Détermination d'une note globale, synthèse d'une évaluation numérique et d'appréciations libres

Application aux études de marché

Mónica Bécue-Bertaut[1], Jérôme Pagès[2],
Ramón Álvarez Esteban[3], José Luis Vázquez Burguete[4]

[1] EIO–UPC– Edifici C5 – Campus Nord Jordi Girona, 1-3 – 08028 Barcelona – Spain
monica.becue@upc.edu

[2]ENSA/ INSFA.– 65 rue de Saint-Brieuc, CS 84215 – F35042 Rennes cedex – France
jerome.pages@agrocampus-rennes.fr

[3]EIO- University of León.– Campus de Vegazana s/n – 24071 León – Spain
dderae@unileon.es

[4] Marketing Depart.- University of León.– Campus de Vegazana s/n – 24071 León – Spain
ddejvb@unileon.es

Abstract

In marketing, the products are usually evaluated by marks and, jointly, free-text comments. We aim at obtaining global marks, taking into account both kinds of valuations, which can be considered as their best synthesis. We start from a proposal by Abdessemed & Escofier to analyse individuals described by sets of continuous variables and one contingency table. Nevertheless, we propose a different solution to weight the rows, taking advantage of the possibility of extending CA to different models and metrics. An application illustrates this methodology.

Keywords : free-text evaluation ; multiple factor analysis ; correspondence analysis ; global score ; lexical table.

1. Introduction

Dans les études de marché, les produits sont habituellement qualifiés au moyen d'évaluations numériques et, conjointement, de commentaires libres. Il est fréquent qu'un utilisateur veuille tenir compte simultanément des deux types d'évaluations pour définir une note globale.

Les évaluations numériques sont considérées comme des variables quantitatives. Les commentaires libres conduisent à construire un tableau lexical, croisant individus et mots (Benzécri, 1981 ; Lebart et Salem, 1994). Ainsi, l'ensemble de l'information constitue des données hétérogènes, se composant d'un groupe de variables quantitatives et d'un groupe de colonnes de type fréquence, correspondant au tableau lexical. Pour analyser ce type de tableau multiple, Abdessemed et Escofier (1996) ont proposé une extension de l'analyse factorielle multiple (AFM) qui détermine les variables (variables générales) qui constituent le meilleur compromis entre les deux groupes de colonnes. Nous partons de cette proposition, mais

présentons une solution différente quant aux poids de lignes, tirant profit de l'extension de l'analyse de correspondances AC à différents modèles et métriques (Escofier, 2003).

La section 2 décrit l'exemple *Vins de Castille et Léon.* La Section 3 présente le tableau à analyser ainsi que les notations employées. Les résultats obtenus en considérant ou bien seulement l'évaluation numérique ou bien seulement les commentaires libres sont exposés Section 4. La Section 5 rappelle brièvement la méthodologie proposée par Abdessemed & Escofier et montre comment le poids initial des individus (en général un poids uniforme) peut être adopté dans l'analyse globale. La section 6 offre les résultats obtenus sur l'exemple.

2. Données "Vins de Castille et Léon"

Le guide gastronomique "*El Mundo*" (El Mundo, 2005a, 2005b) note et commente 522 vins de "Castille et Léon". Cette région (94273 km^2), située au nord-ouest de l'Espagne, comporte cinq désignations d'origine (*El Bierzo, Cigales, Ribera del Duero*, *Rueda* et *Toro*). Chaque vin a reçu une note (entre 70 et 97) et la plupart d'entre eux une évaluation détaillée sous forme de texte libre (la longueur moyenne des commentaires est de 32,4 mots). Nous conservons ici seulement les vins auxquels sont associés des commentaires. Ces vins ont une note égale ou supérieure à 78. La moyenne des notes vaut 84,6 et l'écart-type est égal à 3,92. Quelques informations complémentaires, dont le prix, sont données pour chaque vin.

