



# Nouvelle étape de traitement des données manquantes en analyse factorielle des correspondances multiples dans le système portable d'analyse des données

Habib Benali, Brigitte Escofier

## ► To cite this version:

Habib Benali, Brigitte Escofier. Nouvelle étape de traitement des données manquantes en analyse factorielle des correspondances multiples dans le système portable d'analyse des données. [Rapport de recherche] RT-0085, INRIA. 1987. inria-00071323

**HAL Id: inria-00071323**

**<https://hal.inria.fr/inria-00071323>**

Submitted on 23 May 2006

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**INRIA**

UNITÉ DE RECHERCHE  
INRIA-RENNES

Institut National  
de Recherche  
en Informatique  
et en Automatique

Domaine de Voluceau  
Rocquencourt  
BP 105  
78153 Le Chesnay Cedex  
France  
Tél. (1) 39 63 55 11

Rapports Techniques

N° 85

**NOUVELLE ETAPE DE  
TRAITEMENT DES DONNEES  
MANQUANTES EN ANALYSE  
FACTORIELLE DES  
CORRESPONDANCES MULTIPLES  
DANS LE SYSTEME  
PORTABLE D'ANALYSE DE  
DONNEES**

**Habib BENALI  
Brigitte ESCOFIER**

**AOUT 1987**

Campus Universitaire de Beaulieu  
35042 - RENNES CÉDEX  
FRANCE  
Téléphone: 99 36 20 00  
Télex: UNIRISA 950 473 F  
Télécopie: 99 38 38 32

NOUVELLE ETAPE DE TRAITEMENT DES DONNEES MANQUANTES  
EN ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES  
DANS LE SYSTEME PORTABLE D'ANALYSE DES DONNEES

NEW STEP FOR PROCESSING OF INCOMPLETE DATA IN  
MULTIPLE CORRESPONDENCE ANALYSIS USING S.P.A.D.

Publication Interne n° 370 - Juillet 1987 - 24 Pages

*Habib BENALI*  
*Brigitte ESCOFIER*  
*IRISA - RENNES*

Résumé

La technique que nous présentons dans cet article , permet de neutraliser l'effet de certaines perturbations en analyse des correspondances multiples. Elle résoud de façon "efficace" le problème des données manquantes et des modalités à faibles effectifs.

Summary

The technique we present in this paper allow to avoid the effect of perturbations in the case of qualitative variables. this solves in an efficient way the problem of missing quantities and the problem of unfrequent response modalities .

Mots Clés

Stabilité - Analyse des correspondances multiples - Tableau Disjontif Incomplet - Distance

## TABLE DES MATIERES

- INTRODUCTION

- ASPECT THEORIQUE

I - INTRODUCTION

II - NOTATIONS

III - METHODE MULMD

IV - MISE EN OEUVRE PRATIQUE DE LA METHODE

V - NORME D'ECRITURE DANS SPAD

VI - BIBLIOGRAPHIE

- ASPECT INFORMATIQUE

I - LES FICHIERS

II - ETAPE MULMD

III - DOSSIER DE PROGRAMMATION

- CONCLUSION

## INTRODUCTION

Quelques techniques en Analyse Factorielle des Correspondances traitent les tableaux ayant des données manquantes, elles donnent des résultats assez stables (fonction du nombre de "trous"), pour la plupart elles sont basées sur une méthode de reconstitution des données manquantes, qui est un principe très honnête en temps calcul. Une nouvelle technique proposée en [2], que nous avons développée dans un travail de recherche à l'I.R.I.S.A [1], elle permet de répondre de façon très efficace au double problème : stabilité des résultats et coût calcul. Les algorithmes de cette méthode n'utilisent pas le principe de reconstitution des données. Un intérêt immédiat a été manifesté par certains chercheurs et utilisateurs pour cette méthode, on l'a donc insérée comme une nouvelle étape dans la bibliothèque du système portable d'analyse des données SPAD.

