'agit du fichier ASSUR.SBA fourni avec le logiciel SPAD. La méthode ICT est disponible depuis la version 6.5 de SPAD que nous utilisons pour illustrer cet article. On entend par assuré une personne physique ou morale. Les variables actives concernent le souscripteur et le véhicule :

- L'âge de l'assuré (3 modalités) ;
- Le code usage (2 modalités) ;
- L'année de construction du véhicule (2 modalités) ;
- Le degré de bonus-malus de l'année précédente (2 modalités) ;
- La date d'effet de la police (2 modalités) ;
- Le code postal recodé par arrondissement et regroupé par proximité et caractère urbain, suburbain ou rural (2 modalités) ;
- Le code langue (Français, Néerlandais) ;
- La puissance (2 modalités).

L'objectif de l'analyse est de mettre en évidence les *profils types* des assurés au regard de l'ensemble de ces caractéristiques. L'idée principale étant de regrouper dans une même classe ou segment des assurés dont les caractéristiques sont très proches, et dans des classes/segments différents les assurés qui se différencient le plus. Des paramètres relatifs aux sinistres et aux primes, les variables illustratives, sont utilisés a posteriori pour caractériser les classes obtenues. Ces variables sont :

- Les charges occasionnées par le sinistre en Francs Belges ;
- Les primes d'assurances versées par l'assuré en Francs Belges ;

3.1 Analyse préalable et choix du nombre d'axes à conserver

L'analyse des correspondances multiples effectuée sur les 8 variables actives est illustrée par le premier plan factoriel (Figure 2).

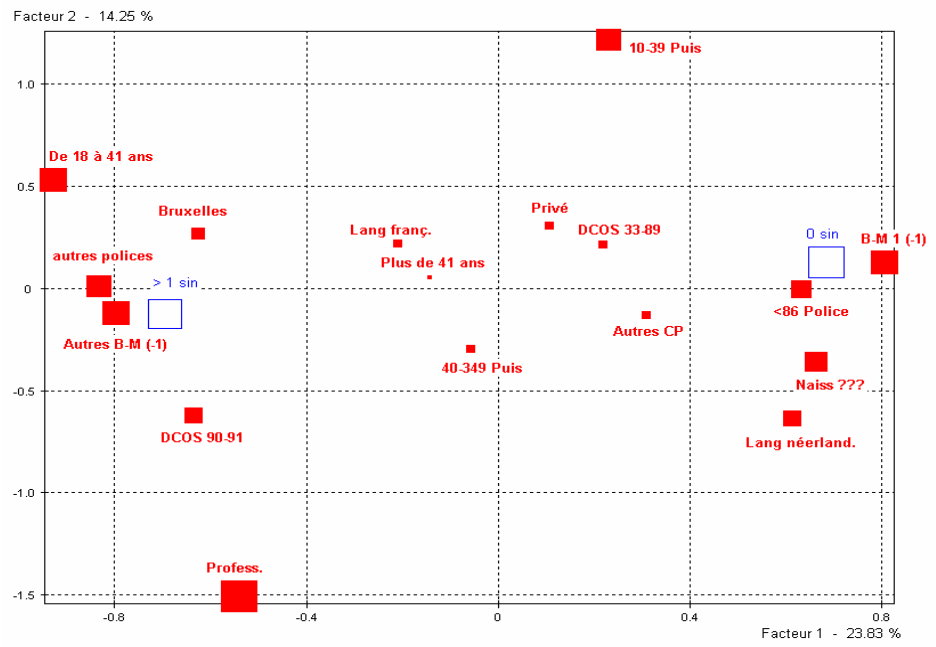


FIG. 2 – *Plan factoriel 1-2 de l'analyse des correspondances multiples*

Dans le cadre de cet exemple, nous proposons de conserver les 7 premiers axes factoriels pour la classification. Ils retiennent 90% de l'information (Tableau 1). Notons que le logiciel ne permet pas une sélection discontinue des axes factoriels.

3.2 Construction de l'arbre et détermination du nombre de classes

Dans cet exemple, ICT prend en compte les 7 premières coordonnées factorielles de l'analyse préalable pour la classification. Les paramètres de construction de l'arbre, présentés dans la figure 3, permettent de définir des critères d'arrêt. Nous retrouvons certains critères d'arrêt des arbres de décision en apprentissage supervisé. Il s'agit bien souvent de critères de bon sens qui évitent à l'algorithme de s'enferrer dans des impasses. C'est le cas par exemple de " l'effectif minimum pour diviser un noeud " ou de " l'effectif d'admissibilité ", l'idée est de contraindre la construction de l'arbre de manière à éviter la formation de groupes de petite

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0,2681	23,83	23,83
2	0,1603	14,25	38,08
3	0,1513	13,44	51,52
4	0,1319	11,72	63,25
5	0,1094	9,72	72,97
6	0,1084	9,64	82,61
7	0,0833	7,40	90,01
8	0,0588	5,23	95,24
9	0,0536	4,76	100,00

TAB. 1 – *Tableau des valeurs propres de l'analyse des correspondances multiples*

taille, peu significatives. Les paramètres sur le nombre de classes et le nombre de niveaux permettent de contrôler le temps de calcul. Les valeurs par défaut suffisent souvent pour initier une première analyse, le dendogramme et les aides à l'interprétation permettent alors au statisticien de mener à sa guise l'exploration.

