Bul.Méthodo.Socio. n°1,oct.83
Ph.Cibois

Méthodes post-factorielles pour le dépouillement d'enquête

A l'origine du concept de méthodes post-factorielles se trouve la grande difficulté rencontrée par beaucoup de chercheurs pour utiliser d'une manière correcte l'analyse factorielle, difficulté due pour une part importante au fait que le graphique factoriel peut être trompeur. En effet on sait que la proximité de deux points sur un plan factoriel n'a de chance d'exprimer une proximité réelle que si ces deux points ont une bonne contribution. Cette contribution ne se trouve pas sous une forme graphique mais sous la forme d'un indicateur numérique à lire par ailleurs.

Fidèle à la logique de l'analyse des données qui veut que la représentation graphique exprime tout ce qui est pertinent, nous faisons l'hypothèse que des indicateurs statistiques simples peuvent être représentés sous une forme graphique fructueuse en se servant, d'une manière auxiliaire, d'une représentation factorielle.

C'est avec cet objectif que nous avons développé, à l'intention des usagers du dépouillement d'enquête un certain nombre de techniques que nous allons maintenant présenter (1).

I La représentation en surface des tableaux croisés

Le principe de la méthode se trouve chez J.Bertin (2) avec ce qu'il appelle la technique des matrices pondérées : on sait que se pose alors le problème de l'ordonnancement des lignes et des colonnes. Cet ordonnancement peut trouver soit des solutions empiriques (permutation manuelles) soit des solutions

---

(1) Post-factoriel signifie que les techniques employées utilisent au préalable l'analyse factorielle.

(2) J.Bertin, La graphique et le traitement graphique de l'information, Paris, Flammarion, 1977.

algorithmiques propres (1) : nous proposons d'utiliser à cet effet le premier facteur d'une analyse des correspondances du tableau de départ. En effet, comme J.P.Benzécri l'a montré (2),s'il existe un ordre pour les lignes et les colonnes d'un tableau de contingence qui maximise les valeurs sur la diagonale, cet ordre se trouve être celui d'un premier facteur d'une analyse des correspondances.

La méthode de la représentation en surface d'un tableau croisé consiste donc d'abord à en faire l'analyse des correspondances, puis à permuter les lignes puis les colonnes de façon qu'elles se trouvent dans l'ordre donné par le premier facteur. Ensuite ce sont les données originelles qui feront l'objet d'une représentation en surface : pour ce faire on représente chaque colonne comme un profil où à chaque case du tableau correspond une surface rectangulaire. Les largeurs sont proportionnelles au poids marginal de chaque ligne et les hauteurs à la différence entre la proportion marginale de la ligne et la proportion marginale toutes lignes confondues. On montre facilement que la surface se trouve de ce fait proportionnelle à l'écart à l'indépendance (3).

L'implémentation de cette méthode a été réalisée par le programme REPFAC (REPrésentation FACtorielle) implanté au CIRCE.

Exemple d'utilisation

Dans leur enquête sur l'orientation en fin de 3e, M.Reuchlin et F.Bacher (4) avaient posé deux questions relatives aux aspirations professionnelles. La première cherchait à enregistrer le désir profond de l'enfant indépendamment des contingences matérielles. C'est le choix idéal :

"En supposant que rien ne s'oppose à vos désirs, quelle profession voudriez-vous exercer plus tard ?"

---

(1) Cf. Alain Leduc, "Chainage automatique des matrices ordonnables", Université de Haute-Normandie.

(2) J.P.Benzécri et al., L'analyse des données, T.1, pp.261-287.

(3) Cet écart peut être représenté sous sa forme brute ou sous sa forme pondérée (contribution au Khi-deux),cf.annexe 1.

(4) M.Reuchlin et F.Bacher, L'orientation à la fin du premier cycle secondaire, Paris, Presses Universitaires de France,1969.

