REVUE DE STATISTIQUE APPLIQUÉE

E. DAMBROISE
Y. ESCOUFIER
P. MASSOTTE

Application de l'analyse de données à l'élaboration de mini-sondages d'opinion

Revue de statistique appliquée, tome 35, n° 1 (1987), p. 9-23
<http://www.numdam.org/item?id=RSA_1987__35_1_9_0>

© Société française de statistique, 1987, 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, 1987, **XXXV** (1), 9-24

APPLICATION DE L'ANALYSE DE DONNÉES A L'ÉLABORATION DE MINI-SONDAGES D'OPINION

E. DAMBROISE et Y. ESCOUFIER

Unité de Biométrie, INRA, 9, place Viala 34060 Montpellier

P. MASSOTTE

IBM Montpellier, BP 1021 34006 Montpellier

RÉSUMÉ

Cet article présente une méthode de réduction du tableau des données en analyse des correspondances.

Cette méthode se réferre au choix de variables en composantes principales et est basée sur la maximisation du coefficient Rv.

Une application portant sur la définition d'un questionnaire réduit puis d'un ensemble réduit de personnes enquêtées est présentée.

Mots clés : *Analyse de correspondances, Analyse de données, Choix de variables, Coefficient Rv, Analyse en composantes principales par rapport à des variables instrumentales.*

ABSTRACT

This paper describes a reduction method of the data array in correspondences analysis.

This method is based upon the choice of variables in principal component analysis and upon the maximization of the Rv coefficient.

This method has been applied in the field of the opinion survey in the industry. This one is based upon the design of a reduced questionnary in the survey. This application is described in details in this document.

Key words : *Principal component analysis on instrumental variables, Choice of variables, Rv coefficient, Correspondences analysis, Data analysis*

1. Introduction

L'introduction de l'informatique dans les entreprises a permis d'étendre les possibilités d'analyses de données et les possibilités d'études dans des domaines très différents.

Dans cet article, nous aborderons celui des sondages d'opinion : il est, en effet, utile de connaître à des périodes données, l'opinion ou même la position du personnel d'une entreprise sur certains sujets d'intérêt général. Grâce aux possibilités de traitement de données offertes, les questionnaires ont pu évoluer pour devenir de plus en plus complets. De même, on a facilité la réalisation de sondages portant sur des populations de plus en plus importantes.

Néanmoins, les exigences des donneurs d'ordres se sont accrues et le problème se pose maintenant de pouvoir réaliser des mini-sondages à intervalles réguliers entre deux recensements, afin de surveiller toute variation ou changement d'opinion d'une population sur certains paramètres. L'objectif est bien entendu de réaliser un mini-sondage au meilleur coût et ayant un niveau de signification à peu près équivalent au sondage complet.

Afin d'atteindre cet objectif, nous proposons une méthode qui, à partir des résultats d'une enquête d'opinion, permet de réduire le nombre des questions et de retenir les groupes d'individus enquêtés fournissant les réponses les plus caractéristiques.

2. Principes de la méthode

2.1. Description du tableau de données

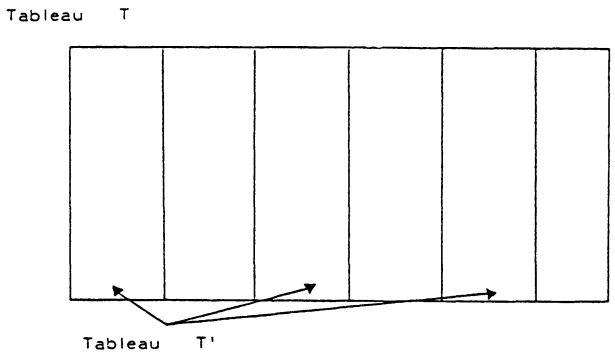
Les données se présentent sous la forme d'une juxtaposition de tableaux de contingence.

Les colonnes du tableau de données correspondent aux différentes modalités prises par les questions.

Les lignes représentent les réponses des groupes d'individus qui ont été constituées dans la population enquêtée.

La méthode proposée est applicable sans changement lorsque les individus ne sont pas regroupés; tous les groupes ont alors un effectif égal à l'unité.

2.2. Réduction du nombre de questions



Il s'agit de rechercher un tableau T′ extrait de T contenant les questions les plus représentatives. Afin de réaliser cette extraction, on utilisera les principes du choix de variables de l'analyse en composantes principales par rapport à des variables instrumentales (ACPVI) mais appliqués dans un contexte d'analyse des correspondances multiples.