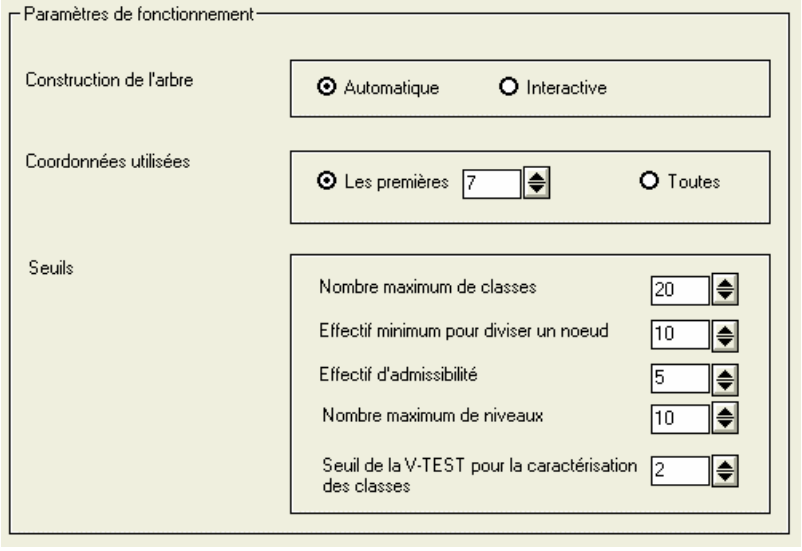


FIG. 3 – *Paramètres de la classification*

La courbe des 20 premiers indices de niveaux liée à la classification descendante produite par ICT montre une rapide décroissance de l'inertie expliquée par chaque segmentation, avec plusieurs paliers (Figure 4).
Le saut associé à la partition en deux classes se retrouve souvent dans les méthodes hiérarchiques. Cette segmentation se démarque car il s'agit tout simplement du premier partitionnement des données. Dans la plupart des cas, il ne correspond pas à la bonne solution. Un autre "

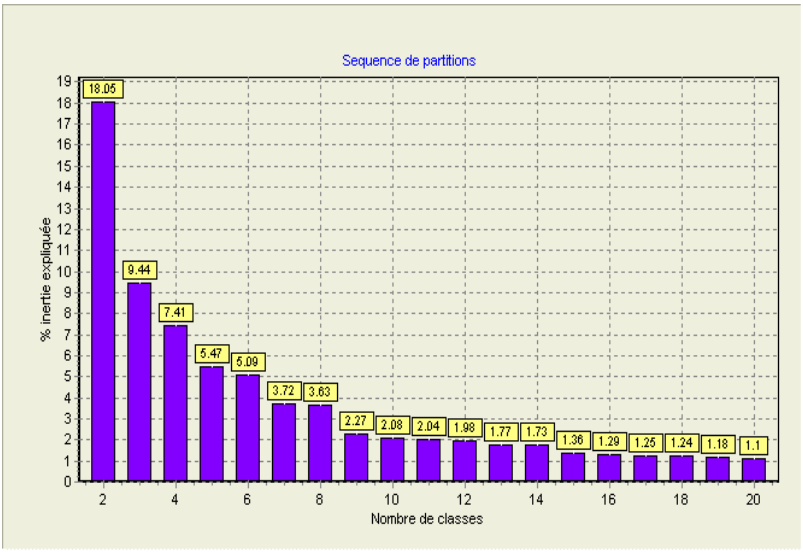


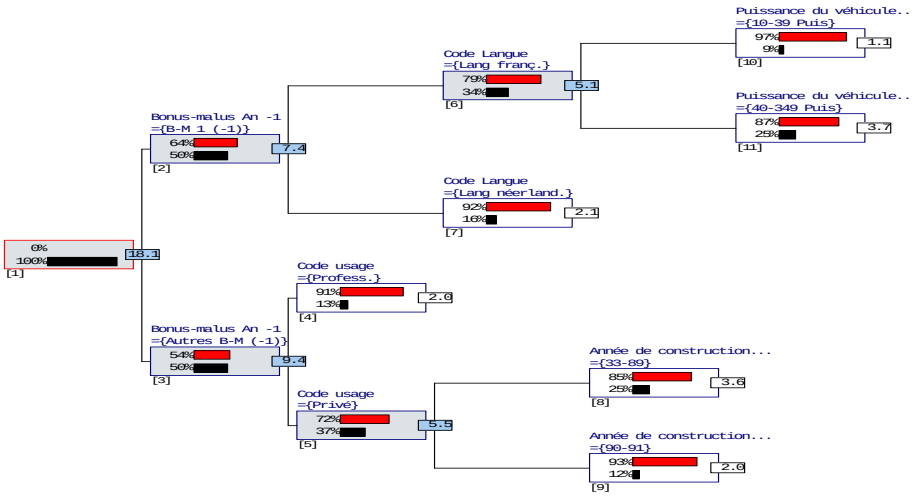
FIG. 4 – *Indices de niveau de la classification descendante*

saut " attire notre attention, il s'agit de la partition en 6 classes. C'est le choix que nous ferons dans cet exemple.

