STRATEGIES DU TRAITEMENT DES DONNEES D'ENQUETES.

Ludovic LEBART
CNRS / Télécom Paris
Dept SES, 46 Rue Barrault
75013 PARIS

Résumé :

Les techniques de traitement des données d'enquêtes ont été profondément modifiées par l'analyse des données (principalement ici: analyse en composantes principales, des correspondances simples et multiples, classification) qui intervient dans une phase préliminaire pour apprécier la qualité de l'information, et orienter la suite des traitements.

Mots-clés :

Enquêtes, Analyse des données

Ce texte tente de résumer la démarche du statisticien lors du dépouillement d'une enquête sur ordinateur avec les outils logiciels maintenant classiques de l'analyse des données. Ce sont des logiciels de ce type que la bibliothèque MODULAD a pour ambition de diffuser très largement auprès des ingénieurs d'études, des universitaires, des étudiants, en mettant à leur disposition une boite à outils adaptable et évolutive.
Les utilisateurs peuvent aussi avoir recours à des logiciels analogues, mais plus spécialisés et intégrés comme SPAD.N ou SICLA. Les enquêtes concernées sont les enquêtes socio-économiques (consommation, marketing, attitudes et opinions) auprès des individus, des ménages ou des entreprises, les enquêtes de type épidémiologique, les campagnes de mesures diverses.

Le dépouillement d'enquête traditionnel met en œuvre des techniques simples, éprouvées, faciles à interpréter: les tris, les tableaux croisés, c'est-à-dire des calculs de pourcentages et de moyennes:

calculs des pourcentages d'individus pour chaque *modalité* d'une *variable nominale* (ces pourcentages seront calculés par rapport à l'échantillon global, mais aussi par rapport à des sous-échantillons),

calculs des moyennes de variables *numériques* ou *quantitatives* qui peuvent être ventilées selon les catégories d'une ou de plusieurs variables nominales,

calculs de corrélations, mesurant l'intensité du lien entre deux variables numériques.

Enfin, des méthodes statistiques plus élaborées viennent parfois compléter ces premiers résultats: régressions, analyses de la variance ou de la covariance, modèles log-lineaires.

Les techniques d'analyse des données (analyses descriptives multidimensionnelles) modifient profondément les premières phases du traitement des données d'enquête : ces techniques ne sont pas des compléments sophistiqués intervenant à la suite des méthodes traditionnelles. Elles vont en fait bouleverser l'enchaînement des tâches, et définir une méthodologie nouvelle.

Dans le cadre de cette méthodologie, les étapes du traitement des données d'enquêtes sont, brièvement, les suivantes:

1) Descriptions élémentaires (tri-à-plat, histogrammes, calculs de statistiques élémentaires, moyennes, écart-types, valeurs extrêmes, quantiles) Retour éventuel aux données de base pour resaisie partielle ou corrections.

2) **Epreuves de cohérence globale par thème, Epreuves d'hypothèses larges.** (Par hypothèses larges, on entend: hypothèses générales permises par les nouveaux outils de description). Structuration des données, typologies, sélection de tableaux croisés. Ces épreuves font appel à un *"enchaînement canonique de méthodes"* que nous décrirons sous le nom de *"thémascope"*, afin d'attirer l'attention du lecteur sur le fait qu'il s'agit d'un instrument d'observation par thème.

3) Epreuves d'hypothèses classiques. (Tests statistiques usuels, Regression, discrimination, analyses de la variance, modèles log-linéaires...)

4) Conclusions : Critique de l'information de base : lacunes dans le choix des variables, déséquilibre de l'échantillon ou du champs d'observation, biais ou erreurs. Choix de modèles, énoncés des résultats, rejets d'hypothèses, suggestions de nouvelles hypothèses.

En fait, c'est la phase 2 qui est relativement nouvelle, et qui est absente de la plupart des logiciels classiques (anglo-saxons). C'est donc de cette phase que nous allons parler.

Lors de cette phase, la cohérence globale du recueil de données peut en effet être éprouvée de façon systématique, des panoramas globaux peuvent être dressés, permettant de critiquer l'information, mais aussi d'orienter la suite des traitements, de choisir les tableaux croisés les plus pertinents.
Les typologies (classification des individus en prenant en compte simultanément plusieurs réponses ou plusieurs caractéristiques de base), les outils de visualisation (plans factoriels) fournissent de nouveaux matériaux d'analyse.