Cette méthode permettra de chercher quelles sont les questions qui assurent une représentation graphique des lignes la plus proche possible de celle issue de l'analyse factorielle des correspondances (AFC) associée au tableau initial. La ressemblance entre l'analyse des données globales (tableau T) et l'analyse effectuée sur le tableau T' sera quantifiée à l'aide du coefficient Rv (dont la définition est rappelée au § 3-1-1).

2.3. Réduction du nombre des groupes d'individus enquêtés

Les techniques habituelles de sondage :

— tirage aléatoire simple,
— stratification,
— méthode des quotas.

sont bien sûr envisageables. Elles reposent sur la connaissance que l'on a de la constitution de la population enquêtée.

L'approche présentée ici est différente : elle veut tenir compte de la connaissance que l'on a des réponses fournies par la population à une enquête préalable. Le but est de retenir les groupes d'individus (les lignés du tableau T) qui structurent le plus les réponses aux questions.

Là encore la méthode est inspirée de l'ACPVI. La qualité de l'extraction pourra être appréciée en comparant les représentations graphiques des modalités des questions telles qu'elles sont fournies par l'analyse du tableau T et par celle du tableau que la méthode lui substitue. Le coefficient Rv permet de quantifier la ressemblance des résultats.

3. Méthodologie

3.1. Réduction du nombre de questions

3.1.1. Préliminaires

a) Rappels

Soient (X, Q, D) et (Y, R, D) deux études statistiques portant sur les mêmes individus munis des mêmes poids définissant la matrice diagonale D; Q et R correspondent aux métriques respectivement associées aux études faites à partir des tableaux X et Y.

Soient $W = XQ^tXD$ et $W' = YR^tYD$ les opérateurs de produit scalaire associés à ces deux études.

On sait (J.P. PAGES et *al.*, 1976) que les ACP des études (X, Q, D) et (Y, R, D) fourniront des représentations identiques (ou homothétiques) des individus si, et seulement si, les opérateurs W et W' sont égaux (ou proportionnels).

On peut d'autre part quantifier la similitude des deux études par le coeffi-

cient :

$$Rv(W, W') = \frac{Tr(WW')}{\|W\| \, \|W'\|}$$

avec $\|W\|^2 = Tr(W^2)$ et $\|W'\|^2 = Tr(W'^2)$.

Il prendra la valeur 1 si les deux opérateurs sont égaux ou proportionnels, la valeur 0 si toutes les composantes principales de l'étude (X, Q, D) sont orthogonales à toutes les composantes principales de (Y, R, D).

b) Notations et définitions

Soient V et $(V_i, i = 1,..., p)$ $p + 1$ variables ayant respectivement m et $(m_i, i = 1, ..., p)$ modalités.

$(T_i, i = 1,..., p)$ est l'ensemble des p tableaux de fréquences obtenus en croisant les modalités de V avec les modalités de chacune des variables $(V_i, i = 1,..., p)$. La somme des éléments de chacun des T_i est donc égale à l'unité.

D sera la matrice diagonale dont les éléments diagonaux sont les fréquences des modalités de V. On l'obtient en faisant les sommes des éléments des lignes de l'un quelconque des tableaux T_i. $(D_{Ji}, i = 1,..., p)$ sera l'ensemble des matrices diagonales associées aux modalités des variables $(V_i, i = 1,..., p)$.

Soit $C = \{1,..., p\}$. On notera $\bigoplus_{i \in C} T_i$ ou plus simplement T le tableau $1/p\,(T_1,..., T_p)$. On peut lui associer la matrice diagonale des poids de ses lignes : $D_I = D$ et la matrice diagonale des poids de ses colonnes :

$$D_J = \frac{1}{p} \begin{bmatrix} D_{J1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & D_{Jp} \end{bmatrix}$$

On sait (Y. ESCOUFIER, 1982) que l'AFC du tableau de nombres positifs T est équivalente à l'ACP de l'étude statistique :

$$(D_I^{-1} T D_J^{-1}, D_J, D_I).$$

On voit simplement que l'opérateur associé à cette étude est

$$W = \frac{1}{p} \sum_{i \in C} D^{-1} T_i D_{Ji}^{-1} {}^t T_i$$

Soit $R \subset C$ et card $R = r$. Notons $T_{(R)}$ le tableau

$$\frac{1}{r} (T_{i1},..., T_{ir}) .$$

L'AFC de ce tableau conduira à la diagonalisation de l'opérateur :

$$W_{(R)} = \frac{1}{r} \sum_{i \in R} D^{-1} T_i D_{Ji}^{-1} {}^t T_i$$

Il s'ensuit que les opérateurs W et $W_{(R)}$ sont D-symétriques, ce qui permet de poser la définition suivante :

Définition

On dira que $T_{(R)}$ est une meilleure approximation de T que $T_{(R')}$ si et seulement si

$$Rv\,(W, W_{(R)}) > Rv\,(W, W_{(R')})$$

3.1.2. Le problème et sa solution

Ayant décidé de retenir un nombre r_0 de p tableaux, on est donc amené à résoudre le problème suivant :

Trouver $R \subset C$, card $R = r_0$ tel que $Rv\,(W, W_{(R)})$ soit maximum.