---- ***Vin 52*** *(note=78)*
Vino limpio y correcto, aunque algo corto en nariz y en boca.*(Vin propre et correct, bien qu'un peu court en narine et en bouche)*
---- ***Vin 507*** *(note=78)*
Fruta al borde de la sobremadurez, con recuerdos de bayas rojas. En boca se muestra delgado, con algo de sequedad al final. *(Fruit presque trop mûr avec un léger goût de fruits rouges, maigre en bouche, avec un peu de sécheresse sur la fin)*
---- ***Vin 53*** *(note=97)*
Impresionante color negro violáceo, balsámicos por doquier, gominolas, maderas dulces, fruta muy jugosa. Mucha jugosidad en boca, chocolate blanco, pimienta blanca y almendros. Muy potente. *(Impressionnante couleur noire-violacé, balsamique, arôme de boule-de-gomme, bois doux, fruit très juteux. Juteux en bouche, saveur de chocolat blanc, poivre blanc et amande. Très puissant)*
---- ***Vin 404*** *(note=97)*
Frío, elegante y estricto ; en nariz disciplinado y en boca estructurado y muy profundo. Se ahorra una golosidad fácil a cambio de un potencial extraordinario de envejecimiento. ¡Un vino serio!
(Froid, élégant et austère, discipliné en narine et structuré en bouche. Très profond. Il donne priorité à son potentiel de vieillissement plutôt qu'à une gourmandise facile. Un vin sérieux!)

Tableau 1. Extrait des l'évaluation des vins : note numérique et commentaires libres

3. Tableau de données et notation

Colonnes Lignes	Commentaires libres J_c=colonnes-mot	Notes J_q= colonnes-note
I produits	*Fréquence du mot dans chaque commentaire* $f_{ij}= k_{ij}/k$ $j=1,\ldots,J_c$	*Valeurs quantitatives* x_{ij} $j=1,\ldots,J_q$

Figure 1. Tableau à analyser

La figure 1 présente le tableau construit à partir des données. Les commentaires libres sont vus comme un tableau de contingence qui contient la fréquence avec laquelle chaque mot est utilisé pour qualifier chacun des vins. f_{ij} indique la fréquence avec laquelle le mot j (j=1,…,J_c)

est utilisé pour qualifier le vin i ; x_{ij} est la note attribuée au vin i relativement à sa caractéristique j (j=1,…,J_q). Dans l'exemple, I=443 (vins), J_c=*250* (mots) and J_q=*1* (évaluation quantitative).

4. Classical strategy to build up a score from multivariate information

Les méthodes factorielles synthétisent une information multidimensionelle au moyen d'une série de variables quantitatives non corrélées. Quand cette information concerne l'évaluation de produits, ces variables quantitatives peuvent être considérées comme des notes qui synthétisent les variables quantitatives initiales. Selon la nature des variables, on applique l'analyse en composantes principales (ACP ; dans le cas de variables quantitatives), l'analyse de correspondances multiples (ACM ; dans le cas des variables qualitatives) ou l'analyse de correspondances (AC ; dans le cas d'un tableau de contingence ou des variables de type fréquence). Quant il y a, comme dans l'exemple, plusieurs groupes de variables de différentes natures, on ne peut considérer comme actif que l'un des groupes de variables, les autres étant considérés comme illustratifs ou supplémentaires.

4.1. Les notes quantitatives actives, les commentaires supplémentaires

Les évaluations numériques sont des variables quantitatives. La méthode de référence est l'ACP. Dans l'exemple, comme il n'y a qu'une évaluation numérique, la note synthétique se confond avec l'évaluation initiale. Pour prendre en compte les commentaires libres en tant qu'illustratifs, nous projetons les mots sur l'axe de l'évaluation numérique, en calculant leur coordonnée au moyen de la formule de transition classique de l'AC (Lebart & Salem, 1994 : 60 ; Escofier & Pagès, 1998 : 65). Ainsi, on peut identifier les mots les plus associés aux vins ayant, respectivement, les notes les plus élevées et les plus faibles.