Ces opérations, intervenant **au début de la chaîne de traitement**, permettent de piloter la suite du dépouillement de l'enquête. Le choix des modèles n'est plus fait de façon aveugle en fonction des hypothèses de base: ces hypothèses pourront souvent être critiquées, d'autres hypothèses pourront être suggérées.

1-Les outils élémentaires: analyse factorielle et classification.

Rappelons tout d'abord les principes communs à toutes les méthodes de statistique descriptive multidimensionnelle :
Chacune des deux dimensions d'un tableau rectangulaire de données numériques permet de définir des distances (ou des proximités) entre les éléments définissant l'autre dimension: ainsi, l'ensemble des colonnes (variables, attributs, mesures) permet de définir à l'aide de formules appropriées des distances entre lignes (individus, observations).
De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes. On obtient ainsi des tableaux de distances, qui sont associées à des représentations géométriques complexes.

Il s'agit de rendre assimilable et accessible à l'intuition ces représentations, au prix d'une perte d'information de base qui doit rester la plus petite possible. Rappelons aussi qu'il existe deux familles de méthodes qui permettent d'effectuer ces réductions:

Les méthodes factorielles (principalement analyses en composantes principales et des correspondances simples et multiples) qui produisent des représentations graphiques sur lesquelles les proximités géométriques usuelles entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes.

Les méthodes de classification qui opèrent des regroupements en classes (ou en familles de classes hiérarchisées) des lignes et des colonnes.

Ces deux familles de méthodes vont être utilisées de façon complémentaire pour décrire de la façon la plus exhaustive possible les grands tableaux numériques constitués par les données d'enquêtes.

Les règles d'interprétation des représentations obtenues à l'issue de ces techniques de réduction n'ont pas la simplicité de celles de la statistique descriptive élémentaire.
L'interprétation des histogrammes, des "camenberts", des graphiques de séries chronologiques est intuitive, alors que dans le cas de l'analyse des correspondances, par exemple, il sera nécessaire de connaître des règles de lectures des résultats plus difficile que ne le laisse croire le caractère souvent suggestif des représentations obtenues. Une formation et une expérience pratique s'avéreront nécessaires.

2- Le modèle de base: Variables actives et illustratives

Travailler au niveau des thèmes, et plus seulement des variables.

L'analyse des correspondances et l'analyse en composantes principales, permettent de trouver des sous-espaces de représentation des proximités entre profils ou entre vecteurs de description

d'observations. Mais elles permettent aussi de positionner dans ce sous-espace des lignes ou des colonnes supplémentaires du tableau de données.

On peut ainsi illustrer les plans factoriels par des informations n'ayant pas participé à la construction de ces plans, ce qui va avoir des conséquences très importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelés éléments actifs ou variables actives: ils doivent former un ensemble homogène pour que les distances entre individus ou observations s'interprêtent facilement. Ils sont en général relatifs à un même thème de l'enquête.

Cette dichotomie entre variables actives et variables illustratives est fondamentale.
Elle est du même ordre que la distinction que l'on établit entre variables endogènes (à expliquer) et exogènes (explicatives) dans les modèles de régression multiple. D'un point de vue géométrique, les deux situations sont d'ailleurs très similaires. Les variables exogènes engendrent un sous-espace sur lequel seront projetées les variables endogènes. Les variables actives engendrent aussi un sous-espace, que l'on va réduire pour le visualiser. C'est sur cet espace réduit que l'on va projeter les variables illustratives.

3- Analyse des Correspondances Multiples.

L'analyse des correspondances multiples fait l'objet d'une mention particulière en raison de l'étendue de son champ d'application.
Elle permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple privilégié: les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers) ; les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions.
Il s'agit en fait d'une extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques.

Cette extension se fonde sur l'équivalence suivante : On dispose pour k individus des valeurs (réponses) prises par deux variables nominales ayant respectivement n et p modalités.
Il est alors équivalent de soumettre à l'analyse des correspondances la table de contingence (n,p) croisant les deux variables, ou d'analyser le tableau binaire Z à k lignes et (n+p) colonnes qui décrit les réponses (chaque ligne de **Z** comporte deux "1" dans les colonnes correspondant aux réponses choisies, et n+p-2 "0").

L'analyse de **Z** est plus coûteuse, mais plus intéressante, car elle se généralise immédiatement au cas de plus de deux variables. Le tableau "disjonctif complet" **Z** comporte q blocs s'il y a q variables. Les réponses des individus peuvent alors être codifiées dans un tableau réduit **R,** qui contient simplement les numéros des modalités. Les procédures de calcul n'utiliseront en fait que ce tableau peu encombrant.

