

REVUE DE STATISTIQUE APPLIQUÉE

P. BRITO

C. CHABANON

Analyse de données numérique et symbolique appliquée au confort automobile

Revue de statistique appliquée, tome 42, n° 4 (1994), p. 39-59

http://www.numdam.org/item?id=RSA_1994__42_4_39_0

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSE DE DONNÉES NUMÉRIQUE ET SYMBOLIQUE APPLIQUÉE AU CONFORT AUTOMOBILE

P. Brito(1), C. Chabanon(2)

(1) Universidade de Aveiro

Dept. de Matemática 3800 Aveiro Portugal

(1) INRIA-Projet CLOREC

Rocquencourt 78153 Le Chesnay

(2) Renault S.A. Centre technique d'Aubevoye

DE/DESC/0865 27940 Aubevoye France

RÉSUMÉ

Une étude portant sur le confort subjectif en automobile est exposée. L'étude comprend deux parties distinctes : la cartographie interne, qui met en relation l'avis des experts avec les mesures pratiquées sur les véhicules, et la cartographie externe, qui tient compte des impressions des passagers. Dans la première partie, nous utilisons la classification hiérarchique pour trouver une bonne partition de l'ensemble des voitures et la segmentation par arbre binaire pour obtenir les règles d'affectation aux classes. Dans la seconde partie, l'étude de la cartographie externe, nous analysons un tableau portant sur 10 versions, à travers l'analyse des correspondances multiples, la classification hiérarchique et la classification pyramidale symbolique. Dans cette étude, la classification pyramidale symbolique a présenté l'avantage, par rapport à la classification hiérarchique classique, de fournir des classes auxquelles nous pouvons associer une interprétation caractéristique en terme de variables actives, ordonner de façon pertinente les 10 versions et faire apparaître des classes qui n'existent pas dans la hiérarchie et qui nuancent l'interprétation.

Mots-clés : Classification pyramidale symbolique , classification hiérarchique, confort automobile.

SUMMARY

In this paper we present a study on car subjectif comfort. This study consists of two distinct parts : the internal cartography, which relates the experts judgement and the measures made on vehicles, and the external cartography, which takes into account the passengers impressions. In the first, we use hierarchical clustering to determine a good partition of the set of cars, and segmentation by a binary tree to obtain assignment rules. On the other hand, the study of external cartography, where a data table on 10 versions is studied, comprehends multiple correspondence analysis, hierarchical clustering and symbolic pyramidal clustering. In this study, symbolic pyramidal clustering presents the advantage, as concerns classical hierarchical clustering, to yield classes to which a characteristic interpretation, in terms of the

active variables, can be associated, to meaningfully order the 10 versions, and to produce classes which cannot be found in the hierarchy and which allow for a graded interpretation.

Keywords : *Symbolic clustering , hierarchical clustering, vehicle comfort.*

1. Introduction

Le confort d'une automobile est quelque chose de très subjectif que l'entreprise Renault mesure depuis de nombreuses années. Le spectre fréquentiel de la réponse de la voiture aux excitations de la route est découpé en bandes de fréquence pour faciliter l'analyse, la mise au point et la compréhension des phénomènes.

La bande des basses fréquences (BF) détaillée dans cet article occupe la plage 0-2 Hz. C'est elle, par exemple, qui occasionne parfois le «mal de mer» sur route. Elle est perçue physiquement comme un déplacement alors que les fréquences plus élevées le seront comme des ondes de pression (si la fréquence f est supérieure à 50Hz).

L'analyse des données a mis à notre disposition un ensemble d'outils qui permet la synthèse multidimensionnelle et la sélection de paramètres pertinents. La reconnaissance des formes nous a apporté les outils d'aide à la décision. Plus récemment, l'analyse de données symboliques a permis d'identifier des références stables en terme de confort subjectif. Cette forme d'interprétation des résultats semble beaucoup mieux appropriée au mode d'analyse des experts de l'automobile.

2. Méthodologie

L'analyse subjective est pratiquée en deux étapes : une première qui construit la cartographie interne des préférences (cf. figure 1) c'est-à-dire celle qui met en relation l'avis subjectif des experts du confort avec les mesures pratiquées sur les véhicules ;

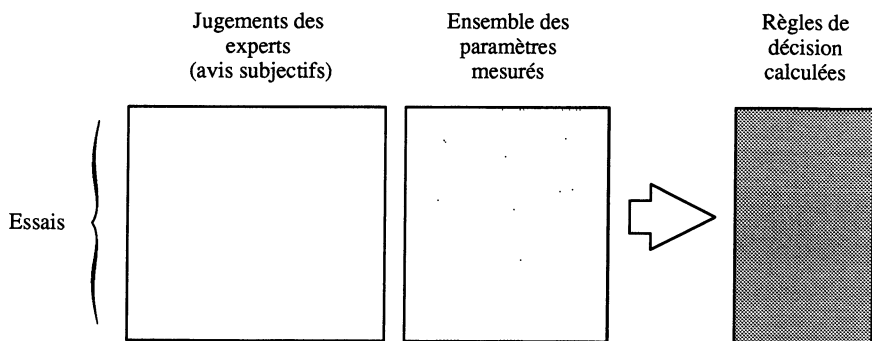


FIGURE 1
Cartographie interne des préférences

Dans la figure 1, chaque ligne du tableau représente un essai sur véhicule et chaque colonne respectivement le jugement des experts et les paramètres mesurés

sur route. L'analyse de ces entrées fournit en sortie les règles de décision. Les avis subjectifs des experts valident les différents résultats d'analyse et les règles calculées. Ils n'interviennent pas dans les calculs. Une seconde analyse établit la cartographie externe des préférences (*cf.* figure 2), c'est-à-dire celle des passagers (clients).

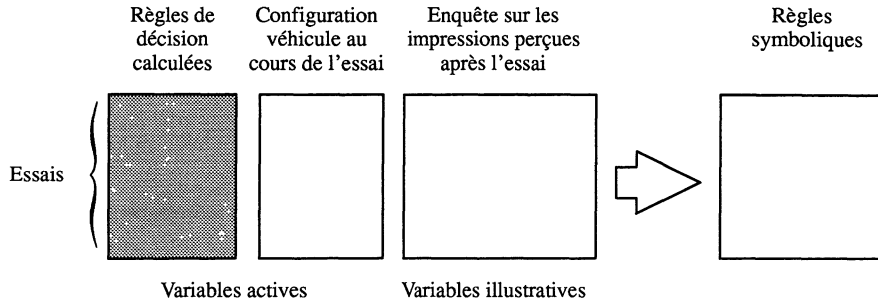


FIGURE 2
Cartographie externe des préférences

Dans la figure 2, chaque ligne du tableau représente un essai, caractérisé par les décisions calculées à l'aide de la segmentation au cours de la première étape, les paramètres de configuration du véhicule (c'est-à-dire les lois d'amortissement de l'avant et de l'arrière) et les réponses sur les impressions perçues au cours de l'essai. En sortie, l'analyse fournit les règles symboliques associant point de vue du client, point de vue de l'expert via les règles calculées et configuration véhicule.

3. Analyse de la cartographie interne des préférences

3.1. Les données

Cette analyse a été effectuée sur un ensemble de 50 véhicules Renault et autres. Les voitures étaient équipées d'accéléromètres verticaux qui procurent les déplacements par double intégration. Ces déplacements sont traités de façon à obtenir les termes caractéristiques du pompage, du tangage et du copiage à partir des mesures réalisées sur une route (*cf.* figure 3). Le pompage est un mouvement vertical généré par le profil longitudinal de la route. Le tangage est un mouvement d'oscillation de l'avant sur l'arrière créé par le même profil. Le copiage (ou roulis) est un mouvement d'oscillation latéral créé par le profil transversal de la route.

