REVUE DE STATISTIQUE APPLIQUÉE

E. M. QANNARI
E. VIGNEAU
PH. COURCOUX

Une nouvelle distance entre variables. Application en classification

Revue de statistique appliquée, tome 46, n° 2 (1998), p. 21-32
<http://www.numdam.org/item?id=RSA_1998__46_2_21_0>

© Société française de statistique, 1998, 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/

Rev. Statistique Appliquée, 1998, **XLVI** (2), 21-32

UNE NOUVELLE DISTANCE ENTRE VARIABLES. APPLICATION EN CLASSIFICATION

E.M. Qannari, E. Vigneau, Ph. Courcoux

ENITIAA/INRA – Unité de Statistique Appliquée à la Caractérisation des Aliments, Domaine de la Géraudière, BP 82225, 44322 Nantes cedex 03 – France

RÉSUMÉ

Une distance euclidienne entre variables quantitatives qui permet de tenir compte aussi bien des variances des variables que de leurs corrélations est discutée. Lorsque les variables sont standardisées, cette distance permet de retrouver un indice usuel de dissimilarité entre variables. De même, cette distance peut s'étendre aux cas de variables qualitatives et d'un mélange de variables qualitatives et quantitatives. Des procédures de classification de variables ainsi que des applications sont étudiées. En particulier, nous montrons comment les résultats de la classification peuvent être mis à profit pour la sélection de variables et pour la détermination de variables synthétiques.

Mots-clés : *classification de variables, sélection de variables, analyse en composantes principales.*

ABSTRACT

We discuss an Euclidean distance between variables which takes account of their correlations as well as of their variances. This distance can be extended to encompass the case of nominal variables and a mixture of nominal and numerical variables. Procedures for clustering of variables and applications are outlined. In particular, we show how cluster analysis of variables can be useful for selecting a subset of variables and for use in connection with principal components analysis.

Keywords : *clustering of variables, selection of variables, principal components analysis.*

1. Introduction

La classification de variables en groupes homogènes revêt un double intérêt : elle peut servir de base pour une stratégie de sélection de variables ou elle peut être vue comme un outil complémentaire de l'analyse en composantes principales (ACP). En effet, la formation de classes homogènes procure à l'utilisateur la possibilité de sélectionner dans les différentes classes une (ou plusieurs) variable(s) en tenant compte de l'importance et de la nature des classes et en faisant intervenir d'autres considérations telles que le coût, la facilité d'évaluation et d'interprétation des

variables ... etc. D'autre part, la classification de variables, parce qu'elle aboutit à une partition des variables telle que dans chaque classe on trouve des variables qui cernent une même tendance du phénomène alors que deux classes distinctes concernent en général des aspects complémentaires, conduit à la définition de classes principales (au même titre que l'ACP conduit à des composantes principales). Il est envisageable d'extraire de ces classes principales des variables synthétiques qui traduisent la structure de l'ensemble des variables et qui présentent l'intérêt d'être beaucoup plus facilement interprétables que les composantes principales. L'objectif visé est dans une certaine mesure similaire à celui recherché en effectuant des rotations obliques sur les composantes principales d'une ACP (Harman, 1976).

Des méthodes de classification de variables sont suggérées dans beaucoup d'ouvrages qui proposent des indices de similarité ou de dissimilarité entre variables. Par exemple, le coefficient de corrélation linéaire r est un indice de similarité usuel entre variables et $(1-r)$ un indice de dissimilarité. Jolliffe (1972) souligne que parmi 8 stratégies de sélection de variables, celles basées sur des méthodes de classification et utilisant l'indice de similarité r conduisent à des résultats satisfaisants. Afin d'éviter de séparer des variables fortement liées mais corrélées négativement, d'autres indices de dissimilarité usuellement considérés sont $1 - r^2$ ou $1 - |r|$. Une autre stratégie proposée par Gallego Pinillia (1980) consiste à considérer simultanément les variables et les variables changées de signe, de façon à classer ensemble les variables dont la corrélation est forte en valeur absolue. Cette procédure semble adaptée pour les méthodes de classification hiérarchique car elle conduit à un arbre symétrique dont il suffit de considérer une moitié pour l'interprétation. Derquenne (1997) souligne l'intérêt de la classification des variables et discute en particulier des procédures de classification de variables qualitatives.