Le *tableau de Burt* **B** est le produit du tableau disjonctif **Z** par son transposé:
Il contient q^2 blocs: les blocs diagonaux sont des matrices diagonales, dont les éléments diagonaux sont les effectifs de réponses correspondant à chaque modalité. Les trois blocs distincts parmi les blocs restant ne sont autres que les trois tables de contingence croisant les trois variables deux à deux.

L'analyse des correspondances du tableau **Z** donne les mêmes axes factoriels normés que celle du tableau $\mathbf{B} = \mathbf{Z'Z}$, c'est-à-dire finalement les mêmes graphiques de proximités, aux échelles près; mais les valeurs propres homologues sont distinctes: à la valeur propre λ de l'analyse de **Z** correspond la valeur propre λ^2 de l'analyse de **Z'Z**.

Dans le cas de deux variables, le tableau **R** n'a que deux colonnes, le tableau **Z** est formé de deux blocs, et **B** en comporte quatre.

Les deux analyses précédentes donnent également la même représentation que l'analyse des correspondances de la petite table de contingence croisant les deux variables (bloc non-diagonal de **B**), avec, cette fois, la valeur propre $(2\lambda - 1)^2$.

Validité de la représentation

On évite en analyse des correspondances multiples d'utiliser les pourcentages de variance pour caractériser les axes: ceux-ci n'ont pas le même sens que lorsqu'il s'agit d'une table de contingence: le codage binaire introduit un "bruit" qui réduit la part d'explication attachée à chaque valeur propre.
En pratique, la stabilité du plan factoriel s'éprouve par des techniques de simulation (pertubations aléatoires du tableau de données) ou de validation par échantillon-test.

Positionnement des variables illustratives

La représentation simultanée des lignes et des colonnes liée à l' analyse des correspondances n'est pas toujours utilisée. Les points-lignes ne sont en général pas tracés pour des raisons d'encombrement graphique, mais surtout parce que les individus sont dans la plupart des applications anonymes.... ils ne présentent de l'intérêt que par l'intermédiaire de leurs caractéristiques. Ce sont précisément les autres informations disponibles sur les individus ou observations qui vont être projetées en éléments illustratifs.
Le fait que le tableau **Z** ne comporte que des "0" et des "1" donne une interprétation particulièrement simple de la disposition relative des colonnes illustratives: les coordonnées des colonnes (réponse illustratives j1,j2,...) sont proportionnelles aux moyennes arithmétiques des coordonnées des individus qui ont choisi les réponses j1,j2,....

4- Complémentarité de la Classification

Dans le cas du traitement statistique des fichiers d'enquêtes en vraie grandeur, la démarche précédente fondée sur des représentations graphiques a deux graves inconvénients:

1) Les visualisations sont limitées à deux , ou en général à très peu de dimensions, alors que le nombre d'axes significatifs peut souvent atteindre 8 ou 10, pour fixer les idées...

2) Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles.

Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats.

L'utilisation conjointe de la classification automatique et des analyses précédentes permet de remédier à ces lacunes. Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes.

Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à dix dimensions.

Autrement dit, l'opération va présenter un double intérêt: **allègement des sorties graphiques** d'une part, **prise en compte de la dimension réelle du nuage de points** d'autre part.

Une fois les individus regroupés en classes, il est facile d'obtenir une description automatique de ces classes: on peut en effet, pour les variables numériques comme pour les variables nominales, calculer des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales; on peut également convertir ces statistiques en **valeurs-test** et opérer un tri sur ces valeurs-test. On obtient finalement, pour chaque classe, les modalités et les variables les plus caractéristiques.

Les algorithmes de classification

L'algorithme de classification qui nous paraît actuellement le plus adapté au partitionnement d'un ensemble comprenant des milliers d'individu est un *algorithme mixte* procédant en trois phases:

4-1 ***Partitionnement initial*** en quelques dizaines de classes (par une technique du type "nuées dynamiques" ou "k-means");

On résumera de façon très sommaire ces techniques en quelques phrases. On commence par tirer au hasard des individus qui seront des centres provisoires de classes. Puis, on affecte tous les individus au centre provisoire le plus proche (proche au sens d'une distance telle que la distance du Chi-2). On construit ainsi une partition de l'ensemble des individus. On calcule de nouveau des centres provisoires, qui sont maintenant des "centres" (points moyens par exemple) des classes qui viennent d'être obtenues, et on itère le processus, autrement dit on affecte de nouveau tous les individus à ces centres, ce qui induit une nouvelle partition, etc... le processus se stabilise nécessairement, mais la partition obtenue dépendra en général du choix initial des centres.