3.3 Interprétation de l'arbre de typologie

La figure 5 est la formalisation par ICT de la typologie sous la forme d'un arbre de décision. Ici, l'arbre est représenté horizontalement. Les 6 classes de la typologie constituent les éléments terminaux de l'arbre. Les segments intermédiaires sont colorés en gris. Le positionnement horizontal des segments de gauche à droite suit l'ordre hiérarchique de la classification. Pour des raisons de lisibilité, les sauts sont réalisés à pas constants et ne respectent pas les indices de niveaux. L'échantillon total des 1106 assurés est dans un premier temps divisé en deux segments avec la variable " Bonus $-$ Malus Année $-$ 1 " : le segment $n^{\circ}2$ regroupe les assurés dont le coefficient Bonus-Malus de l'année précédente est le plus faible (Catégorie$B-M1(-1)$), le segment $n^{\circ}3$ regroupe les assurés de l'autre catégorie (Autres$B-M(-1)$). A ce stade, le choix du segment à subdiviser se fait selon le gain de l'inertie expliquée. Avec un gain de $9,4\%$ contre $7,4\%$, ICT segmente " Autres$B-M(-1)$ " en fonction de l'usage de leur véhicule : à usage privé (segment 5) versus à usage professionnel (segment terminal 4). Dans cet exemple, le gain d'inertie expliquée relatif à une nouvelle subdivision du segment $n^{\circ}4$ est trop faible comparativement aux autres subdivisions possibles. Finalement, ce segment (appelé aussi classe 4) regroupe les assurés dont le coefficient de bonus-malus de l'année précédente est le plus élevé et qui utilisent leur véhicule à des

fins professionnelles. Ce segment représente 13 % de l'échantillon étudié et affiche un indice d'homogénéité de 91%.



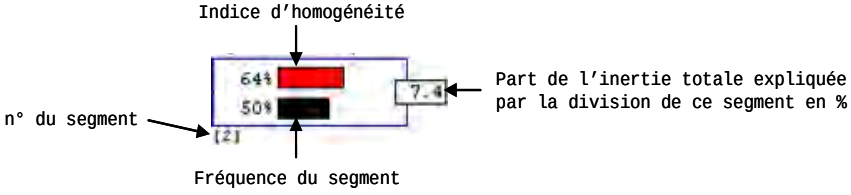


FIG. 5 – *Arbre de typologie en 6 classes.*

Cet indice représente l'homogénéité de la classe du point de vue des variables actives de l'analyse (représentées ici par les 7 premiers axes factoriels). Plus l'indice est élevé, plus la classe

est homogène. Il est calculé comme suit :

$$\text{Indice d'homogénéité} = \left(1 - \frac{\text{Inertie Intra Classe}}{\text{Inertie Totale}}\right) * 100$$

Le caractère monothétique des arbres de décision confère à ICT une grande simplicité de lecture et d'interprétation des classes de la typologie.

3.4 Interactivité

Comme pour les arbres de décision interactifs, ICT permet au praticien d'intervenir dans la construction de l'arbre. Il peut développer ou élaguer l'arbre de typologie à sa guise, choisir les variables les plus judicieuses, modifier les seuils de partitionnement, définir le nombre de classes le plus adéquat en s'appuyant à la fois sur les indicateurs statistiques fournis et sur ses connaissances du domaine. Dans cet exemple, la subdivision du segment 5 fait intervenir la variable " Année de construction de véhicule " en raison de son impact maximal. La liste des variables candidates triées par ordre d'impact décroissant (part de l'inertie totale expliquée) montre que les variables " Age de l'assuré " et " Code postal " sont très proches en termes d'impact (Figure 6).

Distribution | Autres seg.

Variable	Impact
Année de construction du véhicule	5.4683
Age de l'assuré	5.4482
Code postal	5.3878
Puissance du véhicule	4.6588
Code Langue	3.8840
Date effet Police	3.2930
Code usage	0.0000
Bonus-malus An -1	0.0000

FIG. 6 – *Variables candidates à la subdivision du segment 5*

On préférera ici la variable " Age de l'assuré " à "Année de construction de véhicule " pour son aspect plus opérationnel. En effet, on peut imaginer que l'âge est plus caractéristique du comportement de la personne, cette variable est de surcroît plus facile à mesurer sans ambiguïté (cf. figure 7).

