L'Analyse des correspondances et les techniques connexes

Approches nouvelles pour l'analyse statistique des données

Ouvrage édité par
J. Moreau, P.-A. Doudin, P. Cazes

Springer

L'analyse des correspondances multiples: un outil pour la classification de données de cursus

Peter G.M. van der Heijden, Joop Teunissen et Charles van Orlé [1]

[1] Département de Méthodologie et de Statistiques, Faculté des Sciences Sociales, Université d'Utrecht, Pays-Bas

1. Introduction

En sciences sociales, on s'intéresse souvent au cursus des individus. Le concept de cursus peut jouer un rôle soit de variable explicative soit de variable dépendante. C'est par exemple le cas pour des cursus scolaires, professionnels, de patients ou encore de criminels. Ces cursus sont souvent très complexes, ce qui rend leur analyse délicate. Il est alors très important de disposer de mesures qui synthétisent les cursus individuels. Taris (1994) a fourni récemment une vue d'ensemble des outils utiles à une classification ou une quantification des cursus.

Une attention particulière a été portée à l'analyse statistique des parcours individuels (voir, par exemple, Blossfeld et al., 1989; Yamaguchi, 1991). D'une manière générale, on cherche à prédire une transition dans un cursus à l'aide d'un certain nombre de variables explicatives. Par exemple, le laps de temps entre la fin de l'école et l'obtention d'un premier emploi peut être prédit par l'âge, le sexe et des variables contextuelles. Même si cela peut s'avérer utile pour résoudre un certain nombre de problèmes, l'analyse statistique des parcours individuels semble se centrer sur des aspects limités des cursus individuels. Une autre approche consiste à obtenir une description de l'ensemble du cursus. Cette approche est encore dans une phase exploratoire. Récemment, une attention particulière a été portée à des méthodes issues de la biologie (voir Abbott & Hrycak, 1990); ces méthodes sont utilisées pour comparer des informations successives. On compare des suites d'états et on procède à un dénombrement pour mesurer des similarités entre deux suites. Nous reviendrons sur cette méthode à la section 3.

Depuis longtemps déjà, on utilise l'analyse des correspondances (AC) pour analyser les cursus. Dans cette approche, les états du cursus de chaque individu sont définis par des fréquences correspondant au nombre de fois qu'un individu est passé dans chacun des états mutuellement exclusifs. Par exemple, s'il y a 8 états mutuellement exclusifs, on attribue à chacun des individus 8 fréquences qui décrivent le nombre de fois qu'un individu est passé

dans chaque état. Pour un groupe d'individus, on effectue une analyse des correspondances sur l'ensemble de ces fréquences. L'analyse des parcours individuels par l'AC a été proposée par Deville & Saporta (1980, 1983; voir aussi Deville, 1982, et Saporta, 1981, 1985). A leur suite, de Leeuw, van der Heijden & Kreft (1985), van der Heijden (1987), van der Heijden & de Leeuw (1989) ont également développé cette approche. On peut en trouver des applications chez van Buuren & de Leeuw (1992), van der Heijden & van den Brakel (1993), Martens (1994), Taris (1994).

Dans ce chapitre, nous cherchons à promouvoir cette approche qui, de notre point de vue, n'a pas encore reçu toute l'attention qu'elle mérite. A la section 2, nous exposons cette méthode appliquée à un tableau de contingence à deux dimensions. A la section 3, nous montrons pourquoi l'AC est un outil efficace pour traiter les données de cursus. Nous montrons comment l'AC permet d'étudier la dépendance entre le temps et les états du cursus. L'analyse des correspondances multiples (ACM) est introduite à la section 4 comme un cas particulier. A la section 5, nous présentons en détail une application à l'étude de cursus scolaire et montrons comment l'AC et l'analyse des correspondances multiples permettent de déterminer une ou plusieurs quantifications ou classifications de ces cursus scolaires[1].

2. Analyse des correspondances

Nous décrivons tout d'abord l'analyse des correspondances, puis nous présentons différentes applications à des données de cursus, de parcours individuel, etc. Pour une introduction à l'analyse des correspondances, on peut se référer à Benzécri et al. (1973), Nishisato (1980a), Greenacre (1984) et Gifi (1990). Nous évitons ici une description trop technique afin de nous concentrer sur les propriétés des résultats.

On présentera l'analyse des correspondances comme un outil permettant d'obtenir des représentations graphiques des tables de contingences. On considère une table de contingences ayant I lignes ($i = 1, \ldots, I$) et J colonnes ($j = 1, \ldots, J$); k_{ij} désigne les fréquences. Les fréquences marginales sont désignées par $k_{i.} = \sum_j k_{ij}$ et $k_{.j} = \sum_i k_{ij}$ et $k_{..} = \sum_{i,j} k_{ij}$. On notera également k_i pour $k_{i.}$, k_j pouur $k_{.j}$, k pour $k_{..}$ lorsqu'il n'y a pas d'ambiguïté. Les fréquences peuvent être transformées en proportion p_{ij} par $p_{ij} = k_{ij}/k$.

Dans l'analyse des correspondances, on s'intéresse à ce qu'il est convenu d'appeler les profils ligne et les profils colonne. Nous expliquerons d'abord comment étudier les profils ligne. Le profil ligne de la ligne i est défini comme un vecteur de fréquences conditionnelles p_{ij}/p_i dont la somme est égale à 1. Ces valeurs précisent les J fréquences conditionnelles des observations de la ligne i qui tombent dans la colonne j. Il y a I profils ligne, chacun d'entre eux pouvant être représenté par un point dans un espace de dimension J.

[1] Les auteurs remercient Rafaelle Huntjens pour le traitement informatique.

Dans cet espace, les coordonnées de la ligne i sont définies par le vecteur de composante p_{ij}/p_i. Dans cet espace, on associe des poids à chacune des J dimensions de telle façon que les dimensions ayant une plus petite fréquence marginale p_j jouent un rôle relativement plus important dans la définition de la distance entre les I points. Dans ce but, on attribue le poids $1/p_j$ à la dimension j. La distance $d(i, i')$ entre la ligne i et la ligne i' est alors définie par:

$$d^2(i,i') = \sum_{j=1}^{J} \left(\frac{1}{p_j}\right)\left(\frac{p_{ij}}{p_i} - \frac{p_{i'j}}{p_{i'}}\right)^2 . \qquad (2.1)$$

Il s'agit d'une distance euclidienne pondérée entre les profils i et i' avec les poids $(1/p_j)$. La distance du chi-carré (χ^2) (2.1) permet l'interprétation suivante: dans l'espace de dimension J, les lignes i et i' seront proches l'une de l'autre si, pour chaque profil j, les éléments p_{ij}/p_i et $p_{i'j}/p_{i'}$ sont proches. De la même façon, ils seront éloignés quand il y a une grande différence pondérée $p_{ij}/p_i - p_{i'j}/p_{i'}$ entre les lignes i et i'. La différence pour une colonne a une influence d'autant plus grande que p_j est faible. Le profil de la colonne des fréquences marginales d'éléments p_j correspond au centre de gravité 0 du nuage des points ligne. En effet, si l'on considère la moyenne pondérée des profils ligne avec les poids p_i, on obtient $\sum_i p_i(p_{ij}/p_i) = p_j$. La distance χ^2 d'un profil i à l'origine O est petite quand les éléments du profil ligne p_{ij}/p_i sont proches de p_j, et la distance à l'origine est grande quand certains éléments p_{ij}/p_i s'écartent beaucoup de p_j.

Remarquons que, quand la variable ligne est statistiquement indépendante de la variable colonne, c'est-à-dire que $p_{ij} = p_i p_j$, alors pour tout i les éléments du profil p_{ij}/p_i sont égaux à p_j; en d'autres termes tous les profils sont égaux aux profils de la colonne marginale. Ce résultat implique que tous les points sont confondus avec le centre de gravité. Il en résulte que l'étude de la relation entre la variable ligne et la variable colonne n'est utile que lorsque les fréquences s'écartent de l'indépendance. On se propose d'étudier les différences entre les I profils ligne en analysant le nuage des I points dans l'espace de dimension J. C'est une tâche difficile et, pour la simplifier, les I points de l'espace de dimension J sont projetés dans un espace de plus petite dimension. On effectue cette projection de telle façon que le maximum de l'information possible soit conservé sur les premières dimensions. Désignons par r_{ia} la nouvelle coordonnée du profil ligne i sur la dimension a. La projection sur la dimension 1 est déterminée de telle façon que la variance pondérée des distances à l'origine O ($\lambda_1^2 = \sum_i p_i r_{i1}^2$) soit maximisée. Pour la dimension 2, on maximise la variance pondérée des distances à l'origine ($\lambda_2^2 = \sum_i p_i r_{i2}^2$) sous la contrainte que les coordonnées ligne de la deuxième dimension soient orthogonales à celles de la première dimension: $\sum_i p_i r_{i1} r_{i2} = 0$. Il en va de même pour les dimensions suivantes.