Chaque véhicule est étiqueté par l'avis subjectif de l'expert. Le tableau de données croise les véhicules avec les paramètres de déplacements issus de l'analyse spectrale. Chaque déplacement est caractérisé par son amplitude maxi a et la fréquence associée f . Les valeurs d'amplitude maxi sont moyennées sur toute la durée de l'essai. Dix paramètres sont ainsi calculés à partir des quatre accélérations.

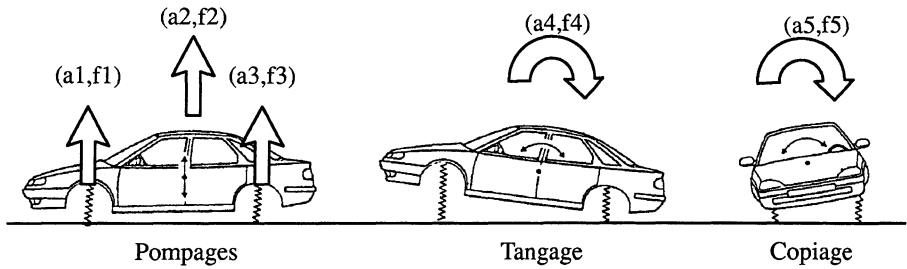


FIGURE 3 : Paramètres extraits

Exemple du tableau de données :

Véhicule	f1	a1	f2	a2	f3	a3	f4	a4	f5	a5
Safrane_1	1.3	7	1.3	6	2	6	1.7	4	1.8	6
R25_1	1.5	9	1.5	6	1	14	1.7	5	1.8	2

Les véhicules mesurés couvrent tous les segments de la gamme Renault (Safrane, R21, Clio, Twingo), Volvo ainsi que quelques étrangères de diverses provenances (cf. figure 4).

La figure 4 montre en exemple l'ensemble des véhicules mesurés en pompage. La fréquence enregistrée f_2 est représentée en abscisse et l'amplitude associée a_2 en ordonnée.

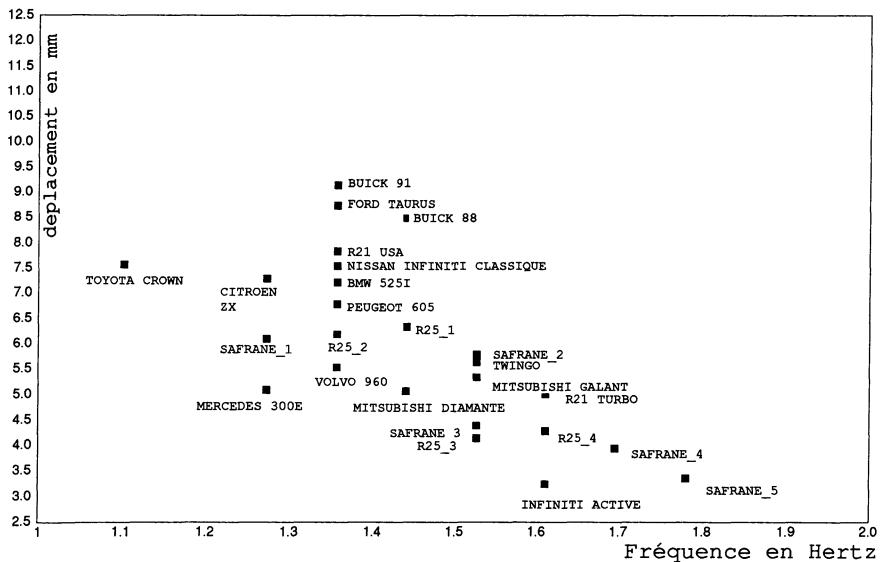


FIGURE 4 : Exemple sur le pompage

Les outils mis en œuvre sont la classification hiérarchique basée sur l'indice de l'accroissement de l'inertie (ou indice de Ward) et la discrimination non paramétrique (ou segmentation par arbre binaire). Pour cette dernière, l'ensemble des données est découpé en un ensemble d'apprentissage (50%) qui permet le calcul des règles et un ensemble test (50%) qui permet de mesurer l'adéquation avec le jugement des experts.

3.2. Recherche d'une partition à l'aide du modèle hiérarchique

Rappelons brièvement que la classification ascendante hiérarchique (C.A.H.) consiste à construire itérativement une suite de partitions emboîtées de moins en moins fines, dont la représentation graphique est l'arbre hiérarchique (ou dendrogramme) (Ward (1963)). (cf. figure 5)

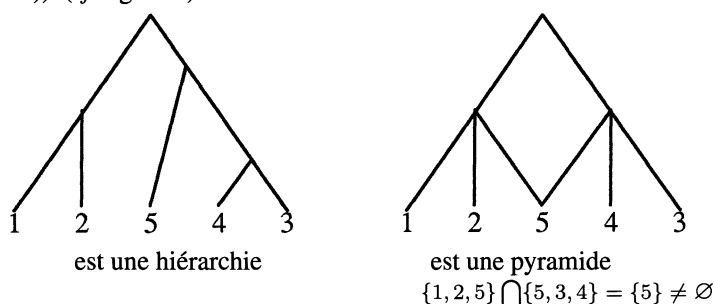


FIGURE 5

Exemple d'une pyramide et d'une hiérarchie

Cette méthode est bien adaptée pour des ensembles de données de petite taille qui génère des représentations de faible encombrement.

Le modèle hiérarchique a été étendu au modèle pyramidal (Diday (1984, 1986)) qui construit une suite de recouvrements emboîtés au lieu de partitions emboîtées. Le modèle pyramidal est repris au paragraphe 4.5.2.1.

Nous avons appliqué le modèle hiérarchique sur le tableau des 50 véhicules décrits par les 10 variables définissant le pompage, le tangage et le copiage.

3.3. Résultats sur les données du confort

L'examen de l'arbre par les experts en confort a été un moyen rapide de déterminer le nombre de classes raisonnables. De cette confrontation entre le savoir-faire dans le métier du confort et la représentation de la C.A.H. nous avons choisi de sélectionner un découpage en quatre classes homogènes. (cf. figure 6). Ces classes sont ensuite considérées comme des références stables en terme de confort. L'ensemble

des classes couvre le confort des véhicules actuels depuis ceux à caractère sportif : la classe 1 à ceux confortables à l'extrême la classe 4.

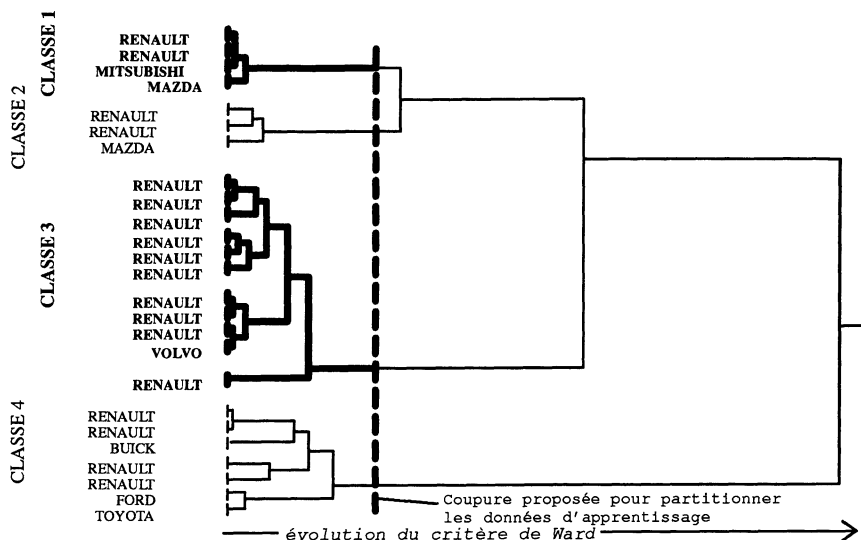


FIGURE 6

Classification hiérarchique sur les données du confort

Les classes 1 à 4 partitionnent les véhicules en tendance sportive, tendance très confortable ou intermédiaires.