3.5 Caractérisation automatique des classes

La procédure fournit automatiquement un rapport synthétique qui présente les traits saillants de chaque classe, à partir de l'ensemble de l'information disponible : les variables actives qui ont servi à calculer les inerties et construire l'arbre ; les variables illustratives que nous utiliserons pour caractériser les classes. La figure 8 est un extrait de ce rapport, elle illustre les classes $4, 10$ et 7. Pour caractériser un groupe, nous procédons de la manière suivante : pour

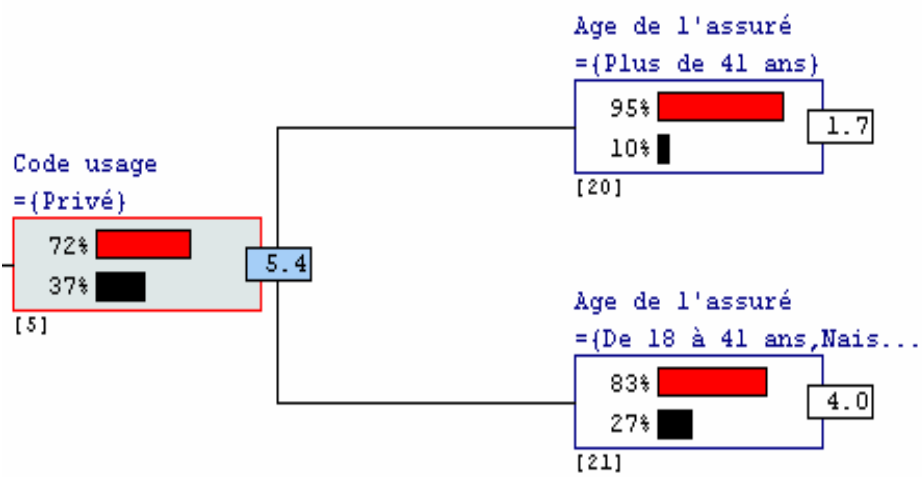


FIG. 7 – *Nouvelle subdivision du segment 5 avec la variable «Age de l'assuré»*

chaque variable, nous calculons et nous comparons un indicateur statistique calculé dans la totalité de l'échantillon puis dans la classe ; nous trions alors les variables selon l'intensité de l'écart. Concrètement, lorsque la variable est numérique, nous procédons à une comparaison de moyennes ; lorsque la variable est catégorielle, nous procédons à une comparaison de proportions. Pour matérialiser l'intensité de l'écart, nous calculons la probabilité critique du test (la p-value) que nous traduisons ensuite en nombre d'écarts types de la loi normale centrée et réduite qu'il faut dépasser pour couvrir cette probabilité. C'est le concept de " valeur test " (Morineau, 1984 ; Lebart et al., 1995). Ainsi lorsque la valeur-test est supérieure à 2 en valeur absolue, l'écart est approximativement significatif au seuil usuel de 5% pour un test bilatéral, 2.5% pour un test unilatéral. Dans SPAD, pour filtrer les résultats et n'afficher que les variables réellement caractéristiques de la classe, nous avons placé en paramètre une valeur test de référence. Nous avons préféré faire porter ce seuil sur la valeur test et non pas sur la probabilité critique car nous avons constaté que lorsque les effectifs deviennent importants, la valeur test atteint des niveaux très élevés, nous obligeant à introduire des seuils de risque très faibles, de l'ordre de 10^-8, qui n'ont plus aucun lien avec les valeurs usuellement utilisées en statistique.

Ne sont listées dans cet exemple que les modalités sur-représentées dans les classes, présentant une valeur test supérieure à 2 en valeur absolue (voir Figure 3, Paramètres de la classification). Nous constatons que les assurés du segment terminal 4 ont des primes d'assurances plus élevées que la moyenne (18297 FB contre 13724 FB), ce qui semble cohérent avec les conditions d'appartenances à ce segment : un coefficient élevé de Bonus-Malus et un usage professionnel de leur véhicule. On note que près de 90% d'entre eux ont eu au moins un sinistre alors que

