

LUDOVIC LEBART

NICOLE TABARD

**Applications des méthodes d'analyse des données à la préparation
des enquêtes auprès des ménages**

Journal de la société statistique de Paris, tome 112, n° 2 (1971), p. 108-121

http://www.numdam.org/item?id=JSFS_1971__112_2_108_0

© Société de statistique de Paris, 1971, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

APPLICATIONS DES MÉTHODES D'ANALYSE DES DONNÉES A LA PRÉPARATION DES ENQUÊTES AUPRÈS DES MÉNAGES

(Morphologie sociale des communes de la région parisienne)

(Communication faite le 16 décembre 1970 devant la Société de statistique de Paris)

RÉSUMÉ

Les méthodes d'analyse des données (réductions automatiques de vastes tableaux de valeurs numériques) permettent d'améliorer la phase de préparation de certaines enquêtes.

L'exploitation de certains des résultats du recensement général de la population de 1968 par l'analyse des correspondances permet de définir la typologie des communes de plus de 10 000 habitants (373 dans la région parisienne et 390 dans le reste de la France) à partir de leurs structures socio-professionnelles caractérisées par 29 variables d'activité pour chaque sexe. Le choix des communes à enquêter se fait alors à partir de la typologie observée et des hypothèses particulières à l'enquête.

Les résultats obtenus sont en réalité plus qu'une simple « aide au sondage », et présentent par eux-mêmes un intérêt sociologique. Pour des raisons de volume, seuls les résultats relatifs à la région parisienne seront évoqués ci-dessous.

PLAN

— Analyse des tableaux (373 × 29) issus du recensement.

a) Objet de l'analyse — Caractéristiques des données.

b) Justification de la méthode : l'équivalence distributionnelle.

- c) Projection des variables « illustratives ».
- d) Résultats et interprétation.
- e) Choix des communes par partition.

a) *Objet de l'analyse — Caractéristiques des données*

Une enquête doit se dérouler en 1974, à la demande de la Caisse nationale d'allocations familiales, dans le cadre d'une recherche sur la politique familiale à laquelle collabore le CREDOC.

Les deux pôles actuels de cette politique : prestation en espèces (allocations) et action sociale (équipement et fonctionnement de services sociaux, crèches, etc.) doivent être étudiés, plus particulièrement en ce qui concerne leur adéquation à l'évolution des structures familiales et des modes de vie.

Le tirage des villes à enquêter (villes de plus de 10 000 habitants) ne pouvait être un tirage aléatoire car les variables d'offre, d'environnement (variables mesurables seulement au niveau d'une agglomération ou d'une zone géographique) font partie des facteurs que l'on désire contrôler et étudier.

Nous avons alors fait l'hypothèse de travail suivante :

« la partition de la population d'une ville en 58 (29 pour chaque sexe) catégories d'activité contient une information suffisante pour caractériser socio-économiquement cette ville, et par conséquent nous fournir des critères adéquats de choix des villes à enquêter ».

Cette hypothèse peut également se formuler de la façon suivante : si deux villes de taille analogues ont des profils socio-professionnels voisins (c'est-à-dire ont sensiblement les mêmes pourcentages de chacune des catégories d'activité) il suffit de tirer dans notre échantillon une seule des deux villes. Le problème du choix des familles dans ces villes ne sera pas étudié ici.

Insistons sur le fait que nous ne cherchons pas une représentativité de notre échantillon vis-à-vis de variables strictement démographiques. Une démarche analogue à celle que nous adoptons ici avait déjà été entreprise lors de l'enquête UNCAF de 1963 ⁽¹⁾, destinée à analyser les déformations du budget en fonction du nombre d'enfants, et pour laquelle un échantillon représentatif de la population française aurait été inutilisable (puisque comportant par exemple, un nombre infime de familles de six enfants).

La méthode la plus adaptée à la description des tableaux qui nous intéressent est l'analyse des correspondances, due au professeur J.-P. BENZECRI.

Nous justifierons au paragraphe b) le choix de cette méthode, et la confiance que l'on peut accorder aux résultats de ce type d'analyse.

Une fois obtenue une représentation visuelle de la typologie des villes, de leurs proximités vis-à-vis de leurs profils socio-professionnels, le choix des villes peut être raisonné (cas d'une bonne représentation plane) ou automatique (programme de partition, avec mise en évidence des centres de classe). Nous écartons toute classification aveugle, comme toute stratification qui privilégierait *a priori* certaines des variables de départ.

Précisons maintenant les caractéristiques des données statistiques utilisées.

Ces données sont extraites de l'exploitation au quart lourd du recensement de la population I.N.S.E.E. 1968, échelon E3.