Remarque

Posant $W_i = D^{-1} T_i D_{Ji}^{-1}\, {}^tT_i$ on a :

$$Rv\,(W, W_{(R)}) = \frac{\sum_{i\in C}\sum_{j\in R} Tr\,(W_i W_j)}{\sqrt{\left[\sum_{i\in C}\sum_{j\in C} Tr\,(W_i W_j)\right]\left[\sum_{i\in R}\sum_{j\in R} Tr\,(W_i W_j)\right]}}$$

La solution du problème :

Lorsque p et r_0 sont grands, on ne peut pas envisager d'énumérer de façon exhaustive tous les sous-ensembles de r_0 tableaux. On propose alors l'algorithme de sélection progressive suivant :

a) Trouver $i \in C$ tel que

$$\forall i' \in C \qquad Rv\,(W, W_{i'}) \leqslant Rv\,(W, W_i)$$

b) Pour tout $r = 1, \ldots, r_0 - 1$; étant donné $R \subset C$ tel que card $R = r$, trouver $i \in C - R$ tel que

$$\forall i' \in C - R \qquad Rv\,(W, W_{(R\cup\{i'\})}) \leqslant Rv\,(W, W_{(R\cup\{i\})})$$

Remarque

Posons $S = R \cup \{i\}$. Les calculs nécessités par l'algorithme ci-dessus sont simplifiés par le fait que :

$$(r + 1)\quad Tr\,(WW_{(S)}) = r\,Tr\,(WW_{(R)}) + Tr\,(WW_i)$$

et

$$\sum_{k\in S}\sum_{j\in S} Tr\,(W_k W_j) = \sum_{k\in R}\sum_{j\in R} Tr\,(W_k W_j) + 2r\,Tr\,(W_{(R)} W_i) + Tr\,((W_i)^2)$$

3.1.3. Interprétation du critère de sélection

Appelons ψ_α la $\alpha^{\text{ème}}$ composante principale de W, et φ_α le facteur associé.

Si A correspond à l'ensemble des valeurs propres non triviales issues de l'AFC de T, on a :

$$W = \sum_{\alpha\in A} \psi_\alpha\, {}^t\psi_\alpha D \qquad \text{avec} \qquad {}^t\psi_\alpha D \psi_\alpha = \lambda_\alpha$$

et

$$\varphi_\alpha = \frac{D_J^{-1}\,{}^tT\psi_\alpha}{\lambda_\alpha}\,, \qquad \varphi_{\alpha_i} = \frac{D_{Ji}^{-1}\,{}^tT_i\,\psi_\alpha}{\lambda_\alpha}$$

φ_{α_i} est le vecteur des composantes de φ_α qui correspondent aux modalités de V_i.

Notons $\varphi_{\alpha_i(j)}$ la composante de φ_α pour la $j^{ème}$ modalité de V_i et $P_{i(j)}$ le poids de cette modalité (l'élément j de la diagonale de D_{Ji}).

On a :

$$1 = \frac{{}^t\varphi_\alpha D \varphi_\alpha}{p} = \frac{\sum_{i\in C}\sum_{j=1}^{m_i} \varphi^2_{\alpha_i(j)} P_{i(j)}}{p}$$

et $\dfrac{\varphi^2_{\alpha_i(j)} P_{i(j)}}{p}$ est la contribution de la modalité j de V_i à l'axe principal α.