SPAD peut être défini de la façon suivante ( "grammaire" ) :

SPAD	::=	programme unique appelant 51 étapes / Enchaînement d'étapes
étape	::=	programme privilégié ayant une certaine cohérence d'un point de vue statistique
Enchaînement	::=	Analyse

L'étape MULMD que nous présentons dans SPAD ( SYSTEME PORTABLE POUR  
L'ANALYSE DES DONNEES ) traite des données manquantes et des modalités à faible effectifs

**Version : 0.0 Avril 1986**

**Auteurs : Habib . BENALI I.R.I.S.A Rennes 1  
Brigitte . Escofier I.R.I.S.A Rennes 1**

ASPECTS THEORIQUES :

I - INTRODUCTION

II - NOTATIONS

III - METHODE MULMD

IV - MISE EN EUVRE PRATIQUE DE LA METHODE

V - NORMES D'ECRITURE

VI - BIBLIOGRAPHIE

## I - INTRODUCTION

Une des sources de perturbation des résultats d'une analyse des correspondances multiples (AFCM) est le problème des non-réponses et des réponses rares. Une variante de cette méthode proposée en [2] et développée en [1], permet de résoudre ce problème. Cette technique a été testée avec d'autres méthodes équivalentes [1], elle trouve un compromis, devant de fortes perturbations du tableau de données, entre coût calcul et stabilité des résultats, qui est très avantageux.

## II - NOTATIONS:

Un questionnaire est formé d'un ensemble Q de questions dont chacune admet J<sub>q</sub> de modalités. on note:

I l'ensemble des individus

J l'ensemble des modalités de réponses à toutes les questions Q

$$K_{IJ} = [ K_{IJ1}, K_{IJ2}, \dots, K_{IJQ} ]$$

$$\text{card I} = n$$

$$\text{card J} = \text{card J}_1 + \text{card J}_2 + \dots + \text{card J}_Q$$

$$j \in JQ \quad K_{ij} = 1 \text{ si l'individu possède la modalité } j$$

$$K_{ij} = 0 \text{ sinon}$$

$$K_{i.} = \sum_{j \in J} K_{ij}$$

$$K_{.j} = \sum_{i \in I} K_{ij}$$

$$K = \sum_{i,j} K_{ij}$$

En A.F.C.M., le tableau  $K_{IJ}$  est disjonctif complet,  $K_{i.} = Q = K / n$ .

les programmes de calcul utilisent le tableau du codage condensé  $C_{IQ}$  du tableau  $K_{IJ}$  où la case (i,q) contient le numéro  $c_{iq}$  de la question q choisie par l'individu i. La mise sous forme disjonctive des données dans notre cas n'est qu'une présentation commode.



Un individu  $i$  est représenté dans  $R_J$  par son profil ligne  $\{ K_{ij} / K_i, j \in J \}$ , et une modalité  $j$  est représenté par son profil colonne  $\{ K_{ij} / K_j, i \in I \}$ .

Le nuage des individus  $N(I)$  est l'ensemble des profils lignes affectés des poids  $K_i / K$ .

le nuage des modalités  $N(J)$  est l'ensemble des profils colonnes affectés des poids  $K_j / K$ .

### III METHODE MULMD

Cette méthode est une variante de l'A.F.C.M, elle résoud simultanément le problème des données manquantes et des réponses rares, en minimisant leur influence sur le résultat, elle traite des tableaux disjonctifs "incomplets" ainsi définis:

#### Tableau disjonctif incomplet $K'_{IJ}$ :

Dans le tableau  $K_{IJ}$ , les non-réponses à la question  $q$  sont codées zéro sur l'ensemble des modalités  $J_q$  de cette question, et les colonnes correspondant aux modalités rares sont supprimées.

#### Choix d'une distance pour l'étude de $K'_{IJ}$ :

##### Distance du KHI-DEUX

L'A.F.C.M classique utilise la distance du khi-deux entre deux profils lignes (resp. colonnes) qui s'écrit,

$$d^2(i, i') = \sum_{j \in J'} (K' / K_j) [ K_{ij} / K_i - K'_{i'j} / K'_{i'} ]^2$$

$$d^2(j, j') = \sum_{i \in I} (K' / K_i) [ K_{ij} / K_j - K'_{ij'} / K'_{j'} ]^2$$

### Inconvénient de cette distance

Dans le tableau disjonctif incomplet  $K'_{IJ}$ , le problème des modalités rares (contribuant fortement à la distance entre deux profils lignes) est résolu. Par contre, si deux individus  $i$  et  $i'$  n'ont pas donné le même nombre de réponses ( $K_{ij} \neq K_{i'j}$ ), une modalité  $j$  choisie simultanément par ces individus augmente leur distance car le terme  $K_{ij} / K_{i'j} - K_{i'j} / K_{ij}$  n'est pas nul, ce qui est un réel problème d'interprétation. Cette métrique est donc inadaptée à l'étude de tableaux disjonctifs incomplets.