3.4. Recherche d'un critère de décision

3.4.1. Le modèle de la segmentation non paramétrique

On désigne par segmentation la famille des méthodes de discrimination non paramétrique qui construisent un arbre de décision généralement binaire (Friedman (1977)). A chaque nœud non terminal de l'arbre, est associée un test qui permet la descente à droite ou à gauche d'un individu dans l'arbre (cf. figure 7). Le principal avantage des arbres de décision réside dans la lisibilité des règles de décision que l'on obtient en définissant les outils suivants :

- un ensemble de questions binaires;
- un critère d'évaluation des coupures engendrées par les questions;
- une règle d'arrêt de construction de l'arbre;
- une règle d'affectation des segments terminaux;
- une règle d'affectation d'un nouvel individu.

Dans le cas présent, on a utilisé comme critère la distance de Kolmogorov-Smirnov (K.S.). Celui-ci a été défini pour la segmentation en deux classes, puis étendu à un nombre quelconque de classes a priori (Celeux, Lechevallier (1982)).

Comme règle d'arrêt de la construction de l'arbre, on a imposé que les nœuds terminaux soient homogènes, c'est-à-dire, qu'ils ne contiennent que des éléments d'une même classe (de la classification obtenue au §3.3).

On résume ci-dessous le critère dans le cas de deux classes P_1 et P_2 .

Pour chaque variable y_j , soient f_1^j et f_2^j les densités de probabilité des groupes P_1 et P_2 et F_1^j et F_2^j les fonctions de répartition correspondantes. Soient π_1 (respectivement π_2) la probabilité *a priori* d'appartenir à P_1 (respectivement P_2) et l_1 et l_2 les coûts de mauvaise classification. On suppose de plus que $\pi * l_1 = \pi * l_2$. Le principe de la méthode repose sur la recherche d'une coupure qui correspond à la minimisation du risque de Bayes $R(z)$ d'une mauvaise classification d'un individu.

On dira que P_1 est la population inférieure si

$$\sup_{z,j} \left(F_1^j(z) - F_2^j(z) \right) > \sup_{z,j} \left(F_2^j(z) - F_1^j(z) \right)$$

sinon on dira que P_2 est la population inférieure.

Soit z la valeur qui définit la coupure; si pour cette variable, l'individu à affecter présente une valeur inférieure à z , on l'affecte à la population inférieure, sinon on l'affecte à la population supérieure.

Si P_1 est la population inférieure, le risque de mauvais classement est

$$R(z) = \pi_1 * l_1 * (1 - F_1^j(z)) + \pi_2 * l_2 * F_2^j(z) = \pi_1 * l_1 * (1 + F_2^j(z) - F_1^j(z)) \quad (1)$$

sinon, ce risque est donné par :

$$R(z) = \pi_2 * l_2 * (1 - F_2^j(z)) + \pi_1 * l_1 * F_1^j(z) = \pi_1 * l_1 * (1 + F_1^j(z) - F_2^j(z)) \quad (2)$$

dans le cas (1), on a $\min_z R(z) \Leftrightarrow \max_z \left(F_1^j(z) - F_2^j(z) \right)$

alors que dans le cas (2), on a $\min_z R(z) \Leftrightarrow \max_z \left(F_2^j(z) - F_1^j(z) \right)$

Comme on ne connaît pas, *a priori*, la population inférieure et celle supérieure, le point qui minimise le risque de Bayes est le point c tel que :

$$\sup_z D(z) = \sup_{z,j} |F_1^j(z) - F_2^j(z)|$$

Minimiser le risque de Bayes revient à maximiser la distance de Kolmogorov-Smirnov $D(z)$ qui mesure de séparabilité de deux fonctions de répartition. En pratique, F_1^j et F_2^j ne sont pas connues, et il faudra les estimer par les fonctions de répartition empiriques correspondantes.

La règle de classification d'un nouvel individu est alors simple : partant du sommet de l'arbre, on effectue à chaque nœud le test qui permet de choisir la branche

à suivre, le parcours se termine lorsqu'on aboutit à un nœud terminal (feuille) de l'arbre; l'individu est alors affecté à la classe correspondante.

3.5. Résultats obtenus sur les données confort

L'arbre construit sur nos données est indiqué sur la figure 7 ci-dessous :

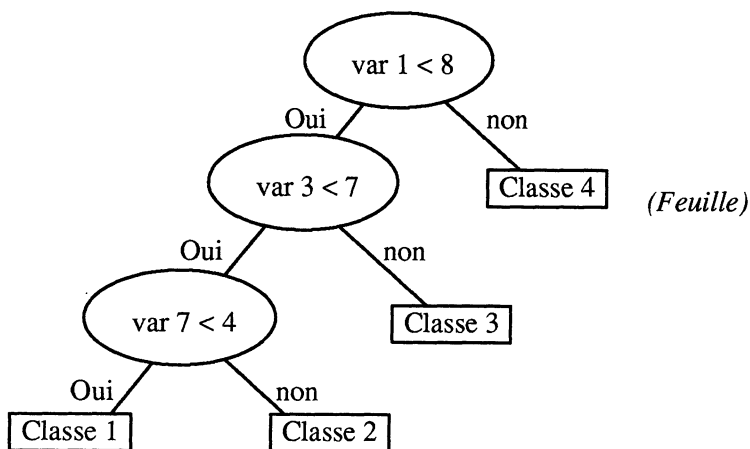


FIGURE 7

Arbre de discrimination calculé à partir des données du confort

L'arbre obtenu est en accord avec la hiérarchie construite avec l'indice de Ward. On constate que la classe 4 est celle qui se sépare en premier lieu dans l'arbre; sur la hiérarchie, c'est celle qui s'agrège le plus tard (donc, à un niveau plus élevé de l'indice) au reste de l'ensemble. Les classes 1 et 2, qui sont les dernières à être séparées dans l'arbre, sont aussi celles qui s'agrègent à un plus bas niveau sur la hiérarchie.

La sélection de 3 paramètres (var 1, var 3 et var 7) parmi ceux présents permet l'établissement de règles de décisions d'une grande lisibilité. Par exemple la règle désignant la classe 3 est la suivante :

Si $(\text{var } 1 < 8)$ et $(\text{var } 3 > 7)$ alors l'individu appartient à la classe 3

Le taux apparent (respectivement réel) de bon classement obtenu sur l'ensemble d'apprentissage (respectivement sur l'ensemble test) a été de 100% (respectivement 96%).

L'analyse de stabilité de l'arbre de décision ci-dessus a été réalisée par une technique de bootstrap sur l'ensemble des 50 véhicules. Cette méthode consiste à tirer au hasard plusieurs fois de suite un sous-échantillon de la population avec la contrainte d'avoir un représentant de chaque classe pour valider le tirage. Environ vingt tirages ont ainsi été conservés. La variabilité des seuils de décision observée sur l'ensemble des simulations permet de calculer une moyenne et un écart-type

bootstrap. Cet écart-type, comparable à une dispersion, permet de moduler les seuils proposés en discrimination.

Pour juger le degré d'appartenance d'un essai à une classe, nous avons arbitrairement écarté les seuils de la discrimination par rapport au seuil moyen bootstrap de $\pm 1\sigma$ et de $\pm 2\sigma$. Ainsi, on crée l'équivalent des limites des cartes de contrôle. Nous considérons qu'au delà de $\pm 2\sigma$ du seuil moyen, il n'y a pas d'ambiguïté de décision. Entre $\pm 2\sigma$ et $\pm 1\sigma$ du seuil moyen, la décision est limite dans la classe; en dessous de $\pm 1\sigma$ du seuil moyen, il y a ambiguïté caractérisée éventuellement par des jugements différents entre les experts. Au cours de la phase de mise au point, ces informations nous ont permis d'évaluer assez précisément les points à améliorer pour atteindre la prestation désirée.