On a donc :

$$Rv\,(W, W_{(R)}) = \frac{Tr\left[\left(\sum_{\alpha\in A} \psi_\alpha\,{}^t\psi_\alpha D\right)\left(\frac{1}{r}\sum_{i\in R} D^{-1} T_i D_{Ji}^{-1}\,{}^tT_i\right)\right]}{\left[\left(\sum_{\alpha\in A}\lambda_\alpha^2\right) Tr\left(\left(\frac{1}{r}\sum_{i\in R} D^{-1}T_i D_{Ji}^{-1}\,{}^tT_i\right)^2\right)\right]^{1/2}}$$

$$= \frac{\sum_{i\in R}\sum_{\alpha\in A} ({}^t\psi_\alpha T_i D_{Ji}^{-1})\, D_{Ji}\, (D_{Ji}^{-1}\,{}^tT_i\psi_\alpha)}{\left[\left(\sum_{\alpha\in A}\lambda_\alpha^2\right)\left(\sum_{i\in R}\|W_i\|^2 + \sum_{\substack{i\in R\\ i\neq i'}}\sum_{i'\in R} Tr\,(W_i W_{i'})\right)\right]^{1/2}}$$

$$= \frac{\sum_{i\in R}\sum_{\alpha\in A}\lambda_\alpha^2 \sum_{j=1}^{m_i}\varphi^2_{\alpha_i(j)} P_{i(j)}}{\left[\left(\sum_{\alpha\in A}\lambda_\alpha^2\right)\left(\sum_{i\in R}\|W_i\|^2 + \sum_{\substack{i\in R\\ i\neq i'}}\sum_{i'\in R} Tr\,(W_i W_{i'})\right)\right]^{1/2}}$$

Pour commenter plus facilement ce résultat, prenons le cas particulier où $T_i = 1/n\ U_i$ avec U_i tableau des variables indicatrices de la variable V_i. On a alors $D = 1/n\ I_{n\times n}$; $n\,{}^tT_i T_i = D_{Ji}$ et on peut vérifier que :

$$\|W_i\|^2 = m_i \qquad Tr\,(W_i W_{i'}) = \frac{\chi^2_{ii'}}{n} + 1$$

où $\chi^2_{ii'}$ est le chi-deux entre les variables V_i et $V_{i'}$.

Il vient donc :

$$Rv\,(W, W_{(R)}) = \frac{\sum_{i\in R}\sum_{\alpha\in A}\lambda_\alpha^2\sum_{j=1}^{m_i}\varphi^2_{\alpha_i(j)} P_{i(j)}}{\sqrt{\left(\sum_{\alpha\in A}\lambda_\alpha^2\right)\left(\sum_{i\in R} m_i + \sum_{i\in R}\sum_{i'\in R}\left(\frac{\chi^2_{ii'}}{n} + 1\right)\right)}}$$

On peut alors énumérer les éléments qui vont intervenir dans le choix des variables :

a) Une variable sera retenue si elle a des contributions fortes $\left(\sum_{j=1}^{m_i} \varphi^2_{\alpha_i(j)} P_{i(j)} \right)$ sur les composantes associées aux fortes valeurs propres (λ_α).

b) De deux variables qui fourniraient le même numérateur, sera préférée celle qui aura le moins de modalités (m_i) et qui sera le moins liée aux variables déjà retenues ($\chi^2_{ii'}$).

3.2. *Réduction du nombre des groupes d'individus enquêtés*

Définitions

1) On considère comme précédemment le tableau T :
$1/p\ (T_1, ..., T_p)$. On appellera q le nombre total des colonnes et n le nombre de lignes du tableau T soumis à l'étude. Ainsi la matrice des poids des lignes de T, D_I est égal à D
et la matrice diagonale des poids des colonnes D_J est :

$$\frac{1}{p} \begin{bmatrix} D_{J1} & . & . & . & O \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ O & . & . & . & D_{Jp} \end{bmatrix}$$

Pour $i = \{1, ..., n\}$, on appellera pseudo-ligne i de T la ligne dont le $j^{ème}$ élément, $j \in \{1, ..., q\}$, est égal à :

$$(D_I)_{ii}\ (D_J)_{jj} = \frac{D_{ii}\ (D_{Ji})_{jj}}{p}$$

La pseudo-ligne est donc la ligne que l'on obtiendrait sous l'hypothèse d'indépendance des lignes et de chacune des colonnes de T.

2) On pose $L = \{1, ..., n\}$, soit $R \subset L$ tel que card $R = r$, $T^{(R)}$ désignera le tableau obtenu en remplaçant les r lignes de T dont les indices appartiennent à R, par r pseudo-lignes.

L'intérêt pratique des tableaux $T^{(R)}$ est d'avoir les mêmes marges que T ce qui permettra les comparaisons des opérateurs W associé à T et $W^{(R)}$ associé à $T^{(R)}$.

Description de la méthode

L'idée est d'éliminer les r lignes de T qui se rapprochent le plus des pseudo-lignes et qui seraient donc représentées par des points proches de l'origine dans les graphiques de l'AFC.