Nous discutons une distance euclidienne entre variables quantitatives qui permet, lorsque les variables sont homogènes, de tenir compte aussi bien des variances des variables que de leurs corrélations. Lorsque les variables sont standardisées, nous retrouvons un indice usuel de dissimilarité entre variables. L'idée centrale exploitée ici est de «plonger» les variables dans une structure euclidienne *via* les opérateurs introduits par Escoufier (1970) à savoir les matrices des produits scalaires entre individus. Par la suite, il est facile d'adapter des stratégies de classification et de mettre en œuvre des logiciels standards de traitement des données. Il est également possible d'étendre cette distance au cas de variables qualitatives et de considérer conjointement des variables quantitatives et des variables qualitatives. Des illustrations à l'aide de données réelles sont présentées.

2. Distance entre variables quantitatives

2.1. Définition de la distance

Soit Ω un ensemble de n individus. On considère ici le cas de variables quantitatives définies sur Ω. A chaque variable x supposée centrée, on associe l'opérateur W_x (de rang 1) défini par :

$$W_x = \frac{1}{n} x \, {}^t x$$

Soit y une autre variable quantitative (centrée) définie sur Ω et W_y, l'opérateur qui lui est associé. En notant $\|\cdot\|$ la norme euclidienne usuelle dans l'espace des matrices carrées d'ordre n, on considère comme distance entre x et y, la distance entre les opérateurs W_x et W_y :

$$D(x,y) = \|W_x - W_y\|$$

Cette distance a été préconisée dans un cadre plus général par Escoufier (1970) pour évaluer la distance entre tableaux de données.

En désignant par $\mathrm{var}(x)$, $\mathrm{cov}(x,y)$ et $r(x,y)$ respectivement la variance de x, la covariance entre x et y et le coefficient de corrélation entre x et y, $D^2(x,y)$ peut s'écrire :

$$D^2(x,y) = \mathrm{var}^2(x) + \mathrm{var}^2(y) - 2\,\mathrm{cov}^2(x,y)$$

ou

$$D^2(x,y) = (\mathrm{var}(x) - \mathrm{var}(y))^2 + 2\,\mathrm{var}(x)\mathrm{var}(y)(1 - r^2(x,y)).$$

Il en découle :

i) $D(x,y) = 0$ si et seulement si $x = y$ ou $x = -y$;

ii) à variances égales, les variables x et y sont d'autant plus distantes que leur coefficient de corrélation est faible en valeur absolue et que les variances sont grandes;

iii) Lorsque x et y sont standardisées, il vient :

$$D^2(x,y) = 2(1 - r^2(x,y)).$$

Plusieurs auteurs, parmi lesquels Krzanowski (1990), préconisent de travailler avec $1 - r^2(x,y)$ comme indice de dissimilarité entre x et y. Nous suggérons ici de travailler avec la racine carrée de cette quantité car ceci induit une structure euclidienne.

2.2. Opérateur centroïde et inertie

Soit un ensemble E formé de p variables $x_1, x_2, \ldots, x_p$. On désigne également par E le tableau des données dont les colonnes sont les variables $x_1,\ x_2, \ldots, x_p$. On associe à E l'opérateur centroïde défini par :

$$W_E = \frac{1}{p}\left(\sum_{i=1}^{p} \frac{1}{n} x_i\, {}^t x_i\right) = \frac{1}{p}\frac{1}{n} E\, {}^t E.$$