### Distance variante du khi-deux

Pour remédier à cet inconvénient, on remplace la marge ( $K_{i\cdot}$ ,  $i \in I$ ) par la marge constante ( $K' / n$ ,  $i \in I$ ) partout où elle intervient : profil et poids des individus, métrique et origine des axes pour le nuage des profils colonne.

Les distances entre profils lignes sont analogues à celle issues du tableau disjonctif complet  $K_{IJ}$  obtenu en supprimant les termes provenant des non-réponses et des modalités rares, et les distances entre profils colonnes sont identiques à celle issues du tableau  $K_{IJ}$ .

### Dualité et formules de transitions

Les facteurs  $F_s$  du nuage  $N(I)$  se déduisent des facteurs  $G_s$  du nuage  $N(J)$  par les formules de transitions suivantes où  $\mu_s$  est la valeur propre d'ordre  $s$ .

$$F_s(i) = n / (\sqrt{\mu_s} \sum_{j \in J} K_{\cdot j}) \sum_{j \in J} K_{ij} G_s(j) - 1 / (\sqrt{\mu_s} \sum_{j \in J} K_{\cdot j}) \sum_{j \in J} K_{\cdot j} G_s(j)$$

$$G_s(j) = (1 / \sqrt{\mu_s}) \sum_{i \in I} (K_{ij} / K_{\cdot j}) F_s(i)$$

Dans la première formule apparaît le terme  $1 / (\sqrt{\mu_s} \sum_{j \in J} K_{\cdot j}) \sum_{j \in J} K_{\cdot j} G_s(j)$  qui représente la coordonnée du centre de gravité  $G$  du nuage  $N(J)$  sur l'axe  $F_s$ , et mesure le décalage du facteur quand l'origine des axes ne correspond pas à  $G$ .

Ce terme en pratique est presque nul, ce qui permet d'interpréter comme en A.F.C.M. classique l'abscisse d'un individu comme le barycentre des modalités de réponses qu'il a choisies. La deuxième formule est exactement celle de l'A.F.C.M.

Element supplémentaire et formule de reconstitution des données

On peut mettre des modalités en élément supplémentaire, particulièrement les modalités non-réponses et les modalités rares. l'abscisse  $G_s(j^+)$  d'une modalité supplémentaire est définie par:

$$G_s(j^+) = (1 / \sqrt{\mu_s}) \sum_{i \in I} (K_{ij^+} / K_{j^+}) F_s(i)$$

A partir des formules de transitions, on peut reconstituer le tableau initial en utilisant la formule:

$$K_{ij} = (K_{.j} / n) (1 + \sum_s (1 / \sqrt{\mu_s}) F_s(i) G_s(j))$$

Nuage N(I) des profils des lignes

La distance entre deux profils lignes  $i$  et  $i'$  est définie par :

$$d^2(i, i') = \sum_{j \in J} K' / K_j [ (n / K') K_{ij} - (n / K') K_{i'j} ]^2 = (n^2 / K') \sum_{j \in J} 1 / K_j [ K_{ij} - K_{i'j} ]^2$$

L'inconvénient rencontré dans la distance du khi-deux disparaît. L'analyse du nuage N(I) est faite à partir de son centre de gravité pris comme origine des axes.

Nuage N(J) des profils des colonnes

La distance entre deux profils colonnes  $j$  et  $j'$  est définie par :

$$\begin{aligned} d^2(j, j') &= \sum_{i \in I} K' n / K' [ K_{ij} / K_j - K_{i'j'} / K_{j'} ]^2 \\ &= n \sum_{i \in I} [ K_{ij} / K_j - K_{i'j'} / K_{j'} ]^2 \end{aligned}$$

Cette métrique possède la propriété intéressante qu'est "l'équivalence distributionnelle".

L'analyse du nuage N(J) est faite à partir du point de coordonnée  $K' / n$  pris comme origine des axes.

Les facteurs sur  $J'$  ne seront pas exactement centrés puisque le centre de gravité de N(J) est  $(K'_i / K', i \in I)$

#### IV MISE EN OEUVRE PRATIQUE DE LA METHODE

##### Le programme

Il s'agit de l'étape MULTM de SPAD adapté aux traitements des questionnaires avec non-réponses et modalités à faibles effectifs. Dans le programme principal, le tableau de Burt (J,J) est calculé à partir du tableau du codage condensé  $C_{IQ}$ , la matrice sur les profils colonne d'ordre (J,J) est calculée, sa dimension ne peut être réduite (3), ce qui est la grande particularité par rapport au programme original MULTM de SPAD.