1. (réf. 6).

Les unités géographiques : l'échelon E3 comporte :

1. Pour les sept départements de la région parisienne (Seine-et-Marne, Yvelines, Essonne, Hauts-de-Seine, Seine-Saint-Denis, Val-de-Marne, Val-d'Oise) : toutes les communes
2. Pour Paris : chaque arrondissement divisé en quatre quartiers.
3. Pour la province : toutes les communes de 10 000 habitants ou plus.

Les analyses effectivement réalisées pour préparer l'enquête concernent séparément :

- a) la région parisienne et Paris, soit 373 communes et quartiers;
- b) la province, soit 390 unités urbaines de 10 000 habitants ou plus et non les communes, partition beaucoup plus fine qui sera analysée ultérieurement.

Seuls, les résultats concernant la région parisienne sont exposés ci-dessous.

Les variables sélectionnées

29 catégories d'activité que nous appelons catégories socio-professionnelles bien qu'il s'agisse d'une répartition de toute la population y compris les inactifs : enfants, femmes au foyer, retraités et autres inactifs qui représentent la moitié de la population étudiée.

Le détail de ces catégories est donné en annexe avec les effectifs correspondants d'hommes et de femmes. Y figurent également les catégories exclues de l'analyse (agriculteurs, armée, police ...) soit environ 3,5 % de la population totale.

En définitive nous analyserons deux tableaux, de dimensions respectives 373×29 et 390×29 , de répartition du nombre de personnes k_{ij} selon leur lieu de résidence i et leur catégorie socio-professionnelle j .

b) *Justification de la méthode : Équivalence distributionnelle*

Nous ne présenterons d'exposé de cette méthode (cf. réf. 1, 2, 3, 4), mais nous insisterons sur les particularités de son adaptation à cette application.

Nous avons en effet, à faire ici à une « correspondance ternaire » les trois modalités étant :

1. Lieu de résidence.
2. Catégorie socio-professionnelle.
3. Sexe.

Rappelons les principes de base de l'analyse des correspondances.

Si k_{ij} désigne le nombre d'individus appartenant à la ville i et à la catégorie socio-professionnelle j , p_{ij} désigne la fréquence relative correspondante :

$$p_{ij} = k_{ij} / \left(\sum_j k_{ij} \right)$$

On définit :

$$p_{i.} = \sum_j p_{ij} \quad \text{et de même} \quad p_{.j} = \sum_i p_{ij}$$

Les villes seront des points de \mathbf{R}^{29} et les communes des points de \mathbf{R}^{373} .

a) Dans l'espace \mathbb{R}^{29} nous construisons un nuage de 373 points, chaque point-ville i ayant pour coordonnées les quantités :

$$\left(\frac{p_{ij}}{p_{i.}}\right) \quad j = 1, \dots, 29$$

et étant affecté de la masse $p_{i.}$

b) Dans l'espace \mathbb{R}^{373} , nous construisons de même un nuage de 29 points, chaque point-C. S. P. j ayant pour coordonnées les n quantités

$$\left(\frac{p_{ij}}{p_{.j}}\right) \quad i = 1, \dots, 373$$

et étant affecté de la masse $p_{.j}$.

L'analyse consiste à projeter les nuages dans des espaces de faible dimension de façon à assurer une déformation minimum des distances.

Le carré de la distance de deux villes est donné par l'expression pondérée :

$$d^2(i, i') = \sum_{j=1}^{29} \frac{1}{p_{.j}} \left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}}\right)^2 \quad (1)$$

Il en est de même dans \mathbb{R}^{373} , pour la distance de deux catégories socio-professionnelles j et j' : nous prenons l'expression analogue :

$$d^2(j, j') = \sum_{i=1}^{373} \frac{1}{p_{i.}} \left(\frac{p_{ij}}{p_{.j}} - \frac{p_{i'j'}}{p_{.j'}}\right)^2 \quad (2)$$

Cette distance pondérée a l'avantage de vérifier le principe « d'équivalence distributionnelle », c'est-à-dire que si deux points i_1 et i_2 sont confondus dans \mathbb{R}^{29} et si on les considère comme un seul point affecté de la somme des masses de i_1 et de i_2 alors les distances entre couples de points dans \mathbb{R}^{29} et dans \mathbb{R}^{373} sont inchangées.

Cette propriété fondamentale explique la stabilité des représentations obtenues :

Le découpage en communes ou en quartiers dans la région parisienne est souvent assez conventionnel, et ne correspond pas toujours à une frontière ayant une signification socio-économique. Il existe d'autre part — en moyenne — une très forte dépendance entre les profils socio-professionnels de deux communes contiguës.