Cet opérateur centroïde est solution du problème d'optimisation suivant :

$$\min_W \sum_{i=1}^{p} \|W_{x_i} - W\|^2.$$

Par la suite, on définit la dissimilarité entre deux groupes de variables comme étant la distance entre les deux opérateurs centroïdes associés. De même, on définit l'inertie relativement à la distance D de l'ensemble des variables constituant le tableau E par :

$$I_E = \frac{1}{p}\sum_{i=1}^{p} \|W_{x_i} - W_E\|^2$$
$$= \frac{p-1}{p^2}\sum_i \mathrm{var}^2(x_i) - \frac{1}{p^2}\sum_i\sum_{j\neq i} \mathrm{cov}^2(x_i, x_j)$$

En particulier, lorsque les variables sont standardisées, il vient :

$$I_E = 1 - \frac{1}{p^2}\sum_i\sum_j r^2(x_i, x_j).$$

Cette dernière quantité est d'autant plus faible que les corrélations entre variables sont élevées (structure vectorielle unidimensionnelle).

3. Classification de variables

3.1. Classification ascendante hiérarchique

Etant donné une partition $E_1, \ldots, E_k$ en k classes de l'ensemble E formé de p variables $x_1, x_2, \ldots, x_p$, l'inertie totale associée à E peut s'écrire comme étant la somme de l'inertie intra- classes et de l'inertie inter-classes :

$$I_{totale} = \sum_{j=1}^{k} \frac{p_j}{p} I_{E_j} + \sum_{j=1}^{k} \frac{p_j}{p} \|W_{E_j} - W_E\|^2,$$

où p_j désigne le nombre de variables de $E_j (j = 1, 2, \ldots, k)$ et W_E (resp. W_{E_j}), l'opérateur centroïde associé à E (resp. E_j).

La classification ascendante hiérarchique basée sur la stratégie de Ward consiste à agréger à chaque étape les deux classes E_i et E_j qui conduisent à l'accroissement, ΔI, minimal de l'inertie intra-classes (voir par exemple Diday *et al.*, 1982; Saporta, 1990) :

$$\Delta I = \frac{p_i\, p_j}{p_i + p_j}\|W_{E_i} - W_{E_j}\|^2 = \left(\frac{1}{p_i} + \frac{1}{p_j}\right)^{-1} D^2(E_i, E_j).$$

Cette stratégie s'apparente à une stratégie de proximité des opérateurs centroïdes mais elle tient également compte des poids respectifs des classes formées à un niveau donné. La tendance à l'agrégation de classes d'effectifs importants sera moins forte que pour une stratégie de proximité des opérateurs centroïdes.

3.2. Classification par partitionnement

L'algorithme des centres mobiles (voir par exemple Diday *et al.*, 1982; Saporta, 1990) peut aussi être adapté afin d'obtenir k classes à partir d'un ensemble de p variables. Dans un premier temps, k noyaux initiaux sont sélectionnés. Ces noyaux peuvent être choisis au hasard parmi les p variables de départ. Au cours d'une première étape, chaque variable est associée au noyau initial le plus proche au sens de la distance D. On forme ainsi k classes initiales $E_j (j = 1, \ldots, k)$. Par la suite, la démarche est itérative. Pour chaque variable $x_i (i = 1, 2, \ldots, p)$, on calcule $D(x_i, E_j)$ pour $j = 1, 2, \ldots, k$ et on affecte x_i à la classe pour laquelle cette quantité est minimale. Ainsi, à chaque étape, de nouvelles classes $E_1, \ldots, E_k$ sont constituées et les opérateurs centroïdes $W_{E_j} = \frac{1}{p_j} \frac{1}{n} E_j \, {}^t E_j$ vont jouer le rôle de nouveaux noyaux. Cet algorithme constitue une heuristique pour minimiser l'inertie intra-classes. Il procure une alternative simple à la procédure VARCLUS qui est implémentée dans le logiciel SAS (SAS/STAT, 1990; Derquenne, 1994). Le choix du nombre de classes de la partition peut se faire en considérant le pourcentage d'inertie expliquée par la première composante principale de chaque classe. Par exemple, l'utilisateur peut imposer que ce pourcentage soit au moins égal à 80% comme cela est généralement préconisée pour la procédure VARCLUS.