Classe	Description des classes			
Segment Terminal 4	Effectif	147		
	Description symbolique	**SI Code usage ={Profess.} et Bonus-malus An -1 ={Autres B-M (-1)} ALORS ST4**		
	Var. continue			
	Variable(s)	**Moyenne Classe**	**Moyenne Ech. total**	**Valeur-test**
	Primes Acquises RC 1991 en francs belges	*18297,6*	*13724,5*	*14,8*
	Var. catégorielle			
	Variable(s) = modalité	**% Classe**	**% Ech. total**	**Valeur-test**
	Code usage = Profess.	100,0%	16,7%	29,0
	Bonus-malus An -1 = Autres B-M (-1)	100,0%	50,4%	12,9
	Sinistralité RC = > 1 sin	*89,1%*	*49,7%*	*10,3*
	Age de l'assuré = Naiss ???	61,2%	44,8%	4,3
	Date effet Police = autres polices	59,2%	43,1%	4,2
	Puissance du véhicule = 40-349 Puis	91,8%	80,4%	3,8
	Année de construction du véhicule = 90-91	36,7%	25,6%	3,3
	Code postal = Bruxelles	44,2%	33,2%	3,1
Segment Terminal 10	Effectif	95		
	Description symbolique	**SI Puissance du véhicule ={10-39 Puis} et Code Langue ={Lang franç.} et Bonus-malus An -1 ={B-M 1 (-1)} ALORS ST10**		
	Var. continue			
	Variable(s)	**Moyenne (Classe)**	**Moyenne Ech. total**	**Valeur-test**
	Charge Sinistre RC 1991 en francs belges	*2269,6*	*20245,5*	*-2,9*
	Primes Acquises RC 1991 en francs belges	*9251,2*	*13724,5*	*-11,4*
	Var. catégorielle			
	Variable(s) = modalité	**% Classe**	**% Ech. total**	**Valeur-test**
	Puissance du véhicule = 10-39 Puis	100,0%	19,6%	20,6
	Bonus-malus An -1 = B-M 1 (-1)	100,0%	49,6%	10,3
	Sinistralité RC = 0 sin	*89,5%*	*50,3%*	*8,0*
	Date effet Police = <86 Police	87,4%	56,9%	6,3
	Code Langue = Lang franç.	100,0%	74,5%	6,0
	Code usage = Privé	98,9%	83,3%	4,3
	Année de construction du véhicule = 33-89	91,6%	74,4%	4,0
	Age de l'assuré = Naiss ???	63,2%	44,8%	3,8
	Code postal = Autres codes	77,9%	66,8%	2,4
Segment Terminal 7	Effectif	178		
	Description symbolique	**SI Code Langue ={Lang néerland.} et Bonus-malus An -1 ={B-M 1 (-1)} ALORS ST7**		
	Var. continue			
	Variable(s)	**Moyenne (Classe)**	**Moyenne Ech. total**	**Valeur-test**
	Charge Sinistre RC 1991 en francs belges	*8174,8*	*20245,5*	*-2,8*
	Primes Acquises RC 1991 en francs belges	*11903,1*	*13724,5*	*-6,6*
	Var. catégorielle			
	Variable(s) = modalité	**% Classe**	**% Ech. total**	**Valeur-test**
	Code Langue = Lang néerland.	100,0%	25,5%	24,9
	Bonus-malus An -1 = B-M 1 (-1)	100,0%	49,6%	14,7
	Sinistralité RC = 0 sin	*87,6%*	*50,3%*	*10,9*
	Code postal = Autres codes	93,8%	66,8%	8,3
	Date effet Police = <86 Police	80,3%	56,9%	6,9
	Age de l'assuré = Naiss ???	64,6%	44,8%	5,8
	Année de construction du véhicule = 33-89	89,3%	74,4%	5,0
	Code usage = Privé	92,7%	83,3%	3,7

FIG. 8 – *Caractérisation des classes 4, 10, 7*

la proportion est de 50% dans l'échantillon. On note un saut au niveau des valeurs-tests des autres caractéristiques. On trouve dans ce segment une majorité d'assurés dont on ignore la date de naissance (probablement lié à l'usage professionnel de leur véhicule). Cette aide à l'interprétation est précieuse car elle permet de mesurer la cohérence métier des classes obtenues

notamment à travers les informations illustratives. En mode interactif, les nouvelles classes créées par l'utilisateur sont automatiquement caractérisées de la sorte. Ce qui permet de juger immédiatement de leur pertinence.

3.6 Comparaison avec les méthodes usuelles

3.6.1 Part d'inertie expliquée par la classification

Les arbres de classification se démarquent des méthodes usuelles de typologie en proposant des solutions interprétables et opérationnelles, il est possible de classer directement un nouvel individu. Ces qualités sont indéniables. Nous pouvons en revanche nourrir quelques inquiétudes quant aux performances de l'approche en termes d'inertie expliquée. En effet, nous introduisons deux biais qui peuvent s'avérer rédhibitoires dans la recherche du partitionnement optimal. Tout d'abord, nous représentons la partition à l'aide d'un arbre de décision. Concrètement, cela veut dire que nous morcelons l'espace de représentation à l'aide d'hyper-rectangles pour constituer les groupes. Il s'agit là d'un a priori qui ne correspond pas forcément, pas du tout même, avec la forme des nuages de points correspondant à chaque classe. Elles peuvent être de forme sphérique ou allongée, ou d'autres formes encore, le choix d'un hyper-rectangle est une contrainte forte de représentation qui peut ne pas correspondre à la réalité. Le second point qui peut poser problème est l'algorithme de recherche du partitionnement. Nous procédons pas à pas, de manière myope, en optimisant localement la segmentation de chaque sommet. La solution proposée au final a de fortes chances d'être sous-optimale, bien en deçà, peut-être, des autres méthodes usuelles de classification. Pour vérifier l'écart de performances entre la méthode ICT et les autres méthodes, nous avons décidé de confronter le partitionnement en 6 classes sur notre fichier de données. Nous avons utilisé les méthodes des centres mobiles (K-Means), la classification ascendante hiérarchique (CAH), et les méthodes mixtes qui combinent ces deux approches (Lebart et al., 1995, chapitre 2 ; Nakache et Confais, 2005, chapitre 5). Dans notre exemple, la méthode mixte $n^{\circ}2$ donne les meilleurs résultats en termes d'explication de l'inertie totale. Cette méthode procède en trois étapes :