Ainsi on se ramène à chercher un tableau $T^{(R)}$ qui approxime au mieux le tableau T, c'est-à-dire tel que les opérateurs de produit scalaire $W^{(R)}$ et W soient les plus voisins possibles.

On regardera la qualité de l'approximation à l'aide du coefficient Rv.

On propose ici un algorithme d'élimination progressive :

a) Trouver $i \in L$, tel que

$$\forall i' \in L \qquad Rv\ (W, W^{(L-\{i'\})}) \leqslant Rv\ (W, W^{(L-\{i\})})$$

b) Pour tout $r = 1, ..., n - 1$; étant donné $R \subset L$ tel que card $R = r$, trouver $i \in L - R$ tel que :

$$\forall i' \in L - R, \qquad Rv(W, W^{(L-R-\{i'\})}) \leqslant Rv(W, W^{(L-R-\{i\})})$$

4. Application

La méthode définie ci-dessus a été appliquée sur un cas réel concernant une étude d'opinion dans une entreprise.

Les données disponibles se présentent sous la forme d'une juxtaposition de 90 tableaux de contingence (un tableau par question) codés avec 5 modalités. Les questions abordées dans l'enquête d'opinion concernent les domaines suivants :

- La compagnie (10 questions).
- Travail : organisation et charge (10 questions).
- Travail : contenu (10 questions).
- Qualité (10 questions).
- Salaire (10 questions).
- Chef de service (10 questions).
- Relations avec le Chef de service (10 questions).
- Encadrement supérieur (10 questions).
- Information (10 questions).

Les individus pour des raisons de confidentialité sont regroupés en catégories socio-professionnelles (61 C.S.P.) qui définissent les lignes des tableaux. Les catégories socio-professionnelles sont constituées en groupes homogènes de tailles voisines. Ces groupes sont établis en fonction du niveau, de l'activité et du service des personnes de la population de référence. Exemples :

- Secrétaires du Service du Personnel.
- Cadres non-Chefs de service du Service Assurance de la Qualité.
- Techniciens directs du Service des Equipements Techniques.

4.1. Réduction du nombre de questions

La tableau T initial a été réduit par étapes successives avec la méthode décrite précédemment. Le résultat de cette transformation est résumé dans la figure 1 dans laquelle les questions repérées par un numéro sont classées par ordre décroissant de signification.

Le Rv figurant sur le tableau pour le numéro d'ordre r est le Rv calculé à partir des r premières questions retenues. Les valeurs élevées du Rv permettent déjà de supposer que la similitude des représentations graphiques obtenues par AFC à partir du tableau initial T et du tableau $T_{(R)}$ est forte.

Remarque

Il est à noter que $\forall i \in \{1, ..., 90\}$, $\forall j \in \{1, ..., 90\}$, $Rv(W_i, W_j)$ est grand, $(Rv(W_i, W_j) \geqslant 0{,}96)$, ce qui permet d'expliquer les très bons résultats obtenus.

N° ordre	Numéro des questions retenues	Rv	N° ordre	Numéro des questions retenues	Rv
1	76	.9864	16	56	.999
2	28	.9926	17	24	.9991
3	73	.9948	18	79	.9992
4	42	.9962	19	21	.9993
5	47	.9968	20	49	.9993
6	3	.9973	21	18	.9993
7	10	.9976	22	6	.9994
8	15	.9979	23	19	.9994
9	59	.9981	24	38	.9994
10	61	.9984	25	9	.9994
11	78	.9985	26	67	.9995
12	43	.9987	27	68	.9995
13	57	.9988	28	34	.9995
14	29	.9989	29	53	.9995
15	14	.999	30	15	.9995

FIGURE 1
Classement des questions retenues.

Sur les figures (2a) et (2b) on peut apprécier la qualité de la ressemblance des représentations graphiques obtenues par AFC à partir du tableau initial T et du tableau $T_{(R)}$, avec card R = 5.

En effet les catégories (lignes du tableau) qui contribuent le plus à la formation des axes factoriels de l'AFC de T, sont les mêmes que dans l'AFC de $T_{(R)}$ (card R=5) : 49 — 45 — 56 — 59 — 6 — 42 — 28 — 10 — 15.