4. Extensions

4.1. Cas des variables qualitatives

Soit X une variable qualitative ayant m modalités. On désigne également par X, le tableau de codage disjonctif complet associé à cette variable. A ce tableau, nous associons l'opérateur de projection $W_X^* = \frac{1}{n}(XV^{-1}{}^t X - j\,{}^t j)$, où $V = \frac{1}{n}\,{}^t XX$ et j est le vecteur $(n \times 1)$ dont toutes les composantes sont égales à 1 (Cazes *et al.*, 1976; Saporta, 1976). Nous savons que $\|W_X^*\| = \sqrt{m-1}$. L'opérateur de projection normé associé à X est par conséquent $W_X = \frac{1}{\sqrt{m-1}} W_X^*$.

Soit Y une autre variable qualitative ayant q modalités et soit W_Y l'opérateur normé qui lui est associé. La distance entre X et Y est définie par :

$$D(X, Y) = \|W_X - W_Y\|$$

Nous pouvons vérifier (Cazes *et al.*, 1976; Saporta, 1976) que :

$$D^2(X, Y) = 2(1 - T^2(X, Y))$$

où $T(X, Y)$ est le coefficient de Tschuprow défini par :

$$T^2(X, Y) = \frac{1}{\sqrt{m-1}\sqrt{q-1}} \sum_{i=1}^{m} \sum_{j=1}^{q} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}};$$

$(p_{ij})(i = 1,2,\ldots,m; j = 1,2,\ldots,q)$ étant la distribution conjointe de X et Y; $(p_{i+})(i = 1,2,\ldots,m)$, la distribution marginale de X et enfin $(p_{+j})(j = 1,2,\ldots,q)$, la distribution marginale de Y.

L'inertie relativement à D d'un ensemble de p variables qualitatives $X_1, X_2, \ldots, X_p$ peut être définie par :

$$I = \frac{1}{p}\sum_{i=1}^{p} \|W_i - W\|^2 ,$$

où W_i est l'opérateur normé associé à la $i^{ème}$ variable et $W = \frac{1}{p}\sum_{i=1}^{p} W_i$.

Il vient :

$$I = 1 - \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} T_{ij}^2$$

où T_{ij} est le coefficient de Tschuprow entre les variables X_i et X_j.

4.2. Cas des variables qualitatives et quantitatives

Soit x une variable quantitative centrée et normée. Il s'ensuit que l'opérateur $W_x = \frac{1}{n}x^t x$ est lui aussi normé. Soit Y une variable qualitative à q modalités et soit W_Y l'opérateur de projection normé associé à Y, défini comme au paragraphe précédent. La distance $D(x,Y)$ entre la variable quantitative x et la variable qualitative Y est donnée par la distance entre les opérateurs W_x et W_Y. Nous pouvons vérifier que (Saporta, 1976) :

$$D^2(x,Y) = 2\left(1 - \frac{R^2(x,Y)}{\sqrt{q-1}}\right).$$

où $R^2(x,Y)$ est le rapport de corrélation entre x et Y défini comme étant le rapport de la variance inter-classes et de la variance totale de x. Cette distance fait apparaître l'indice de similarité entre x et Y défini par :

$$S(x,Y) = \frac{R^2(x,Y)}{\sqrt{q-1}}$$

L'intérêt de cet indice est qu'il fait intervenir la dimension de la variable qualitative au même titre que le coefficient de Tschuprow fait intervenir les dimensions des variables qualitatives. L'idée ici est que, toute chose étant égale par ailleurs, la similarité entre x et Y est d'autant plus faible que la dimension de Y est grande. *A contrario*, la similarité $S(x,Y)$ atteint son maximum qui est égal à 1 si, et seulement si, x prend seulement deux valeurs, Y a deux modalités et les individus ayant la même modalité de Y prennent aussi la même valeur pour x. Dans ce cas, en effet, le numérateur et le dénominateur de $S(x,Y)$ sont égaux à 1.