4. Analyse de la cartographie externe des préférences

4.1. Les données

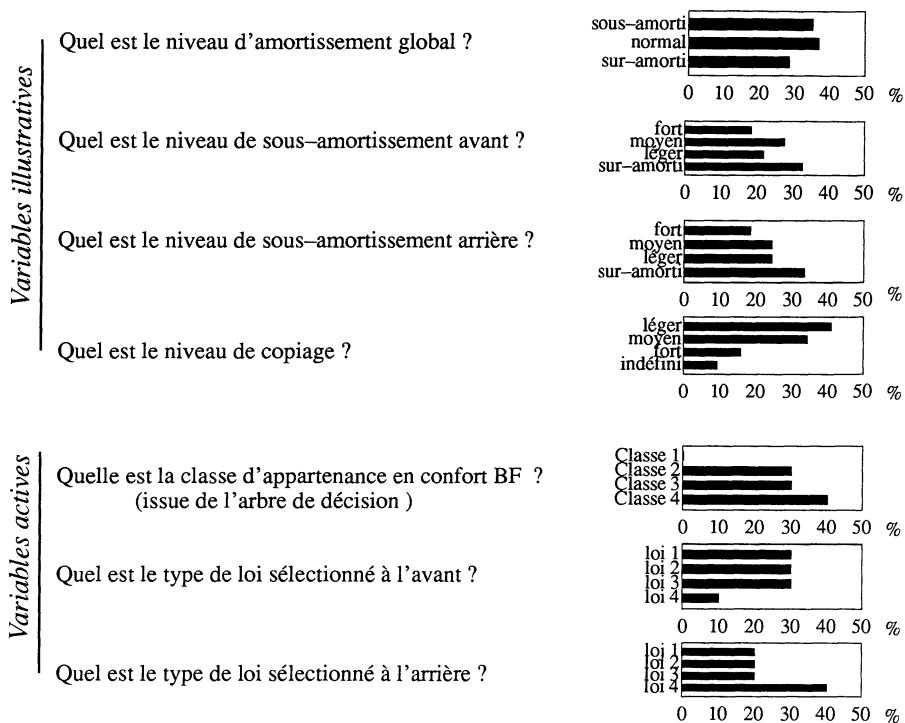
Cette analyse s'intéresse à la perception du confort BF à «chaud» après une conduite sur route; elle est réalisée à l'aide d'un questionnaire de type fermé. Au cours de l'essai, les véhicules subissent du pompage, du tangage et du copiage (cf. figure 3 pour la définition). Douze personnes, hors experts confort, ont été interrogées. La campagne d'essais a été menée sur une Renault 25 qui comporte quatre lois programmables sur chaque essieu : loi 1, loi 2, loi 3 et loi 4. Dix versions parmi les seize possibles ont été proposées aux enquêtés pour apprécier le confort. Chaque personne a essayé dans un ordre pré-établi chaque version de la plus caricaturale à la plus fine, ultime mélange d'un bon confort. La classe 1 est absente des tests.

Le questionnaire proposait quatre questions, possédant chacune quatre modalités équiréparties sur l'échelle des valeurs subjectives et ordonnées de manière naturelle du très fort sous-amortissement («quasi- absence de l'amortisseur») à un sur-amortissement («quasi- absence de suspension»).

Les règles calculées au paragraphe précédent complètent les variables actives sur la configuration essayée en désignant la classe d'appartenance en confort.

Trois outils ont été mis en œuvre pour analyser ce tableau : l'analyse des correspondances multiples (A.C.M.) (L. Lebart (1977)) complétée par les analyses élémentaires, la C.A.H. et enfin la classification pyramidale symbolique exposée en détail dans la suite de l'article.

4.2. Histogrammes



La première question montre que les véhicules présentés couvrent équitablement les trois niveaux d'amortissement. Aux deux questions suivantes, les personnes ont répondu, en général, sur-amorti de l'avant et de l'arrière. Pour le copiage, ils ont souvent répondu léger et moyen, avec toutefois un niveau de non réponse élevé (modalité «indéfinissable»).

4.3. Résultats issus de L'A.C.M

L'analyse est effectuée sur le tableau des variables illustratives ci-dessous :

Libellé des variables		Versions présentées										Total
		Vers 1	Vers 2	Vers 3	Vers 4	Vers 5	Vers 7	Vers 9	Vers 10	Vers 13	Vers 14	
Niveau d'amortissement global?	sous-amorti	0	1	11	9	0	4	4	2	10	1	42
	normal	0	7	1	2	6	8	8	10	2	0	44
	sur-armorti	12	4	0	1	6	0	0	0	0	11	34
Niveau du sous amortissement avant?	fort	0	0	7	7	0	0	1	0	7	0	22
	moyen	0	3	5	4	1	1	4	4	5	6	33
	léger	1	5	0	1	3	4	4	3	0	5	26
	sur-armorti	11	4	0	0	8	7	3	5	0	1	39
Niveau du sous amortissement arrière?	fort	0	0	5	3	0	5	3	1	5	0	22
	moyen	0	1	4	3	3	4	5	6	3	0	29
	léger	3	6	2	4	2	2	3	2	3	2	29
	sur-armorti	9	5	1	2	7	1	1	3	1	10	40
Niveau de copiage?	léger	3	4	5	3	5	5	6	7	9	2	49
	moyen	4	3	5	2	3	6	6	4	3	5	41
	fort	5	3	1	2	3	1	0	0	0	4	19
	indéfinissable	0	2	1	5	1	0	0	1	0	1	11

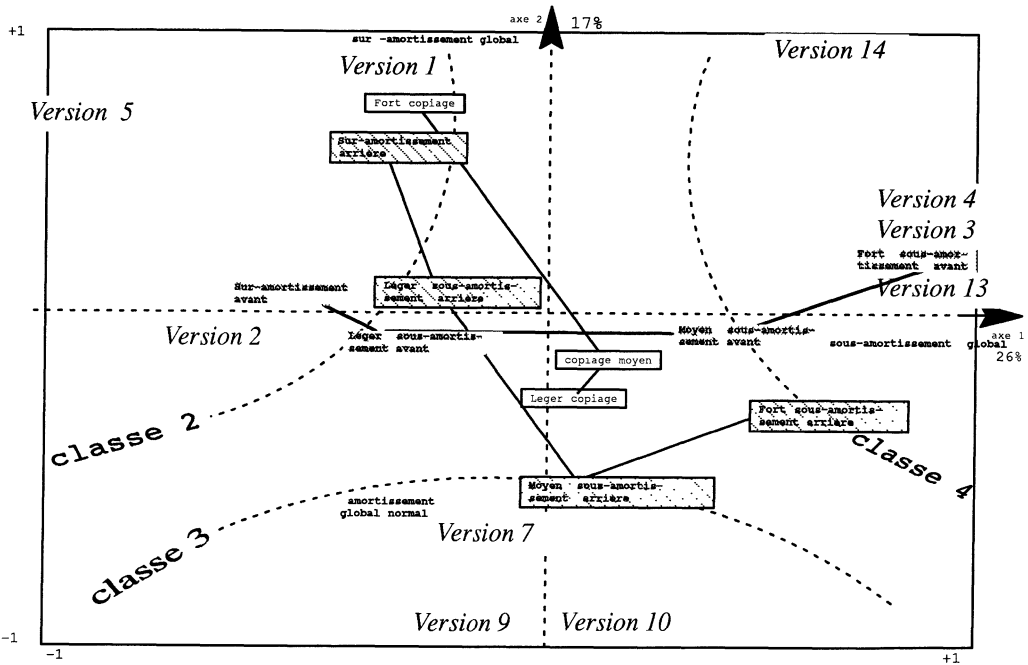


FIGURE 8
Plan de projection 1-2 de l'A.C.M.