##### Cas des réponses rares

On fixe au seuil NMIN à partir duquel les modalités d'effectifs inférieurs à NMIN sont considérés comme rares, elles sont affectées à un fichier IMOD où elles sont éliminées des calculs.

##### Cas des non-réponses

Les non-réponses sont codées sur le fichier initial des données comme les autres modalités ; leur nombre est fixé par un paramètre NABAND, elles sont lues puis éliminées des calculs.

#### V - NORMES D'ECRITURE

Le programme MULMD doit respecter les normes imposées dans SPAD qui sont :

- La portabilité
- Homogénéité et la transparence des notations
- modularité
- programme commenté

## VI - BIBLIOGRAPHIE

- 1 - Habib BENALI (1985) : Stabilité de l'analyse en composantes principales et de l'analyse des correspondances multiples en présence de certains types de perturbations  
- méthode de dépouillement d'enquêtes-  
Thèse de troisième cycle - Université de Rennes 1 .
  
- 2 - Brigitte ESCOFIER (1981) : Traitement de questionnaires avec non-réponses et analyse des correspondances avec marge modifiée et analyse multicanonique  
avec contraintes . IRISA Publication interne N 146.
  
- 3 - Ludovic LEBART - Alain MORINEAU (1982) : SPAD = système portable d'analyse des données  
C.E.S.I.A

ASPECT INFORMATIQUES

DOSSIER DE PROGRAMMATION

LES FICHIERS

I - CATALOGUES DES FICHIERS

- FICHIERS\_UTILISATEUR

- FICHIERS\_ARCHIVES

- FICHIERS\_SAUVEGARDABLES

- FICHIERS\_DE\_TRAVAIL

II - DESCRIPTION DES FICHIERS

CATALOGUE DES FICHIERS

Les fichiers sont repérés par leur nom ( NDICA ,NDONA,...) , auquel correspond un numero logique définissant l'unité de lecture , ce numero est défini dans le programme SPADF gérant les étapes .

Fichiers\_utilisateur

Ces des fichiers créés par l'utilisateur a l'exterieur de SPAD

NDONZ : fichier des données

NDICA : fichier des libellés des variables

Fichier\_archives

Ces des fichiers créés par SPAD , à partir des fichiers utilisateur , et contiennent les fichiers de référence pour toutes les analyses.

NDONA : fichier archive des données

NDICA : fichier du dictionnaire

Fichier\_sauvegardables

Ces des fichiers de transmission des résultats entre les étapes . Ils ont une structure bien définie , et peuvent être lus à l'extérieur de SPAD.

NDIC : fichier du dictionnaire utile                      numéro logique 08

NDON : fichier des données utiles                      numéro logique 09

NGUS : fichier des coordonnées factorielles                      numéro logique 11

Fichier\_de\_travail

Il s'agit de fichiers de travail

NSAV : fichier écrit sans format                      numéro logique 14

NBAND : fichier écrit sans format                      numéro logique 15



DESCRIPTION DES FICHIERS

On ne donnera que la description du fichier NDIC , pour les autres fichier on se reportera à la documentation sur spad [3].

FICHIERS NDIC ( numéro logique 8 )

Ce fichier contient les libellés des colonnes d'un tableau , il est associer au fichier des données NDON , créé par l'étape LILEX.

enregistrement 1 : (ltitr(l) , l = 1,20 ) , ldic

ltitr (l) , l=1,20 : Titre en 80 caractères écrit sans format  
ldic : Variable contenant les 4 caractères "NDIC"

enregistrement 2 : nqfin , ngr(l);l =1,10 , icard , isup , jfin , jmax , mmax , nid

nqfin : Nombre total de variables  
ngr(l) : Nombre de variables dans chaqu'un des 10 groupes  
icard : Nombre total d'individus  
isup : Nombre de lignes illustratives  
jfin : Somme de toutes les modalités et des variables continues  
jmax : Nombre cumulé de modalités des variables nominales du groupe des variables nominales actives.  
mmax : Maximum parmi les variables actives du nombre cumulé de leur modalités  
nid : Longueur de l'identificateur de ligne , comptée en "A4"

enregistrement 3 : (mcum(l) , l = 1,nqf1) , (( idq (l,k) , l = 1,nqfin ) , k =1,20)

mcum(l) : Cumul progressif des nombres de modalités pour les nqfin variables  
nqf1 = nqfin + 1  
idq (l,k) : Tableau des libellés en 60 caractères pour les nqfin variables

enregistrement 4 : (idj (j) , j =1,jfin) , (( libel (l,j) ,l =1,jfin) , j = 1,5)

idj(j) : Libellés en 4 caractères des modalités  
libel(l,j) : Tableau des libellés en 20 caractères pour les modalités

enregistrement 5 : (numq(l) , l = 1,nqfin)

numq(j) : Places occupées par les variables actuelles sur le fichier-archive NDONA  
numq(l) = numéro archive de la l-ème variable du fichier NDON