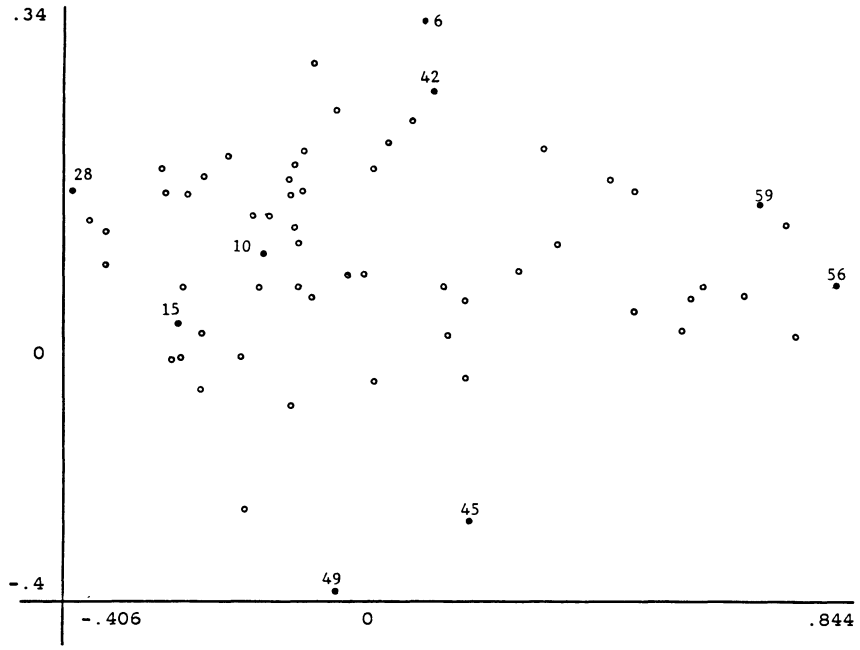
(Points marqués sur les graphiques) et de plus il est à remarquer que ces points se retrouvent projetés approximativement de la même manière dans ces 2 plans factoriels.

Validation

Une simulation a été effectuée pour différents cas correspondants à un nombre donné de questions. Dans chacun de ces cas, 1 000 tirages aléatoires de questions ont été effectués. Dans chaque simulation le Rv maximum trouvé était inférieur au Rv calculé à partir du choix optimal de questions proposé par la méthode (Fig. 3).

Nombre de questions retenues	Rv moyen observé dans la simulation	Rv maximum dans la simulation	Rv proposé
5	.9921	.9964	.9968
10	.996	.9980	.9984
15	.9973	.997	.999
20	.998	.9992	.9993
25	.9983	.9994	.9994
30	.9985	.9994	.9995

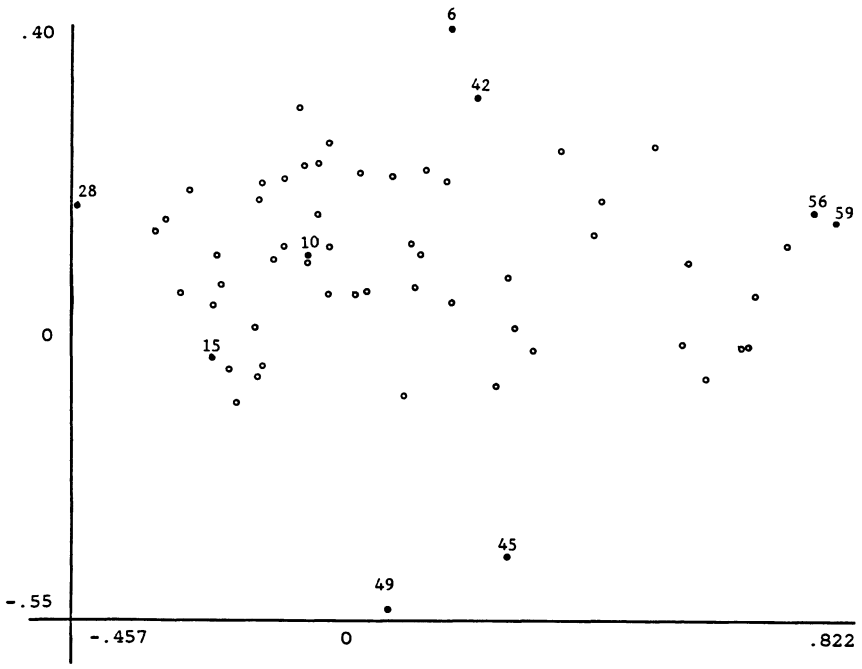
FIGURE 3
Simulation.



PLAN I.2.
Représentation graphique des catégories professionnelles issues du tableau T.