L'A.C.M. fait ressortir sur ce premier plan factoriel l'évolution du niveau d'amortissement avant principalement expliqué par l'axe 1 et l'évolution du niveau d'amortissement arrière et du copiage principalement sur l'axe 2. Une évolution sur l'essieu avant ou sur l'essieu arrière n'aura pas les mêmes effets sur le confort subjectif des passagers.

La version 1 (loi 1 Av/ loi 1 Ar) est analysée globalement sur-amortie par l'ensemble des enquêtés. De même, la version 7 (loi 1 Av/ loi 4 Ar) semble avoir un amortissement global normal. Enfin les versions 3 (loi 3 Av/ loi 3 Ar) et 13 (loi 3 Av/ loi 4 Ar) sont jugées globalement sous- amorties.

Subjectivement, les interrogés ont trouvé la version 7 comme un bon compromis d'amortissement global, qu'ils définissent par un léger sous-amortissement de l'avant et un moyen sous-amortissement de l'arrière. Au delà, on tend sur des voitures extrêmes (raides ou moles).

Ils considèrent que le copiage s'atténue au fur et à mesure que le sous-amortissement arrière augmente. Cette sensation a peu de relation avec le sous-amortissement avant.

De même, ils considèrent que toutes les versions de la classe 4 de la classification BF (versions 3,4,13,14) se caractérisent par un fort sous-amortissement avant. Par contre, le fort sous-amortissement arrière est perçu à la fois dans les classes 3 et 4.

L'analyse de la stabilité du premier plan factoriel a été réalisée à partir de 20 sous-échantillons de 9 à 10 enquêtes parmi les 12 disponibles. Elle confirme la position des modalités dans le plan.

4.4. Résultats de la classification hiérarchique

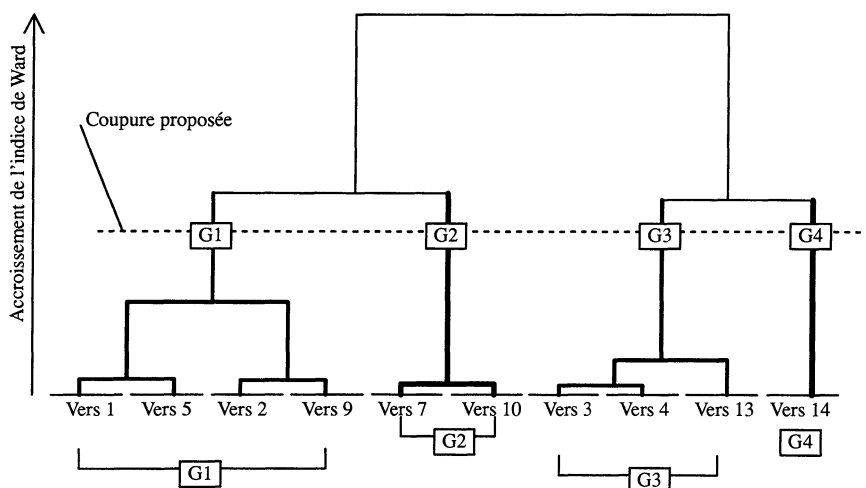


FIGURE 9

Hierarchie sur variables actives de la représentation externe

La construction de la hiérarchie suivant l'indice de Ward s'effectue sur les variables actives : la classification BF (issue de l'arbre de décision) et les lois avant et arrière.

La coupure a été établie à partir d'une interprétation visuelle des regroupements proposés aux différents niveaux de la représentation figure 9. Elle isole la version 14 qui était une voiture très caricaturale.

La représentation figure 9 ci-dessus échelonne les versions du sur-amortissement à gauche sur le graphique au sous amortissement à droite et un bon compromis au centre du graphique. Comme le critère permet des permutations des individus (les versions) les uns par rapport aux autres, l'échelonnement des versions de gauche à droite n'est pas forcément unique. Par exemple, si Vers 2 et Vers 9 sont permutées, le critère est inchangé par contre l'interprétation subjective est contrariée.

De plus, le critère d'agrégation n'assure pas qu'un élément d'une autre classe, ne vérifie pas les propriétés associées à une classe donnée. Il y a absence de recouvrement des groupes G_1, G_2, \dots ce qui gêne considérablement l'expert dans les choix d'interprétation. Ceci nous a conduit à employer la classification pyramidale symbolique.

4.5. Classification pyramidale symbolique (C.P.S)

Le but de l'approche proposée est celui d'obtenir une structure d'héritage sur les données, dont les classes sont associées à un ensemble de propriétés sur les valeurs prises par les variables. Pour obtenir une telle structure, la méthode de classification pyramidale symbolique est utilisée. Le modèle pyramidal constitue une extension de la classification hiérarchique, présentant des recouvrements emboîtés au lieu de partitions, de plus chaque classe d'une pyramide doit être un intervalle d'un ordre total sur les individus classifiés. La C.P.S., qui relève des méthodes dites «de classification conceptuelle», utilise le modèle pyramidal pour obtenir une structure où chaque classe est associée à une conjonction de propriétés portant sur les valeurs des variables descriptives.

4.5.1. Le langage de représentation

Le langage de représentation utilisé est basé fondamentalement sur la notion d'objet symbolique (Diday (1987, 1991), Brito (1991)). Un objet symbolique peut être assimilé à une conjonction de propriétés portant sur les valeurs prises par les variables : on parle alors de définition en intention de l'objet symbolique. Ces propriétés peuvent porter soit sur les individus pris séparément (cas des objets assertion) soit sur les relations entre individus (cas des objets horde et de synthèse). Un autre moyen de définir un objet symbolique est l'énumération exhaustive de tous les individus (ou tous les k -uplets d'individus) qui satisfont la conjonction de propriétés : on parle alors de définition en extension. Par la suite, nous n'envisagerons plus que le cas simple des objets assertion.

Soit Ω l'ensemble des objets observés, $\Omega = \{w_1, \dots, w_n\} \subseteq \prod$, où \prod est la population étudiée (ensemble des observables). Chaque élément de \prod est décrit par

p variables $y_i : \prod \rightarrow O_i$. Soit $y = (y_1, \dots, y_p)$. Nous avons alors :

$$Y : \prod \rightarrow O_1 \times \dots \times O_p$$

$$w \rightarrow (y_1(w), \dots, y_p(w))$$

Le point $(y_1(w), \dots, y_p(w)) \in \mathcal{O} = O_1 \times \dots \times O_p$ est appelé *description de w* , et $\mathcal{O} = O_1 \times \dots \times O_p$ l'*espace des descriptions*.

Un *objet assertion* est formé d'une conjonction d'événements élémentaires de la forme $[y_i = V_i], y_i \in \{y_1, \dots, y_p\}, V_i \subseteq O_i$, sur un ensemble $Y_1 \subseteq \{y_1, \dots, y_p\}$. Sans perte de généralité, on peut considérer $Y_1 = \{y_1, \dots, y_q\}, q \leq p$. Un objet assertion « a », noté :

$$a = [y_1 = V_1] \wedge \dots \wedge [y_q = V_q], V_i \subseteq O_i, i = 1, \dots, q$$

exprime la condition «la variable y_1 prend des valeurs dans V_1 et ... et la variable y_p prend des valeurs dans V_q ».

Pour chaque variable, on donne un ensemble de valeurs qui représente sa *variabilité* dans la classe considérée. Remarquons que ce modèle contient le modèle tabulaire usuel comme cas particulier, en associant à chaque individu-ligne du tableau la fonction indicatrice du point qu'il représente.