1. Partitionnement préliminaire : recherche de classes stables par croisement de 2 partitions de 10 classes construites par la méthode des centres mobiles.
2. Agrégation hiérarchique (CAH) des classes stables obtenues
3. Partition finale en 6 classes et consolidation par la méthode des centres mobiles

Méthode de classification	% Inertie expliquée par la partition en 6 classes
ICT	45,4
K-Means : 6 centres tirés au hasard	49,6
Mixte 1 : K-Means + CAH	51,6
Mixte 2 : K-Means + CAH + K-Means	53,1
CAH	47,3
Mixte 3 : CAH + K-Means	50,7

TAB. 2 – *Tableau comparatif des méthodes de classification*

Nous constatons surtout que la part d'inertie expliquée par ICT est proche de ce que l'on obtient avec les méthodes usuelles, bien que légèrement inférieure. Ces résultats ont été observés à maintes reprises sur d'autres fichiers. Les craintes que nous avons émises précédemment n'ont

pas lieu d'être. Même si effectivement, à cause des contraintes de représentation et d'exploration, il paraît difficile avec ICT de surpasser en termes d'inertie expliquée les autres méthodes telles que les centres mobiles, nous ne notons pas une dégradation notable des performances.

3.6.2 Affectation des individus dans les classes

Outre l'interactivité inhérente à ICT, son autre avantage est la formalisation automatique de la typologie sous la forme d'un arbre de décision. Ce qui simplifie considérablement l'affectation des assurés au cours du temps : ceux dont le profil a changé et les nouveaux assurés. Lorsque la typologie provient de méthodes de classification usuelles, on réalise dans un second temps une discrimination par arbre de décision pour permettre la réaffectation. La variable à prédire est la typologie ou partition obtenue et les variables explicatives sont les variables actives de la classification. Il est alors fréquent d'observer que certaines classes sont " oubliées " par l'arbre de décision et que les taux d'erreur de ré-affectation peuvent être importants. Dans notre exemple, nous avons utilisé l'algorithme CART (Breiman et al., 1984), de la version 6.5 de SPAD en sortie des méthodes Mixte $n^{\circ}2$ et Mixte $n^{\circ}3$.

Fichier ASSUR 8 var actives-7 axes conservés	Inertie expliquée	6 classes Arbre Nb Segments	Erreur moyenne d'affectation	Arbre Nb Segments	Erreur moyenne d'affectation
ICT	45,4	6	-	6	-
Mixte 2 : K-Means + CAH + K-Means	53,1	26	4,2 %	7	7,7 %
Mixte 3 : CAH + K-Means	50,7	15	6,3 %	9	9,8 %
Paramètres spécifiques		NXIND = 5 Seuil de spécialisation = 1		NXIND = 50 Seuil de spécialisation = 1	

TAB. 3 – *Comparaison des méthodes de réaffectation*

Deux types de tests ont été réalisés selon le paramètre NXIND : Effectif minimum pour diviser un segment. Nous avons ensuite mesuré l'erreur d'affectation sur la base de 10 échantillons tests tirés aléatoirement. Avec un seuil de 5 pour NXIND, les arbres de décision associés aux méthodes mixtes $n^{\circ}2$ et $n^{\circ}3$ présentent une erreur moyenne d'affectation très faible (respectivement $4,2$ et $6,3\%$) mais sont en revanche très denses (respectivement 26 et 15 segments terminaux contre 6 pour ICT). Cette forte densité complique bien évidemment l'interprétation de la réaffectation proposée par les arbres de décision comme l'illustre la figure 9 à suivre.

Mixte n°2 et Mixte n°3

Avec un seuil de 50 pour NXIND, on gagne en lisibilité (respectivement 7 et 9 segments terminaux) mais l'erreur moyenne d'affectation s'accroît.