Evolution du choix professionnel

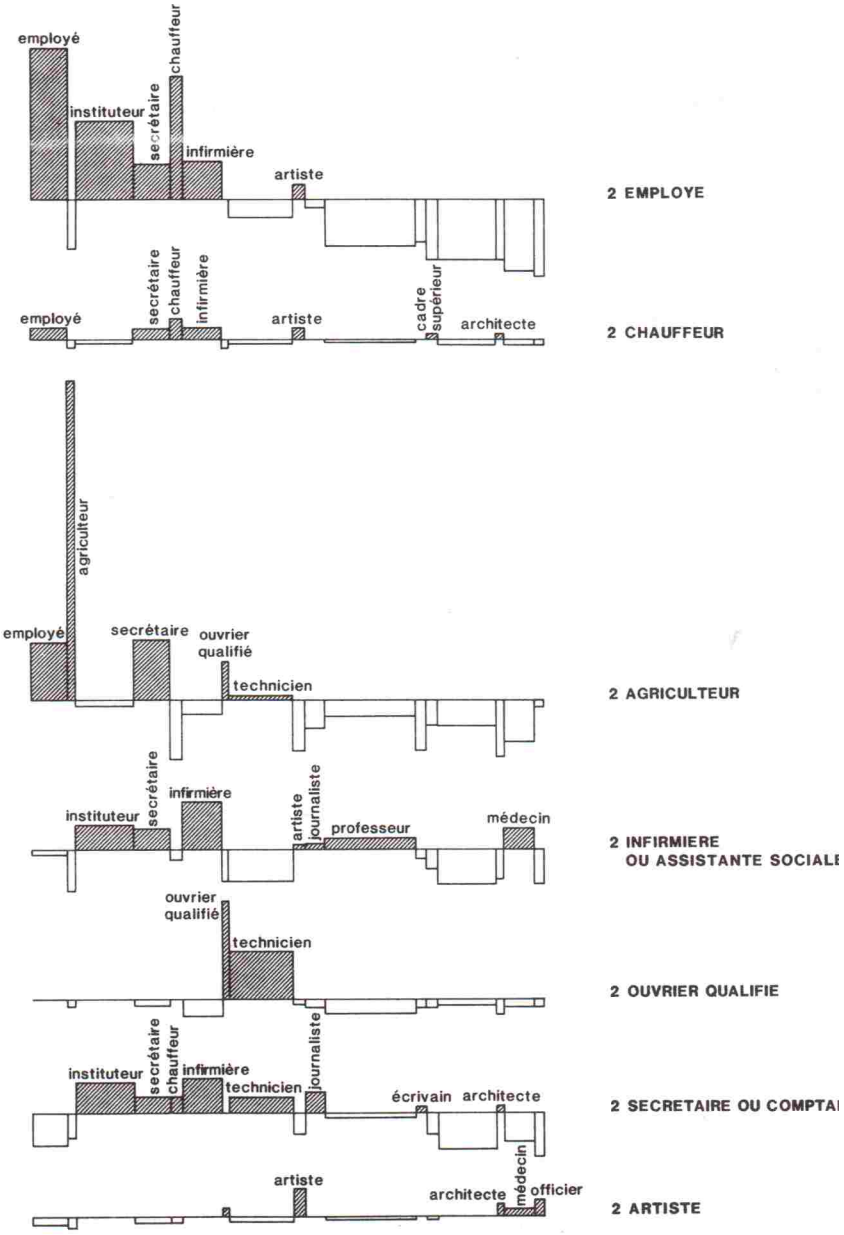
2ème choix / *1er choix*	Agriculteur	Médecin	Professeur	Ingénieur	Cadre supérieur	Architecte	Ecrivain	Instituteur	Infirmière (1)	Technicien	Secrétaire (2)	Journaliste	Employé	Ouvrier qualifié	Chauffeur	Artiste	Officier	*Total*
Agriculteur	36	0	0	0	0	0	0	0	0	1	3	1	3	1	0	0	2	47
Médecin	11	15	56	33	0	0	3	26	26	11	10	7	5	2	0	7	7	219
Professeur	70	17	80	18	11	5	22	175	72	55	69	26	51	4	6	12	7	700
Cadre supérieur	6	4	16	9	9	0	0	2	3	9	5	6	3	1	2	1	1	77
Ingénieur	34	17	64	34	5	13	11	32	4	121	15	9	18	8	0	8	27	420
Architecte	1	3	10	3	3	3	1	2	1	13	6	0	2	0	1	2	1	52
Ecrivain	3	4	23	5	1	2	4	6	5	8	10	1	7	1	1	2	1	84
Instituteur	53	2	11	3	1	1	1	24	58	31	72	13	137	13	4	10	4	438
Infirmière(1)	31	1	12	1	1	0	4	35	53	10	52	10	69	2	9	6	2	298
Technicien	66	6	17	16	2	5	2	20	9	131	66	3	60	56	3	8	4	474
Secrétaire(2)	71	1	13	4	0	1	1	13	35	22	38	0	68	7	9	5	0	288
Journaliste	10	1	21	5	0	1	3	12	12	15	20	6	19	2	2	3	3	135
Employé	62	0	2	0	0	0	0	12	19	8	12	2	118	9	10	3	1	258
Ouvrier qualifié	10	0	1	0	1	0	0	2	1	8	5	0	8	11	0	2	0	49
Chauffeur	1	0	4	0	1	0	1	5	4	5	8	0	24	2	3	1	1	60
Artiste	3	1	15	3	0	1	3	8	8	3	6	10	17	2	3	7	1	91
Officier	9	2	15	10	1	0	1	5	1	15	1	3	1	1	0	4	8	77
Total	477	74	360	144	36	32	57	379	311	466	398	97	610	122	53	81	70	3767