Le choix de ces distances nous permet d'avoir des résultats stables si l'on agrège ou si l'on redécoupe différemment le tissu urbain.

Si les 373 zones géographiques retenues sont homogènes socio-professionnellement, l'équivalence distributionnelle nous permet d'affirmer que les résultats ne seraient pas meilleurs si le découpage était plus fin.

Tout ce qui peut être dit à propos des communes est également valable pour les catégories socio-professionnelles, que l'on agrégerait ou que l'on découperait plus finement.

Par contre, si l'on agrège deux catégories socio-professionnelles ayant des poids importants et des profils géographiques très différents, la typologie des communes sera probablement modifiée.

Les 29 postes retenus pour chaque sexe constituent un découpage suffisamment fin pour être assez homogène, et donc pour remplacer une nomenclature beaucoup plus fine, qui tiendrait compte de toutes les professions, par exemple.

Montrons que la distance choisie vérifie bien les propriétés ci-dessus :

a) Invariance des distances dans \mathbb{R}^{29} .

Si les communes i_1 et i_2 sont confondues dans \mathbb{R}^{29} , on a, pour tout j :

$$\frac{p_{i_1 j}}{p_{i_1}} = \frac{p_{i_2 j}}{p_{i_2}} = \frac{p_{i_0 j}}{p_{i_0}} \quad (3)$$

On a donc en particulier :

$$\frac{p_{i_1 j} + p_{i_2 j}}{p_{i_1} + p_{i_2}} = \frac{p_{i_0 j}}{p_{i_0}}$$

D'où, puisque les dénominateurs sont égaux, pour tout j :

$$p_{i_1 j} + p_{i_2 j} = p_{i_0 j}$$

Les calculs des quantités $p_{.j} = \sum_i p_{ij}$ ne sont donc pas affectés, et donc les distances $d^2(i, i')$ données par la formule (1) sont invariantes.

b) Invariance des distances dans \mathbb{R}^{373} .

La distance $d^2(j, j')$ donnée par la formule (2) contient deux termes $A(i_1)$ et $A(i_2)$ correspondant aux indices i_1 et i_2 :

$$A(i_1) + A(i_2) = \frac{1}{p_{i_1}} \left(\frac{p_{i_1 j}}{p_{.j}} - \frac{p_{i_1 j'}}{p_{.j'}} \right)^2 + \frac{1}{p_{i_2}} \left(\frac{p_{i_2 j}}{p_{.j}} - \frac{p_{i_2 j'}}{p_{.j'}} \right)^2$$

Ces deux termes sont remplacés par un seul terme $A(i_0)$ tel que :

$$A(i_0) = \frac{1}{p_{i_0}} \left(\frac{p_{i_0 j}}{p_{.j}} - \frac{p_{i_0 j'}}{p_{.j'}} \right)^2$$

Remarquons par exemple que :

$$A(i_1) = p_{i_1} \left(\frac{p_{i_1 j}}{p_{i_1} p_{.j}} - \frac{p_{i_1 j'}}{p_{i_1} p_{.j'}} \right)^2$$

$A(i_2)$ et $A(i_0)$ s'écrivent de la même façon, et les quantités entre accolades sont égales, d'après les relations (3), à un même nombre B .

On a donc $A(i_1) + A(i_2) = p_{i_1} B + p_{i_2} B = p_{i_0} B = A(i_0)$.

D'où l'invariance de la distance $d^2(j, j')$.

La démonstration serait la même dans le cas de l'agrégation de deux catégories socio-professionnelles j_1 et j_2 .

c) *Projection de variables « illustratives »*

Une fois déterminé le sous-espace sur lequel le nuage se projette sans trop de déformation (par exemple le sous-espace à deux dimensions constitué par les graphiques ci-après), on peut projeter des points supplémentaires, afin d'illustrer la représentation obtenue. Ainsi, dans \mathbb{R}^{373} , on a un point « téléphone » (dont la $i^{\text{ème}}$ coordonnée est le pourcentage de téléphones que représente la commune « i » dans l'ensemble des communes) que l'on peut projeter.

Le tableau effectivement analysé est le tableau 373×29 où figurent les individus sans distinction de sexes. Afin de faire apparaître les différences entre sexes, on projettera comme variables illustratives les profils géographiques et socio-professionnels correspondant aux sexes séparés. (L'équivalence distributionnelle nous permet d'affirmer que la typologie obtenue en analysant le tableau global 373×29 est très voisine de celle obtenue en analysant le tableau 373×58 , où les catégories socio-professionnelles correspondant à des sexes différents sont considérées comme des catégories socio-professionnelles différentes — cette similitude a été effectivement vérifiée. Il revient donc au même de projeter les catégories socio-professionnelles par sexe séparé comme variables illustratives que d'analyser le tableau désagrégé.)

Voyons comment se réalisent pratiquement ces projections :

Soient F_1, F_2 les deux axes factoriels qui engendrent le sous-espace à deux dimensions de déformation minimum de la projection du nuage de \mathbb{R}^{373} (nuage de 29 points). Soient t_{j_1}, t_{j_2} les coordonnées de la j -ème catégorie socio-professionnelle sur ces axes. Soient G_1, G_2 les deux axes factoriels analogues dans \mathbb{R}^{29} (où le nuage est formé de 373 points) et soient u_{i_1}, u_{i_2} les coordonnées de la i -ème commune sur ces axes.

On a les relations (cf. réf. 2, 3, 4), pour $k = 1, 2$:

$$t_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{373} \left(\frac{p_{ij}}{p_{.j}} \right) u_{ik}$$

$$u_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{29} \left(\frac{p_{ij}}{p_{i.}} \right) t_{jk}$$

$\lambda_k = k$ -ème plus grande valeur propre extraite au cours de l'analyse.

Si l'on néglige dans un premier temps les coefficients $1/\sqrt{\lambda_k}$, ces relations nous prouvent que les points-catégories socio-professionnelles sont des barycentres des points-communes, chaque point commune i étant affecté du poids : $p_{ij}/p_{.j}$ (qui caractérise la part de la commune i dans la catégorie socio-professionnelle j).

De même, les points-communes de notre représentation (graphique ci-après) seraient des barycentres des points-catégories socio-professionnelles, chaque point catégories socio-professionnelles « j » étant affecté de la masse $p_{ij}/p_{i.}$ (qui caractérise l'importance de la catégorie socio-professionnelle j dans la commune i). Ces relations justifient la représentation simultanée des communes et des catégories socio-professionnelles.

Les coefficients $1/\sqrt{\lambda_k}$, supérieurs à 1 en général (car $\lambda_k \leq 1$) assurent un déplacement excentrique des barycentres et permettent la réciprocité des relations liant u et t .

Les quantités $(u_{ik}/\sqrt{\lambda_k})$ [$i = 1, \dots, 373$] constituent l'opérateur-projection du profil $(p_{ij}/p_{.j})$ [$i = 1, \dots, 373$] de la catégorie socio-professionnelle j sur l'axe F_k .

Si l'on dispose de la répartition communale d'une variable supplémentaire : par exemple : nombre k_i de familles ayant 3 enfants dans la commune i — posons :

$$f_i = \frac{k_i}{\sum_1^{373} k_i}$$

La position du point « 3 enfants » dans le plan F_1, F_2 est donnée par les coordonnées :

$$e_k = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{373} f_i \cdot u_{ik} \quad k = 1, 2$$

De même, si l'on dispose de la répartition socio-professionnelle d'une variable, par exemple : profil socio-professionnel des femmes d'une commune particulière (k'_j femmes appartiennent à la catégorie j dans cette commune).

Le point aura pour coordonnées, dans le plan G_1, G_2 :

$$e'_k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{29} \left(\frac{k'_j}{\sum_j k'_j} \right) t_{jk} \quad k = 1, 2$$

d) Résultats et interprétation

Le graphique ci-dessous représente le plan des deux premiers facteurs F_1 et F_2 d'un espace, que nous avons fait coïncider avec le plan des deux premiers facteurs G_1 et G_2 de l'autre espace.

La proximité entre deux communes i et i' signifiera que les profils socio-professionnels de ces communes sont voisins, alors que la proximité de deux catégories socio-professionnelles signifiera que leurs profils communaux sont voisins.

Les relations « λ . barycentriques » évoqués plus haut nous permettent d'interpréter les positions respectives des points-communes et des points-catégories socio-professionnelles.

Ces deux premiers facteurs expliquent 70 % de la variance totale (50 % pour le seul premier facteur).

Rappelons que l'analyse ne porte pas sur les seules activités professionnelles, mais sur le statut de toutes les personnes actives ou non : au total 29 variables d'effectif inégal : 3 concernant les retraités, 3 les sans-emploi, 1 les étudiants, ces 7 catégories représentant la moitié de la population, les 22 autres variables représentant une structure socio-professionnelle relativement détaillée des personnes actives.

Deux observations préliminaires doivent être faites avant d'interpréter les graphiques :

— La première intéresse surtout les analyses portant sur la région parisienne : il s'agit de la correspondance entre le *type d'activité* (C. S. P.) et le lieu de *résidence*, non entre cette activité et le lieu où elle s'exerce. Les deux se confondent pour les enfants et peut-être les élèves, certainement pour les personnes retraitées et retirées des affaires; mais pour les actifs, résidence et lieu de travail peuvent être distincts lorsque la partition des communes est suffisamment fine (non au niveau des grandes agglomérations). Il serait intéressant de caractériser les communes par leur densité et leur type d'entreprises.

— La correspondance porte sur les hommes et les femmes, non sur les ménages. Or l'activité est individuelle, tandis que la résidence concerne le couple et la famille. L'analyse est plus riche mais plus complexe; elle exclut, non explicitement (on n'a même pas le statut matrimonial des personnes) l'homogamie ou l'exogamie qui mériterait une analyse particulière.

Pour des raisons de lisibilité et d'encombrement, les 373 communes de la région étudiée n'ont pas pu figurer sur les graphiques.

Seules quelques-unes d'entre elles, réparties de façon approximativement uniforme, ont été sélectionnées.

La carte socio-professionnelle de la région parisienne ainsi obtenue va servir de trame aux projections de variables illustratives.

1. Le premier graphique nous montre quelles sont les professions qui cohabitent, et dans quelles communes habitent les gens qui travaillent de la même façon.

Le premier axe factoriel (abscisse) est clairement un axe de ségrégation sociale (à l'exception de la cohabitation « fonctionnelle » des professions les plus favorisées avec les gens de maison).

Le second axe peut être qualifié au prime abord d'axe « d'expansion » (du haut vers le bas).

De façon anecdotique, on peut retenir que le premier axe oppose Gennevilliers à Auteuil (les manœuvres aux professions libérales), que le second oppose les vieux quartiers du centre de Paris aux villes-dortoirs ou résidentielles (les retraités, petits commerçants, aux enfants, aux ingénieurs).

2. Ce graphique 1 nous donne également les projections des points-catégories socio-professionnelles pour les femmes (et non plus pour les deux sexes).

En moyenne, les femmes se situent à droite par rapport aux hommes de la même profession en ce qui concerne les actifs. C'est surtout vrai à gauche du graphique. Ça ne l'est pas pour les inactifs.

Ce décalage semble intéressant : ce n'est pas le statut professionnel de la femme (souvent activité d'appoint d'ailleurs) qui détermine le statut de la famille et donc son lieu de résidence, mais le statut du mari, lequel doit être souvent supérieur. D'une manière plus générale, le travail des femmes est une preuve d'émancipation, favorisée dans les milieux aisés.

Les points-catégories socio-professionnelles « hommes » n'ont pas été tracés. Ils sont sensiblement symétriques des points-C. S. P. « femmes » par rapport aux points moyens représentant les deux sexes.

On vérifie ainsi, sur ces données pourtant assez agrégées et globales, que le fait d'être ingénieur pour une femme, est un signe de haut statut social ou d'émancipation plus net que chez l'homme.

3. Nous venons de projeter des points de \mathbb{R}^{373} (catégories socio-professionnelles caractérisées par leurs profils communaux) dans le plan des deux premiers facteurs de l'analyse.

Notre représentation simultanée nous permet également de projeter sur ce même graphique des points de \mathbb{R}^{29} , c'est-à-dire des points représentant des « profils sociaux » : de même que les femmes d'une catégorie socio-professionnelle ont un profil communal sensiblement différent de celui de l'ensemble des individus de cette catégorie (et *a fortiori* des hommes de cette catégorie), les femmes d'une commune ont un profil socio-professionnel différent de celui de l'ensemble des individus de cette commune.

Nous projeterons ainsi sur le graphique 1 373 points résumant les profils socio-professionnels féminins des communes.

Pour des questions de lisibilité et d'encombrement graphique, ces points n'ont pas été portés. Indiquons brièvement quels sont les résultats de ces projections.

Les « communes-femmes » sont également décalées vers la droite par rapport aux « communes totales » (les communes-hommes se situeraient au contraire à gauche).

Ce décalage est beaucoup plus important sur la gauche du graphique. Le profil socio-professionnel des femmes d'une commune ne peut être voisin de celui des hommes si des professions à dominante masculine prédominent (manœuvres, ouvriers spécialisés, qualifiés).

Inversement, la catégorie « non actifs de 17 à 64 ans » comprend essentiellement des femmes.

Le décalage moyen vers la droite signifie néanmoins que le statut socio-professionnel moyen des femmes d'une commune est supérieur à celui des hommes de cette même commune.

Ceci peut paraître en contradiction avec le point 2, où l'on a noté que pour une même catégorie socio-professionnelle, la résidence des femmes correspond à un statut supérieur au leur.

Une des interprétations possibles de ce phénomène semble être ceci :

Le travail des femmes dans une commune est dans une certaine mesure optionnel. Elles choisissent entre avoir une activité rémunérée ou rester au foyer, et opèrent donc une sélection dans l'éventail des professions (aussi faible que soit la latitude de cette sélection).

Sur le graphique 2 ont été projetées des variables décrivant la structure par âge des populations actives et non actives.

On a également porté les variables illustratives suivantes :

— 6 variables caractérisant le nombre de familles ayant respectivement 0, 1, 2, 3, 4, 5 et plus, enfants.

— 4 variables caractérisant l'âge des habitations :

avant 1914
de 1915 à 1948
de 1949 à 1961
1962 et au-delà.

Les « itinéraires » suivis par les actifs et les non-actifs sont très parlants : seuls les individus appartenant aux couches sociales les plus modestes commencent à travailler tôt (Le point « Actif 15-19 ans » est à l'extrême gauche du graphique 2), et seules, les couches les plus aisées peuvent continuer à travailler tard (Le point « Actif 70-74 ans est à droite du graphique 2). Par contre, les non-actifs entre 20 et 30 ans se recrutent dans les couches aisées. Il s'agit essentiellement d'étudiants. On peut constater que la taille des familles et l'âge des immeubles ont des répartitions communales assez liées : ces deux variables dépendent des deux facteurs mis en évidence. Le second facteur peut d'ailleurs être qualifié d' « axe de migration quotidienne », ou de « ségrégation du lieu de résidence et du lieu de travail », alors que le premier facteur extrait, plus de deux fois plus important, nous était apparu comme un axe de ségrégation selon la fonction exercée.

e) *Choix des communes par partition*

Après cette analyse, le choix des communes repose sur une information plus copieuse.

Pour ne pas se limiter aux distances dans le plan des deux premiers facteurs, qui n'expliquent pas toute la dispersion, et en l'absence de critères de choix, nous avons procédé à une partition automatique de l'ensemble des communes à partir de leurs distances (distributionnelles) dans tout l'espace \mathbb{R}^{20} . Les centres des classes obtenus seront retenus pour constituer l'échantillon des communes.

Le programme de partition en n classes (n étant fixé à l'avance) que nous avons utilisé a été conçu et réalisé par E. Diday. Le principe, fort simple, en est schématiquement le suivant : on tire aléatoirement n communes. La relation « être plus proche de la commune i que de toutes les autres » induit alors une partition en n classes. On recalcule le centre de

chaque classe (caractérisé par exemple par 3 communes), la même relation induit une partition, ainsi de suite jusqu'à convergence (rapide) de l'algorithme. On tire ensuite n nouvelles communes pour vérifier la stabilité des classes obtenues.

L'analyse ci-dessus nous permet alors de contrôler la pertinence de ce choix, et au besoin de corriger le fonctionnement aveugle de l'algorithme.

