Validité des visualisations de données textuelles

Ludovic Lebart

CNRS - ENST – 46, rue Barrault – 75013 Paris
lebart@enst.fr

Abstract

Recent trends of research show that exploratory techniques (principal axes methods and unsupervised clustering as well) are widely used and have great potential in the processing of textual data. However, the problem of the validation of the obtained results is seldom addressed. We will focus our paper on the use of meta-data for external validation, and on the assessments of visualizations. The examples of application concern the statistical processing of responses to open-ended questions in the context of an international survey.

Résumé

Les méthodes de statistiques exploratoires (analyses en axes principaux, classification et cartes de Kohonen), contribuent à la description des données textuelles, description venant d'ailleurs en complément, et non en substitution, des approches plus manuelles ou traditionnelles. On peut regretter que dans la pratique, les visualisations obtenues soient rarement accompagnées de procédures de validation. On évoque ici les méthodes de *validation externe*, utilisant la meta-information disponible, mais on insiste surtout sur les méthodes de *validation interne* qui font intervenir les techniques de rééchantillonnage (principalement : *bootstrap*). Les exemples d'application concernent des réponses à une question ouverte dans une enquête internationale.

Mots-clés : lexicométrie, analyse de données textuelles, validation, *bootstrap*.

1. Introduction

Les méthodes d'analyse descriptive des données multidimensionnelles qui constituent le cœur de la méthodologie de l'analyse statistique exploratoire des données se perfectionnent en s'efforçant d'utiliser d'une part la méta-information disponible (informations que l'on possède sur le tableau de données que l'on analyse, et qui ne figurent pas dans le tableau lui-même) d'autre part les techniques de validation modernes essentiellement fondées sur les méthodes de rééchantillonnage (Monte-Carlo, validation croisée, Bootstrap). On peut schématiser ces récents développements en disant qu'ils tentent de répondre aux deux questions :

– *Comment utiliser ce que l'on sait pour essayer d'en savoir plus ?*

– *Comment valider ou généraliser les résultats obtenus ?*

2. La complexité des données textuelles : point de vue d'un statisticien

Les concepts de variables et d'observations sont beaucoup plus complexes que ceux de la statistique usuelle. Montrons-le en six points.

2.1. *Des variables construites*

Les variables ne sont pas définies *a priori*, mais dérivées à partir des textes. On peut considérer comme variables (qualitatives) les formes graphiques, les lemmes, les segments, mais des

seuils de fréquences interviendront qui modifieront le nombre et la nature de ces variables, pour un même recueil de textes.

2.2. *Des unités statistiques diverses*

Les observations, ou individus peuvent être des documents, des abstracts ou des brevets, des répondants (dans une enquête comportant des questions ouvertes, par exemple), des paragraphes, des phrases, des discours, des auteurs… et, comme nous le verrons plus bas, plusieurs niveaux peuvent coexister dans un même recueil, ce qui ne manquera pas de poser des problèmes lorsqu'il s'agira de procéder à des inférence statistiques.

2.3. *Des tableaux creux*

Les données numériques résultantes seront très volumineuses, clairsemées (*sparse*), et très bruitées. Ce dernier terme, emprunté à la théorie du signal, signifie en bref que les structures qui nous intéressent ne constitueront qu'une faible part de la variabilité (variance) totale.

2.4. *Une méta-information considérable*

Il existe un volume important de méta-information, c'est-à-dire d'informations externes aux tableaux analysés. Pour un statisticien, la meta-information disponible dans le cas de données textuelles est quelque chose de démesuré, si on compare cette meta-information à celle des recueils de données numériques ou qualitatives habituels (fichiers d'enquête, par exemple). Chaque mot utilisé, même si c'est un mot grammatical (mot vide, ou encore mot-outil) a droit à plusieurs lignes, voire plusieurs pages dans un dictionnaire…

2.5. *Des processus, et non des vecteurs*

Le recueil de données est globalement séquentiel (analyse d'un très grand texte) ou localement séquentiel (analyse de documents, de réponses). Prenons le cas de la réponse d'un individu i à une question ouverte. Si $\{ s_1, s_2, \ldots , s_v \}$ désigne un ensemble de v éléments différents (le vocabulaire), un individu statistique i ($i \leq n$) pourra être caractérisé par une séquence ordonnée de longueur variable $\gamma(i)$: $\{ s_{r(i,1)}, s_{r(i,2)}, .. , s_{r(i,\gamma(i))}\}$, où : $1 \leq r(i,k) \leq v$, et $1 \leq k \leq \gamma(i)$. Un mot peut apparaître plusieurs fois dans une réponse. La quantité r(i, k) est alors l'index repérant le k-ième mot dans la réponse de l'individu i.