Les procédures de classification discutées pour les variables quantitatives peuvent s'étendre sans difficulté au cas de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

5. Etudes de cas

5.1. variables quantitatives

Les données illustrant la classification de variables quantitatives concernent l'évaluation sensorielle de 18 mousses de poisson selon 17 descripteurs (variables sensorielles) dont la liste est donnée au tableau 1. L'évaluation consiste en une note entre 0 et 9. Précisément, pour chaque mousse de poisson et chaque variable sensorielle, nous avons considéré la valeur moyenne de tous les juges qui ont participé à l'expérience. Nous discutons ici les résultats de la classification des variables standardisées et nous discuterons ultérieurement les différences obtenues lorsque les variables ne sont pas standardisées.

TABLEAU 1
Description des 17 descripteurs et statistiques sommaires.

Variable	Libellé court	Moyenne	Ecart-type
«Lisse» au toucher	TLISS	6,37	1,51
Humidité au toucher	THUMI	5,52	1,80
«Collant» au toucher	TCOLL	6,00	2,36
Fermeté au toucher	TFERM	6,05	2,73
«Déformable» au toucher	TDEFO	4,83	2,48
«Cassant» au toucher	TCASS	5,70	1,18
«Gras» au toucher	TGRAS	6,27	1,52
Fermeté au couteau	CFERM	6,28	2,19
«Collant» au couteau	CCOLL	4,06	2,31
Humidité en bouche	BHUMI	3,85	1,48
«Huile» en bouche	BHUIL	4,06	1,13
«Gras» en bouche	BGRAS	5,59	1,73
«Mousse» en bouche	BMOUS	5,39	1,35
Fermeté en bouche	BFERM	5,06	2,76
«Morceaux» en bouche	BMORC	5,05	2,24
«Râpeux» en bouche	BRAPE	4,80	1,22
«Collant» en bouche	BCOLL	6,21	2,40

Le dendrogramme (figure 1) de la classification hiérarchique ascendante selon le critère de Ward suggère d'effectuer une partition de l'ensemble des variables en 3 classes. La qualité de cette partition peut être évaluée par le rapport de l'inertie inter-classes et de l'inertie totale. Ce rapport est ici égal à 48,69%. Ce rapport ainsi que les inerties des classes (tableau 2) indiquent qu'une partition plus fine des descripteurs serait souhaitable, conduisant notamment à une division de la classe E_2 en deux classes dont l'une contiendrait le seul descripteur TCASS (le rapport de l'inertie inter-classes et de l'inertie totale est dans ce cas égal à 58,32%).

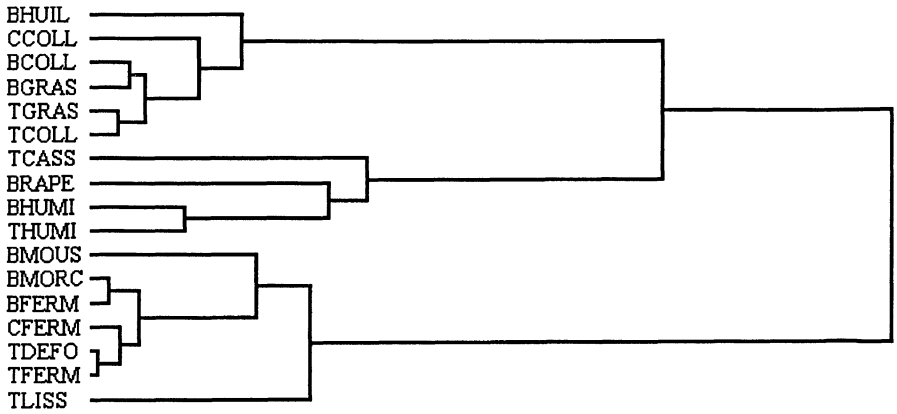

FIGURE 1
Dendrogramme de la classification (stratégie de Ward) des 17 descripteurs.

TABLEAU 2
Caractérisation de la partition des 17 descripteurs en 3 classes.