Une présentation analogue peut être donnée pour les colonnes de la table de contingence. Les éléments d'un profil colonne j sont égaux à p_{ij}/p_j. On utilise ces éléments pour représenter le profil j comme un point d'un espace de

dimension I. On détermine ainsi J profils colonne dans cet espace. Des poids $1/p_i$ sont associés respectivement à chacune des I dimensions. La distance du χ^2 entre la colonne j et la colonne j' s'écrit alors:

$$d^2(j,j') = \sum_{i=1}^{I} \left(\frac{1}{p_i}\right)\left(\frac{p_{ij}}{p_j} - \frac{p_{ij'}}{p_{j'}}\right)^2 . \qquad (2.2)$$

Elle a une interprétation analogue à celle donnée plus haut pour les lignes. Les J points de l'espace de dimension I sont projetés dans un espace de plus petite dimension afin d'en simplifier l'analyse.

Désignons par c_{ja} la coordonnée du profil colonne j sur la dimension a. Sur la dimension 1, cette projection est déterminée de telle façon que la variance pondérée des distances à l'origine O ($\lambda_1^2 = \sum_j p_j c_{j1}^2$) est maximisée. Pour la dimension 2, on maximise la variance pondérée des distances à l'origine $\lambda_2^2 = \sum_j p_j c_{j2}^2$ sous la contrainte que les coordonnées colonne de la deuxième dimension soient orthogonales à celles de la première dimension: $\sum_j p_j c_{j1} c_{j2} = 0$. Il en va de même pour les dimensions suivantes.

Par conséquent, l'AC conduit à une solution pour les profils ligne et à une solution pour les profils colonne. Il s'agit d'une technique symétrique pour l'analyse d'un tableau de contingences. En effet, l'analyse d'un tableau conduit au même résultat que l'analyse du tableau transposé. Un des aspects intéressants de l'analyse des correspondances est que la solution pour les profils ligne est reliée étroitement à la solution obtenue pour les profils colonne. Les scores ligne r_{ia} peuvent être déduits des scores colonne c_{ja} et vice versa par les expressions suivantes:

$$r_{ia} = \lambda_a^{-1} \sum_{j=1}^{J} \frac{p_{ij}}{p_i} c_{ja} \qquad (2.3)$$

et

$$c_{ja} = \lambda_a^{-1} \sum_{i=1}^{I} \frac{p_{ij}}{p_j} r_{ia} . \qquad (2.4)$$

La relation (2.3) montre qu'à un coefficient près λ_a^{-1}, le profil de la ligne i est la moyenne pondérée des points colonne, avec les éléments du profil de la ligne i comme coefficient. De la même façon la relation (2.4) montre qu'au coefficient λ_a^{-1}, le profil de la colonne j est la moyenne pondérée des points ligne avec les éléments du profil de la colonne j comme coefficient. En d'autres termes, les éléments des profils déterminent où sont situés les points ligne et les points colonne dans leur représentation respective. En fait, quand nous comparons la représentation des profils ligne et des profils colonne, une ligne i est attirée dans la direction des colonnes pour lesquelles $p_{ij}/p_i > p_j$ et le profil d'une colonne j est attiré dans la direction des lignes pour lesquelles $p_{ij}/p_j > p_i$.

Les relations (2.3) et (2.4) prouvent également l'égalité des dimensions de la représentation des lignes et celle des colonnes qui est égale au minimum de

$I-1$ et $J-1$. On désignera par inertie la distance totale à l'origine $\sum_a \lambda_a^2$. On peut montrer que cette mesure est égale au χ^2 de Pearson divisé par la taille de l'échantillon k:

$$\sum_{a=1}^{J} \lambda_a^2 = \sum_{i,j} \frac{(p_{ij} - p_i p_j)^2}{p_i p_j} = \frac{\chi^2}{k} .$$

On peut utiliser l'inertie pour évaluer la part de la distance totale prise en compte par chaque dimension en calculant $\lambda_a^2 / \sum_a \lambda_a^2$ pour la dimension a.

3. Analyse des correspondances des données de cursus

A la section 2, nous avons donné une description technique de l'analyse des correspondances des tables de contingences. Nous précisons ici son intérêt pour l'analyse des données de cursus.

On peut se demander dans quelle situation l'analyse des correspondances est un outil adapté pour l'analyse des données. Une des réponses à cette question pourrait être la suivante: dans tous les cas où l'on peut construire une matrice de données pour laquelle il serait utile d'étudier la différence entre les profils ligne, les profils colonne ou entre les deux. Remarquons que c'est une question beaucoup plus générale que la description technique de la section 2 qui prend son origine dans l'analyse des tables de contingences à deux dimensions. Beaucoup d'autres types de données peuvent déterminer des matrices dont l'analyse des correspondances a un sens. Des exemples comme les données d'incidence, les données de préférence, les données quantitatives sont développés dans les références bibliographiques citées à la section 2.

Il est légitime d'utiliser l'analyse des correspondances dans le cadre des données de cursus. En effet, chaque cursus individuel peut être codé en données de fréquence. Les fréquences décrivent la durée qu'un individu a passée dans chacun des différents états mutuellement exclusifs. Donnons un exemple: supposons que nous nous intéressions au temps passé dans les différents états suivants: école, en emploi, sans emploi. Supposons que nous disposions de ces renseignements pour une cohorte d'individus qui sont suivis de 12 à 24 ans. Nous pouvons compter le nombre d'années qu'un individu a passées à l'école, en emploi et sans emploi. On pourrait alors obtenir les données présentées au tableau 2.1.

On voit par exemple que l'individu 1 a passé 6 ans à l'école, 6 ans en emploi et n'a jamais été sans emploi. L'individu 2 a passé 5 ans à l'école, 5 ans en emploi et 2 ans sans emploi. L'individu 3 a passé 8 ans à l'école, n'a jamais eu d'emploi et a passé 4 ans sans emploi. Les profils des individus sont respectivement 6/12, 6/12, 0 pour l'individu 1, 5/12, 5/12, 2/12 pour l'individu 2, 8/12, 0, 4/12 pour l'individu 3. L'analyse des correspondances étudie les différences entre ces profils en donnant une représentation bi-dimensionnelle des individus et des états. On montre, d'une part, quels sont les individus

qui ont des fréquences semblables (ou dissemblables) et, d'autre part, quels sont les états semblables (ou dissemblables), des états étant semblables ou dissemblables selon que des individus qui ont passé du temps dans un état vont vraisemblablement ou non passer du temps dans l'autre état.

Tableau 2.1. Exemple d'un codage de cursus individuel

		école	en emploi	sans emploi	total
individu	1	6	6	0	12
	2	5	5	2	12
	3	8	0	4	12
	etc.				

Remarquons qu'en changeant l'unité de mesure (mois, semaines, jours), on obtiendrait une augmentation importante des fréquences, mais cela ne modifierait que relativement peu les profils des individus. Remarquons également que nous perdons l'ordre des différents états. Afin de résoudre ce problème, on juxtapose deux matrices semblables à celles définies plus haut pour deux périodes de 6 ans. Nous obtenons ainsi le tableau 2.2.

Tableau 2.2. Décomposition du tableau 1 en deux périodes de 6 ans

		Premières six années			Six années suivantes		
		école	en emploi	sans emploi	école	en emploi	sans emploi
individu	1	6	0	0	0	6	0
	2	5	0	1	0	5	1
	3	6	0	0	2	0	4
	etc.						

Chaque profil possède maintenant 6 éléments. Les trois premiers états sont mesurés avant les trois suivants, mais l'analyse des correspondances n'utilise pas cette information. On peut pourtant utiliser cette information dans l'interprétation en distinguant ces états par des labels différents (voir les exemples présentés à la section 5). L'analyse des correspondances de la matrice avec seulement 3 états peut être considérée comme une version avec contrainte de l'analyse des correspondances avec 6 états, la contrainte étant l'égalité des scores colonne sur les unités de temps (voir van der Heijden, 1987; van Buuren & de Leeuw, 1992). Ceci montre également l'utilité de décomposer la période de 12 ans en 2 périodes de 6 ans quand les profils des catégories correspondantes diffèrent considérablement sur les deux périodes de 6 ans.

Il est aussi possible de coder les données en 12 périodes, 1 pour chaque année. Les données sont alors définies par des 0 et des 1 qui précisent si un individu est dans tel état pour une année particulière ou non. Dans l'exemple précédent, on obtiendrait les données présentées au tableau 2.3.

Tableau 2.3. Exemple de supermatrice d'indicateurs

		1	2	3	4	5	6	7	8	9	10	11	12
individu	1	100	100	100	100	100	100	010	010	010	010	010	010
	2	100	100	100	100	100	001	001	010	010	010	010	010
	3	100	100	100	100	100	100	100	100	001	001	001	001
	etc.												

Pour chaque année, on a des patterns 100, 010 et 001 indiquant dans quel état se situe l'individu. On désigne par matrice indicatrice la matrice définie pour chaque unité de temps et la juxtaposition de telles matrices par supermatrice d'indicatrices. Une AC de cette matrice conduirait à une solution avec 3×12 points état-année. L'AC de la matrice à 3 états, présentée plus haut, peut encore être considérée comme une version réduite de l'AC de la supermatrice d'indicatrices, la restriction étant que la quantification des catégories correspondantes est égale sur l'ensemble des 12 unités temporelles.

On appelle analyse des correspondances multiples l'AC de la supermatrice d'indicatrices. Ce type d'AC sera développé plus loin. Auparavant nous voudrions préciser que même si l'interprétation des résultats est plus simple quand les états sont mutuellement exclusifs, ces états n'ont pas nécessairement à être exhaustifs. En effet, on peut perdre la trace d'un individu pour une raison quelconque, par exemple durant les 2 dernières années. Dans le premier tableau, les fréquences pour le premier individu seraient alors 6, 4, 0. Il reste cependant utile de déterminer son profil (soit .6, .4, 0) et de le comparer aux autres profils. On résout ainsi le problème des données manquantes pour l'analyse des correspondances. Une autre approche consiste à définir une nouvelle catégorie appelée "perdue de vue". On obtiendrait alors les profils suivants: 6/12, 4/12, 0, 2/12 pour l'individu 1; 5/12, 5/12, 2/12, 0 pour l'individu 2 et 8/12, 0, 4/12, 0 pour l'individu 3. A la section 5, nous donnerons un exemple de ces deux approches et de leurs conséquences possibles. Avant d'effectuer une analyse des correspondances, on doit s'interroger sur la manière de construire la matrice des données. La question principale étant "comment comparer les cursus des individus après avoir calculé la distance du χ^2?", on posera un certain nombre de questions classiques dans le cas de deux cursus.

Est-ce que dans l'échelle de temps, les deux cursus montrent une bonne correspondance année par année de telle façon que, par exemple en 1992, les états soient comparés par le calcul de la distance du χ^2?

Ou bien utilisons-nous l'âge des répondants de telle façon que les états d'un individu âgé de 12 ans en 1993 soient comparés avec les états d'un autre individu âgé de 12 ans en 1990?

Ou encore comparons-nous les deux cursus en ajustant année par année leurs différents états (par exemple un individu commençant l'université en 1992 serait comparé avec un autre individu commençant en 1990)?

Bien entendu, le choix dépendra des questions que le chercheur se pose. Les différents choix induiront des patterns de données manquantes au début ou à la fin des données de cursus. Rappelons que les deux choix présentés dans la section précédente sont les plus appropriés pour résoudre ces problèmes de données manquantes. Le choix de la solution n'est pas évident. Il est préférable d'essayer chacune des solutions et de comparer les solutions données par l'analyse des correspondances. Il peut se faire également que, étant donné la problématique sous-jacente, la manière dont on doit ajuster les données de cursus dépende des données elles-mêmes. Par exemple, supposons que l'un des cursus présente la suite des états suivants $a-b-c-d$ et un autre la suite $b-c-d-a$; on peut alors faire correspondre exactement la suite $b-c-d$ (voir tableau 2.4).

Tableau 2.4. Exemple de correspondance des suites

a	b	c	d	manquant
manquant	b	c	d	a

On peut utiliser à cet effet une technique d'ajustement optimal (voir Abbott & Hrycak, 1990). Après avoir fait correspondre les suites, on suggère de ne pas calculer une mesure de distance en dénombrant les différences entre les cursus, mais d'appliquer l'AC pour obtenir une représentation graphique des cursus. A notre connaissance, il n'existe pas d'exemple d'une telle approche.

4. Analyse des correspondances multiples des données de cursus

L'analyse des correspondances multiples des données de cursus peut être interprétée en termes de distance du χ^2. Mais cela présente des difficultés que nous présenterons ci-dessous. Selon la distance du χ^2 (2.1), des cursus seront proches ou éloignés si les individus utilisent ou non leur temps de la même façon.

L'origine O est définie par le profil d'éléments p_j qui correspond ici aux fréquences relatives avec lesquelles les états sont utilisés à chaque unité de temps. Plus le profil des individus est différent du profil moyen et plus ils

sont éloignés de l'origine. L'analyse des correspondances multiples peut être utilisée pour localiser des groupes d'individus qui se distinguent de l'origine de la même façon. On peut définir une matrice croisant les unités de temps (ici 12) par les états (ici 3) à l'aide des éléments correspondants f_j. Cette matrice montre quels sont les états les plus fréquents et pour quelle unité de temps. Il est important d'avoir une idée de cet aspect des données, car l'analyse des correspondances multiples met seulement l'accent sur l'écart avec cette moyenne. Les groupes sont définis en fonction de leur écart avec cette moyenne. Van der Heijden & de Leeuw (1989) ont proposé alors d'effectuer trois analyses dans le cas de données de cursus ("event history data"), à savoir les analyses de correspondances des matrices suivantes:

i la matrice des marges f_j montrant quels sont les états qui sont utilisés et à quelle unité de temps;

ii la supermatrice d'indicatrices montrant comment les individus s'écartent de la moyenne étudiée sous (i);

iii la matrice où les cursus ne sont pas décomposés en différentes matrices d'indicateurs, ce qui peut être considéré comme une version restreinte de l'analyse précédente (ii).

Si la configuration des profils de cursus est très voisine dans les analyses (iii) et (ii), alors on ne gagne pas beaucoup d'informations avec la procédure (iii). Cependant, on gagne beaucoup de stabilité en choisissant la procédure (iii) au lieu de la procédure (ii). A la section 5, on donne un exemple de ces trois analyses.

On considère en général que ce n'est pas une bonne idée d'interpréter l'analyse des correspondances multiples en termes de distance du χ^2. C'est essentiellement dû au fait qu'il y a des contraintes dans le pattern complet des 0 et 1, c'est-à-dire qu'à chaque unité de temps, seule une valeur 1 peut se produire. Cela conduit à des dimensions qui peuvent être considérées comme artificielles. Dans l'espace total, la distance du χ^2 présente aussi des éléments artificiels. Ceci dépassant le cadre de ce chapitre, nous n'en dirons pas plus. Pour des développements sur ce point, on peut se référer à Israëls (1987). Cependant, il est clair que d'autres arguments peuvent légitimer le recours à l'AC. Par exemple, l'interprétation de l'analyse des correspondances comme une analyse en composantes principales de données qualitatives (de Leeuw & van Rijckevorsel, 1980) sera développée ci-après.

4.1. L'analyse des correspondances multiples considérée comme une analyse en composantes principales de données qualitatives

L'une des façons de définir l'analyse en composantes principales (ACP) est la suivante: soit une matrice X de n lignes et m colonnes de mesures quantitatives; alors la première composante principale z_1 est l'un des scores qui maximise $\phi_1 = \sum_i (\mathrm{cor}(z_1, x_i))^2$, où ϕ_1 est la première valeur propre. Donc la première composante principale z_1 résume ce que les m X-

variables ont en commun. La seconde composante principale z_2 maximise $\phi_2 = \sum_i (\text{cor}(z_2, x_i))^2$, sous la contrainte que $\text{cor}(z_1, z_2) = 0$, et ainsi de suite de z_3 à z_m. Plaçons-nous maintenant dans le cas où l'information dans la matrice X est qualitative; par exemple, dans le cas des données de cursus précédentes, nous avions 12 variables, chacune d'elles comprenant 3 états: école, emploi et sans emploi. Supposons que nous souhaitions effectuer une ACP, mais que nous ayons besoin de mesures quantitatives pour remplacer les différents états. Si nous utilisons les scores colonne c_{j1} obtenus pour la première dimension dans l'analyse des correspondances multiples, alors les scores ligne correspondant à la première dimension rassemblée sous r_1 sont les scores qui maximisent la première valeur propre $\phi_1 = \sum_i (\text{cor}(r_1, x_i))^2$. On voit donc que la première dimension dans l'ACM peut être interprétée en termes d'ACP de données qualitatives. Pour les dimensions plus élevées, on observe également une relation entre ACP et ACM (voir Gifi, 1990, section 3).

4.2. Analyse des correspondances multiples et tableau de Burt

La relation existant entre la supermatrice d'indicatrices et ce qu'il est convenu d'appeler le tableau de Burt est intimement liée à l'interprétation de l'ACM en termes d'ACP de variables qualitatives. Soit G la supermatrice d'indicatrices, on peut démontrer que la solution de l'ACM peut être obtenue en effectuant une décomposition en valeurs singulières d'une fonction de la matrice $G'G$, celle-ci étant une matrice carrée symétrique que l'on appelle le tableau de Burt.

Chaque carré de cette matrice représente une sous-matrice. Sur la diagonale, nous trouvons des matrices diagonales avec les fréquences marginales des trois états pour chaque unité de temps; en dehors de la diagonale nous trouvons les tables de contingences croisant des unités de temps t pour les lignes avec t' pour les colonnes. Ces tables de contingences sont les matrices de transition montrant ou précisant combien d'individus de l'état j au moment t vont dans l'état j' au moment t'. Le point précédent sur l'ACM et l'ACP de données qualitatives illustre en fait le point suivant: en quantifiant les états pour chaque unité de temps, chaque matrice peut être résumée par un coefficient de corrélation. Comme le tableau de Burt possède toutes les informations nécessaires pour déterminer la configuration d'une colonne dans l'ACM, il devient évident que seules des suites de deux unités de temps sont utilisées dans une analyse et que l'information concernant des transitions de plus de deux unités de temps est négligée. On trouvera des remarques concernant la relation entre l'ACM et les modèles de chaînes de Markov chez van der Heijden & de Leeuw (1989).

5. Exemple: cursus scolaires et professionnels d'une cohorte d'élèves dans une zone d'éducation prioritaire

Nous poursuivons plusieurs objectifs avec cet exemple: tout d'abord, nous voulons montrer l'utilité de l'AC et de l'ACM pour analyser des données de cursus; ensuite, on veut montrer dans quelles circonstances on peut obtenir des quantifications ou des classifications des cursus utiles pour des analyses ultérieures; enfin, nous voulons comparer les résultats obtenus en fonction de différents choix méthodologiques pour traiter des données manquantes.

5.1 Présentation du système scolaire hollandais

Après 8 ans d'école primaire (beaucoup d'élèves migrants de par leur arrivée tardive n'ont pas fait le cycle complet), les élèves reçoivent une éducation secondaire obligatoire jusqu'à l'âge de 16 ans (la figure 2.1 présente la structure scolaire hollandaise). Il existe trois divisions secondaires. Cependant, pour être orientés dans une de ces divisions, la plupart des élèves fréquentent pendant un ou deux ans un cycle d'orientation ("bridge-classes" - bk1 ou bk2). Les trois divisions secondaires sont:

a école professionnelle préparatoire (vbo); cette division dure 4 ans et consiste en une formation professionnelle élémentaire de base (lbo) et individualisée (ibo);

b formation générale (avo); elle peut prendre deux formes: moyenne (mavo) et haute (havo), qui durent respectivement 4 ou 5 ans;

c formation scientifique (vwo); elle peut également prendre deux formes: moderne (atheneum) et classique (gymnasium), chacune durant 6 ans.

Après l'école secondaire, la plupart des élèves ont de 16 à 18 ans et continuent leur cursus scolaire dans différentes formations supérieures. Cependant, une minorité d'élèves choisissent une activité professionnelle.

La plupart des élèves de vbo et mavo continuent leur cursus scolaire en suivant une formation professionnelle (mbo) qui dure 4 ans ou une formation professionnelle plus courte (kmbo) de 2 ans. Les élèves peuvent également choisir une formation mixte "travail-étude" (apprentissage). Les élèves de avo peuvent poursuivre leurs études en suivant une formation professionnelle supérieure (hbo) qui dure de 2 à 5 ans. vwo prépare à l'université, mais un nombre important de ces élèves se destinent également à la formation professionnelle supérieure. Sur la figure 2.1, on peut aussi remarquer que le passage d'une filière à une autre est toujours possible.

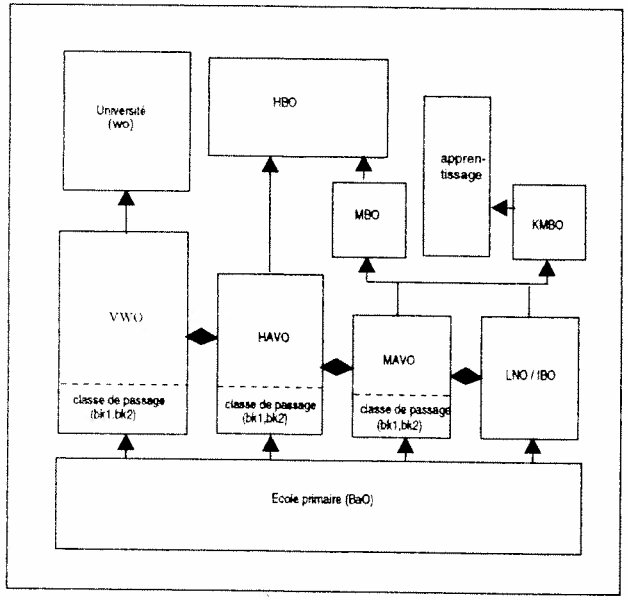

Fig. 2.1. Présentation succincte du système scolaire hollandais. Les flèches indiquent des changements entre des types de formation qui ne sont pas rares.

5.2 Échantillon

688 élèves du 8e degré appartenant à 31 écoles primaires ont participé à cette recherche. Ces écoles sont situées dans trois grande villes (La Haye, Rotterdam et Utrecht) de l'ouest et du centre des Pays-Bas. Ces écoles sont organisées de manière plus ou moins identiques. Elles se situent toutes dans des zones d'éducation prioritaire. Ces zones ont été sélectionnées par le gouvernement hollandais car de nombreux élèves y sont en difficulté scolaire. Les écoles de ces zones reçoivent des subsides spéciaux pour engager des enseignants supplémentaires et ainsi diminuer le nombre d'élèves par classe.

La population scolaire dans des zones d'éducation prioritaire est caractérisée par une forte proportion d'élèves migrants et d'élèves hollandais issus d'un milieu socio-économique défavorisé. Par exemple, dans la moitié de ces écoles, on trouve plus de 33% d'élèves migrants. On retrouve cette même diversité ethnique dans notre échantillon: au total 43% d'élèves hollandais et 57% d'élèves migrants. Parmi eux, 18% viennent de Turquie, 16% du Surinam et 14% du Maroc. Cela correspond aux proportions habituelles dans les zones d'éducation prioritaire.

Après avoir sélectionné les 31 écoles primaires, on considère la cohorte définie par tous les élèves du 8e degré. On établit année après année le cursus scolaire de ces élèves au cours de leur formation secondaire et éventuellement supérieure (Figure 2.1). Ces informations nous ont été fournies par les établissements secondaires. Dans certains cas, on a dû recueillir des informations directement auprès de ces enfants. Bien sûr certains élèves disparaissent de notre échantillon originel lorsqu'ils sortent du système scolaire, par exemple pour travailler, retourner dans leur pays d'origine, se marier, effectuer leur service militaire, etc.

Nous avons codé le cursus scolaire de la manière suivante: à partir de l'année scolaire 1985/86, chaque élève est classé dans un des 38 niveaux, pour la plupart des niveaux scolaires (voir les lignes du tableau 2.5). La plupart des niveaux trouvent leur explication dans la figure 2.1. Bao est la dernière année de l'école primaire. On voit que 13 élèves n'ont pas passé du 8e degré au 1er degré de l'école secondaire. bk1 et bk2 correspondent aux classes d'orientation. Il y a 4 niveaux pour ibo, 4 pour lbo, 4 pour mavo, 5 pour havo (mais havo1 n'apparaît pas dans nos données), 6 pour vwo.

Il y a ensuite un niveau pour tous les autres degrés primaires, comme le "Middenschool". Nous passons ensuite aux niveaux qui suivent vwo, havo, mavo et lbo/ibo. Nous trouvons kmbo, quatre niveaux pour mbo et trois niveaux pour hbo/wo. Les niveaux restants sont "travail", "rééducation" (i.e. choisir un autre type d'éducation afin d'accroître ses chances de trouver un emploi), apprentissage. On doit tenir compte aussi d'un certain nombre d'enfants pour lesquelles l'information manque. La catégorie "autre" concerne des élèves qui se répartissent grossièrement de la façon suivante: 75% des élèves migrants retournent au Maroc ou en Turquie, 3% des élèves font leur service militaire, 20% des élèves sont inscrits mais ne fréquentent plus l'école et 3% des élèves sont décédés. La catégorie "autre" s'accroît constamment, ce qui est spécifique à ce type de données. L'accroissement du nombre des élèves manquants est en relation avec le fait que des élèves quittent l'école sans fournir d'informations sur la suite de leur cursus.

Tableau 2.5. Types de formation par année; dans une cellule une fréquence correspond au nombre d'individus fréquentant tel type de formation pour une année particulière. Les abrévations dans les lignes sont expliquées à la figure 2.1.

	85/6	86/7	87/8	88/9	89/0	90/1	91/2	92/3	93/4
Bao	13	1	-	-	-	-	-	-	-
bk1	362	61	5	-	-	-	-	-	-
ibo1	40	7	1	-	-	-	-	-	-
ibo2	-	46	8	2	-	-	1	-	-
ibo3	-	-	46	14	3	-	-	-	-
ibo4	-	-	-	35	8	3	-	-	-
lbo1	118	8	-	-	-	-	-	-	-
lbo2	-	142	43	5	-	-	-	-	-
lbo3	-	1	167	87	10	1	-	-	-
lbo4	-	-	-	126	49	12	1	-	-
mavo1	101	13	1	-	-	-	-	-	-
mavo2	1	149	58	5	-	-	-	-	-
mavo3	-	-	141	86	8	-	-	-	-
mavo4	-	-	-	117	55	6	2	-	-
havo2	-	18	10	-	-	-	-	-	-
havo3	-	-	31	23	-	-	-	-	-
havo4	-	-	-	28	37	14	3	-	-
havo5	-	-	-	-	9	25	14	1	-
vwo1	2	-	-	-	-	-	-	-	-
vwo2	-	27	1	-	-	-	-	-	-
vwo3	-	-	39	2	-	-	-	-	-
vwo4	-	-	-	29	3	1	-	-	-
vwo5	-	-	-	-	16	4	7	1	-
vwo6	-	-	-	-	-	14	4	7	2
o.pri	24	22	19	18	4	1	1	-	-
kmbo	-	-	2	-	21	20	13	4	-
mbo1	-	-	-	-	20	23	7	4	-
mbo2	-	-	-	-	-	10	13	5	-
mbo3	-	-	-	-	-	-	6	9	4
mbo4	-	-	-	-	-	-	-	4	-
hbo/wo1	-	-	-	-	-	-	1	3	-
hbo/wo2	-	-	-	-	-	-	-	1	2
hbo/wo>5	-	-	-	-	-	-	1	-	-
travail	-	-	-	2	13	29	37	42	42
ch.éduc	-	-	-	-	2	7	2	3	-
apprentissage	-	-	-	2	15	19	13	7	3
autre	27	37	61	84	111	183	223	247	261
manquant	0	0	3	17	304	316	339	350	374

Comme nous nous occupons d'une cohorte d'élèves qui fréquentaient le dernier degré de l'école primaire en 1984/85, on constate dans le tableau 2.5 un grand nombre de fréquences nulles pour l'année 1985/86. Le schéma de la figure 2.1 montre qu'après l'école primaire, tous les élèves commencent en

vwo1, o.pri, mavo1, lbo1, ibo1 ou bk1. Par exemple, 101 élèves commencent en mavo1 en 1985/86. En 1986/87, il y a 13 élèves en mavo1 (probablement des élèves qui ont échoué le passage en mavo2); en 1987/88, il reste seulement 1 élève dans ce type de formation; pour les années suivantes, ce niveau n'est plus fréquenté pour des raisons évidentes. En ce qui concerne la formation de type mavo, le cursus scolaire habituel pour les 4 premières années est soit mavo1-mavo2-mavo3-mavo4 ou bk1-mavo2-mavo3-mavo4. Cela explique le nombre important d'élèves en mavo2 pour 1986/87, en mavo3 en 1987/88, et en mavo4 en 1988/89. Bien que ces cursus soient les plus habituels, il est clair que beaucoup d'élèves ne suivent pas cette voie en raison d'échec ou de changement d'orientation.

5.3 Analyses

On étudie les cursus comme un cas particulier de données de parcours individuels. Pour ce type de données, van der Heijden & de Leeuw (1989) (voir aussi van der Heijden, 1987) suggèrent de se centrer sur 3 types d'analyses que nous discuterons ci-dessous.

La première analyse proposée par van der Heijden & de Leeuw (1989) est l'analyse des correspondances (AC) de la table de contingences croisant les catégories par le temps (voir tableau 2.5). Dans cet exemple, les lignes de la table de contingences sont les 38 degrés scolaires, et ses colonnes les 9 années de scolarité. Ce tableau présente les fréquences d'élèves par niveau et par année. L'analyse des correspondances de ce tableau ne semble pas fournir beaucoup plus d'informations que celles que l'on peut obtenir du tableau par une simple inspection visuelle. Par conséquent, nous ne présenterons pas les résultats d'une telle analyse.

5.3.1 Analyse des correspondances multiples sans les données manquantes

Le tableau 2.5 fournit des informations sur la fréquentation des degrés scolaires au cours du temps. Il ne donne pas d'information sur les cursus individuels. L'étude de ces cursus révèle notamment quels types de réorientation se sont produits entre les différents types de formation, comme ibo, lbo, mavo, havo, vwo, mbo et hbo/wo. Nous étudions ces cursus en effectuant une analyse des correspondances multiples. Dans ce but, nous considérons un tableau dont les lignes sont les 688 élèves, et les colonnes les niveaux scolaires par année. Bien que 38 niveaux scolaires soient fréquentés durant 9 années, le nombre de colonnes est inférieur à 9×38 car, durant les premières années, certains niveaux solaires ne sont pas encore fréquentés; de même par la suite il n'y a plus de données pour certaines années. En fait, dans le tableau 2.5, seules des combinaisons d'années et de niveaux qui ont des fréquences supérieures à 0 ont une colonne. Ainsi, pour l'année 1985/86, il y a 9 colonnes, pour l'année 1986/87, il y a 14 colonnes, etc. Le tableau final indique par (1) ou (0) si

un élève appartient à un niveau scolaire donné et pour une année donnée ou pas. Si une information est manquante pour un élève et pour une année particulière, on lui attribue la valeur 0 pour chacun des 38 niveaux scolaires. On effectue alors une analyse des correspondances multiples sur ce tableau.

Les quatre premières valeurs propres issues de l'analyse des correspondances multiples sont .72, .65, .62 et .58. La figure 2.2 présente un graphique qui illustre la quantification des élèves obtenue pour les deux premières dimensions.

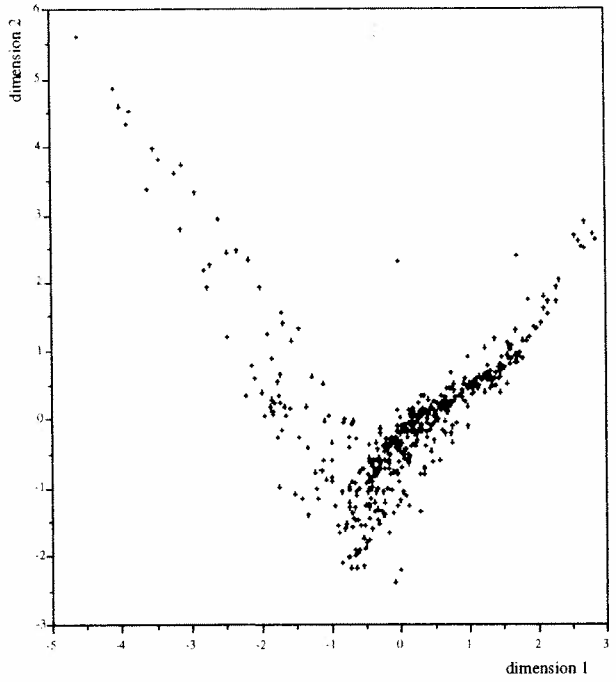

Fig. 2.2. Projection de 688 cursus sur les deux premières dimensions de l'analyse de correspondances avec 38 niveaux du cursus scolaire et professionnel

On remarque que le nuage des points des élèves a la forme parabolique caractéristique de l'effet Guttman ("effet fer à cheval") sur le premier plan factoriel. Dans une telle situation, la première dimension reflète la structure principale des données et il n'est pas très utile d'étudier les dimensions supérieures (voir Schriever, 1983, 1986; van Rijckevorsel, 1986). Par conséquent, en ce qui concerne les niveaux scolaires, nous retenons leur quantification sur la première dimension. Dans la représentation graphique qui en est donnée à la

figure 2.3, les niveaux scolaires annuels qui ne contiennent que deux élèves sont éliminés afin d'en simplifier l'interprétation.

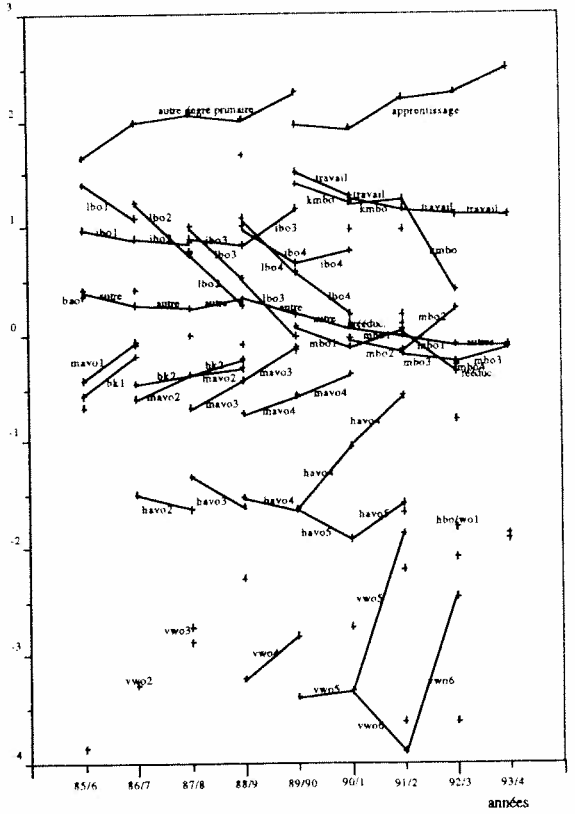

Fig. 2.3. Projection de 688 cursus sur les deux premières dimensions de l'analyse des correspondances avec 38 niveaux du cursus scolaire et professionnel

Pour interpréter les figures 2.2 et 2.3, il est utile de considérer les formules de transition (voir équations 2.3 et 2.4) et les formules de la distance du χ^2 (2.1 et 2.2). Ces formules fournissent le cadre général suivant pour l'interprétation (voir Gifi, 1990, pp. 118-120):

i dans la figure 2.2, les élèves qui sont proches représentent des cursus scolaires très semblables. Des élèves qui sont éloignés représentent des cursus scolaires présentant peu de similarités. Ces remarques se déduisent de l'équation de la distance du χ^2 entre les profils ligne;

ii dans la figure 2.3, les combinaisons année-cursus scolaire, dont les quantifications sont proches, sont fréquentées simultanément par un grand

nombre d'élèves; par exemple havo2 en 1986/87 et havo3 en 1987/88 sont proches, donc les élèves qui étaient en havo2 en 1986/87 ont une probabilité plus grande que la moyenne d'être en havo3 en 1987/88. En d'autres termes, la probabilité conditionnelle d'être en havo3 en 1987/88, sachant qu'on est en havo2 en 1986/87, est plus grande que la probabilité non conditionnelle d'être en havo3 en 1987/88. Ceci résulte de l'équation de la distance du χ^2 entre les profils colonne;

iii dans la figure 2.2, les cursus scolaires sont proches des combinaisons type de scolarité-année dont ils sont composés (figure 2.3) et vice versa. Cela résulte des formules de transition. On peut alors interpréter la figure 2.3 en relation avec la figure 2.2 de la façon suivante:

a les cursus scolaires à l'extrême gauche de la figure 2.2 (sur la première dimension) sont définis principalement par des années vwo. En allant de l'extrême gauche jusqu'à l'extrême droite, les carrières scolaires résultent essentiellement de vwo, havo, mavo, bk1 and bk2, mbo, lbo and ibo, kmbo, emploi, degrés primaires et enfin apprentissage. La dimension 1 ordonne les types de scolarité en fonction de ce qui peut être considéré comme des critères de réussite ou d'échec;

b la propriété ii permet l'interprétation suivante: les cursus scolaires qui ont des années vwo comprennent également d'autres types de scolarité qui sont pour la plupart des années havo. Pour les carrières scolaires qui ont surtout des années havo, les autres années possibles sont des années vwo et mavo. Les carrières qui se terminent par des années mbo commencent surtout par des années mavo et lbo. Les carrières qui se terminent par des formations en cours d'emploi ont commencé surtout par d'autres degrés primaires, lbo et ibo;

c comme on l'a vu ci-dessus, les cursus scolaires s'ordonnent de haut en bas à la figure 2.3 (de gauche à droite à la figure 2.2), allant de la réussite à une moins bonne réussite. Ceci permet d'obtenir des *résultats inattendus*. Par exemple, la quantification pour vwo4 en 1988/89 caractérise plus qu'en 1989/90 des cursus de réussite. Ces résultats sont cohérents car les cursus scolaires qui ont vwo4 en 1989/90 ont vwo4 en 5e année; par conséquent, les élèves concernés n'ont pas dépassé ce niveau. On observe un phénomène contraire dans le cas des années havo. Par exemple, des élèves qui, en 1989/90, ont havo3 en 4e année réussissent mieux que ceux qui, en 1987/88, ont havo3 en 3e année! On observe les résultats attendus pour mavo et bk mais pas pour lbo.

Pour havo3, nous avons comparé les 31 cursus qui ont havo3 en 1987/88 avec les 23 cursus qui ont havo3 en 1988/9 (Tab. 2.5). Nous avons calculé le nombre d'années que les deux groupes ont passées en vwo et en vwo-havo. Pour l'ensemble des élèves qui se caractérisent par havo3 en 1988/89 (niveau que les élèves ne dépassent pas), on voit qu'ils ont passé 12 des 162 années

non manquantes en vwo (proportion 7.4%) et 94 années en vwo ou havo (proportion 58%). Pour l'ensemble de ceux qui ont havo3 en 1987/88 (élèves qui réussissent), il n'y a que 3 années des 179 en vwo (1.7%) et 94 des 179 en vwo-havo (52.5%). On voit alors pourquoi les cursus avec havo3 en 1987/88 (élèves qui réussissent) sont plus proches des carrières d'échec et que les havo3 en 1988/89 (élèves qui échouent) sont plus proches des carrières de réussite.

Pour mavo3, nous comparons les 141 carrières qui ont suivi mavo3 en 1987/88 avec les 86 qui ont suivi mavo3 en 1988/9 (Tab. 2.5). Nous calculons le nombre d'années, passées par les deux groupes en havo-vwo-hbo/wo et en bk-mavo-mbo-havo-vwo-hbo/wo. On constate que, pour les cursus ayant mavo3 en 1988/89 (élèves qui ne dépassent pas ce niveau), 11 des 588 années non manquantes sont en havo-vwo-hbo/wo (1.87%) et 400 des 588 années non manquantes sont en bk-mavo-mbo-havo-vwo-hbo/wo (68%). Pour les cursus ayant mavo3 en 1987/88 (cursus de réussite), 30 des 842 années non manquantes sont en havo-vwo-hbo/wo (3.56%) et 94 des 179 en bk-mavo-mbo-havo-vwo-hbo/wo (72.2%). Cela explique pourquoi les cursus ayant havo3 en 1987/88 (réussites) sont plus proches des carrières de réussite et que les cursus ayant havo3 en 1988/89 (échecs) sont plus proches des carrières d'échec.

Pour lbo2, nous comparons les 142 cursus qui ont suivi lbo2 en 1986/87 avec les 43 qui ont suivi lbo2 en 1987/88 (Tab. 2.5). Pour chacun des groupes, nous déterminons le nombre d'années passées en mbo et en mbo-mavo-bk. Nous remarquons que, pour les cursus lbo2 en 1987/88 (cursus d'échec), 7 des 281 années non manquantes sont en mbo (2.49%) et 48 années en mbo-mavo-bk (17.1%). Pour les cursus lbo2 en 1986/87 (carrières de réussite), il y a 17 des 861 années en mbo (2%) et 57 des 860 années en mbo-mavo-bk (6.6%). On comprend pourquoi les cursus lbo2 en 1986/87 (réussites) sont plus proches des carrières d'échec, alors que les cursus lbo2 en 1986/87 (échecs) sont plus proches des carrières de réussite.

On voit, à la lumière de cette interprétation, que l'on peut utiliser les quantifications sur la première dimension comme un score résumant le cursus scolaire de ces élèves. L'étape suivante dans l'analyse consiste alors à établir une relation entre les quantifications obtenues par l'analyse des correspondances multiples et certaines variables quantitatives et qualitatives caractérisant l'élève. Dans le cas des variables quantitatives, on calculera leurs corrélations avec les autres variables. Pour les variables qualitatives, on calculera la moyenne des scores pour chacune de leurs modalités.

Une première variable intéressante est celle que l'on désigne par le score "Dutch CITO" qui correspond a une évaluation des acquisitions scolaires avant l'entrée dans l'école secondaire. On dispose de ce score pour 382 élèves. On tient compte de ce score pour orienter l'élève dans une école secondaire de type professionnel ou général. Pour ces 382 élèves, la corrélation entre le score au CITO et la première dimension de l'analyse des correspondances multiples est de .58. De plus, pour certaines variables qualitatives pouvant avoir un lien avec la carrière scolaire, nous avons calculé les moyennes des quantifications

des cursus scolaires sur le premier axe de l'analyse des correspondances multiples (voir les résultats à la figure 2.4). Les différences des moyennes entre garçons et filles ne sont pas significatives. Les cursus scolaires de réussite sont liés aux écoles catholiques romaines et protestantes, aux garçons européens, aux garçons originaires du Surinam et aux enfants dont le père est un employé (peu qualifié et qualifié). Les cursus scolaires de moindre réussite sont liés aux écoles publiques, aux élèves d'origine marocaine et turque et aux enfants dont le père est un ouvrier non qualifié.

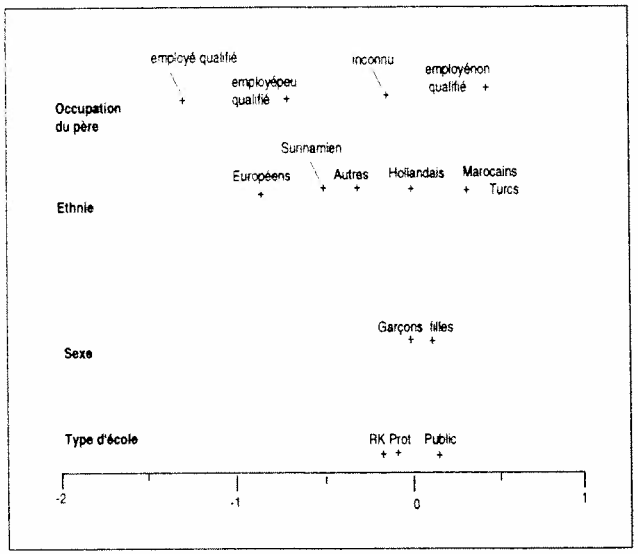

Fig. 2.4. Scores moyens sur le première dimension de l'analyse des correspondances pour quatre variables signalétiques: occupation du père, éthnie, sexes et types d'école

5.3.2 Analyse des correspondances multiples avec des contraintes d'égalité

Dans l'analyse précédente, nous avons vu à la figure 2.3 que certains changements d'orientation des lignes correspondent à certains aspects intéressants des données. Par exemple, des élèves qui échouent en mavo obtiennent un faible score de carrière scolaire, alors que les élèves qui échouent en havo et en lbo obtiennent un score plus élevé. Cependant, certaines variations au cours du temps pourraient être légèrement instables en raison de la petite taille de l'échantillon.

Une autre analyse proposée par van der Heijden & de Leeuw (1989) permet d'obtenir des résultats plus stables. Il s'agit de l'analyse des correspondances (AC) de la matrice croisant les 688 élèves avec les 38 niveaux du cursus scolaire et professionnel. Cette analyse est équivalente a une analyse des correspondances multiples avec une contrainte. En effet, les quantifications de niveaux identiques doivent être égales au cours du temps (van der Heijden & de Leeuw, 1989; van Buuren & de Leeuw, 1992). Autrement dit, sur la figure 2.3, les lignes doivent être horizontales. Les 688 cursus sont représentés sur les deux premières dimensions dans la figure 2.5. On observe à nouveau un effet Guttman. De plus, la corrélation entre les quantifications des cursus scolaires obtenues par cette analyse et les quantifications des cursus scolaires obtenues par l'analyse des correspondances multiples est égale à .9927. Les quantifications des cursus scolaires obtenues par les deux analyses peuvent donc être considérées comme identiques.

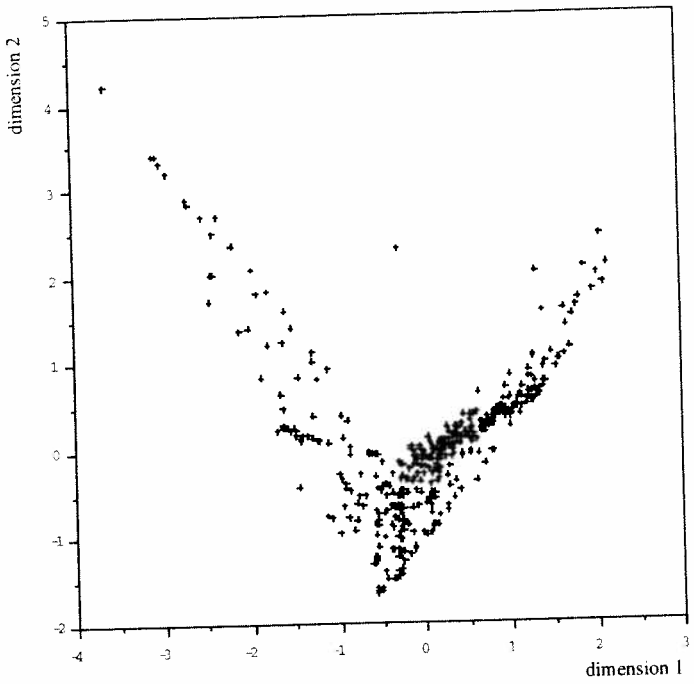

Fig. 2.5. Projection de 688 cursus sur les deux premières dimensions de l'analyse de correspondances utilisant 38 niveaux du cursus scolaire et professionnel

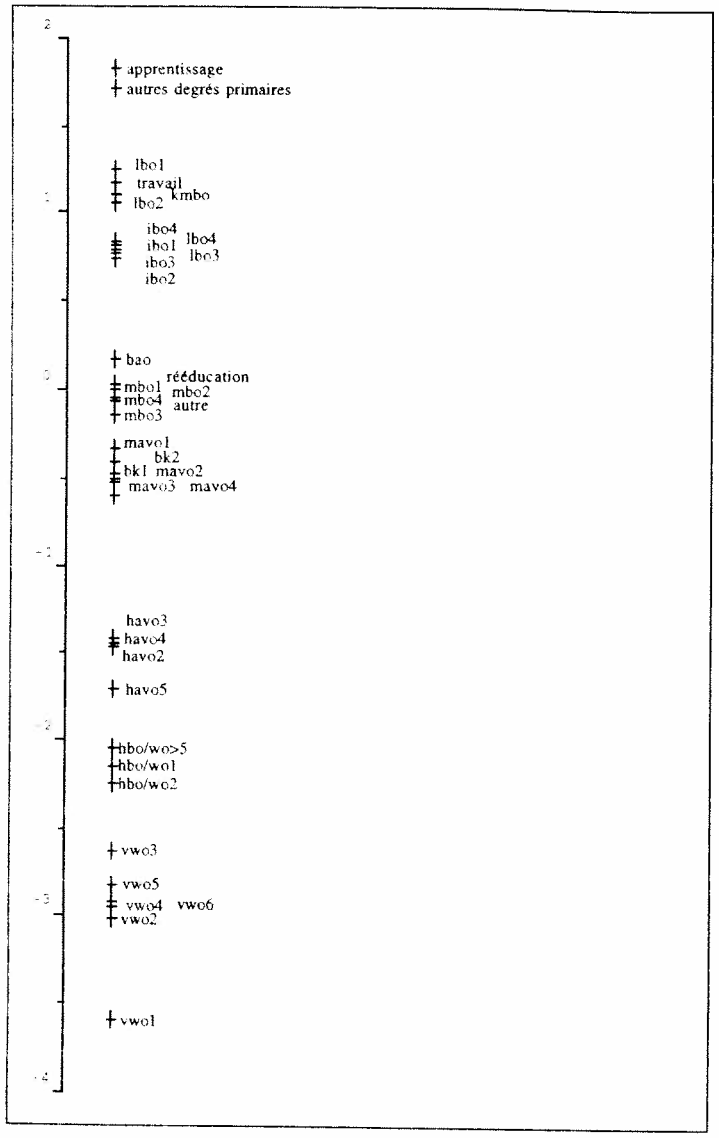

Fig. 2.6. Projection de 688 types de formation sur la première dimension de l'analyse des correspondances utilisant 38 niveaux du cursus scolaire et professionnel

Sur la figure 2.6, on présente les niveaux de quantification obtenus dans cette analyse. Il est clair que ces quantifications sont très proches des moyennes des lignes correspondantes à la figure 2.3.

5.3.3 Analyse des correspondances multiples avec des données manquantes

La façon de traiter les données manquantes est un des points cruciaux lors de l'application de l'analyse des correspondances multiples. Dans le cas de cette étude, on a déjà rencontré deux types de données manquantes qui ont été traitées différemment:

i l'information est connue mais ne concerne pas les carrières scolaires. On lui a attribué une modalité "autre". Elle concerne principalement les élèves faisant leur service militaire, rentrant au Maroc ou en Turquie, ou encore décédés. Cette modalité est une modalité active dans l'analyse et, pour chaque année, elle se situait près de l'origine du premier plan. Ceci indique qu'elle n'avait pas de relation claire avec un cursus scolaire de réussite ou d'échec. On voit dans le tableau 2.5 que le nombre de garçons tombant dans cette modalité croît constamment, allant de 27 la première année à 261 la dernière année (parmi les 688 élèves);

ii l'information n'est pas encore connue. Au tableau 2.5, elle est indiquée par "manquant". Ce phénomène concerne beaucoup d'élèves à partir de leur 5e année. Ils ont achevé des cursus comme ibo, lbo et mavo, et comme ils ne suivent plus une formation, ils sont plus difficiles à suivre. Cette option "manquant" est la façon la plus courante de traiter l'information manquante dans une analyse des correspondances multiples: en termes de supermatrice d'indicatrices, si un élève est "manquant" pour une année, on lui attribue des zéros pour toutes les modalités (pour plus de détails, voir Meulman, 1982; van der Heijden & Escofier, 1988; Gifi, 1990; van Buuren & van Rijckevorsel, 1992).

Les données manquantes du type ii seront également traitées en utilisant une modalité distincte, et nous montrerons l'influence d'une telle approche sur les solutions dans l'analyse des correspondances multiples.

Les résultats de l'analyse des correspondances multiples font ressortir les 4 premières valeurs propres suivantes: .5703, .5687, .5041 et .4514. On voit que la première et la seconde dimensions ne se distinguent pas clairement. Les valeurs propres pouvant être interprétées comme des variances, il en résulte qu'une rotation du nuage de points dans le premier plan conduirait approximativement aux mêmes variances.

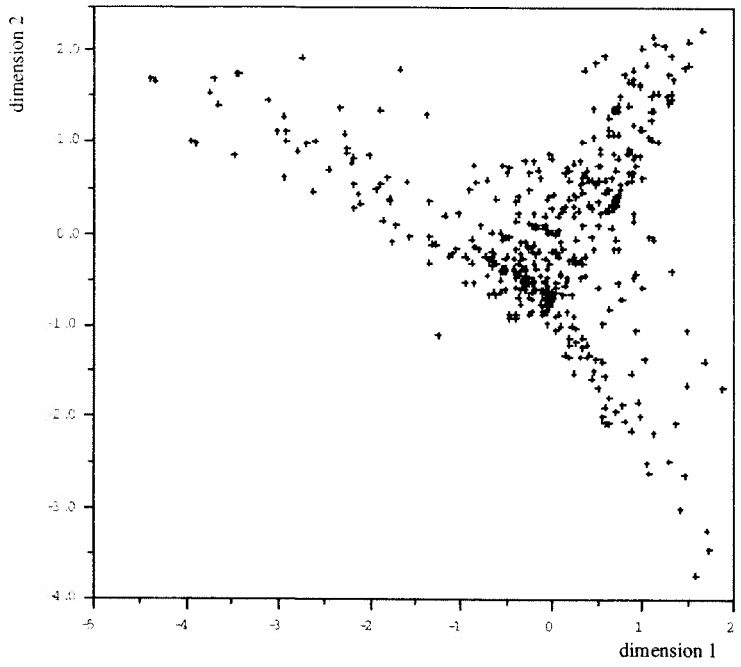

Fig. 2.7. Projection de 688 cursus sur les deux premières dimensions de l'analyse des correspondances multiples utilisant 39 niveaux du cursus scolaire et professionnel

La figure 2.7 montre la projection de 688 carrières et la figure 2.8 la progression des niveaux scolaires sur deux dimensions. Il n'y a pas d'effet Guttman pour les cursus scolaires (figure 2.7). Il n'est donc pas possible de résumer les cursus scolaires par un seul score. Sur la figure 2.8, les niveaux scolaires ne sont pas tous désignés par des labels différents. Nous voulons seulement étudier l'influence d'une autre modalité "manquant" sur les résultats et montrer comment les cursus scolaires peuvent être liés au passé de l'élève. Depuis en haut à gauche jusqu'en bas à droite, les modalités sont ordonnées de l'échec à la réussite, plus ou moins de la même façon que sur la figure 2.5. La spécificité de la figure 2.7 est le nuage de points en haut à droite de l'origine. On voit sur la figure 2.8 qu'il s'agit de cursus scolaires qui utilisent la modalité manquante "autre" avec ibo et bk. Donc la modalité "autre" apparaît plus souvent que la moyenne avec ce type d'école. De plus, la modalité "autre" s'oppose à l'autre modalité notée "manquant". Cette modalité se retrouve plus souvent que la moyenne avec les cursus mbo. Nous ne sommes pas étonnés d'une opposition entre les modalités "autre" et "manquant". En effet, d'après les fréquences marginales, nous pouvons penser que lorsque la

modalité "autre" apparaît dans un cursus scolaire, elle demeure la modalité pour les unités de temps suivantes. Lorsque la modalité "manquant" apparaît dans un cursus scolaire, elle demeure également la modalité pour les unités de temps manquantes suivantes. Les deux modalités "autre" et "manquant" sont les modalités finales: quand vous y êtes vous n'en sortez plus. Mais des cursus qui présentent à la fois des modalités "autre" et "manquant" sont rares et donc ces carrières sont très éloignées l'une de l'autre. Un des objectifs principaux de l'analyse des correspondances multiples des cursus scolaires est de résumer ces cursus par des mesures qui peuvent être utilisées facilement dans des analyses ultérieures. Dans l'analyse des correspondances multiples sans la modalité "manquant", il s'est avéré facile d'associer la quantification obtenue pour la première dimension avec des variables du passé de l'élève en déterminant les corrélations ou les moyennes. Tel n'a pas été le cas pour les cursus présentés sur la figure 2.7: deux scores sont nécessaires pour résumer les cursus. Il est difficile de choisir entre ces deux scores en raison de la configuration du nuage de points.

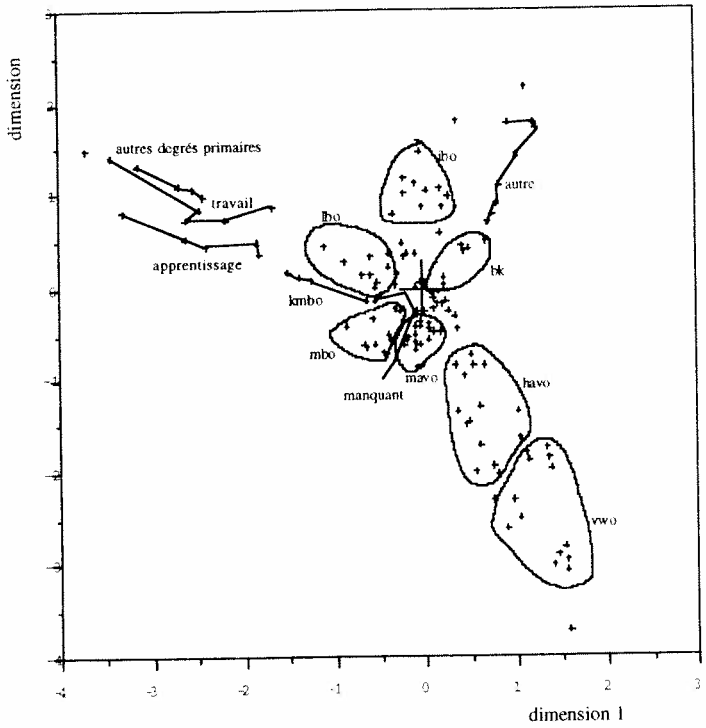

Fig. 2.8. Projection des types de formation sur les deux premières dimensions de l'analyse des correspondances multiples utilisant 39 niveaux du cursus scolaire et professionnel

Il est beaucoup plus simple de déterminer une classification des cursus en un certain nombre de groupes. Dans ce but, on peut notamment utiliser certaines techniques comme des méthodes de classification sur les coordonnées de la figure 2.8.

Dans ce dernier cas, on voit que l'on peut déterminer au moins quatre groupes (les cursus sur la gauche, en bas à droite, en haut à droite et au centre). Cette classification peut alors être reliée à d'autres variables du passé de l'élève en construisant des tables de contingences ou en déterminant des moyennes pour chacun des groupes de la classification.

6. Conclusion

Les données de cursus (par exemple scolaires) peuvent être codées de telle façon que l'on puisse leur appliquer l'analyse des correspondances multiples avec profit. On obtient une quantification ou une classification de chaque cursus. La quantification semble être le résumé le plus utile si l'analyse des correspondances produit seulement une seule dimension interprétable ou si l'interprétation de la première dimension peut être distinguée de celle de la deuxième. Par ailleurs, lorsqu'il est difficile de donner une interprétation différente aux dimensions successives, il est plus intéressant d'obtenir une classification des cursus. Ces classifications permettent alors de résumer les cursus par une variable catégorielle.

Il est licite d'utiliser l'analyse des correspondances même dans le cas où le nombre d'états pour chaque unité de temps est élevé (dans notre exemple 38 et 39). On trouve un exemple avec 25 états dans van der Heijden & de Leeuw (1989). Bien sûr la stabilité des résultats peut décroître avec le nombre de modalités: une petite perturbation dans les données provoque des modifications assez importantes des résultats. Afin d'étudier la stabilité des résultats, on pourra appliquer le "bootstrap" (voir Giffi, 1990, pour des exemples et Markus, 1994, pour son efficacité). On peut augmenter la stabilité en diminuant le nombre de périodes utilisées dans la matrice des données (voir la section 3).

Analyse statistique de réponses ouvertes: application à des enquêtes auprès de lycéens

Mónica Bécue Bertaut [1] et Ludovic Lebart [2]

[1] Faculté d'Informatique, Université de Barcelone, Espagne
[2] Ecole Nationale Supérieure des Télécommunications, Paris, France

1. Introduction

Ce chapitre présente une application de l'analyse des correspondances à des tableaux de données lexicales construits à partir de réponses à des questions ouvertes. Poser une question ouverte ou bien fermée, voilà un choix qui doit être fait lors de la construction du questionnaire d'une enquête. Et, bien sûr, ce choix sera guidé, entre autres raisons, par les différentes manières de traiter les réponses obtenues. Les traitements statistiques proposés ici partent du texte brut, sans précodage ni intervention manuelle, facilitant ainsi une appréhension des réponses qui relègue la subjectivité au stade ultérieur de l'interprétation des résultats obtenus. Cette méthodologie opère au moyen de comptages de mots ou de segments répétés et permet ainsi un traitement systématique du contenu et de la forme des réponses. On verra que la déconstruction du texte ainsi effectuée peut faciliter la mise en évidence de signes sociaux transparents à une lecture plus classique.

2. L'apport spécifique des questions ouvertes dans les enquêtes

De nombreux travaux ont mis en évidence la spécificité des réponses aux questions ouvertes. On peut trouver une présentation de ces travaux dans Lebart & Salem (1994). Il faut rappeler ici que, comme des études comparatives l'ont montré, le questionnement ouvert et le questionnement fermé ne peuvent apporter la même information. En particulier, les items proposés comme réponses possibles à une question fermée induisent les réponses des répondants. Ceux-ci peuvent être amenés à choisir une réponse considérée a priori comme "correcte" car présente, être parfois peu enclins à admettre qu'aucune des réponses prévues ne corresponde à la leur, ou encore se montrer soucieux de répondre à l'attente de l'enquêteur en choisissant une réponse préétablie... Si de plus l'on tient compte du fait que les items proposés peuvent