Valeurs propres	% inertie
0.077	0.227
0.023	0.069
0.016	0.048
0.013	0.039
0.011	0.033
0.010	0.030

FIGURE 2a



PLAN I.2.
Représentation graphique des catégories professionnelles issues du tableau $T^{(R)}$, avec card $R = 5$.

Valeurs propres	% inertie
0.065	0.222
0.024	0.087
0.015	0.051
0.015	0.050
0.013	0.044
0.010	0.034

FIGURE 2b

4.2. Réduction du nombre des groupes d'individus enquêtés

La méthode décrite dans cet article a été utilisée pour rejeter les catégories socio-professionnelles les moins caractéristiques. Elle permet de calculer le Rv après remplacement par les pseudos-lignes (Fig. 4).

Nombre de catégories rejetées	Numéro des catégories rejetées	Rv
1	31	0.916
2	15	0.891
3	29	0.853
4	30	0.811
5	8	0.804
.	.	.
.	.	.
.	.	.
16	12	0.518
17	49	0.481
18	53	0.449
19	48	0.432

FIGURE 4
Les 19 premières catégories rejetées.

Les figures (5a) et (5b) permettent de noter que les représentations graphiques issues de l'AFC sur le tableau initial, T, comprenant 61 catégories et le tableau $T^{(R)}$, card R = 19, ramené à 42 catégories, sont là aussi ressemblantes.

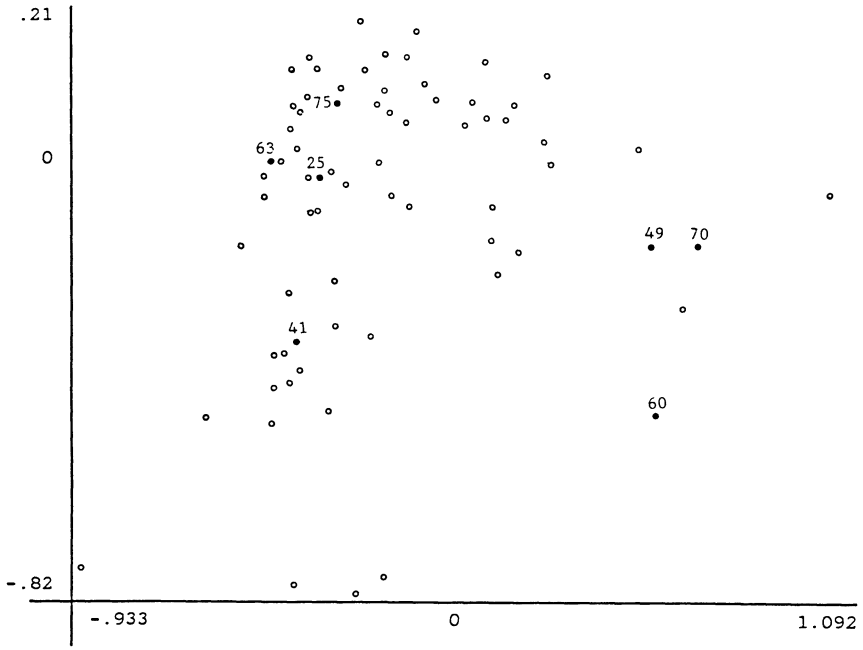
Comme dans le paragraphe précédent, les points (modalité des questions) qui contribuent le plus à la formation des axes sont les mêmes dans l'analyse du tableau T et dans celle du tableau $T^{(R)}$ (n° des colonnes ayant une forte contribution absolue : 70 — 49 — 60 — 63 — 75 — 41 — 25).

Ces points se projettent approximativement de la même manière.

Il est à noter que pour certains points la reconstruction graphique ne s'est pas très bien opérée, par exemple : 15, 37, 48, 69.

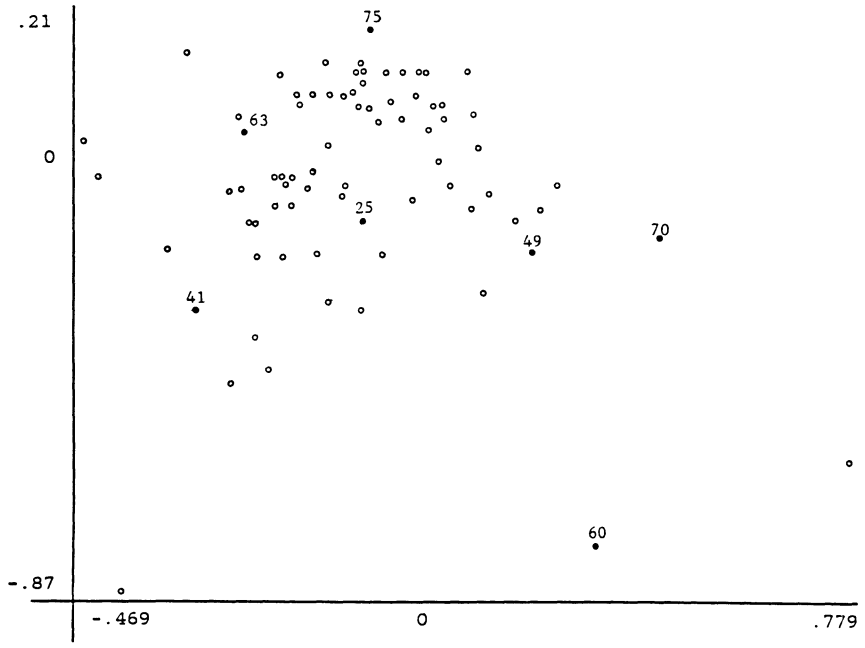
Ces points avaient notamment de très faibles contributions sur l'axe 2. En effet, les effectifs de chacun de ces points font partie des 7 plus faibles et ne représentent respectivement que 1 %, 1.2 %, 0.2 %, 1.3 % de l'effectif de la population totale.

Une validation, sur les mêmes bases que précédemment a été effectuée, elle a permis de vérifier la qualité de la réduction (1 000 simulations effectuées, figure 6).



PLAN I.2.
Représentation graphique des modalités des questions issues du tableau T.

FIGURE 5a



PLAN I.2.
Représentation graphique des modalités des questions issues du tableau $T^{(R)}$, avec card R = 19.

Valeurs propres	% inertie
0.015	0.094
0.015	0.091
0.012	0.071
0.011	0.067
0.009	0.057
0.009	0.053

FIGURE 5b

Nombre catégories rejetées	Rv moyen observé par simulation	Rv maximum observé	Rv proposé
5	0.789	0.791	0.804
10	0.594	0.608	0.615
19	0.411	0.430	0.432

FIGURE 6
L'écart entre les valeurs propres des deux études est illustré en grande partie par les valeurs peu importantes du Rv (Rv = 0.432).

5. Conclusion

La méthode de réduction des variables et/ou des individus décrite dans cet article est une méthode efficace qui a été appliquée de façon satisfaisante à un cas réel.

Cette méthode souffre toutefois d'un inconvénient : la quantité de calculs nécessaires peut être importante; elle est fonction de la taille du tableau initial T et du niveau de réduction désiré. L'évolution des moyens informatiques et de l'architecture des ordinateurs estompera à court terme cet inconvénient.

La confiance qui peut être accordée aux résultats de cette méthode sera d'autant plus grande que les précautions habituelles concernant la quasi-égalité des effectifs des lignes et la quasi-égalité des effectifs des colonnes du tableau T auront été prises.

Enfin on voit aisément que la méthode peut être d'un usage intéressant dans d'autres domaines d'application que celui qui a été présenté ici.

On peut envisager par exemple :

— le choix de variables indicatrices dans le suivi de performance d'un matériel et le contrôle d'un processus de fabrication;
— le choix de critères de décision dans un système d'aide automatique au diagnostic médical;
— le choix de paramètres dans un système d'alerte.

Bibliographie

[1] BONIFAS L., ESCOUFIER Y., GONZALEZ P.L., SABATIER R. (1984). — Choix de variables en analyse des données — *RSA (Revue Statistiques Appliquées),* Vol. XXXII, n° 2, p. 5-25.

[2] CAILLIEZ F. et PAGES J.P. (1976). — Introduction à l'analyse des données — *SMASH,* 9, rue Duban, 75016 Paris.

[3] DAMBROISE E. (1984). — *Rapport de DEA : Etude d'opinion* — Université des Sciences et Techniques du Languedoc, Montpellier.

[4] ESCOUFIER Y. (1982). — L'analyse des tableaux de contingence simples et multiples. *Metron,* Vol. XL, n° 1-2, p. 53-77.

[5] GOURIEROUX C. (1981). — *Théorie des sondages* — Economica, 49, rue Hericart, 75015 Paris.

[6] MASSOTTE P. (1981). — *Rapport technique : Echantillonnage dans les populations finies.* Compagnie IBM FRANCE, Usine de Montpellier.

[7] PAGES J.P., ESCOUFIER Y., CAZES P. (1976). — Opérateurs et analyse des tableaux à plus de deux dimensions. *Cahier du BURO,* n° 25, p. 61-89.