4-2 ***Agrégation des classes obtenues;***
L'agrégation hiérarchique est assez coûteuse si elle s'applique à des milliers d'individus, c'est pourquoi il faut réduire la dimension du problème en opérant un regroupement préalable en quelques dizaines de classes.
Le principe de l'agrégation hiérarchique est également simple: partant de n individus, on agrège les deux individus les plus proches, et l'on considère le couple agrégé comme un nouvel individu (ce qui peut être fait de différentes façons, d'où un nombre important de "variantes" de cette technique); on est ramené au problème précédent avec maintenant (n-1) individus... on agrège à nouveau les deux plus proches, et l'on itère le processus jusqu'à n'avoir plus qu'un seul individu, qui sera le sommet de "l'arbre hiérarchique". L'interêt de cet arbre (ou dendrogramme) est qu'il peut donner une idée du nombre de classes existant effectivement dans la population. Chaque coupure d'un tel arbre va donner une partition, ayant d'autant moins de classes que l'on coupe près du sommet.

4-3 ***Choix du nombre de classes par coupure de l'arbre*** (en général après une inspection visuelle) et optimisation de la partition obtenue par réaffectations.

Plus on agrège de points, autrement dit plus on se rapproche du sommet de l'arbre, plus la distance entre les deux classes les plus proches est grande. On peut "indexer" les "noeuds" de l'arbre par cette plus petite distance de façon à ce que de grandes branches correspondent à des augmentations importantes de cette distance. En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité, car les individus regroupés auparavant étaient proches, et ceux regroupés après la coupure seront nécessairement éloignés, ce qui est la définition d'une bonne partition...

Mais une telle partition, en k classes par exemple, n'est pas la meilleure possible, car l'algorithme de classification hiérarchique n'a malheureusement pas la propriété de donner à chaque étape une partition optimale.
On peut encore améliorer la partition obtenue par réaffectation des individus. Malgré la relative complexité de la procédure, on ne peut être assuré d'avoir trouvé la "meilleure partition en k classes".

4-4 ***Description automatique des classes***

Un tableau récapitulatif décrira les classes de façon précise, en comparant les pourcentages de réponses internes aux classes aux pourcentages globaux, puis en sélectionnant les modalités les plus caractéristiques

Parmi les modalités les plus caractéristiques d'une classe figurent également des modalités illustratives, qui n'ont pas participé à la fabrication des classes.

Pour chacune des classes, les modalités les plus caractéristiques sont rangées suivant les valeurs décroissantes d'un critère. Ce critère, ou valeur-test, est, brièvement, l'analogue de la valeur absolue d'une variable normale centrée réduite, qui est significative au seuil 5% si elle dépasse la valeur 1.96. On voit assez bien les avantages que présentent les partitions pour décrire des ensembles multidimensionnels:

La notion de classe est intuitive (groupes d'individus les plus semblables possibles). La description des classes fait appel à des classements de libellés complets et donc facile à lire, ces classements étant fondés sur des simples comparaisons de pourcentages. *Il est donc plus facile de décrire des classes qu'un espace continu.*

Mais c'est l'analyse des correspondances qui permet de visualiser les positions relatives des classes dans l'espace, et aussi de mettre en évidence certaines variations continues ou dérives dans cet espace qui auraient pu être masquées par la discontinuité des classes.

Les deux techniques sont donc complémentaires, et se valident mutuellement.

5- Le themascope

L'"enchaînement canonique de méthodes" annoncé sous le nom de *thémascope* comprend en définitive les étapes suivantes:

1) Choix d'un thème, c'est-à-dire d'une batterie homogène d' éléments actifs. Ce thème définira un *point de vue* pour la description. On peut décrire les individus du point de vue de leurs caractéristiques de base, mais aussi à partir d'un thème particulier: habitudes de consommation, opinions politiques, etc.....).

2) Etablissement d'une typologie bidimensionnelle de la population à partir des variables actives (Analyse des correspondances simples ou multiples ou analyse en composantes principales, selon la nature des éléments actifs).