On associe à un objet assertion deux applications :

D'une part $a : \mathcal{O} \rightarrow \{\text{vrai}, \text{faux}\}$ telle que $a(x_1, \dots, x_p) = \text{vrai}$ ssi $\forall i = 1, \dots, q, x_i \in V_i$, a est donc la fonction de caractéristique de l'ensemble $V_1 \times \dots \times V_q \times O_{q+1} \times \dots \times O_p$; et d'autre part, $a^* : \Omega \rightarrow \{\text{vrai}, \text{faux}\}$ telle que $a^*(w) = a(Y(w))$: $a^*(w) = \text{vrai}$ ssi sa description appartient à l'ensemble $V_1 \times \dots \times V_q \times O_{q+1} \times \dots \times O_p$. On définit l'*extension* de a (sur Ω) par :

$$\text{ext}_\Omega a = \{w \in \Omega : a^*(w) = \text{vrai}\} = a^{*-1}(\text{vrai})$$

et son *extension virtuelle* (sur \mathcal{O}),

$$\text{ext}_\mathcal{O} a = \{x \in \mathcal{O} / a(x) = \text{vrai}\} = a^{-1}(\text{vrai}).$$

La notion d'extension virtuelle permet de munir l'espace des objets assertion d'une relation d'ordre partielle, qui est simplement la relation d'inclusion dans l'espace de description : un objet a est *plus général* qu'un objet assertion b (b est *plus spécifique* que a) si $\text{ext}_\mathcal{O} b \subseteq \text{ext}_\mathcal{O} a$. Cette relation d'ordre, appelée aussi «spécialisation- généralisation» est présente dans les travaux de plusieurs auteurs sous des formes équivalentes (Mitchell (1982), Gascuel (1989), etc.).

Des opérations d'union et d'intersection ont été définies, permettant de généraliser ou de spécifier les propriétés explicitées par deux objets assertion. Si

$$a_1 = [y_1 = V_1] \wedge \dots \wedge [y_p = V_p], a_2 = [y_1 = W_1] \wedge \dots \wedge [y_p = W_p]$$

(où l'on peut avoir $V_i = O_i$ et/ou $W_i = O_i$ pour quelque i) alors :

$$a_1 \cup a_2 = [y_1 = V_1 \cup W_1] \wedge \dots \wedge [y_p = V_p \cup W_p] \text{ et}$$

$$a_1 \cap a_2 = [y_1 = V_1 \cap W_1] \wedge \dots \wedge [y_p = V_p \cap W_p]$$

On verra facilement que $a_1 \cup a_2$ est l'objet assertion le plus spécifique qui est plus général que a et b ; de même, $a_1 \cap a_2$ est l'objet assertion le plus général qui est plus spécifique que a et b . Il est aussi facile de démontrer que $\text{ext}_\Omega(a_1 \cup a_2) \supseteq (\text{ext}_\Omega a_1 \cup \text{ext}_\Omega a_2)$, mais $\text{ext}_\Omega(a_1 \cap a_2) = (\text{ext}_\Omega a_1 \cap \text{ext}_\Omega a_2)$.

Complétude

Dans le cadre d'une méthode de classification, la notion d'objet symbolique, associant à une classe d'individus une représentation en intention, permet d'obtenir une interprétation des classes en termes des variables descriptives. Une classe donnée d'individus peut être représentée par différents objets assertion : on peut tenir compte ou non de toutes les variables, considérer un objet assertion plus ou moins général, etc. La notion de complétude (Diday (1987), Brito (1991)) permet d'associer à une classe un objet assertion – dit complet – qui exprime la dualité intention-extension. Un objet assertion est dit complet s'il explicite toutes les propriétés de son extension et s'il est minimal (pour l'ordre introduit) à vérifier cette condition. Ce choix correspond à chercher des descriptions complètes tout en évitant des sur-généralisations. La notion de complétude a été formalisée dans le cadre de la théorie des Correspondances de Galois (Brito (1991)), généralisant l'approche de Wille (1982) déjà proposée par Barbut et Monjardet (1970).

4.5.2. La méthode de classification

4.5.2.1 Le modèle pyramidal

Le modèle de la représentation pyramidale (Diday (1984, 1986); Bertrand (1986); Bertrand, Diday (1990)) généralise celui des arbres hiérarchiques en présentant des recouvrements emboîtés au lieu de partitions emboîtées.

Soit Ω l'ensemble (fini) des objets à classer. Une pyramide est une famille P de sous-ensembles non vides de Ω (appelés aussi «paliers»), telle que

- 1) $\Omega \in P$
- 2) $\forall w \in \Omega, \{w\} \in P$
- 3) P est stable pour l'opération d'intersection, c'est-à-dire, $\forall p, p' \in P, p \cap p' = \emptyset$ ou $p \cap p' \in P$;
- 4) il existe au moins un ordre total sur Ω tel que chaque $p \in P$ est un intervalle de cet ordre. Un ordre satisfaisant 4) est dit *compatible* avec P .

Généralisant le modèle hiérarchique en permettant la présence de classes non-disjointes, la classification pyramidale produit une structure beaucoup plus simple

que la structure en treillis puisque chaque classe est un intervalle d'un ordre total Θ (en conséquence, les pyramides ne présentent pas de croisement au niveau de la représentation graphique).

Comme pour les hiérarchies, une pyramide induit un indice de dissimilarité sur l'ensemble Ω classifié. Dans le cas pyramidal, cet indice, appelé indice pyramidal (ou encore dissimilarité de Robinson) vérifie les propriétés suivantes :

- (i) $d(x, y) = 0 \Rightarrow x = y$
- (ii) Il existe un ordre Θ tel que $x\Theta y\Theta z \Rightarrow d(x, y) \geq \max\{d(x, y), d(y, z)\}$

Le premier algorithme de classification ascendante pyramidale a été proposé par Diday (1984), par analogie à la méthode de classification ascendante hiérarchique. Bertrand (1986) expose un algorithme qui vérifie les principes énoncés dans Diday (1984) et où les choix à effectuer lors de l'implémentation informatique sont précisés. Un deuxième algorithme, que l'on démontre être équivalent au premier, est aussi exposé, ayant l'avantage de réduire la complexité de l'étape de recherche du couple de paliers à agréger.

4.5.2.2. Classification pyramidale symbolique

La méthode de classification pyramidale symbolique (Brito (1991)) construit une pyramide par un algorithme ascendant. Les critères de formation des classes sont ici de type logique, basés sur les notions d'intention/extension, ce qui conduit à une structure de classes interprétées.

La méthode construit une pyramide utilisant un algorithme de type ascendant. Chaque palier formé sera représenté par un objet assertion. Soit p un palier candidat à être formé, par agrégation de p_1, \dots, p_m , représentés, respectivement, par a_1, \dots, a_m . On associe alors à p l'objet assertion $a = a_1 \cup \dots \cup a_m$ (remarquons que $a_1 \cup \dots \cup a_m$ est l'objet assertion le plus spécifique qui généralise a_1, \dots, a_m). On démontre dans Brito (1991) que l'objet assertion ainsi défini est complet. Le critère principal qui guide le choix de l'agrégation à effectuer est la convergence intension-extension : le palier $p = p_1 \cup \dots \cup p_m$ est formé si $\text{ext}_{\Omega} a = p_1 \cup \dots \cup p_m = p$. Ainsi, seules les classes admettant une description conjonctive sont formées. Les paliers dont les objets assertion associés sont plus spécifiques sont d'abord créés. L'ordre de «spécialisation-généralisation» étant partiel, un indice de généralité d'un objet assertion, compatible avec cet ordre a été défini. La pyramide obtenue en fin d'algorithme présente alors une structure d'héritage, en d'autres termes, chaque palier hérite des propriétés possédées par ses prédécesseurs.

En pratique, l'algorithme commence par chercher la meilleure agrégation binaire (c'est-à-dire celle qui agrège deux paliers); s'il n'y a pas d'agrégation binaire vérifiant les conditions imposées, l'algorithme cherchera alors le couple de paliers agrégeables qui, par union avec un palier voisin, permet de former un nouveau palier vérifiant ces conditions (on effectue alors une agrégation non-binaire).

4.5.3 Résultats sur les données du confort

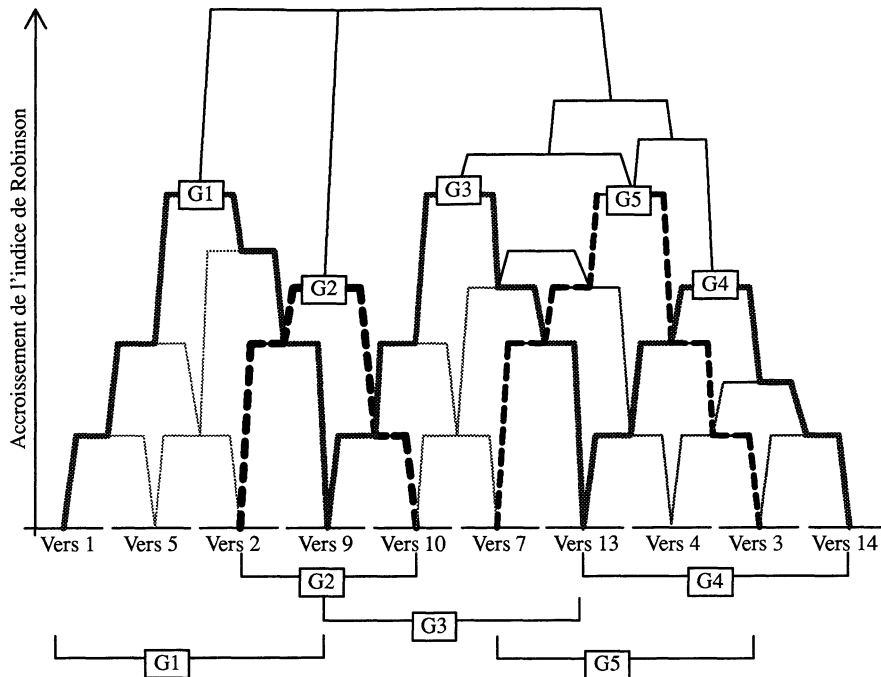


FIGURE 10

Pyramide sur les variables actives de la représentation externe

La méthode de classification pyramidale symbolique a été appliquée à l'ensemble des 10 versions, décrites par les 3 variables actives (la classification BF et les lois avant arrière) du tableau des préférences externes figure 2.

La figure 10 montre la pyramide obtenue, ainsi que les classes considérées intéressantes.

Du sous-amortissement (à droite) au sur-amortissement (à gauche), les versions s'échelonnent et s'agrègent dans plusieurs groupes distincts (G1,...,G5). L'emboîtement des groupes présente ainsi les *nuances d'appréciations* : Par exemple le groupe G5 recouvre des jugements de G3 et de G4.

A partir des paliers de la pyramide et de la structure d'héritage, les variables actives et illustratives nous fournissent l'ensemble des règles d'interprétation suivantes :

Groupe 1 : versions 1,2,5,9

$G1 = [classe = classe\ 2, classe\ 3] \wedge [loi\text{-}av = loi\ 1, loi\ 2] \wedge [loi\text{-}ar = pas\ de\ loi\ 4] \wedge [am\text{-}global = normal, sur\text{-}armorti] \wedge [sous\text{-}a\text{-}av = léger, sur\text{-}amorti] \wedge [sous\text{-}a\text{-}ar = léger, sur\text{-}amorti] \wedge [copiage = n'est\ pas\ indéfinissable]$

Ce groupe agrège des versions avec une tendance au sur-amortissement de l'avant (variable sous-a-av) et de l'arrière (variable sous-a-ar). Certains enquêtés

jugent globalement ce type d'amortissement normal (variable am-global). Le copiage n'est pas clairement identifié.

Groupe 2 : version 2,9,10

$$G2 = [classe = classe\ 2, classe\ 3] \wedge [loi\text{-}av = loi\ 2] \wedge [loi\text{-}ar = pas\ de\ loi\ 1] \wedge [am\text{-}global = normal] \wedge [sous\text{-}a\text{-}av = pas\ fort] \wedge [sous\text{-}a\text{-}ar = pas\ fort] \wedge [copiage = léger, moyen]$$

Ce groupe agrège des versions jugées globalement normales qui ne sont pas fortement sous-amorties de l'avant et de l'arrière. Le copiage n'est pas fort.

Groupe 3 : versions 7,9,10,13

$$G3 = [classe = classe\ 3, classe\ 4] \wedge [loi\text{-}av = pas\ de\ loi\ 4] \wedge [loi\text{-}ar = loi\ 3, loi\ 4] \wedge [am\text{-}global = sous\text{-}amorti, normal] \wedge [copiage = léger, moyen]$$

Ce groupe est jugé bon et sous-amorti globalement. Aucune indication ne peut être donnée sur le sous-amortissement avant et arrière car les avis sont trop partagés. Toutefois le copiage n'est pas fort.

Groupe 4 : versions 3,4,13,14

$$G4 = [classe = classe\ 4] \wedge [loi\text{-}av = loi\ 3, loi\ 4] \wedge [loi\text{-}ar = pas\ de\ loi\ 2] \wedge [sous\text{-}a\text{-}av = fort, moyen] \wedge [copiage = léger, moyen]$$

L'ensemble des enquêtés trouvent le sous-amortissement avant trop prononcé. Le copiage est toujours jugé moyen ou léger. Les avis sont très contrastés sur l'amortissement global et sur le sous-amortissement arrière; aucune tendance ne se détache pour ces deux points.

Groupe 5 : versions 3,4,7,13

$$G5 = [classe = classe\ 3, classe\ 4] \wedge [loi\text{-}av = pas\ de\ loi\ 2] \wedge [loi\text{-}ar = loi\ 3, loi\ 4] \wedge [am\text{-}global = normal, sous\ amorti] \wedge [sous\text{-}a\text{-}av = fort, moyen] \wedge [copiage = léger, moyen]$$

Comme dans le groupe 4, le jugement global est normal et sous-amorti, principalement sur l'essieu avant.

En reportant les groupes construits par la pyramide figure 10 sur le plan 1-2 de l'ACM figure 8, nous faisons apparaître l'évolution des versions du sur-amortissement au sous-amortissement ainsi que les nuances d'appréciation subjective suivant qu'on aime une version plus sur-amorti ou plus sous-amorti.

La liaisons entre les versions, marquées par un trait fort, et telles qu'elles apparaissent lors de la visualisation de la pyramide figure 10, montrent ainsi la capacité de la méthode à retrouver un ordre implicite.

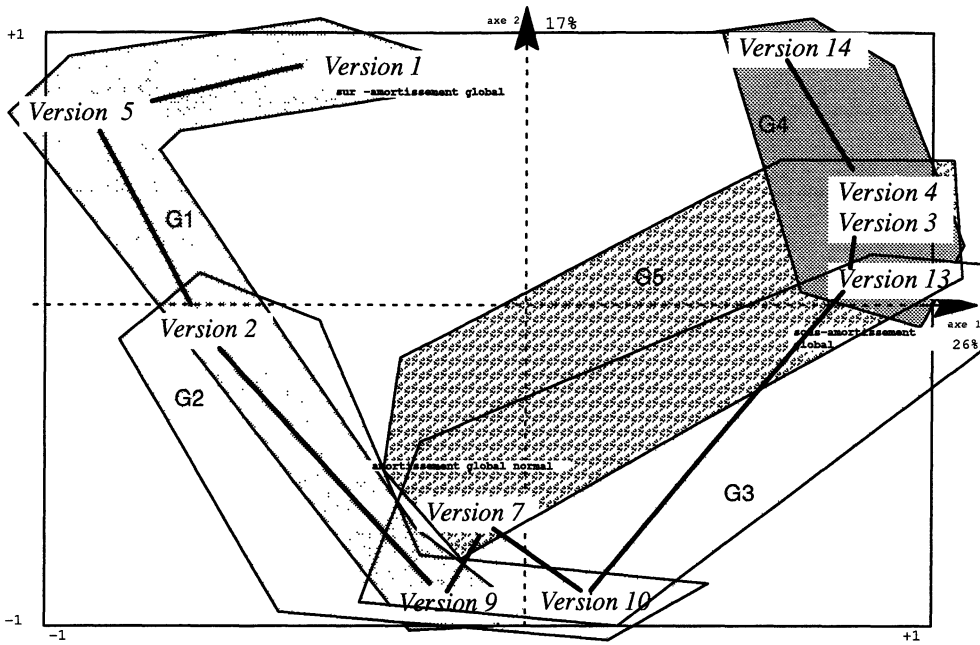


FIGURE 11

Classes de la pyramide sur le plan de projection 1-2 de l'A.C.M.