Classes	Taille	variables	Inertie	Distance au centre de gravité : $D(E_i, E)$
E_1	7	TLISS, TFERM, TDEFO, CFERM, BFERM, BMORC, BMOUS	0,259	0,506
E_2	4	THUMI, TCASS, BHUMI, BRAPE	0,569	0,620
E_3	6	TCOLL, TGRAS, CCOLL, BHUIL, BGRAS, BCOLL	0,237	0,562

La classification par partitionnement a aussi été mise en oeuvre et a conduit à une solution légèrement différente (Le descripteur TCASS n'est plus associé aux descripteurs THUMI, BHUMI et BRAPE; il rejoint la classe de descripteurs associés aux notions de «collant» et «gras»).

Le problème de sélection d'un sous ensemble de variables «représentatives» a suscité l'intérêt de plusieurs auteurs (Jolliffe, 1972; McCabe, 1984; Bonifas *et al.*, 1984; Schlich *et al.*, 1987; Krzanowski, 1987; Gonzalez *et al.*; 1990). La classification de variables peut être un outil intéressant dans ce contexte. Nous illustrons ici deux stratégies de sélection de trois variables parmi les 17 variables originales en nous basant sur les résultats de la classification hiérarchique. La première stratégie consiste à sélectionner dans chacune des trois classes, la variable la plus proche du centroïde de cette classe. Ceci conduit au choix $F = \{\text{TFERM, THUMI, BGRAS}\}$. La dissimilarité $D(E, F)$ entre l'ensemble E de toutes les variables et la sélection F est égale à 0,264. La deuxième stratégie consiste à évaluer toutes les combinaisons

de trois variables à raison d'une variable par classe (soit 168 combinaisons) et à choisir la sélection F pour laquelle $D(E, F)$ est minimale. Cette procédure réduit considérablement le nombre de combinaisons à évaluer lorsque les cardinaux de E et F sont relativement élevés rendant matériellement impossible une inspection exhaustive de toutes les combinaisons. Cette deuxième stratégie conduit à la sélection $F' = \{\text{TFERM, THUMI, BCOLL}\}$ et nous avons $D(E, F') = 0{,}239$. Le nombre de toutes les combinaisons de trois variables parmi 17 étant «raisonnable», nous avons évalué, à titre de vérification, toutes les combinaisons (au nombre de 680) et nous avons obtenu la même solution, F'.

Afin d'illustrer comment la classification peut être utilisée pour déterminer des variables factorielles synthétiques, une ACP normée a été réalisée dans chacune des trois classes définies précédemment (stratégie de Ward) et à chaque fois, la première composante principale a été retenue. La première composante de E_1, $C_1^{(1)}$, représente 85,6% de l'inertie associée à E_1. $C_1^{(2)}$ qui est la première composante principale pour la classe E_2, représente 59,3% de l'inertie associée à E_2. Nous avions déjà noté que cette classe était relativement hétérogène (inertie relativement élevée). La variable $C_1^{(3)}$ qui est la première composante principale pour la classe E_3 représente 87,1% de l'inertie de la classe E_3. Bien que non orthogonales, ces trois composantes synthétiques rendent cependant compte d'aspects complémentaires relatifs au phénomène étudié. Leur intérêt par rapport aux composantes principales de l'ensemble des 17 variables est qu'elles sont faciles à interpréter car chacune est une combinaison linéaire d'un ensemble réduit de variables relativement homogènes.

Les descripteurs étant évalués sur une même échelle, une classification des variables non standardisées a aussi été effectuée. La classification en trois classes fait apparaître une première classe de variables fortement liées à la fermeté (TFERM, TDEFO, CFERM, BFERM, BMORC), une deuxième classe de variables exprimant les caractères «gras» et «collant» (TCOLL, TGRAS, CCOLL, BGRAS, BCOLL) et une troisième classe constituée des autres descripteurs (TLISS, BMOUS, THUMI, TCASS, BHUMI, BRAPE, TGRAS, BHUIL). L'interprétation de cette dernière classe n'est pas aisée car les variables s'y regroupent principalement du fait de leurs variances relativement faibles (tableau 1).