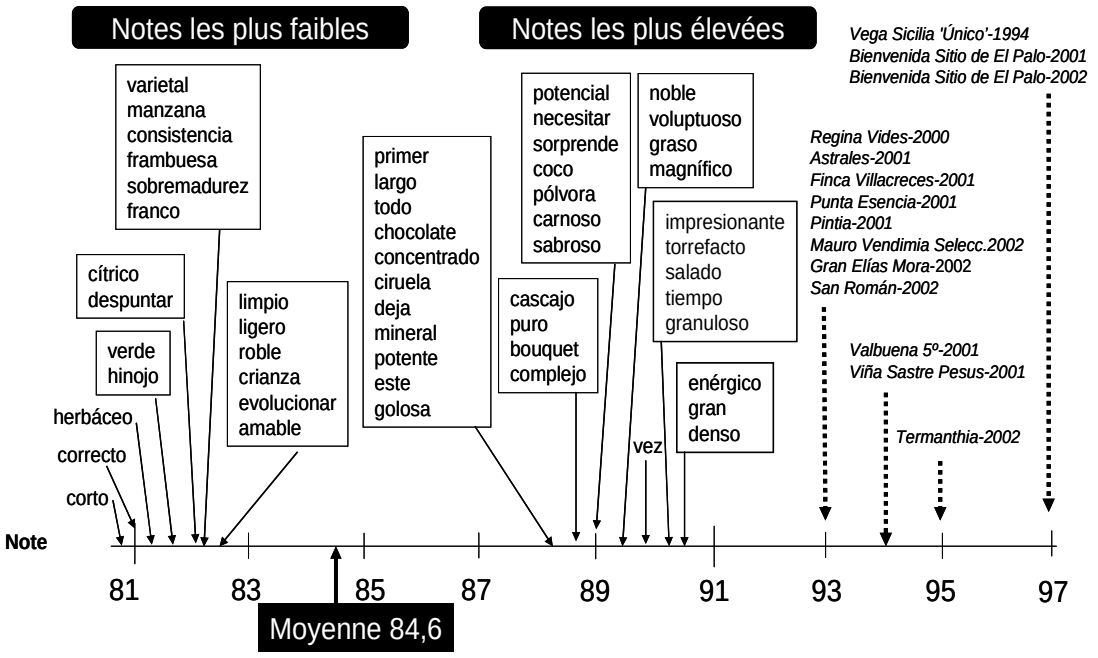

Figure 2. Extrait des mots projetés sur l'axe de l'évaluation numérique

Glossaire :
amable *(aimable)*, bouquet, carnoso *(goût de viande)*, cascajo *(goût de grillé)*, chocolate *(chocolat)*, ciruela *(prune)*, cítrico *(citrique)*, coco *(noix de coco)*, complejo *(complexe)*, concentrado *(concentré)*, consistencia *(consistence)*, correcto *(correct)*, corto *(court)*, crianza *(vieillissement en barrique)*, deja ((il) *(laisse)*, denso *(dense)*, despuntar *(se distinguer)*, enérgico *(énergique)*, este *(cet)*, evolucionar *(évoluer)*, frambuesa *(framboise)*, franco *(franc)*, golosa *(gourmand)*, gran *(grand)*, granuloso *(granuleux)*, graso *(gras)*, herbáceo *(herbacé)*, hinojo *(fenouil)*, impresionante *(impressionnant)*, largo *(long)*, ligero *(léger)*, limpio *(propre)*, magnífico *(magnifique)*, manzana *(pomme)*, mineral (minéral), necesitar *(nécessiter)*, noble, pólvora *(poudre)*, potencial *(potentiel)*, potente *(puissant)*, primer *(premier)*, puro *(pur)*, roble *(chêne)*, sabroso *(savoureux)*, salado *(salé)*, sobremadurez *(trop mûr)*, sorprende *(surprend)*, tiempo *(temps)*, todo *(tout)*, torrefacto *(torréfié)*, varietal,, verde *(vert)*, vez *(fois)*, voluptuoso *(voluptueux)*.

Ce procédé ne prend pas en compte les commentaires libres dans une note globale. Néanmoins, les mots sont ordonnés, ce qui peut être utile pour suggérer quelles sont les caractéristiques des meilleurs vins (respectivement, des vins les moins bien notés).

4.2. Les commentaires libres actifs, la note supplémentaire

Une autre approche part des commentaires libres. Ces commentaires sont d'abord lemmatisés et seuls les mots utilisés au moins 8 fois sont conservés. En outre, les articles, les prépositions, les conjonctions et les adjectifs possessifs sont éliminés. Ainsi, 250 mots différents sont conservés. On applique alors une AC au tableau lexical croisant les 443 vins et les 250 mots (Figure 3).