## ETAPE MULMD

I - OBJET DE L'ETAPE

II - LISTE DES COMMANDES

III - FICHIERS NECESSAIRE A L'EXECUTION

IV - CALCUL DE LA RESERVATION MEMOIRE

V - DEFINITION DES CARTES PARAMETRES DE COMMANDES

VI - UN EXEMPLE

OBJET DE L'ETAPE

L'étape MULMD gère les calculs , et effectue les principales éditions pour une A.F.C.M en cas de données manquantes et de modalités à faibles effectifs.

Comme dans l'étape MULTC le choix des individus et des modalités à faibles effectifs , doit être effectué au préalable par l'étape LILEX qui crée les fichiers (NDON et NDIC) pour MULMD

MULMD crée le fichier NGUS des coordonnées factorielles

LISTE DES COMMENDES

- a) MULMD
- b) carte de 9 paramètres : NFAC NMIN LIST3 NTEXT NPAGT NLIGHT NCOR NTAB NABAND
- c) carte de deux paramètres : NUMQ NUMM

FICHIERS NECESSAIRE A L'EXECUTION

En entrée    NDIC / 8 / fichier du dictionnaire ( crée par LILEX )  
               NDON / 9 / fichier des données    ( idem        )

En sortie    NGUS / 11 / fichier des coordonnées factorielles .

De travail    NSAV / 14 / fichier temporaire  
               NBAND / 15 / (idem)

CALCUL DE LA RESERVATION MEMOIRE

$$\text{MOTS} > \text{JFIN} * ( 12 + \text{MMAX} + \text{JMAX} ) + 18 * \text{NQFIN} + \text{MMAX} * \text{MMAX}$$

- NQFIN : nombre total de variables dans l'analyse
- JFIN : somme des deux quantités suivantes : nombre de modalités de toutes les variables nominales et nombre de variables continues.
- JMAX : nombre total de modalités des variables nominales actives
- MMAX : maximum du nombre de modalités des variables nominales active.

Définition des paramètres de commandes

carte b) 9 paramètres

- NFAC : nombre d'axes factorielles à calculer
- NMIN : effectif critique pour une modalités active : si l'effectif est inférieur ou égal à NMIN la modalités est écartée pour le calcul des axes factorielles.  
si NMIN = 0 seuls les modalités d'effectifs nuls sont éliminées.
- LIST3 : paramètres pour l'édition concernant les individus  
LIST3 = 0 pas d'édition  
LIST3 = 1 édition des coordonnées et contributions des lignes sur les 6 premiers facteurs
- NTEXT : nombre de facteurs successifs pour lesquels on demande une édition des rangement des modalités , pour une caractérisation rapide des facteurs.
- NPAGT : nombre de pages ( 1, 2 ou 3 ) définissant la largeur d'édition de chacun des NTEXT facteurs
- NLIGT : nombre de modalités à retenir pour chaque édition de facteur  
si NLIGT = 0 toutes les modalités seront retenues.
- NCOR : paramètres pour l'édition des corrélations des variables continues avec les 6 premiers facteurs  
si NCOR = 0 pas d'édition  
= 1 édition
- NTAB : paramètres pour l'édition du tableau de contingence multiples ( tableau de Burt )  
si NTAB = 0 pas d'édition  
= 1 édition des effectifs seuls  
= 2 édition des effectifs et des profils  
= 3 édition des profils seuls
- NABAND: nombre de modalités à éliminer ( modalités correspondant aux données manquantes et des modalités d'effectifs inférieur à NMIN si elles sont connues de l'utilisateur sinon le programme les élimines grace à NMIN)

Carte c) 2 paramètres

- NUMQ : numéro de la question où il y a la modalité à éliminer
- NUMM : numéro de la modalité à éliminée ( il sagit de son numéro dans le fichier de donnée)  
autant de carte c) qu'il y a de modalités à éliminer.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique