JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

FERIDOUN S. AGHILI

Construction de quelques indices démographiques robustes basés sur des méthodes d'analyse multivariées et applications

Journal de la société statistique de Paris, tome 133, n° 1-2 (1992), p. 58-71

<http://www.numdam.org/item?id=JSFS_1992__133_1-2_58_0>

CONSTRUCTION DE QUELQUES INDICES DÉMOGRAPHIQUES ROBUSTES BASÉS SUR DES MÉTHODES D'ANALYSE MULTIVARIÉES ET APPLICATIONS

par Feridoun S. AGHILI
Chef des Projets de Recherches en Statistiques
DPSA (État de Vaud) et Université de Genève

Résumé : On introduit à l'aide des méthodes robustes d'estimation de la covariance plusieurs indices démographiques. Ces indices nous permettent de caractériser les profils des populations sociales. On applique ces résultats pour identifier les profils démographiques des enfants atteints des troubles du comportement, de la personnalité et du langage.

Abstract : We introduce several demographics indices based on robust covariance estimation. These indices allow us to characterize the profile of social populations. These results are applied to characterize the demographic profils of children suffering "behaviour", "personality" and "language" troubles.

Mots-clés : Estimation robuste, analyse discriminante, indices démographiques, distance de Mahalanobis, troubles : du comportement, de la personnalité, du langage.

Codes AMS : 62-070.

1. Introduction

Dans ce travail on définit plusieurs indices nous permettant de caractériser les profils démographiques des populations sociales.

Ces indices sont définis à partir des propriétés de l'estimation robuste de la matrice de covariance et de l'analyse discriminante.

Elles nous permettent de mesurer d'une part le degré de cohésion démographique d'une population et d'autre part le niveau de discrimination par rapport à une population de référence.

Le premier indice qu'on appellera «degré de cohésion démographique» nous permet de comparer la cohésion interne d'une population à celle de la population de référence.

Le second permet d'identifier «l'écart démographique» d'une population par rapport à la population de référence.

Enfin la notion de «sphère démographique» permet d'identifier des populations ayant des profils semblables.

Ces méthodes peuvent s'utiliser en sociologie médicale et en psychiatrie, afin d'établir des correspondances entre les types de maladies (ou troubles psychiatriques) et les profils sociaux et démographiques des populations.

Actuellement on assiste à une augmentation de ce type de travaux depuis les premières recherches effectuées sur le sida en 1980.

Nous appliquerons ces résultats pour identifier les profils démographiques des populations des enfants atteints des troubles de «comportement», de «langage» et de la «personnalité» dans la canton de Vaud (Suisse), entre 1982-83 et 1988-89, ceci pour les variables sexe, nationalité et classes d'âges.

Les résultats obtenus dans cette application sont concluants. Ils permettent d'une part d'avoir une projection graphique des profils des enfants sur un plan discriminant et d'autre part de mettre en relief les profils de ces différentes populations.

2. Notations et tableau de données

On considère $n+1$ populations $I_0, I_1, I_2, ..., I_n$. Sur chacune de ces populations, on effectue une succession de mesures sur des périodes temporelles $t_1, t_2, ..., t_r$.

On désigne par $I_{jk}(v_1, v_2, ..., v_q)$ la $k^{\text{-ième}}$ $(1 \leq k \leq r)$ mesure de la population I_j, sur les variables $v = (v_1, v_2, ..., v_q)$ de l'espace R^q.

Par abus de langage, on désigne par I_j l'ensemble des r observations concernant la population I_j :

$$I_j = \{I_{jk}(v)\}_{1 \leq k \leq r}.$$

La population I_0 est notre population de référence, elle nous permet d'avoir un élément de comparaison. Ceci est par exemple utile dans le cas où l'on veut comparer la structure démographique d'une population atteinte d'un trouble (ou d'une maladie) à la structure démographique de la population globale.

Notre modèle est donc constitué de $(n+1) \times r$ observations, et q variables quantitatives (en général des pourcentages), et $n+1$ variables qualitatives (permetant d'identifier chacune des populations).

3. Quelques indices démographiques

La construction d'indices démographiques permet d'identifier et différencier des populations.

Ces indices doivent satisfaire à un certain nombre de propriétés, en particulier :

Exhaustivité (permettre de définir le plus complètement chacune des populations) ;

Robustesse (c'est-à-dire ne pas être influencé par des données aberrantes) ;

Equivariance linéaire (c'est-à-dire de ne pas être modifié par des translations linéaires).

La robustesse et l'équivariance linéaire de ces indices dépendent intimement des distances (ou mesures de dissimilarités), choisies.

Les méthodes d'analyse multivariées, nous fournissent une variété de distances et de dissimilarités, dont certaines ne vérifient pas les propriétés requises ci-dessus.

Par exemple nous trouvons un grand choix de distances associé aux méthodes de la classification hiérarchique. Le choix de ces distances dépend de la nature des données [c.f. (4)].

En analyse des correspondances la distance de X^2, sur les profils lignes (ou colonnes), nous permet également de définir des indices appropriés (quand les données sont constituées de tableaux de contingence, c.f. (1)).

Enfin la distance de Mahalanobis, utilisée en analyse discriminante, permet d'obtenir une cohésion maximum des nuages de données, et son utilisation permet en outre d'obtenir des projections sur des plans discriminants, qui ont la propriété, d'une part de séparer les différentes populations, d'autre part de compactifier chacun des nuages associés aux populations.

Il faut cependant noter, que malgré les bonnes propriétés de cette distance, elle se base sur la moyenne arithmétique et la matrice de covariance, qui sont en général sensibles aux données aberrantes.

C'est pour cette raison que dans la suite de ce travail nous utiliserons la version robuste de la distance de Mahalanobis, basée sur la matrice de covariance robuste.

3.1. Estimateurs de covariances robustes

Afin de construire des indices démographiques vérifiant les propriétés mentionnées au paragraphe 3, nous utiliserons les outils de la statistique robuste.

On rappelle brièvement les diverses méthodes, permettant la construction d'estimateurs de la matrice de covariance robuste.

Il existe diverses approches c.f. (10), (12), pour robustifier la matrice de covariance, et les plus importantes sont :

1) Robustification de la matrice élément par élément.

2) Généralisation des M-estimateurs aux matrices de covariances.

3) L'approche de Rousseeuw (Utilisation de l'ellipsoïde de volume minimal).

Bien que la première approche permette dans certains cas de donner des résultats intéressants, nous nous intéresserons plus particulièrement aux approches basées sur les M-estimateurs (introduite par Maronna 1976 (10)), et celle plus récente introduite par Rousseeuw 1983, 1986 (12).

Ces méthodes ont en effet l'avantage d'être plus performantes dans la détection des points aberrants.

La méthode basée sur les M-estimateurs (introduite par de Maronna (10)), consiste à robustifier les distances de Mahalanobis en remplaçant la moyenne $\overline{X}$ et la covariance V par des estimateurs robustes $T(X)$ et $C(X)$, obtenus en résolvant un système itératif :

$$1/n \sum w_1(d_i) \cdot (X_i - T(X)) = 0$$

(3.1.1)

$$1/n \sum w_2(d_i^2) \cdot (X_i - T(X))' (X_i - T(X)) = C(X)$$

où d_i est la distance de Mahalanobis robustifiée :

(3.1.2) $$d_i = [(X_i - T(X))\ C(X)^{-1} (X_i - T(X))']^{1/2}.$$

Et $T(X)$ est un M-estimateur de position dans R^p :

(3.1.3) $$T(X) = \sum w_1(d_i) \cdot X_i / \sum w_1(d_i).$$

Remarque 3.1.1. :

Campell a proposé une variante de $C(X)$, pour obtenir un estimateur sans biais de la covariance (cette matrice a notamment servi à Campell à introduire une nouvelle version de l'analyse discriminante robuste (3)).

Le système itératif (3.1.1) est résolu, une fois que les fonctions poids $w_1(d)$ et $w_2(d^2)$ sont choisies (les poids les plus utilisés sont : les poids de Huber, de Cauchy et enfin les poids de la distance moyenne c.f. (12)).

L'approche de Rousseeuw (12), consiste à trouver l'estimateur de l'ellipsoïde de volume minimale «MVE», défini par (3.1.4) :

$T(X)$ = centre de l'ellipsoïde de volume minimal contenant au moins h observations de X, où $h = [n/2] + 1$;

$C(X)$ = matrice de covariance des h observations de l'ellipsoïde de volume minimal.

On multiplie $C(X)$ par un facteur de correction qui assure la convergence de l'estimateur dans l'hypothèse de normalité des variables.

Les étapes de calcul de l'estimateur MVE, sont les suivantes :

Premièrement, on tire un sous-échantillon de $p+1$ observations parmi l'ensemble X des r observations. Soit L le sous-ensemble :

$$L = \{i_1, i_2, ..., i_{p+1}\}.$$

Pour le sous-échantillon associé à L, le vecteur ligne des moyennes arithmétiques, et la matrice de covariances débiaisée sont donnés par :

$$(3.1.5) \qquad \overline{X}_L = 1/p + 1 \sum X_i \quad \text{et} \quad C_L = 1/p \sum (X_i - \overline{X}_L)'(X_i - \overline{X}_L)$$

où C_L doit être non singulière, dans le cas échéant on retire $p + 1$ points au hasard. L'ellipsoïde associé à L est alors gonflé (ou déflaté) pour contenir exactement h points, ceci en fait revient à multiplier C_L par m_L^2 où

$$(3.1.6) \qquad m_L^2 = \underset{i}{\text{médiane}} (X_i - \overline{X}_L) C_L^{-1} (X_i - \overline{X})'$$

m_L^2 étant le facteur d'amplification exacte de l'ellipsoïde. Le volume de l'ellipsoïde modifié, c'est-à-dire correspondant à $m_L^2 C_L$ est proportionnel à :

$$(3.1.7) \qquad [\det(m_L^2 C_L)]^{1/2} = [\det(C_L)]^{1/2} m_L^p.$$

Cette procédure est effectuée pour tous les sous-échantillons de taille $p + 1$. Le sous-échantillon qui minimise la fonction d'objectif (3.1.7) est alors retenu. Les estimateurs $T(X)$ et $C(X)$ sont donc :

$$(3.1.8) \qquad T(X) = \overline{X}_L \qquad \text{et} \qquad C(X) = k^{-1} m_L^2 C_L$$

où k est la valeur médiane de la distribution du *Khi* carré à p degré de libertés. Ce facteur intervient pour assurer la convergence de $C(X)$ lorsque les variables sont tirées d'une population multinormale.

Remarque 3.1.1 :

L'estimateur obtenu par la méthode MVE est linéairement équivariant, i.e. elle satisfait aux deux propriétés (3.1.9) :

$$(i) \qquad T_n(X_1 + b, ..., X_n + b) = T_n(X_1, ..., X_n) + b$$

$$(ii) \qquad T_n(cX_1, ..., X_n) = cT_n(X_1, ..., X_n).$$

Remarque 3.1.2 :

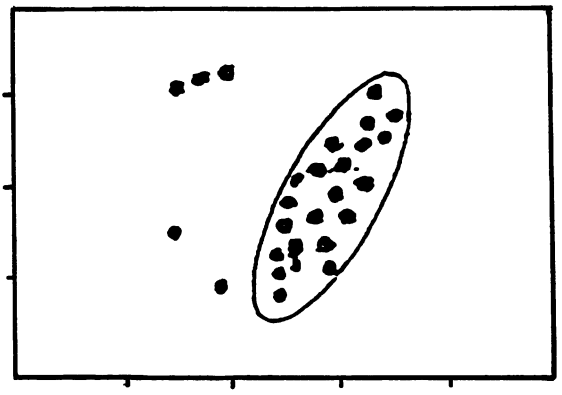
Cet estimateur a un point de rupture égale à $([n/2] - p + 1)/n$ qui est asymptotiquement maximal.

Remarque 3.1.3 :

L'inconvénient majeur de l'estimateur MVE, est qu'il converge comme $n^{-1/3}$, alors que la plupart des estimateurs convergent comme $n^{-1/2}$.

Exemple 3.1.1 :

Ellipsoïde de volume minimal en dimension deux.



3.2. Choix d'une mesure robuste de cohésion démographique

On définira des mesures de cohésion d'une population I_j, en fonction de la méthode de robustification de la matrice de covariance.

Dans le cas où on utilise les M-estimateurs de la matrice de covariance, la mesure de cohésion de la population I_j est définie par :

$$m(I_j) = \sum (X_i - \overline{X})\, C^{-1} [(X_i - \overline{X})']^{1/2} \tag{3.2.1}$$

où $\overline{X}$ et C sont des estimateurs robustes de $T(X)$ et $C(X)$ (c.f. (3.1.1)), de la population I_j.

Cette définition est encore plus simple dans le cas, où l'on utilise la méthode de robustification par des ellipsoïdes de volume minimal «MVE».

Soit T_j le centre des ellipsoïdes de volume minimal V_j associé à la population I_j, $(1 \leqslant j \leqslant n)$, on a alors :

$$m(I_j) = V_j. \tag{3.2.2}$$

Remarque 3.2.1 :

On rappelle que le volume d'un ellipsoïde de dimension p :

$$(X - T_j)\, C^{-1} (X - T_j)' < p$$

est proportionnel à $(\det\ C)^{1/2}$, et ceci facilite le calcul de V_j.

Remarque 3.2.2 :

On peut définir d'autres mesures de cohésion à l'aide d'une distance quelconque d. Par exemple en calculant la somme des $d^2(x, y)$, où x et y appartiennent à I_j,

ou encore par la somme des $d^2(x, m)$ (m étant le centre de gravité de I_j et x un élément de I_j). Cependant ces mesures sont sensibles aux données atypiques (ou aberrantes).

3.3. Indice de cohésion démographique

On définit l'indice de cohésion démographique de la population I_j (relativement à I_0), qu'on désignera par $CD\ (I_j)$, comme le rapport entre les mesures de cohésion de I_0 et I_j :

$$CD\ (I_j) = m\ (I_0)/m\ (I_j). \tag{3.3.1}$$

La mesure $m\ (I_j)$ peut se définir à partir d'une distance d quelconque, cependant le choix d'une mesure de cohésion robuste permet d'éviter l'effet des données atypiques.

L'utilisation de la distance de Mahalanobis robustifiée, permet d'utiliser les méthodes d'analyse discriminante développées par Campell (4).

Propostion 3.3.1 : Soient I_0 la population de référence et I_j, une des populations de notre étude, si $m\ (I_0) \leqslant m\ (I_j)(1 \leqslant j \leqslant n)$, alors :

$$0 \leqslant CD\ (I_j) \leqslant 1. \tag{3.3.2}$$

Démonstration : On a toujours $CD\ (I_j) \geqslant 0$, car la distance d (ou le volume V_j) est positive. Comme $m\ (I_0) \leqslant m\ (I_j)$, alors $CD\ (I_j) \leqslant 1$.

L'hypotèse $m\ (I_0) \leqslant m\ (I_j)$ est souvent vérifiée dans la pratique, car la population de référence a un profil plus compact que celle des autres populations.

Cet indice est une des caractéristiques des populations, il permet de mesurer la compacité de I_j, par rapport à une population de référence.

Par exemple si $CD\ (I_j)$ est proche de 1, alors I_j a une cohésion proche de la population de référence.

Par contre si $CD\ (I_j)$ est proche de 0, alors I_j est très dispersé par rapport à I_0.

Comme nous le verrons dans l'application du paragraphe 4, le manque de cohésion démographique d'une population indique une variation des profils à travers les temps.

Ce manque de stabilité temporelle ne permet donc pas de tirer des conclusions sociologiques sur les correspondances entre les profils d'une population et les types de maladie ou troubles de cette population.

Cet indice nous permet en outre de comparer la cohésion de plusieurs populations entre elle.

Remarque 3.3.1 :

Le cas $m\ (I_0) = 0$, peut correspondre à la population I_0 réduit à un point, et dans ce cas $CD\ (I_j) = 0$, quelque soit la valeur de $m\ (I_j)$. Par contre dans les applications concrètes, on a rarement $CD\ (I_j) = 1$, car cela signifie que $m\ (I_0) = m\ (I_j)$.

3.4. Ecart démographique

Ce deuxième indice permet de mesurer l'écart d'une population par rapport à la population de référence.

On désigne par D une distance entre les populations (par exemple une distance euclidienne). Soit g_i le centre de gravité robuste de la population $I_j (0 \leqslant i \leqslant n)$, obtenu par la méthode de l'estimation robuste du paragraphe 3.1.

On définit l'écart démographique «*ED*» entre les populations I_0 et I_k par :

$$ED\,(I_0, I_k) = D\,(g_0, g_k). \tag{3.4.1}$$

Le choix de la distance D, dépend de la méthode robuste de discrimination choisie. En effet, il est préférable d'obtenir directement ces distances dans les «outputs» de l'analyse discriminante.

Remarque 3.4.1 :

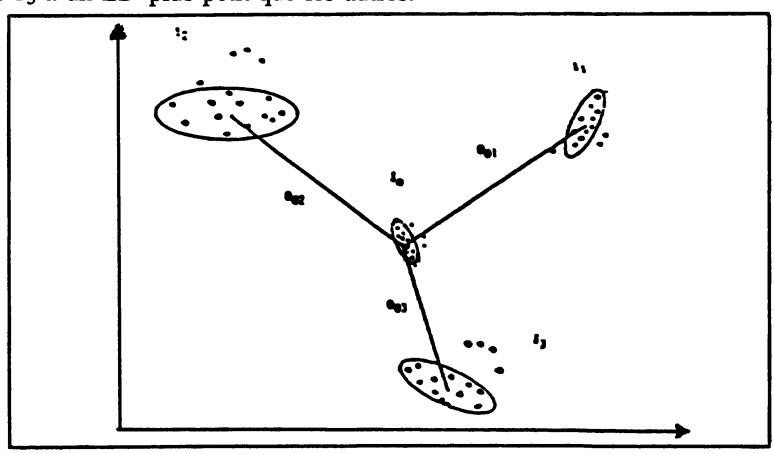
Les logiciels classiques d'analyse de données (par exemple SAS), fournissent les distances de Mahalanobis entre les différents groupes, elles permettent donc de calculer les écarts démographiques.

Une des utilisations possible de cet indice, est d'identifier les populations ayant des profils très différents de la population de référence. En particulier une population ayant un *ED* assez grand (une population marginale), peut représenter une population à «risque» (par exemple en médecine ou en psychiatrie).

Exemple 3.4.1. :

On considère 3 populations I_1, I_2, I_3 en dimension 2, ou I_0 est la population de référence. L'écart démographique entre les populations I_0 et I_k est désigné par D_{0k}.

On remarque que la population I_1 a un *CD* plus grand que les autres populations et que I_3 a un *ED* plus petit que les autres.



Graphique 3.4.1 : Exemple de 3 populations en dimension 2

3.5. Notion de «sphère démographique» et application à l'identification des populations

Une des utilisations possible des indices définis plus haut est d'établir des correspondances entre les populations et les troubles (ou maladies). Les indices définis plus haut ne permettent pas de caractériser complètement les populations, en effet la proximité entre $ED(I_0, I_k)$ et $ED(I_0, I_j)$ et celle de $CD(I_k)$ et $CD(I_j)$, n'implique pas toujours que I_k et I_j, aient des structures démographiques semblables (i.e. que I_k soit proche de I_j).

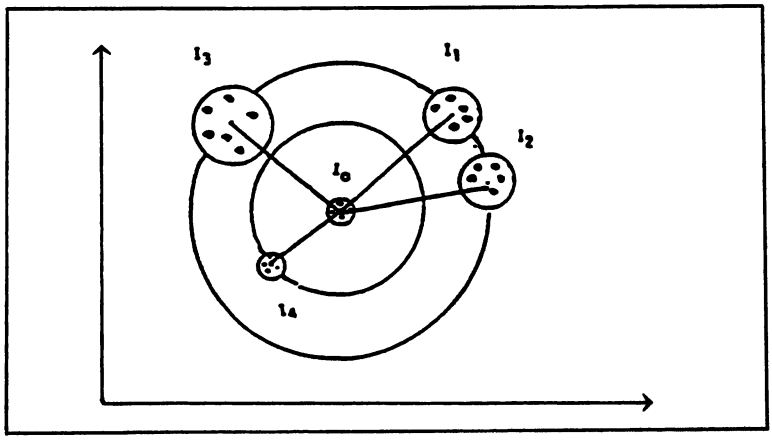
Afin d'identifier une population de manière plus précise, on définit la notion de «sphère démographique» :

Définition 3.5.1. : Soit R^q l'espace des variables de notre modèle et D_{0k} l'écart démographique entre I_0 et I_k, la sphère démographique de la population I_k est définie comme une sphère de centre g_0 (centre de gravité de I_0) et de rayon D_{0k} dans R^q.

Une des propriétés de cette sphère est qu'elle permet d'identifier des populations ayant des profils semblables. En effet si deux populations sont voisines sur la même sphère (avec des «*CD*» proches), alors elles ont des structures démographiques semblables.

En outre on peut associer à chaque point de cette sphère des caractéristiques de chacune des variables (par l'intersection des axes de l'espace R^q avec la sphère). Ainsi deux points situés dans deux parties diférentes de la même sphère représentent des particularités spécifiques par rapport aux variables du modèle. Ce type de représentation est utile pour obtenir une «typologie» des profils démographiques des populations, ainsi que leurs correspondances avec des troubles (ou maladie).

Exemple 3.5.1. Dans cet exemple on a 4 populations situées sur deux sphères. Les populations I_1 et I_2 ont des profils démographiques très proches, en effet elles sont voisines et sur la même sphère avec des «*CD*» très proches. Par contre le profil de I_3 est distinct de I_1 et I_2.



Graphique 3.5.1 : Cercle démographique de 4 populations en dimension 2

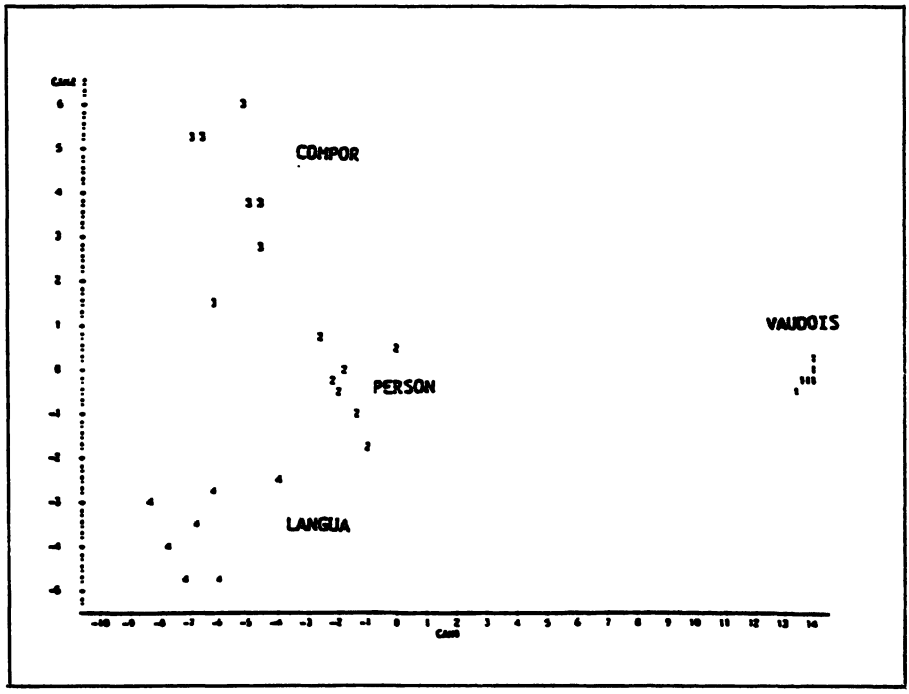
4. Application : Comparaison des profils démographiques des enfants atteints des troubles du comportement, de la personnalité et du langage

Dans cette partie, nous utiliserons les résultats du paragraphe 3 pour différencier les profils démographiques des populations des enfants atteints des troubles de la personnalité, du comportement et du langage (identifiés sur les graphiques par respectivement 2, 3, 4). La population des enfants de 0 à 19 ans du canton de Vaud sera notre population de référence (identifiée sur les graphiques par 1).

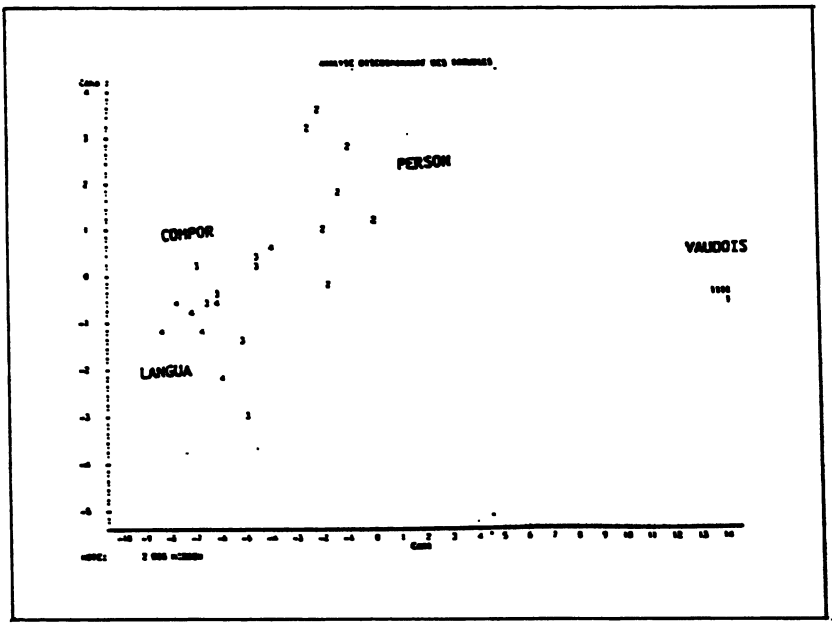
Les variables utilisées sont le pourcentage (par rapport à chacune des populations) : de garçons, des classes d'âges (0 à 4 ans, 5 à 14 ans, 15 à 19 ans) et de la nationalité (suisse, étrangère).

Pour chacune des populations, nous avons 7 observations, obtenues sur les périodes de 1982 à 1988.

Le but de cette application est d'une part d'identifier si chacune des populations a des profils spécifiques (par exemple si il y a une concentration d'étrangers d'une classe d'âge particulière, etc.), d'autre part de mesurer le degré de cohésion associé à chacune de ces populations à travers le temps (i.e. il y a des grandes variations des profils d'une période à l'autre), ceci en comparaison avec la population de référence. On obtient la projection des nuages associés aux quatre populations sur les plans discriminants :



Graphique 4.1 : Projection des populations sur les axes discriminant 1 et 2



Graphique 4.2 : Projection des populations sur les axes discriminant 1 et 3

Les indices de cohésion et d'écart démographiques («*CD*», et «*ED*») de ces trois populations calculés à partir de la distance de Mahalanobis sont :

Troubles	*CD*	*ED*
comportement	**.16**	**19.85**
de la personnalité	**.29**	**15.58**
du langage	**.25**	**20.76**

Ces résultats nous permettent de constater que la structure démographique des enfants étiquetés sous **«trouble du comportement»** a un *CD* très faible donc une très grande dispersion. Cela signifie en particulier que le profil de ces enfants varie à travers le temps, il n'est donc pas très stable. Par contre le profil des enfants étiquetés sous **«trouble de la personnalité»** est plus stable que les deux autres troubles.

En ce qui concerne l'écart et la sphère démographique «*ED*», on remarque que les troubles du comportement et du langage se trouvent pratiquement sur la même sphère (avec des *ED* de 19.85 et 20.76 respectivement).

Par contre ces deux populations sont distinctes (leur distance de Mahalanobis est de 7.72). Il en est de même entre les troubles du comportement et de la personnalité et celle de la personnalité et du langage (leurs distances de Mahalanobis est respectivement de 6.42 et 6.64).

Les résultats de l'analyse discriminante nous confirment les conclusions précédentes, elles nous permettent en outre, de mettre en évidence la structure démographique de chacune de ces populations.

précédentes, elles nous permettent en outre, de mettre en évidence la structure démographique de chacune de ces populations.

Nous résumons les constatations obtenues par cette méthode :

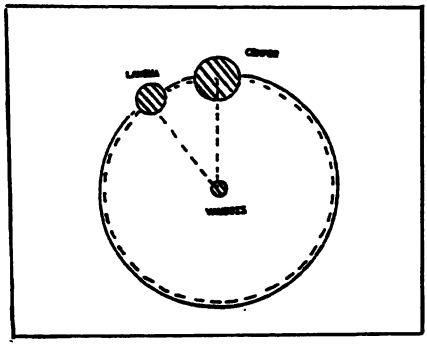
La discrimination la plus importante entre la population globale des enfants du canton de Vaud et les trois troubles sont dans l'ordre d'importance :

1. La proportion de garçons suisses de 0 à 4 ans est plus faible dans la population des enfants atteints des trois troubles que dans la population de référence.
2. La proportion de garçons étrangers de 5 à 14 ans est plus faible dans la population Vaudoise que dans les populations des enfants atteints des trois troubles (ceci indique une surreprésentation de cette catégorie d'enfants dans les trois troubles, en particulier dans la catégorie des troubles du comportement).
3. La proportion de garçons suisses de 15 à 19 ans est plus forte dans la population des enfants atteints des trois troubles que celle des «vaudois» (en particulier dans la catégorie des «troubles du comportement»).
4. La proportion de garçons suisses de 5 à 14 ans est plus forte dans la catégorie des «troubles de la personnalité» que dans les autres troubles.
5. En ce qui concerne les «troubles du langage», on remarque que la discrimination avec les autres troubles se fait par des proportions plus fortes des enfants suisses et étrangers de «5 à 14 ans» et par une proportion plus faible de garçons suisses et étrangers de «0 à 4 ans» que dans les autres troubles.

Ces quelques constatations nous permettent de mieux situer la correspondance entre les profils démographiques des populations et les différents troubles.

On peut également situer ces différentes structures démographiques sur les «sphères démographiques» associées aux troubles.

Par exemple on représente les profils des populations des enfants atteints des troubles du langage et du trouble du comportement sur le cercle :



Graphique 4.3 : Représentation des profils des enfants atteints par les troubles du comportement et du langage sur le cercle démographique

Cette représentation est schématique, les rayons des cercles sont proportionnels aux *CD* des populations (la population de référence a un rayon égale à l'unité) et les rayons des sphères démographiques sont proportionnels aux *ED*.

5. Conclusion

L'étude des profils socio-démographiques des populations atteintes de troubles psychiatriques (ou maladies), permet à de nombreux chercheurs en sociologie médicale, d'obtenir des correspondances entre les profils des populations et types de maladies (le SIDA a été le déclencheur de ce genre de recherches).

Cette correspondance permet de mieux situer la typologie des maladies (ou troubles) par rapport à la problématique socio-démographique des populations.

La construction d'indices multidimensionnels, permet de caractériser les similarités et la stabilité temporelle des profils démographiques des populations.

La complexité du Champ social, ne permet pas d'obtenir une information exhaustive à l'aide d'un seul indice démographique. C'est pourquoi nous avons défini plusieurs indices pour caractériser chacune des populations.

Ces indices ont de bonnes propriétés (robustesses et équivariances linéaires). Il est cependant clair que les indices définis dans ce travail dépendent de la sélection des variables (donc de la modélisation).

L'exemple traité dans la partie 4 de ce travail, permet de définir certaines caractéristiques des enfants atteints des troubles du comportement, de la personnalité et du langage.

Il faut cependant compléter cette étude par un plus grand nombre d'informations sur ces populations.

6. Bibliographie

(1) BENZÉCRI J.P., *L'analyse des données*, Dunod 1982.

(2) CAMPELL N.A., 1980, Robust Procedure in Multivariate Analysis I: Robust Covariance Estimation, *Applied Statistics*, 29, 231-237.

(3) CAMPELL N.A., 1982, Robust Procedure in Multivariate Analysis II : Robust Canonical Variate Analysis, *Applied Statistics*, No 1, pp1-8.

(4) DIDAY E. et Coll., 1980, *Optimisation en classification automatique*, INRIA, Rocquencourt, France.

(5) DONOHO D.L., 1982, "*Breakdown Properties of Multivariate Location Estimators*", qualifying paper, Harvard University.

(6) HAMPEL F.R., RONCHETTI E., ROUSSEEUW P.J., STAHEL W., 1986, *Robust Statistics: the Approach based on Influence Function*, New York : John Wiley & Sons.

(7) HUBER P.J., 1981, *Robust Statistics*, New York : John Wiley & Sons.

(8) KSHIRSAGAR A.M., 1972, *Multivariate Analysis*, New York : Marcel Dekker.

(9) LECOUTRE J.P. et TASSI Ph., 1987, *Statistique non paramétrique et robustesse*, Economica, Paris.

(10) MARONNA R.A., 1976, "Robust M-estimators of Multivariate Location and Scatter", *Ann. Statist.*, 4, 51-67.

(11) ROUSSEEUW P.J., 1984, "Least Median of Squares Regression", *J. Am. Statist. Assoc.* 79, 871-880.

(12) ROUSSEEUW P.J. and VAN ZOMEREN B.C., 1988, "*Unmasking Multivariate Outliers and Leverage Points by Means of Robust Covariance Matrices*", Exposé à l'Université de Genève.

(13) RAO C.R., 1973, *Linear Statistical Inference*, New York : John Wiley & Sons.

(14) S. AGHILI, F., 1987, Instabilité hiérarchique d'un ensemble de données économiques et applications, *Jour. Soc. Statist. de Paris*, No 1, 1987.

(15) SAPORTA G., 1980, *Théorie et méthodes de la statistique*, Technip, Paris.

(16) TUKEY J.W., 1977, *Exploratory Data Analysis*, Addison-Wesley.