3) Positionnement des éléments illustratifs. On projettera ainsi toute l'information disponible susceptible d'aider à comprendre ou à interpréter la typologie induite par les éléments actifs. En fait, c'est l'ordinateur lui même qui sélectionnera les variables supplémentaires ayant des coordonnées significatives sur les axes factoriels, ce qui permet d'envisager des explorations systématiques, avec de nombreux croisements de variables.

4) Partition de l'ensemble des individus ou observations, en utilisant par exemple la procédure mixte suggérée plus haut.

5) Positionnement sur les graphiques précédents des centres des principales classes (une partition définit toujours une variable nominale particulière, dont les modalités sont les réponses à la question :"à quelle classe appartenez-vous?"). Ces modalités peuvent donc elles aussi être projetées en éléments illustratifs sur les plans factoriels.

6) Description systématique des classes obtenues par les modalités et les variables les plus caractéristiques.

En somme, cet enchaînement décrit un thème (multidimensionnel par nature) par la conjonction des deux techniques disponibles (Réduction de dimension d'une part, regroupement d'autre part) et plonge ce thème dans le contexte général de l'enquête, grâce

aux deux techniques de projection de variables illustratives et de description automatique des classes.
La sélection des éléments les plus significatifs sur les plans factoriels et lors de la description des classes est faite automatiquement (elle est pilotée par des seuils). Les éléments sélectionnés peuvent être décrits par des libellés longs et explicite. Le lecteur des résultats dispose donc d'une information filtrée et parfaitement lisible.

6- Applications aux premières phases du traitement des enquêtes

Structure de base d'un échantillon et sélection visuelle des tableaux croisés.

Cette application illustre une utilisation assez courante de l'analyse des correspondances multiples pour le traitement des données d'enquêtes. Prenons l'exemple d'une enquête national représentative. Etant donnée la structure actuelle de la population, les caractéristiques de base (sexe, niveau de vie, statut matrimonial, niveau d'instruction....) ne sont pas indépendantes, d'où l'idée de décrire le réseau d'interrelations entre toutes ces caractéristiques de base, puis de positionner les autres thèmes de l'enquête en tant qu'éléments illustratifs.

Les caractéristiques des personnes qui répondent sont alors visibles immédiatement dans un cadre qui tient compte des interrelations existant entre ces caractéristiques. Les consultations classiques (sans visualisation factorielle préalable) de tableaux croisés sont en effet hypothéquées par le fait "qu'une variable peut en cacher une autre"..... elles sont de plus largement redondantes lorsque les caractéristiques successives sont liées entre elles. Le système de projection de variables supplémentaires permet donc d'économiser du temps et d'éviter des erreurs d'interprétation. Chaque variable illustrative fournit une information qui ne pourrait être acquise que par la lecture de dizaines de tableaux croisés.

Les Situations-types ou Noyaux factuels

On veut maintenant obtenir un petit nombre de groupes d'individus les plus homogènes possible vis-à-vis de leurs caractéristiques de base.
L'idée en est brièvement la suivante: on aimerait croiser des caractéristiques tels que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout-à-fait comparables entre eux du point de vue de leur situation objective (réaliser, autant que faire se peut, le "toutes choses égales par ailleurs"). Mais de tels croisements conduisent vite à des dizaines de milliers de modalités, dont on ne sait que faire lorsqu'on étudie un échantillon lui-même de l'ordre de quelques milliers d'individus.

De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre ces caractéristiques: certaines sont évidentes a priori (il n'y a pas de "moins de 30 ans" retraités, d'hommes enceintes), d'autres sont également connues a priori, mais peuvent souffrir des exceptions (il y a peu d'étudiants veufs), d'autres enfin ont un caractère plus statistique (il y a plus de femmes dans les catégories employés et veufs).
Une classification des individus décrits par la batterie active des caractéristiques de base va permettre de regrouper les individus ayant, dans l'échantillon, le maximum de caractéristiques en commun. En pratique, elle fournira des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2000 individus. De telles classes sont appellées situations-types ou encore noyaux factuels.
Le tableau croisant une des variables nominales de l'enquête avec la partition en noyaux factuels résume pratiquement tous les tableaux obtenus en croisant cette même variable avec chacune des caractéristiques de base. De plus, certaines interactions indécelables à partir de ces tableaux binaires peuvent être détectés.

Les analyses par thème

Ce qui vient d'être dit pour le thème "structure de base" est vrai pour n'importe quel thème de l'enquête. Mais alors que l'analyse de la structure de base, peu intéressante en elle-même, est

surtout un lieu d'accueil des thèmes de l'enquête (comme éléments illustratifs), les analyses par thème pourront mettre en évidence des dimensions sous-jacente inconnues ou des regroupements d'individus ou d'observations originaux. Bien entendu, les informations de base et les autres thèmes pourront illustrer avec profit ces dimensions ou ces groupements.

Conjectures sur les non-réponses

Les non-réponses sont des modalités comme les autres, qui peuvent être positionnées dans les espaces des thèmes, comme dans les espaces de la structure de base.
Le traitement des non-réponses qui se prête mal aux tests statistiques usuels reçoit ici une importante contribution, dans la mesure où l'on peut étudier le contexte de ces refus ou lacunes, soit en termes de caractéristiques des répondants, soit à partir des réponses effectives à d'autres thèmes.

Le positionnement des variables techniques

Dans l'espace des caractéristiques de base ou dans celui des principaux thèmes, que ceux-ci soient résumés par un plan factoriel ou par une partition en noyaux factuels, il est possible de placer les modalités de variables nominales dites "techniques" telles que: numéro ou nom de l'enquêteur, caractéristiques diverses de l'enquêteur, heure de l'interview, lieu et durée de l'interview, appréciation de l'enquêteur ou de l'enquêté sur l'interview, etc....
On obtient ainsi un panorama de la fabrication de l'information, permettant de rapprocher globalement les circonstances des interviews et les caractéristiques des personnes interrogées. Cette confrontation permet souvent d'apprécier la validité des données de base et de nuancer l'interprétation des résultats.

Bibliographie sommaire

Benzécri J-P. & coll. (1973) - *La Taxinomie*, Vol. I, *L'Analyse des Correspondances*, Vol. II, Dunod, Paris.
Bouroche J.M., Saporta G. (1980) - *L'Analyse des Données*, coll."Que sais-je", n°1854, P.U.F. Paris .
Burt C. (1950) - The factorial Analysis of Qualitative Data. *British J of stat. psychol.* vol 3, n°3, p166-185.
Caillez F., Pagès J.P. (1976) - Introduction à l'Analyse des Données. S.M.A.S.H., Paris
Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H.(1989) - Classification Automatique des Données: Environnement Statistique et Informatique Dunod, Paris.
Cibois P. (1984) - *Analyse des Données en Sociologie*. P.U.F. Paris.
Diday E. (1971) - La Méthode des Nuées Dynamiques. *Revue Stat. Appl.*,vol 19, n°2, p 19-34.
Escofier B., Pagès J. (1988) - Analyses factorielles Multiples. Dunod, Paris.
Escoufier Y. (1985) - L'Analyse des Correspondances, ses Propriétés, ses Extensions. *Bull of the Int. Stat. Inst* , 4, 28-2.
Fénelon J.P. (1981) - *Qu'est-ce-que l'Analyse des Données?* Lefonen, Paris.
Guttman L. (1941) - The Quantification of a Class of Attributes: a Theory and Method of a Scale Construction. in "*The prediction of personal adjustment* (P Horst,ed.) p 251 -264, SSCR New-York.
Hayashi C. (1956) - Theory and Examples of Quantification (II) *Proc. of the Institute of Stat. Math.* 4 (2) p 19-30.
Lebart L. (1975) - L'Orientation du Dépouillement de Certaines Enquêtes par l'Analyse des Correspondances Multiples. Consommation, n°2, p 73-96. Dunod.
Lebart L., Morineau A., Warwick K.(1984) - *Multivariate Descriptive Statistical Analysis* . J. Wiley and sons. New York.
Lebart L., Salem A. (1988) - *Analyse Statistique des Données Textuelles*, Dunod, Paris.
Nishisato S.(1980) - *Analysis of Categorical Data. Dual Scaling and its Application.* Univ. of Toronto Press.
Nishisato S. (1986) - *Quantification of Categorical Data: A Bibliography* 1975-1986. Microstats, Toronto.
SICLA *(Système Interactif de Classification Automatique)* (1988) - voir G. Celeux et alii.
SPAD.N (1987) - *Système Portable pour l'Analyse des Données*. CISIA Ed., 2bis Rue Jules Breton, 75013,Paris.
Tenenhaus M., Young F. W. (1985) - An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other Methods for Quantifying Categorical Multivariate Data. *Psychometrika*, vol 50; p 91-119.
Volle M. (1980) - *Analyse des Données*. Economica, Paris.