4 Extensions de la méthode

Dans un premier temps, nous nous sommes placés dans le cadre *stricto-sensu* de la classification pour présenter la méthode ICT. Les variables qui servent à calculer l'homogénéité des classes ont été utilisées pour constituer les classes en tant que variables de segmentation dans l'arbre de classification. En réalité, ce cadre n'est pas inhérent à la méthode. Il est tout à fait possible de calculer l'inertie des groupes à partir d'une série de variables , $(Y_i, i = 1, \ldots, I)$ et construire l'arbre à l'aide de variables de segmentation, qui servent à caractériser les groupes, $(X_j, j = 1, \ldots, J)$ différentes. Nous sommes toujours dans un processus de classification

mais avec un positionnement différent, il s'agirait alors d'une extension des arbres de décision ou de régression. Certains auteurs parlent alors de " Predictive Clustering Trees " (Blockeel et al., 1998) ou d' " Arbres de Décision Multi-Cibles " (Meyer et Clerot, 2006). Les approches diffèrent selon que l'on travaille sur des variables à prédire continues ou catégorielles. L'approche ICT simplifie l'approche, elle a le mérite de la cohérence. Si les Y sont toutes numériques, nous réalisons une analyse en composantes principales ; si elles sont catégorielles, nous procédons à une analyse des correspondances multiples. Nous pouvons ensuite reproduire la démarche décrite dans la section 2 pour construire l'arbre de classification. Reprenons notre exemple ci-dessus. Nous voulons maintenant construire des groupes d'individus homogènes du point de vue de leurs caractéristiques financières pour la compagnie d'assurances (sinistres déclarés, primes d'assurances versées). Ce sont les variables précédemment illustratives. Nous voulons construire ces groupes à l'aide des variables caractérisant les assurées (les variables précédemment actives). Le traitement dans le logiciel SPAD produit alors le résultat suivant (Figure 10).

Segment	Effectif	Montant moyen des Sinistres en FB	Valeur-test Sinistre	Prime d'assurance moyenne en FB	Valeur-test Prime
6	139	0	-4,0	9 467	-13,4
7	417	0	-8,3	12 915	-5,2
9	131	33 113	2,5	18 315	13,9
5	410	40 107	8,0	14 505	5,0
8	9	178 863	7,6	14 626	0,7
TOTAL	1106	20 246	–	13725	–

TAB. 4 – *Caractéristiques financières des classes*

L'exemple est relativement simpliste dans la mesure où nous ne disposons ici que de deux variables " à expliquer ". Il permet néanmoins d'illustrer le principe de la construction de l'arbre par d'autres variables que celles dont on mesure l'inertie. Nous distinguons principalement deux familles de classes : les assurés qui n'ont pas de sinistre et qui paient une prime d'assurance faible, nous pourrions à cet égard fusionner les segments [6] et [7] c.à.d. élaguer les deux feuilles pour revenir au segment [2] ; et les assurés qui ont des sinistres, ils paient une prime plus ou moins élevée par rapport à la moyenne, ce sont les segments [5], [8] et [9].

Cette utilisation de la méthode ICT s'apparente aux méthodes prédictives. Elle permet d'expliquer un phénomène multidimensionnel, voire de le prédire, de la même façon que les arbres de décision. Elle est d'autant plus intéressante que les variables à prédire sont peu corrélées dans l'échantillon initial, elle permet de mettre en évidence des sous-groupes où justement ces variables présentent des évolutions concomitantes ou opposées.

La prédiction multidimensionnelle avec les arbres de classification constitue une alternative intéressante aux méthodes usuelles de prédiction ou de scoring. En effet, bien souvent les concepts de valeurs-client, de risque-client, de satisfaction, sont par nature multidimensionnels, il est difficile de les synthétiser sous la bannière d'une seule variable qui résumerait le comportement des clients que l'on voudrait expliquer ou prédire.

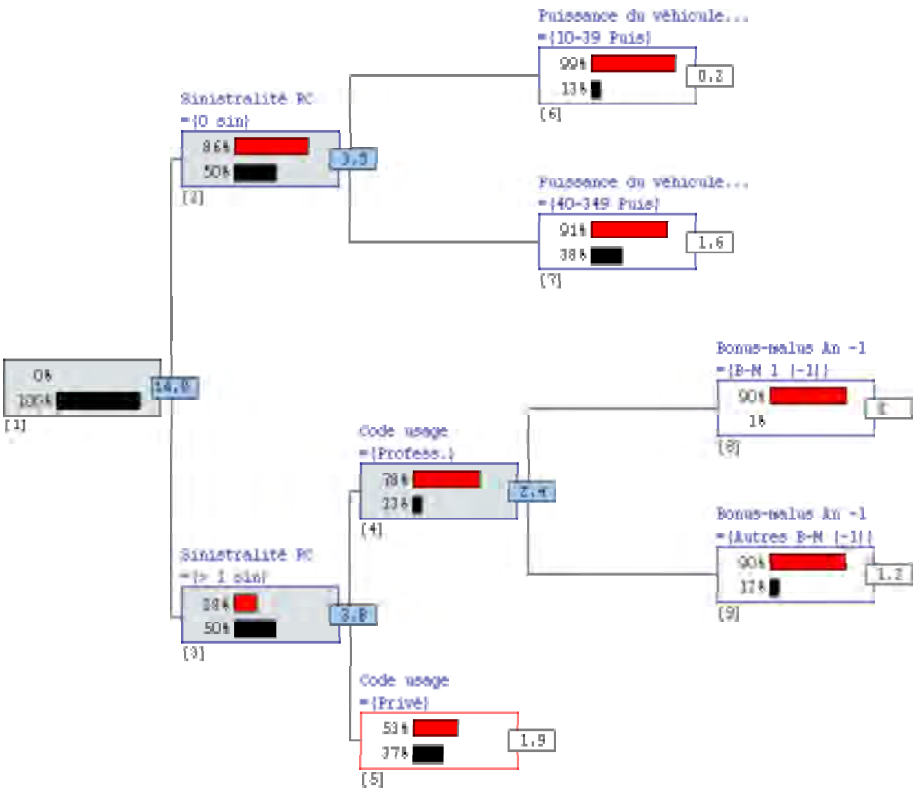


FIG. 9 – *Segmentation financière en 5 classes*

5 Conclusion

Dans le secteur Banque - Assurance, l'interprétation est au moins aussi importante que la performance brute. Pouvoir expliquer un résultat permet de le valider, de le faire comprendre et d'emporter l'adhésion des décideurs. Les arbres de classification répondent parfaitement à cette spécification, la lecture des résultats est immédiate, elle est accessible à tout un chacun, y compris aux personnes totalement étrangères aux techniques de fouilles de données. Le déploiement, l'exploitation industrielle, se résume à produire les règles logiques d'appartenance aux groupes, leur exportation vers les systèmes informatiques est extrêmement simple. Autre aspect important, sa complexité de calcul est bien maîtrisée, elle est identique aux arbres de décision classiques, nous pouvons mettre en oeuvre la technique sur des bases comportant des

centaines de milliers d'observations sur des ordinateurs de bureau. Cette propriété renforce le potentiel en matière d'exploration interactive des données. Si la technique a été essentiellement définie pour l'analyse non supervisée, nous pouvons l'étendre naturellement à la prédiction multi-supervisée : l'idée est de produire des groupes homogènes selon une série de descripteurs, décrivant le comportement de clients face à une série de produits par exemple, à partir d'autres variables, décrivant les caractéristiques des clients. Le rôle dévolu à ce qu'on appelle les " variables illustratives ", couramment utilisées dans l'analyse typologique, est élargi et renforcé. Enfin, les arbres de classification intègrent également une caractéristique qui a déjà largement fait la popularité des arbres de décision dans le domaine de la prédiction depuis une dizaine d'années, elle offre à l'expert du domaine la capacité d'intervenir dans le processus d'exploration des connaissances.

Références

Blockeel H., De Raedt L., Ramon J., (1998). Top-down induction of clustering trees, *Proceedings of the 15th International Conference on Machine learning*, 55-63.

Breiman L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). Classification And Regression Trees. New York : Chapman and Hall.

Chavent M. (1998). A monothetic clustering method, Pattern *Recognition Letters*, 19, 989-996.

Chavent M., Guinot C., Lechevallier Y., Tenenhaus M. (1999). Méthodes divisives de classification et segmentation non supervisée : recherche d'une typologie de la peau humaine saine, *Revue de Statistiques Appliquées*, XLVII (4), 87-99. Meyer F.,

Nakache J.P., Confais J., (2005). Approche pragmatique de la classification : arbres hiérarchiques et partitionnements. Paris :Technip.

Lebart L., Morineau A., Piron M. (1995). Statistique exploratoire multidimensionnelle. Dunod, Paris.

Meyer F., Clerot F., (2006). Arbres de décision multimodes et multi-cibles, *Actes de EGC'2006*, 541-546.

Morineau A. (1984). Note sur la caractérisation statistique d'une classe et les valeurs-tests. *Bull. Techn. du Centre de Statist. et d'Infor. Appl*., 2 , 20-27.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique. Cahiers de l'analyse des données 3, 187-198.

Saporta G. (1990). Probabilités, analyse des données et statistiques. Paris : Technip.

Tuffery, S. (2006). Data Mining et statistique décisionnelle. Paris : Technip.

Volle, M. (1976). Analyse des données. Economica.

Zighed D., Rakotomalala R., (2000). Graphes d'induction : Apprentissage et Data Mining, Hermès.

Summary

We present a new unsupervised classification method, particularly adapted to huge datasets and to the needs of the Insurance, Banking and Retail sectors. This descendent hierarchical clustering method represents the final segmentation as a decision tree where the cluster assignment depends on logical rules based on the variables of the analysis. Thus, the method inherits the properties of the decision trees (interactivity, choice of the cutting variables, splitting / cutting). It confers the method simplicity, easiness of interpretation and operational cluster assignment for assigning a new case to its segment. At last, the method gives the possibility of building the tree using other variables than the ones involved in the global inertia. In this sense, we can consider the method as a generalization of the multi-targets case - many variables to predict simultaneously - of the supervised methods.