5.2. Variables quantitatives et qualitatives

Le tableau de données considéré est extrait de la thèse de Langron (1981). Il concerne l'évaluation sensorielle des caractéristiques aromatiques de 27 lots de pommes de la variété Cox (14 descripteurs quantitatifs) sous différentes conditions de stockage (3 variables qualitatives). Le tableau 3 donne les libellés et une description succincte des variables. Les notes sensorielles considérées pour chaque lot correspondent à la moyenne des notes attribuées par 8 juges. Pour la suite de l'étude, le lot de pommes 18 a été écarté car il présentait des caractéristiques aromatiques très atypiques.

La figure 2 donne le dendrogramme de la classification ascendante hiérarchique des 17 variables (critère de Ward). Elle suggère d'effectuer une partition de l'ensemble des variables en 2 classes, voire 4 classes si on s'intéresse à un niveau de détail plus fin.

TABLEAU 3
Liste des variables (14 variables quantitatives et 3 variables qualitatives).

Descripteurs d'arôme		Conditions de stockage	
n°	libellé	n°	libellé
1	ALCO alcoolisé	15	SAC stockage atmosph. contrôlée
2	VERT vert		SAC1 0 semaine
3	ACID acide		SAC2 10 semaines
4	PARF parfumé		SAC3 18 semaines
5	COX typique de la variété Cox		SAC4 25 semaines
6	BANA banane		SAC5 31 semaines
7	FRUI fruité	16	SAL stockage à l'air libre
8	ESYN ester synthétique		SAL1 0 jours
9	ECIR ester cireux		SAL2 6-10 jours
10	SUCR sucré		SAL3 15-19 jours
11	EPIC épicé		SAL4 25-32 jours
12	FSEC feuilles sèches		SAL5 39, 48 ou 51 jours
13	GRAI graisseux	17	TEMP température (air libre)
14	RANC rance		TEMP1 11°C
			TEMP2 17°C

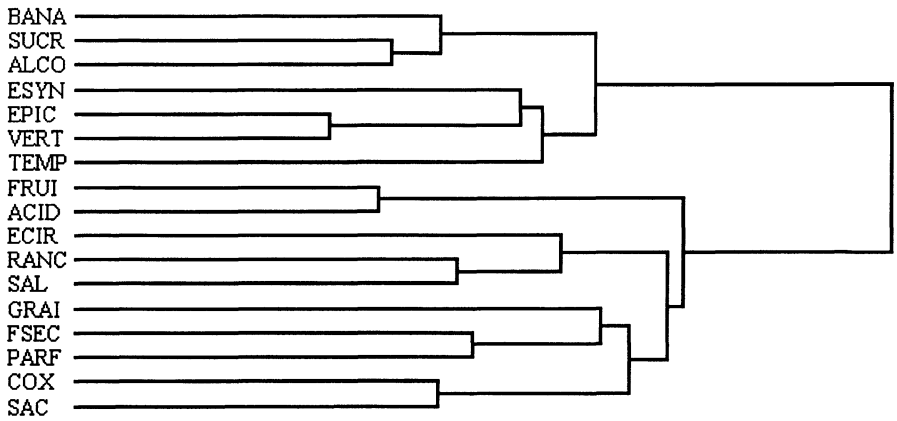

FIGURE 2
Dendrogramme de la classification (stratégie de Ward) des 14 descripteurs aromatiques et des 3 variables qualitatives de stockage.

La variable qualitative SAC (stockage en atmosphère contrôlée) se situe dans la même classe que les variables quantitatives COX, PARF (parfumé), FSEC (feuilles sèches) et GRAI (graisseux) car d'une part, les variables quantitatives sont fortement corrélées (positivement ou négativement) et d'autre part, la variable SAC a un fort impact sur ces descripteurs. Cette affirmation est étayée par des analyses de la variance

à un facteur sur les descripteurs en prenant les modalités de la variable SAC comme niveaux du facteur considéré. Il ressort entre autres que la variable COX est, parmi toutes les variables, la plus discriminante ($F = 7{,}12$, niveau de signification : 0,09%) : l'arôme COX est particulièrement fort sans stockage (première modalité de SAC). Des interprétations similaires peuvent être faites dans les autres classes pour expliquer le regroupement des variables. Ainsi, le descripteur RANC (rance) est lié à la variable de stockage SAL ($F = 5{,}20$, niveau de signification : 0,45%). La température pendant le stockage à l'air libre (TEMP) semble avoir une influence sur les caractéristiques d'arôme VERT, EPIC, ESYN, ALCO, SUCR, BANA.

Les résultats de cette classification pourraient être complétés en effectuant une classification des variables quantitatives et de toutes les modalités des variables qualitatives, considérées comme étant des variables nominales binaires (absence, présence).

6. Conclusion

Les distances entre variables et les différentes stratégies de classification discutées dans ce papier présentent plusieurs avantages :

– dans le cas de variables quantitatives, elles procurent à l'utilisateur la possibilité de tenir compte des corrélations entre les variables et éventuellement de leurs variances;

– leur mise en œuvre peut s'effectuer à l'aide de logiciels standards;

– elles englobent le cas de variables qualitatives et un mélange de variables qualitatives et quantitatives.

Plus généralement, il apparaît que la classification de variables constitue un outil intéressant pour explorer la structure d'un tableau de données, sélectionner un sous ensemble de variables et compléter les résultats d'une ACP.

Références bibliographiques

BONIFAS L., ESCOUFIER Y., GONZALEZ P. L., SABATIER R. (1984). Choix des variables en Analyse en Composantes Principales. Revue de Statistique Appliquée, 32(2), 5-15.

CAZES P., BONNEFOUS S. BAUMERDER A., PAGES J.-P. (1976). Description cohérente des variables qualitatives prises globalement et de leurs modalités. Statistique et analyse des données, 2 et 3, 48-62.

DERQUENNE C. (1994). VALITYPO : une macro pour valider statistiquement une typologie. SEUGI/club SAS'94. Strasbourg, France.

DERQUENNE C. (1997). Classification de variables qualitatives. XXIX^{e} Journées ASU, Carcassonne.

DIDAY E., LEMAIRE J., POUGET J., TESTU F. (1982). Eléments d'analyse de données. Dunod, Paris.

ESCOUFIER Y. (1970). Echantillonnage dans une population de variables aléatoires réelles. Thèse de doctorat d'Etat. Université de Montpellier.

GALLEGO PINILLIA J. (1980). Un codage flou pour l'analyse des correspondances. Thèse de 3e cycle, université Paris VI.

GONZALEZ P. L., CLÉROUX R., RIOUX B. (1990). Selecting the best subset of variables in principal components analysis. Compstat, Physica-Verlag Heidelberg – IASC, 115-120.

HARMAN H. H. (1976). Modern factor analysis. Third edition, Chicago : University of Chicago Press.

JOLLIFFE I. T. (1972). Discarding variables in a principal component analysis. I : artificial data. Appl. Statist., 21, 160-173.

KRZANOWSKI W. J. (1990). Principles of multivariate analysis, a user's perspective. Oxford statistical science series.

KRZANOWSKI W. J. (1987). Selection of variables to preserve multivariate data structure using principal components. Appl. Statist., 36 (1), 22-33.

LANGRON S. P. (1981). The statistical treatment of sensory analysis data. Ph.D. thesis, Univ. of Bath.

McCABE G. P. (1984). Principal variables. Technometrics, 26 (2), 137-144.

SAPORTA G. (1976). Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives. Statistique et analyse des données, 1, 38-46.

SAPORTA G. (1990). Probabilités, analyse de données et statistique. Ed. Technip, Paris.

SAS/STAT (1990). User's guide. Version 6, Vol. 2.

SCHLICH P., ISSANCHOU S., GUICHARD E., ETIEVANT P. et ADDA J. (1987). RV coefficient : a new approach to select variables in PCA and to get correlations between sensory and instrumental data. Flavour Science and Technology, Martens, Dalen and Russwurm Eds, Wiley & Sons Ltd, 469-474.