Calculer des similarités ou des distances entre séquences ordonnées de longueurs variables avec répétitions n'est pas une tâche courante de la statistique exploratoire multidimensionnelle. L'introduction d'unités composites, comme les segments, permet une prise en compte partielle de l'ordre des mots dans les phrases. On peut aussi utiliser les informations données par un arbre syntaxique (dont l'obtention ne peut être entièrement automatique) dans les calculs de similarités entre phrases ou réponses. Notons que des procédures d'analyse statistique multidimensionnelle, comme l'analyse des correspondances d'une table de contingence croisant des mots (en ligne) et des parties de textes (en colonne) ne font appels qu'à des modèles statistiques rudimentaires, mais peu contraignant. La méthode est une sorte de bilan global, qui ne prend pas en compte la nature séquentielle du texte, mais qui peut intégrer toutes les nouvelles variables construites après une analyse linguistique.

2.6. *Particularités statistiques des réponses aux questions ouvertes*

Dans les réponses aux questions ouvertes, le statut de la fréquence des mots est ambigu. Les fréquences lexicales observées sont pour une large partie artificielles, car la même question est parfois posée à des centaines ou des milliers de personnes... la juxtaposition des réponses constitue un « texte » redondant par construction. Il existe en fait deux niveaux d'individus

statistiques : les personnes interrogées, qui sont les individus habituels des statisticiens d'enquêtes, et les mots utilisés, qui sont, les individus les plus familiers pour les linguistes quantitatifs. Certains tests statistiques seront significatifs au niveau des comptages de mots, mais pas au niveau des comptages d'individus, et donc ne permettront pas une inférence des résultats à la population interrogée. Les analyses de réponses regroupées seront assez voisines des analyses de « vrais textes » (littéraires, politiques, historiques), alors que les analyses directes de réponses seront voisines des traitements effectués en recherche documentaire.

3. Les analyses statistiques ; les outils de base

3.1. *Analyse des correspondances et classification*

Les analyses des correspondances (A.C.) et les méthodes de classification peuvent décrire les tables de contingence croisant les réponses et les formes graphiques (Benzécri, 1981), ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser sous forme de séries de cartes planes (ou de dendrogrammes, ou de cartes auto-associatives de Kohonen) les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories. Il est toujours intéressant de comparer les analyses sur textes bruts et sur textes lemmatisés, analyses qui, selon les cas, se confirment ou se complètent.

Toutes ces visualisations graphiques, parce qu'elles sont souvent suggestives, sont habituellement produites sans critère de validation, dans un premier temps faute de procédures disponibles, puis, probablement, par absence de demande pressante de la part des utilisateurs.

3.2. *Un exemple : visualisation simultanée mots-catégories par cartes de Kohonen*

Nous illustreront cette brève évocation des outils usuels par un exemple de visualisation à l'aide de cartes de Kohonen (Kohonen, 1989) qui, appliquées au système de distances de l'A.C., permettent de représenter simultanément les lignes et les colonnes de la table lexicale.

Contexte de l'étude : une question ouverte a été posée lors d'une enquête multinationale (cf. Hayashi *et al*., 1992) avec le libellé suivant "*Que signifie pour vous la culture de votre pays?*". C'est le sous-échantillon relatif à la France qui est traité ici (effectif : 1009 personnes représentatives des personnes de 15 ans ou plus). L'ensemble des réponses à cette question représente 14742 occurrences de 2248 mots (formes graphiques) distincts. Sont traités ici les 111 mots apparaissant au moins 20 fois, qui représentent quand même 10005 occurrences, auxquels on a retranché 40 mots-outils ou mots vides usuels, ce qui porte à 71 le nombre définitif de mots gardés. Les réponses sont ici regroupées en neuf catégories obtenues par croisement de l'âge en trois classes (moins de 30 ans, 30-55 ans, plus de 55 ans) et du niveau d'éducation en trois classes (bas, moyen, haut).

La carte de Kohonen utilise ici le même système de distance que celui de l'analyse des correspondances de la table lexicale qui croise les occurrences des 71 mots les plus fréquents et les neuf catégories précédentes (les identificateurs des catégories sont abrégés). On obtient ainsi une synthèse assez comparable à celle de l'analyse des correspondances qui figure plus loin (figure 2). Cette synthèse nous permet de repérer assez rapidement des cooccurrences de mots dans une même catégorie, et des proximités lexicales entre catégories. Les premiers avantages de la représentation de la figure 1 dans le cas textuel concernent le format d'édition, composé de fenêtres rectangulaires particulièrement aptes à recevoir des listes de mots.

littérature grands art +55/HAUT	niveau esprit comme -30/HAUT	france	écrivains surtout histoire cinéma	peut grand	patrimoine
vivre faire arts 30-55/HA		langue	français chose		notre nos
tous musique	pense aussi	peu pays par française culture		enfants autres	même 30-55/MO
être trop gens +55/MOYE	très ont fait	monde	temps	beaucoup	mais
peinture	vie faut	sont nous bien	pour plus jeunes ils +55/BAS	pas	livres
théâtre quand moins cultivés -30/BAS	choses assez 30-55/BA	mal façon bonne		école télévision sais	sait rien lecture -30/MOYE

Figure 1. Représentation par carte de Kohonen des proximités entre mots et entre catégories

S'agissant d'une classification (avec contraintes, pour les classes, de figurer dans une grille), on peut penser que les regroupements prennent en compte des informations qui vont au-delà de celles données par les deux dimensions d'un premier plan factoriel.

Parmi les inconvénients de cette technique de visualisation, citons le caractère « non déterministe » (comme les méthodes de classification autour de centres mobiles ou *k-means* usuelles) des algorithmes qui produisent, dans ce cas particulier, des cartes différentes en fonction de l'initialisation aléatoire de départ. Enfin, il n'existe pas vraiment, dans les implémentations disponibles, de méthodes de validation permettant de décrire la confiance à accorder à la fois à la composition et aux positions relatives des différentes cases de la grille.

La seule validation possible, commune avec les méthodes de classification du type *k-means* ou nuées dynamiques, consiste à opérer divers démarrages aléatoires de l'algorithme, et à observer visuellement la stabilité des résultats. De simples clics de souris génèrent instantanément de nouvelles cartes, qui peuvent différer à la fois par les regroupements à l'intérieur des cellules et par les positions relatives des cellules.

4. La validation externe pour les analyses « en axes principaux »

Le principe de la validation externe consiste à confronter le corpus, ainsi que toute structure déduite exclusivement du corpus, avec des informations externes concernant des composantes du corpus. Un exemple courant est fourni par, d'une part un corpus de réponses à une question ouverte, et d'autre part un tableau des caractéristiques des répondants à la question.

4.1. *Les valeurs-test*

L'information externe se manifeste souvent sous forme de « variables supplémentaires » (ou « variables illustratives » qui sont positionnées (en fait : projetées) *a posteriori* sur les axes principaux. Pour chaque axe et chaque variable, une « valeur-test » exprime la signification statistique de la coordonnée de la variable sur l'axe (sous l'hypothèse d'indépendance de la variable et des variables actives de l'analyse).

Supposons qu'une modalité supplémentaire j concerne k_j individus (répondants d'une enquête, documents, etc.). Si ces k_j individus sont tirés au hasard (hypothèse nulle H_0) parmi les k individus analysés (tirage supposé sans remise), la moyenne de k_j coordonnées $\psi_{\alpha i}$ tirées au hasard dans l'ensemble fini des k valeurs est une variable aléatoire $X_{\alpha j}$, avec pour espérance (moyenne) :E $(X_{\alpha j})= 0$,(origine de l'axe α au point moyen) et pour écart-type : $v(j, k_j)$.

Le carré de cet écart-type vaut $v(j,k_j)^2 = \frac{1}{k_j}\left(\frac{k-1}{k-k_j}\right)\lambda_\alpha$ (*cf.*, *e.g.*, Lebart *et al.*, 1998).

Alors, $t_\alpha(j) = \varphi_{\alpha j} / v(j, k_j)$ est une variable centrée réduite ($\varphi_{\alpha j}$ est la coordonnées du point-modalité j sur l'axe α). La quantité $t_\alpha(j)$ mesure en nombre d'écart-types la distance entre la modalité j, c'est-à-dire le quasi-barycentre des k_j individus, et l'origine sur l'axe factoriel α. On appelle cette quantité « valeur-test ». On montre que sa distribution tend vers une loi normale centrée réduite lorsque k augmente indéfiniment. La valeur-test est un critère qui permet d'apprécier rapidement si une modalité a une position « significative » sur un axe. On considère comme occupant une « position significative » les modalités dont les valeurs-test sont supérieures à *2* en valeur absolue, correspondant approximativement au seuil *5 %*. On note que les valeurs-test n'ont de sens que pour les modalités supplémentaires.

4.2. *Le problème des comparaisons multiples*

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples,* bien connu des statisticiens. Supposons que l'on projette *100* modalités supplémentaires qui soient vraiment tirées au hasard. Les valeurs-test attachées à ces modalités sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes. Dans ces conditions, en moyenne, sur *100* valeurs-test calculées, *5* seront en dehors de l'intervalle [*-1.96* , *+1.96*], et *5* dépasseront la valeur *1.65* (test unilatéral). Le seuil de *5 %* n'a de sens en fait que pour un seul test, et non pour des tests multiples. On résout de façon pragmatique cette difficulté en choisissant un seuil plus sévère. Plusieurs solutions ont été proposées pour ce faire : la méthode de *Bonferroni* recommande de diviser le seuil statistique par le nombre de tests (ici, le nombre de valeurs-test calculées). Cette réduction du seuil est souvent considérée comme étant trop sévère (Hochberg, 1988 ; Perneger, 1998). Des études approfondies de ce problème se trouvent dans Hsu (1996), Westfall et Young (1993).

5. Validation interne : le re-échantillonnage.

Les visualisations provenant des analyses en axes principaux [décomposition aux valeurs singulières (SVD), composantes principales (ACP), correspondances (AC)] n'ont de sens que si elles sont accompagnées de la confiance que l'on peut accorder à la position de chaque point. Or ces visualisations sont d'une grande complexité analytique, et il est exclu de procéder à des évaluations selon les méthodes de la statistique classique.

5.1. *Principe des zones de confiance "bootstrap"*

La technique de *bootstrap* (cf. Efron et Tibshirani, 1993) va permettre de tracer des zones de confiance (ellipses ou enveloppes convexes de réplications) autour des points représentés sur les plans principaux, que ces points représentent des mots ou des textes. La méthode consiste à construire *n* "réplications" de l'échantillon par tirage *avec remise* des unités statistiques. Dans une réplication, certaines unités apparaîtront ainsi deux fois ou plus, d'autres n'apparaîtront pas. On crée ainsi une variabilité autour du tableau de données de départ. Sur chacune des *n* réplications, on peut faire les calculs les plus complexes, comme ceux que nécessitent les techniques de visualisation. Sous des hypothèses faibles, on montre que la variabilité observée sur les *n* réplications est de l'ordre de grandeur de celle que l'on aurait observée dans la population. Autrement dit, on peut disposer de *n* réplications de paramètres complexes, et donc obtenir des intervalles de confiance pour ces paramètres. Cette propriété est valable pour les vecteurs propres, les réplications des valeurs propres étant biaisées.

5.2. *Cas des composantes principales et de la SVD*

Dans le cas des composantes principales, on relève plusieurs variantes de la méthode : le *bootstrap total* consiste à refaire une analyse complète pour chaque réplication. Cette procédure engendre une difficulté : les axes répliqués ne sont pas forcément homologues d'un réplication à une autre, il peut y avoir des interversions d'axes, voir des rotations. Il faut alors faire coïncider par des techniques d'*analyses procustéennes* les axes homologues (Markus, 1994 ; Milan et Whittaker, 1995). Le *bootstrap partiel* permet de lever cette difficulté. Il part de la constatation que le tableau initial est plus proche de la réalité observée que tous les tableaux répliqués, qui en sont des perturbations. L'analyse et les plans principaux de ce tableau initial servent de référence pour la projection de tous les tableaux répliqués (lignes et colonnes) en tant qu'éléments supplémentaires. Des expérience intensives (cf. Lebart *et al.*, 2003) ont montré l'efficacité de cette méthode.

5.3. *Cas de l'analyse des correspondances*

Les travaux de pionniers furent ceux de Gifi (1981), Greenacre (1984). Dans le cas d'une table lexicale, la technique consiste à tirer avec remise les occurrences de formes retenues. Ce tirage se fait selon un schéma multinomial comportant autant de catégories que la table a de cellules, et dont les fréquences théoriques sont celles des cellules. Les lignes et colonnes des tables répliquées sont alors projetées comme éléments supplémentaires sur les plans factoriels de l'analyse de la vraie table lexicale (*bootstrap partiel*). Des ACP des « nuages de réplications » correspondant à chaque point-élément (ligne ou colonne) fournissent les ellipses de confiance cherchées. Dans le cas d'un *bootstrap total*, on retrouve les difficultés déjà mentionnées pour les valeurs propres (voir, par exemple, Alvarez *et al*., 2002).