Ludovic LEBART et Nicole TABARD
CREDOC

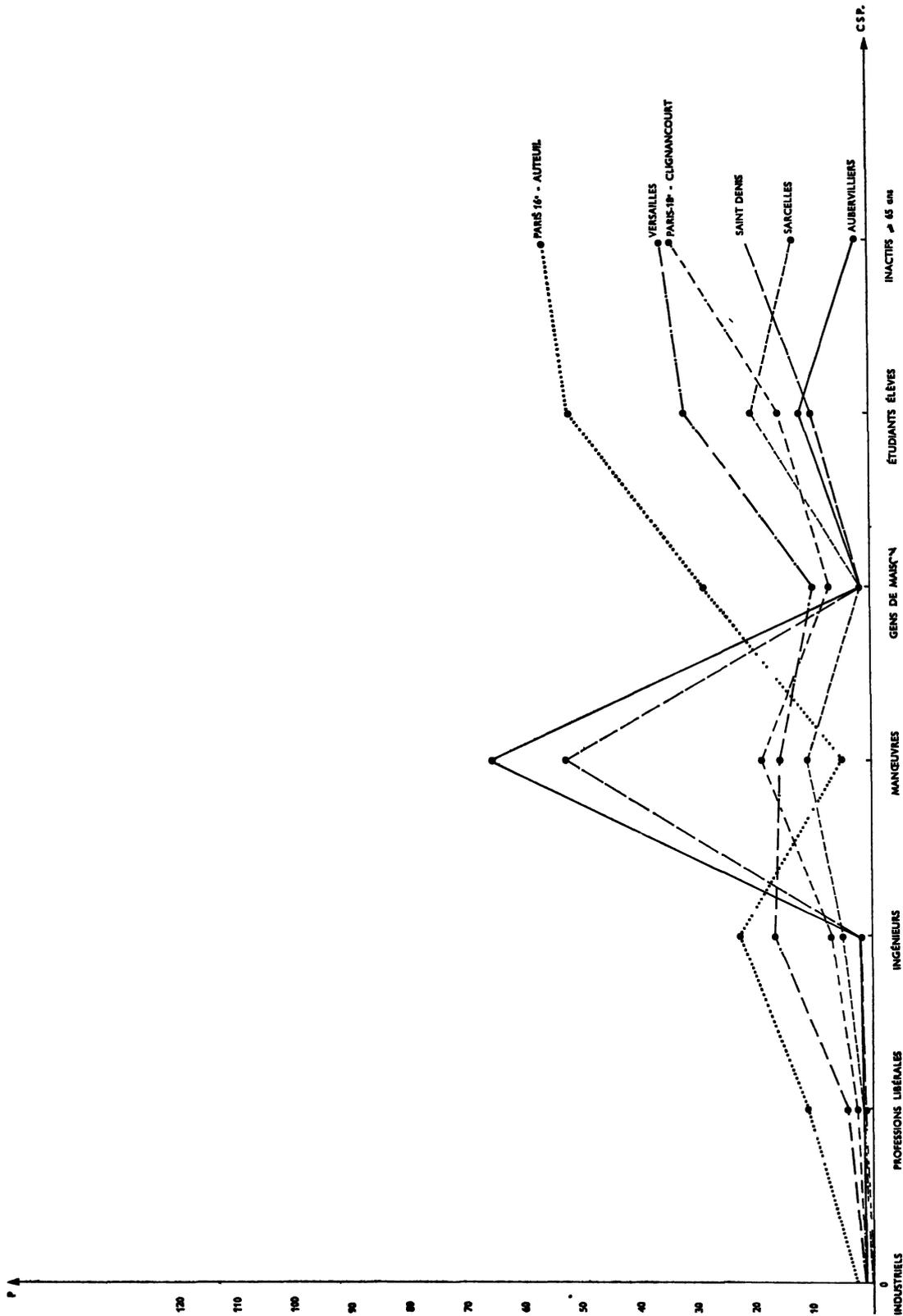
ANNEXE

Exemples de profils socio-professionnels

Profils d'activité de 6 communes choisies parmi les 373 du tableau initial
(Région parisienne)

		Aubervilliers		Saint-Denis		Sarcelles		Paris-18 ^e Clignancourt		Paris-16 ^e Auteuil		Versailles	
		Hab.	o/oo	Hab.	o/oo	Hab.	o/oo	Hab.	o/oo	Hab.	o/oo	Hab.	o/oo
1 Industriels	H	60	1	28	—	24	—	112	1	224	3	108	1
2	F	16	—	4	—	0	—	8	—	32	—	8	—
3 Artisans	H	504	7	468	5	228	5	876	11	504	6	388	4
4	F	100	1	80	1	44	1	316	4	240	3	88	1
5 Gros commerçants	H	132	2	232	2	60	1	428	5	788	9	316	4
6	F	76	1	132	1	24	—	228	3	300	4	124	1
7 Petits commerçants	H	744	10	888	9	204	4	1 452	18	692	8	604	7
8	F	632	9	824	8	192	4	1 280	15	568	7	604	7
9 Professions libérales	H	88	1	124	1	64	1	288	3	924	11	372	4
10	F	8	—	28	—	24	—	76	1	200	2	96	1
11 Professeurs, Prof. lit., sc.	H	60	—	100	1	120	2	800	4	972	12	476	5
12	F	68	1	108	1	108	2	368	4	960	11	444	5
13 Ingénieurs	H	168	2	232	2	284	5	572	7	1 932	23	1 524	17
14	F	0	—	4	—	4	—	32	—	156	2	44	—
15 Cadres adm. supérieurs	H	424	6	484	5	756	15	1 292	16	4 328	52	2 328	26
16	F	84	1	100	1	60	1	448	5	764	9	308	3
17 Instituteurs, Prof. int. div.	H	168	2	240	2	256	5	428	5	580	7	460	5
18	F	260	4	340	4	416	8	704	8	972	12	888	10
19 Services médicaux sociaux	H	20	—	28	—	20	—	56	1	60	1	48	—
20	F	188	3	292	3	148	3	440	5	568	7	576	7
21 Techniciens	H	1 132	16	1 500	15	1 040	21	1 572	19	916	11	1 800	20
22	F	192	3	224	2	108	2	316	4	284	3	244	3
23 Cadres administratifs moyens	H	668	9	982	10	1 068	21	1 868	23	2 000	24	1 888	16
24	F	652	9	772	8	492	10	1 476	18	1 528	18	848	10
25 Employés bureaux	H	1 916	26	2 732	28	2 108	42	3 548	43	1 704	20	2 348	27
26	F	4 396	61	5 836	60	3 220	64	7 184	87	4 276	51	5 204	59
27 Employés com.	H	688	10	616	6	376	6	1 044	13	500	6	616	7
28	F	788	11	948	10	432	9	1 512	18	908	11	1 104	13
29 Contremaîtres	H	828	11	1 096	11	436	9	612	7	200	2	596	7
30	F	68	1	80	1	64	1	156	2	112	1	36	—
31 Ouvriers qualifiés	H	6 232	86	8 472	87	2 772	55	5 224	63	1 480	18	4 488	51
32	F	984	14	1 144	12	444	9	1 300	16	504	6	528	6
33 Ouvriers spécialisés	H	5 324	74	6 824	70	1 688	34	3 184	38	1 016	12	2 588	29
34	F	1 852	26	2 172	23	356	7	1 240	15	308	4	864	10
35 Apprentis ouvriers	H	168	2	256	3	96	2	124	1	44	—	280	3
36	F	24	—	40	—	16	—	12	—	4	—	32	—
37 Manœuvres	H	4 776	66	5 244	54	548	11	1 580	19	436	5	1 512	17
38	F	1 756	24	2 304	24	468	9	968	12	360	4	760	9
39 Gens de maison	H	12	—	20	—	0	—	44	1	144	2	52	1
40	F	152	2	188	2	84	2	564	7	2 436	29	920	10
41 Femmes de ménage	H	8	—	0	—	0	—	16	—	24	—	8	—
42	F	304	4	456	5	116	2	980	12	792	9	784	9
43 Autres personnes service	H	660	9	836	9	296	6	1 600	19	592	7	564	6
44	F	948	13	1 316	14	620	12	2 992	36	1 852	22	1 152	13
45 Étudiants élèves	H	748	10	1 372	14	1 160	23	1 640	20	3 540	42	2 888	33
46	F	888	12	1 100	11	1 000	20	1 344	16	3 556	47	2 856	32
47 Retirés des affaires	H	156	2	184	2	88	2	332	4	428	5	200	2
48	F	228	3	248	3	76	2	468	6	420	5	300	3
49 Retraités Services publics	H	400	6	956	10	284	6	1 132	14	1 008	12	1 120	13
50	F	392	5	640	7	268	5	944	11	860	10	756	9
51 Anciens salariés secteur privé	H	1 824	25	2 460	25	416	8	3 352	40	2 316	28	1 584	18
52	F	2 588	36	3 504	36	680	14	5 321	71	3 080	37	2 448	28
53 Inactifs 0 à 16 ans	H	9 488	191	12 888	131	9 720	194	6 712	81	6 440	77	11 228	128
54	F	9 104	126	12 304	127	9 380	187	6 612	80	6 460	77	11 092	126
55 Inactifs 17 à 64 ans	H	484	7	1 052	11	156	3	1 148	14	964	12	800	9
56	F	7 540	104	9 520	98	6 204	124	8 356	101	10 872	130	10 828	123
57 Inactifs ≥ 65 ans	H	76	1	92	1	44	1	340	4	384	5	200	2
58	F	160	2	2 100	22	716	14	2 828	34	4 689	56	3 204	36
POPULATION TOTALE		72 399	1 000	97 064	1 000	50 076	1 000	82 877	1 000	83 600	1 000	88 004	1 000

FIGURE 3
RÉSUMÉS DES PROFILS SOCIO-PROFESSIONNELS DE QUELQUES COMMUNES
(P = proportions pour 1 000 hommes)



BIBLIOGRAPHIE

- [1] — *Leçons sur l'analyse statistique des données multidimensionnelles*, par J.-P. BENZECRI. L. S. M. (I. S. U. P.).
- [2] — *Distance distributionnelle et métrique du chi-deux en analyse des correspondances*, par J.-P. BENZECRI. L. S. M. (I. S. U. P.).
- [3] — *Introduction à l'analyse des données — « Consommation »*. Annales du CREDOC, nos 3 et 4, 1969.
- [4] — *Statistique et Informatique appliquées*, par L. LEBART et J.-P. FENELON. DUNOD édit., 426 pages (janvier 1971).
- [5] — *Méthodes statistiques de l'Économétrie*, par E. MALINVAUD. Dunod, 2^e éd., 1970.
- [6] — *Les conditions de vie des familles*, par N. TABARD et alii. CREDOC-UNCAF, 1967, 597 pages.