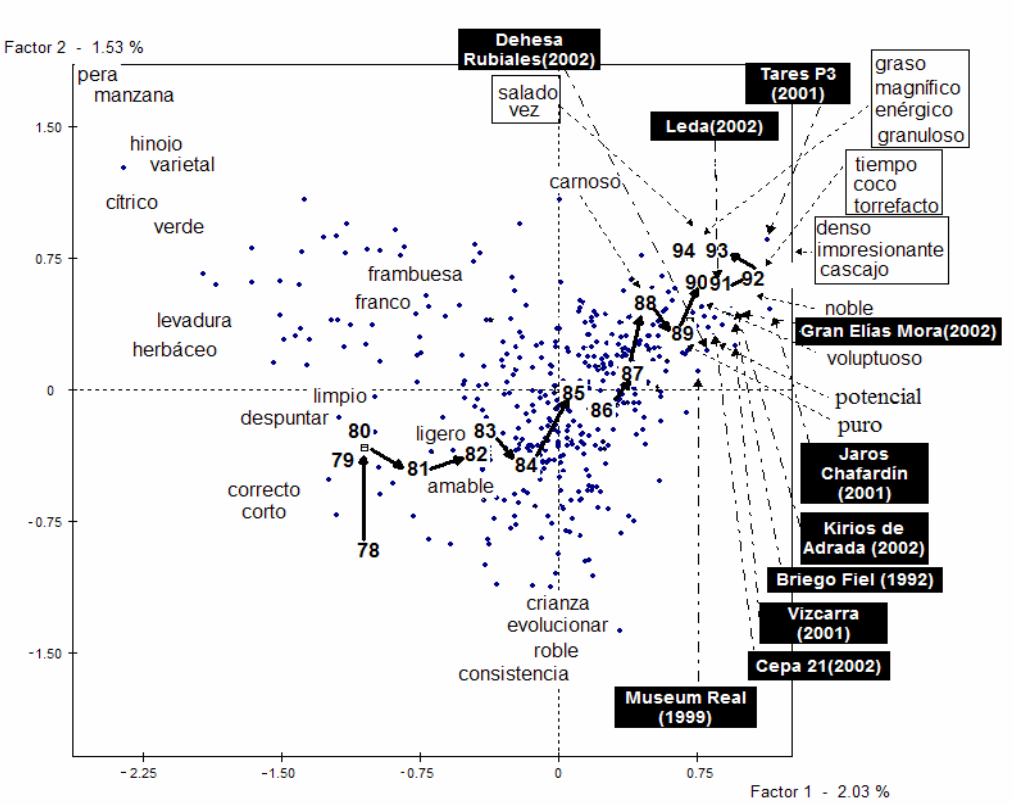

Figure 3. Premier plan factoriel fourni par l'AC du tableau lexical

(les notes 95 (un seul vin) et 97 (seulement 3 vins) ne sont pas représentés)

Le premier axe montre une opposition entre les mots assez proches de celle que montrait la figure 2. Par exemple, *impresionante* (impressionnant) et *salado* (salé) sont opposés à *cítrico* (citrique) et à *green* (vert) comme sur cette figure 2. En calculant la corrélation entre les deux

premiers facteurs et l'évaluation numérique (respectivement, 0,61 avec le premier et le 0,38 avec le second), nous pouvons conclure qu'il y a une forte relation entre la note et les commentaires libres, bien que cette forte relation ne conduise pas à obtenir un facteur qui puisse être utilisé comme une note globale. Si l'on considère les vins ordonnés selon leur coordonnée sur le premier axe, nous trouvons un seul vin (*Gran Elias Mora-2002*) qui soit aussi l'un des 15 vins les mieux notés initialement. L'évaluation numérique (considérée comme variable qualitative avec autant de modalités que de valeurs différentes) est projetée sur le premier plan factoriel, ce qui permet de confirmer la forte relation entre la note et ce plan. Quelques mots caractérisent bien les vins les moins bien notés, et d'autres mots caractérisent les meilleurs. Néanmoins, nous pouvons noter que les mots utilisés dans les commentaires ne différencient pas les vins ayant une évaluation de 90 ou plus.

Ce procédé ne permet pas de définir une note quantitative globale interprétable en tant qu'opposition entre les bons et les moins bon vins, mais offre une mesure de la relation existante entre les commentaires libres et l'évaluation numérique, en calculant la corrélation entre cette dernière et les facteurs issus de l'AC du tableau lexical.