5.4. *Un exemple de zones de confiance*

La figure 2 représente de telles zones de confiances dans le premier plan factoriel de l'AC de la table lexicale (déjà utilisée section 3.2) qui croise les occurrences des 71 mots et les 9 catégories précédentes. Comme il n'est pas possible de représenter sur un même graphique toutes les ellipses, on a choisi quatre catégories (encadrées) en excluant les âges et les niveaux d'éducation moyens. On a de même choisi cinq mots (formes graphiques, en fait) de façon à recouvrir les différentes zones du graphique *(arts, chose, choses, cinéma, télévision*). Remarquons que les zones de confiance des catégories sont relativement petites, indiquant des positions (et des positions relatives) assez significatives pour ces catégories (la plus grande concerne les jeunes peu instruits, catégorie dont l'effectif est le plus faible (13 réponses dans l'échantillon). On voit qu'à la mention du mot *télévision* correspond une longue ellipse de

confiance, qui fait osciller la position du point entre les catégories –30/BAS (jeunes peu instruits) et +55/BAS (seniors peu instruits) mais reste du côté de ces classes de niveaux d'éducation. On note le comportement inverse du mot *arts*. Enfin, on a représenté les positions de deux formes graphiques (*chose* et *choses*) du même lemme *chose*, pour montrer qu'il peut arriver que les positions de telles formes soient significativement distinctes (les ellipses ne se recouvrent pas) (il semble que *chose* au singulier soit souvent associé à la réponse *pas grand chose*, qui exprime une opinion critique ou désabusée vis-à-vis de la culture). Comme souvent, la forme graphique brute contient des informations de contexte. Il est intéressant de noter que l'augmentation du nombre de mots consécutive au choix d'un seuil plus bas a en général pour effet de diminuer la taille des ellipses relatives aux catégories, malgré la prise en compte de données évidemment plus bruitées (mots plus rares).