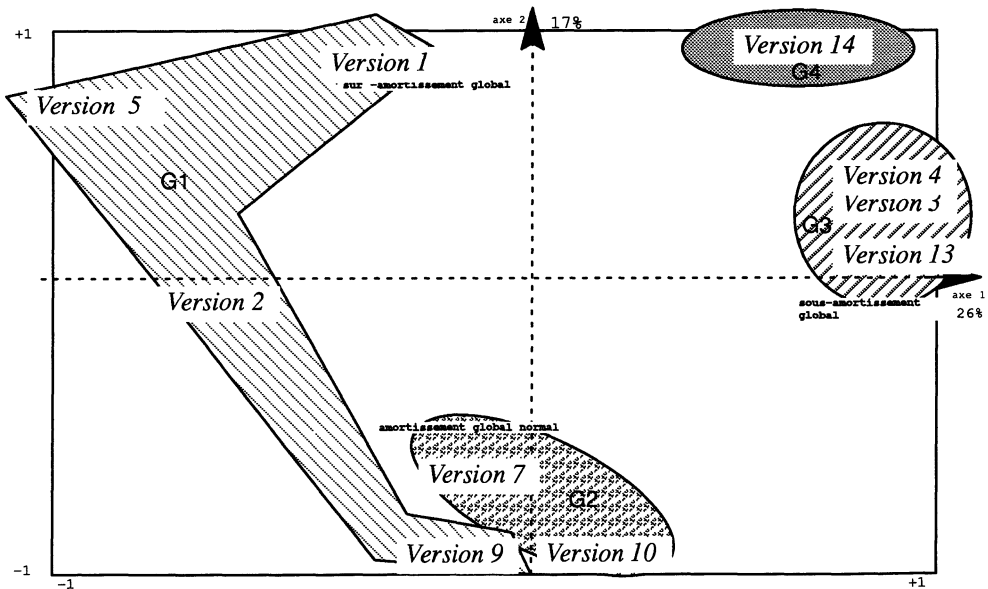


FIGURE 12

Classes de la hiérarchie sur le plan de projection 1-2 de l'A.C.M.

Ce recouvrement, comparé à la partition de la hiérarchie figure 12 ci-dessus, montre le groupe formé des versions «sur- amortissement global» de façon identique, et prend en compte les nuances autour du compromis «amortissement global normal». De plus, il n'isole pas la version 14 extrême dans le sous-amortissement.

Finalement, il suggère que chaque passager, identifié à un groupe G_1, \dots, G_5 *a posteriori*, répond à la question «ce véhicule est sur-amorti ou sous-amorti globalement?» de manière relative dans l'ordre implicite. Par exemple les personnes identifiées dans G_1 (respectivement G_3) trouveront sur-amorti la version 1 (respectivement la version 9) et sous-amorti la version 9 (respectivement la version 13).

5. Conclusion

L'analyse des données permet la simplification du nombre de paramètres nécessaires à la compréhension d'un phénomène puisque 3 variables (var1 var3 var7) parmi les 10 ont été sélectionnées par la discrimination par arbre binaire. Quelque soit la qualité des paramètres : judicieux, redondants ou pas, cette opération de réduction est souvent observée, l'idéal étant de se ramener au plan pour effectuer une décision de façon simple ou visuelle.

La discrimination par la méthode des arbres binaires et la classification pyramidale symbolique apportent la synthèse des résultats sous forme de règles et des nuances d'appartenance à telle ou telle classe *priori*. La manipulation de symboles et de règles sont plus compréhensibles par les experts de l'automobile que des plans factoriels. La confrontation de leurs connaissances avec les interprétations issues des méthodes employées a validé le processus de traitement utilisé et les quatre classes de confort basse fréquence proposées.

L'interprétation de la C.P.S. complétée par l'A.C.M met en relief les nuances d'appréciation suivant que l'on aime une voiture sportive ou très confortable. Si les avis divergent quelque peu suivant les goûts de chacun, tous acceptent une version idéale à 9 ou 7. De plus, le plan factoriel montre la capacité de la méthode de classification pyramidale à retrouver un ordre implicite dans le tableau de données; en effet, les versions s'échelonnent d'une version sur- amortie, qualifiée de sportive voir «karting» à une version très très souple voire «bateau».

Cet article montre la démarche adoptée pour caractériser de manière objective le confort dans le domaine des basses fréquences et l'avis du conducteur qu'il soit expert ou non.

Bibliographie

- M. BARBUT, B. MONJARDET (1970). Ordre et classification, Algèbre et combinatoire; Tome I et II, Hachette, Paris.
- P. BERTRAND (1986). Etude de la représentation pyramidale, Thèse de 3 ème cycle, Université de Paris-IX Dauphine.

- P. BERTRAND, E. DIDAY (1985). A visual representation of compatibility between an order and a dissimilarity index : The Pyramids; *Computational Statistics Quarterly*, Vol 2, Issue 2, p.31-42.
- P. BERTRAND, E. DIDAY (1990). Une généralisation des arbres hiérarchiques : les représentations pyramidales; *Rev. Statistique Appliquée*, vol XXXVIII (3), p. 53-78.
- P. BRITO (1991). Analyse de données symboliques - pyramides d'héritage; Thèse de Docteur en sciences Université Paris IX - Dauphine.
- G. CELEUX et Y. LECHEVALLIER (1982). Méthodes de segmentation non paramétriques. *RSA*, vol 30, n°4.
- G. CELEUX et Coll. (1990). Analyse discriminante sur variables continues; Collection Didactique INRIA.
- E. DIDAY (1984). Une représentation visuelle des classes empiétantes : les pyramides; Rapport de Recherche INRIA n°291, Avril.
- E. DIDAY (1986). Une représentation visuelle des classes empiétantes : les pyramides; *RAIRO*, Vol 20, n°5, p. 475- 526.
- E. DIDAY (1987). The symbolic approach in clustering and related methods of data analysis : the basic choices; *Proceedings of the first conference of the International Federation of Classification Societies*, ed H. Bock.
- E. DIDAY (1991). Des objets de l'Analyse des données à ceux de l'Analyse des connaissances; *Induction Symbolique- Numérique à partir de données*, eds. Y. Kodratoff, E. Diday; Cepadues.
- J. H. FRIEDMAN (1977). A recursive partitioning decision rule for non parametric classification. *IEEE Trans. on comp.* C26-4, p. 404-408.
- O. GASCUEL (1989). Inductive learning, numerical criteria and combinatorial optimization, some results; *Proceedings of the conference on data analysis, Learning Symbolic and Numerical Knowledge*, Ed. E. Diday, Nova Science Publishers, Inc., New York.
- D. GRANGÉ, P. GRETH, J. B. DANION (1991). Application des méthodes de la classification à la construction d'un test de mémoire, *Revue Modulad N 8* Décembre.
- L. LEBART, A. MORINEAU, N. TABARD (1977). *Technique de la description statistique*, DUNOP.
- T. MITCHELL (1982). Generalization as search; *Artificial Intelligence*, Vol 18, p. 203- 226.
- WARD (1963). Hierarchical grouping to optimize an objective function; *JASA* vol 58.
- WILLE (1982). Restructuring lattice theory : an approach based on hierarchies of concepts, in : *Proc. of the Symp. on Ordered Sets*, Ed. I. Rival, Reidel, Dordrecht-Boston, p. 445-470.