La figure 2 montre aussi qu'il existe une direction de dispersion dans le nuage des vins, tel qu'il est défini par le tableau lexical, liée à l'évaluation numérique. Cette observation suggère de combiner celle-ci et les commentaires dans une nouvelle note globale. Dans ce but, la section suivante propose une méthodologie originale.

5. Combiner la note et les commentaires libres dans une note globale

Nous résumons d'abord la proposition d'Abdessemed et Escofier pour analyser simultanément un ou plusieurs groupes de variables quantitatives et un tableau de contingence (ou plusieurs tableaux de contingence, à condition qu'ils aient la même marge sur les lignes). Puis, nous présentons comment définir des poids pour les lignes différents des poids induits par les marges du tableau de contingence. Finalement, nous présentons la méthode obtenue en adoptant un poids uniforme pour les produits-ligne

5.1. Equivalence entre el AC et une PCA particulière

Les résultats d'une AC d'un tableau de contingence peuvent être obtenus en appliquant une ACP non normalisée au tableau dont le terme général est : (Escofier et Pagès, 1998 : 96) :

$$\left(f_{ij} - f_{i.} \cdot f_{.j}\right) \Big/ \left(f_{i.} \cdot f_{.j}\right) \qquad (1)$$

avec $\{f_{i..}\ ;\ i = 1,\ldots, I\}$ comme poids des lignes (et métrique dans et métrique dans l'espace des lignes).

5.2. Analyse globale

La méthode proposée par Abdessemed et Escofier pour analyser un tableau juxtaposant un tableau de contingence (avec J_c colonnes) et un tableau quantitatif (avec J_q colonnes) est une analyse factorielle multiple spécifique (AFM) et consiste à :

- appliquer une ACP au tableau multiple juxtaposant le tableau issu du tableau de contingence, dont le terme général est donné par (1) et le tableau de variables quantitatives ;
- donner aux lignes le poids induit par l'AC (i.e. $\{f_{i.}\ ;\ i = 1,\ldots, I\}$) ;

- donner aux colonnes-fréquence le poids induit par l'AC $\{f_{.j}$; $j= 1, \ldots, J_c\}$ et aux colonnes quantitatives un poids unitaire, mais en divisant, dans les deux cas, ce poids par la première valeur propre obtenue dans les analyses factorielles séparées – ici AC ou ACP – appliquées aux tableaux correspondants. Cette repondération correspond à la solution adoptée en AFM pour équilibrer l'influence des différents tableaux et normalise à 1 la plus grande inertie axiale de chaque tableau (Escofier & Pagès, 1998 : 132).

5.3. Généralisation de l'AC à d'autres modèles et métriques

L'AC peut être généralisée (Escofier, 2003, chap. 6) à des modèles $\{m_{ij}$; $i=1,\ldots,I$; $j= 1, \ldots,J_c)$ et métriques (p_i, $i=1,\ldots,I$ avec $\sum_{i=1} p_i = 1$; p_j, $j=1,\ldots,J$ avec $\sum_{i=1} p_j = 1$) quelconques. Les résultats de cette généralisation peuvent être obtenus en appliquant une ACP au tableau de terme général :

$$\frac{f_{ij} - m_{ij}}{p_i p_j} \tag{2}$$

avec $\{p_i$; $i = 1,\ldots,I\}$ comme poids pour les lignes (et métrique dans l'espace des colonnes) et$\{p_j$; $j= 1, \ldots, J_c\}$ comme poids des colonnes (et métrique dans l'espace des lignes).

5.4. Modèle et poids des colonnes classiques de l'AC et poids uniforme pour les lignes

Dans notre cas, le modèle d'indépendance et les poids classiques pour les colonnes nous paraissent appropriés tandis que donner un poids uniforme aux produits-ligne semble plus adapté au but recherché.

Ainsi l'analyse globale que nous proposons consiste à :

- appliquer une ACP au tableau juxtaposant le tableau issu du tableau de contingence, dont le terme général est donné par (3), et le tableau des variables quantitatives (éventuellement standarisées) ;

$$\left(f_{ij} - f_{i.} \cdot f_{.j}\right) \Big/ \left(\frac{1}{I} \cdot f_{.j}\right) \tag{3}$$

- donner aux lignes un poids uniforme (i.e. $1/I$; $i = 1,\ldots, I\}$;
- donner aux colonnes-fréquence le poids $\{f_{.j} / \lambda_1^c$; $j= 1,\ldots,J_c)$ et aux colonnes quantitatives le poids $1/\lambda_1^q$. λ_1^c est la première valeur propre obtenue dans l'analyse séparée du tableau issu du tableau de contingence, dont le terme général est donné par (3) ; λ_1^q est la première valeur propre obtenue dans l'analyse séparée du tableau quantitatif.