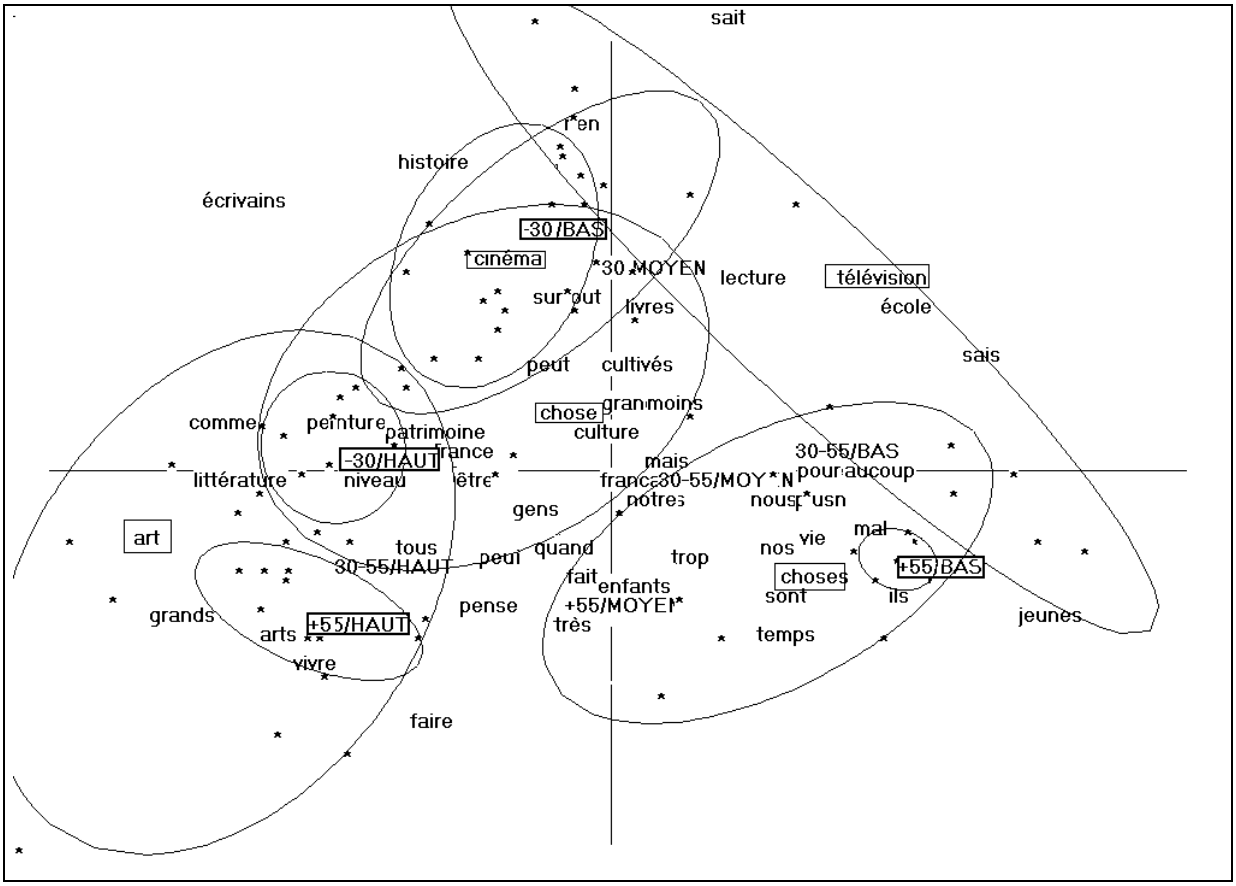

Figure 2. Zones de confiance pour 4 catégories de répondants et pour 5 mots.

5.5. *Les différentes unités statistiques*

Nous avons vu, section 2.2, qu'il existait plusieurs niveaux d'unités statistiques. Un exemple est donné par les questions ouvertes : l'individu interrogé constitue l'unité classique des statisticiens, alors que l'occurrence des mots est celle des lexicomètres (cf. Tuzzi et Tweedie, 2000). La souplesse de la procédure de bootstrap permet d'adapter les réplications aux deux préoccupations (tirage avec remise des répondants, ou des occurrences de mots), et obtenir ainsi des zones de confiance adaptées aux deux hypothèses de départ. On retrouve la même situation en *Web-log mining*, (mots, pages, sites) ou encore en recherche documentaire (mots, titres, abstracts, textes complets).

6. Conclusion

Jusqu'à une date récente, les matériaux textuels, complexes, volumineux, riches en informations externes, étaient mal adaptés aux formalisations et aux épreuves de validité de la statistique classique. Les techniques de ré-échantillonnage fournissent un outil de validation souple, permettant de transformer une visualisation plaisante en un document scientifique.

[Les calculs et figures de cet article ont été réalisés avec le logiciel académique DTM]

References

Alvarez R., Bécue M., Lanero J.J. et Valencia O. (2002). Results stability in Textual Analysis: its Application to the Study of the Spanish Investiture Speeches (1979-2000). In *Actes des JADT 2002* : 1-12.

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

Efron B. et Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Efron B. (1979). Bootstraps methods : another look at the Jackknife. *Ann. Statist*., vol. (7) : 1-26.

Gifi A. (1981). *Non Linear Multivariate Analysis*. Department of Data theory, University of Leiden. (Nouvelle version, même titre : Wiley J., Chichester, 1990).

Greenacre M. (1984). *Theory and Applications of Correspondence Analysis*. Academic press.

Hayashi C., Suzuki T. et Sasaki M. (1992). *Data Analysis for Social Comparative Research: International Perspective*. North-Holland.

Hochberg Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, vol. (75) : 800-803.

Hsu J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall.

Kohonen T. (1989). *Self-Organization and Associative Memory*. Springer Verlag.

Lebart L., Piron M. et Steiner J.-F. (2003). *La Sémiométrie*. Dunod.

Lebart L., Salem A. et Berry E. (1998). *Exploring Textual Data*. Kluwer Ac. Publisher.

Markus M.Th. (1994). Bootstrap Confidence Regions for Homogeneity Analysis; the Influence of Rotation on Coverage Percentages. In Dutter R. et Grossmann W. (Eds), *COMPSTAT 1994*. Physica Verlag : 337-342.

Milan L. et Whittaker J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Appl. Statist*., vol. (44) : 131-149.

Perneger T.V. (1998). What is wrong with Bonferroni adjustments. *British Medical Journal*, vol. (136) : 1236-1238.

Tuzzi A. et Tweedie F.J. (2000). The best of both worlds: Comparing Mocar and Mcdisp. In *Actes des JADT 2000* : 271-276.

Westfall P.H. et Young S.S. (1993*). Resampling Based Multiple Testing: Examples and Methods for p-values Adjustment.* Wiley.