Les propriétés de cette méthode peuvent être facilement déduites des propriétés générales de l'ACP, en tenant compte des expressions spécifiques des données et des poids.

5.5. Note globale

Les coordonnées des produits sur le premier axe principal, ré-échelonnées de façon à obtenir le même rang de variation que l'évaluation numérique initiale, sont considérées comme les valeurs de la note globale, ou note combinant l'évaluation numérique et les commentaires

libres. Dans certains cas, on peut conserver plusieurs axes et ainsi définir différentes notes globales correspondant à différents aspects des produits.

En outre, la relation de transition qui permet de calculer la coordonnée d'un produit-ligne sur chacun des axes en fonction des coordonnées des mot-colonnes et l'évaluation numérique (formule (4)) se déduit facilement de la formule générale de l'ACP. Cette relation de transition permet de calculer la contribution à la note globale de chacun des mots, ainsi que de l'évaluation numérique.

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \left[\frac{1}{\lambda_1^1} \sum_{j \in J_c} f_{ij} G_s(j) + \frac{1}{\lambda_1^2} \sum_{j \in J_q} \left(\frac{x_{ij} - \mu_j}{\sigma_j} \right) G_s(j) \right] \quad (4)$$

6. Résultats

Dans l'exemple, le groupe quantitatif consiste en une seule colonne, ce qui ne pose pas de problème spécifique. Le tableau à analyser comporte 443 vins-ligne et 251 colonnes (250 mots-colonne et l'évaluation-colonne)

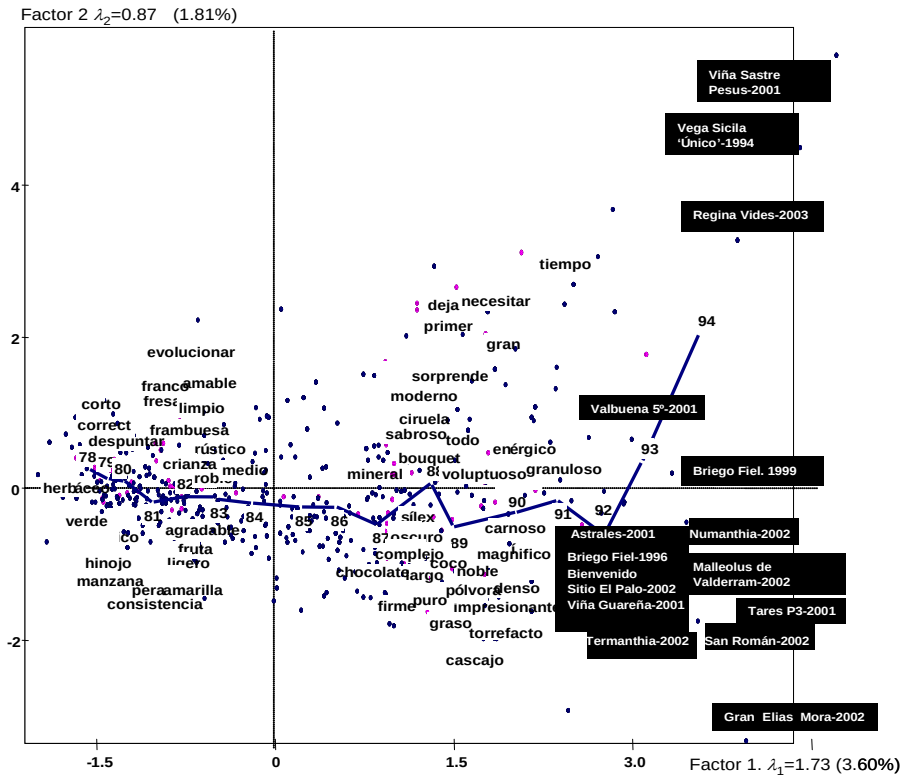

Figure 4. Premier plan factoriel issu de l'AFM appliquée au tableau juxtaposé (les notes 95 (un seul vin) et 97 (seulement 3 vins) ne sont pas représentées)

6.1. Note globale

Le premier facteur, dont l'échelle a été modifiée de façon à ce que le rang de variation des coordonnées soit 78-97, constitue la note globale.

$$Z(i) = 83.98 + F_1(i) \cdot 2.88$$

Les deux groupes de colonnes, les colonnes-mot, d'une part, et la colonne-évaluation, d'autre part, contribuent d'une manière très équilibrée à l'inertie du premier axe (avec, respectivement, 48,3% et 51,7%) et donc à la variance de la note globale.

La note *Z* est fortement corrélée avec l'évaluation numérique (0,95) et présente aussi une forte corrélation avec le premier facteur de l'analyse séparée du tableau lexical (0,66) : cette note constitue donc un bon compromis entre l'évaluation numérique et les commentaires libres. La figure 4 visualise le nuage des vins sur le premier plan factoriel. Les 15 vins les mieux positionnés (donc, ayant les 15 meilleures notes globales) présentent une évaluation numérique égale ou supérieure à 90 (avec une moyenne égale à 93,3). Parmi ces vins, 8 appartiennent à l'ensemble des 15 vins les mieux notés lorsque l'on considère seulement l'évaluation numérique initiale (section 4.1). On peut remarquer que les commentaires sur les meilleurs vins présentent une plus grande variation le long du deuxième axe, non prise en considération dans note globale étant donné que cette partie des commentaires n'a aucune corrélation avec l'évaluation numérique.

6.2. Contribution des deux groupes à la note globale

Dans cet exemple, la formule (4) adopte l'expression particulaire suivante :

$$F_s(i) = \frac{1}{\sqrt{1.73}} \left[\frac{1}{0.38} \sum_{j \in J_c} \frac{k_{ij}}{k} G_1(j) + \left(\frac{x_{ij} - \mu_j}{\sigma_j} \right) \cdot 0.95 \right] \quad (5)$$

où k_{ij} note la fréquence du mot j dans le commentaire sur le produit i et k la longueur de l'ensemble des commentaires (k=8315). La note globale peut s'interpréter comme la moyenne des deux évaluations partielles (qui correspondent aux deux parties de la formule 5). La présence dans le commentaire libre d'un mot de coordonnée (fortement) positive $G_1(j)$ sur le premier axe augmente (beaucoup) la note globale, tandis que la présence d'un mot ayant une coordonnée (fortement) négative diminue (beaucoup) cette note. Ainsi, les coordonnées des mots sur le premier axe peuvent s'interpréter comme la mesure de leur contribution à la note globale.

7. Conclusion

Les commentaires libres sont partie prenante de l'évaluation, destinée à apporter des nuances à au classement issu de l'évaluation numérique. La méthodologie décrite les intègre dans une note globale et fournit une série d'outils pour contraster et critiquer les résultats obtenus. Dans l'exemple "*vins de Castille et Léon*" la note globale ainsi construite est proche de l'évaluation numérique initiale mais est aussi fortement liée aux commentaires libres.

Software note

La méthodologie décrite dans ce travail peut être appliquée en utilisant l'étape AFM de SPAD 6.0 (2005), à condition de transformer les matrices et les poids de façon adéquate (SPAD-Groupe Test & Go, Paris, France. http://www.decisia.fr).

Références

Abdessemed L. and Escofier B. (1996). Analyse factorielle multiple de tableaux de fréquences **;** comparaison avec l'analyse canonique des correspondances. *Journal de la Société de Statistique de Paris*, vol.(137-2) : 3-18.

Benzécri J.-P. (1981). *Pratique de l'analyse des données.* T.3, Linguistique & Lexicologie, Dunod.

El Mundo (2005a). *Guía de catas 2005. Vinos de Castilla y León*. Biblioteca la Posada.
El Mundo (2005b). *Vinos de España. Catálogo de bodegas 2005*.
Escofier B. (2003). *Analyse des correspondances*. Dunod.
Escofier B. and Pagès J. (1998). *Analyses factorielles simples et multiples*. Dunod.
Lebart L. and Salem A. (1994). *Statistique textuelle*, Dunod.