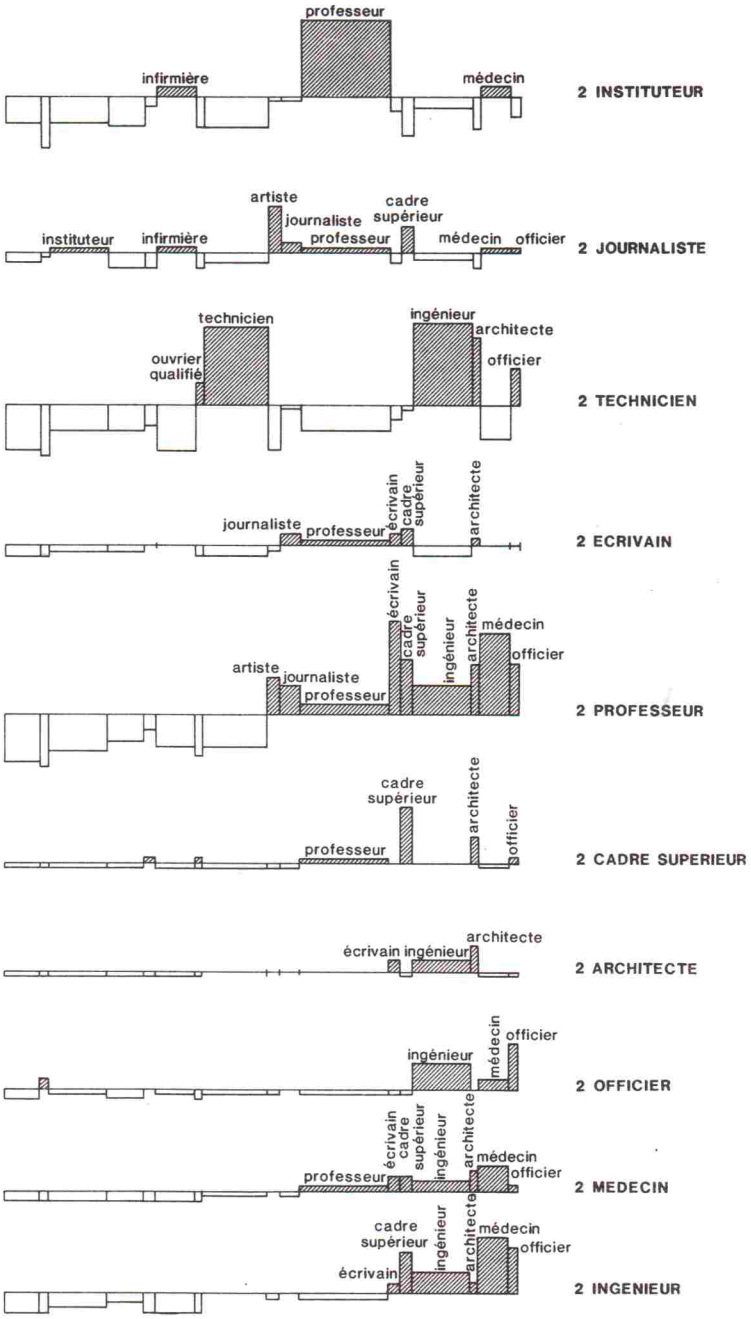
(1) ou assistante sociale
(2) ou comptable ou clerc de notaire

Figure 1

EVOLUTION DU CHOIX PROFESSIONNEL

FIGURE 2




professeur
infirmière
médecin
2 INSTITUTEUR
artiste
journaliste
cadre supérieur
instituteur
infirmière
professeur
médecin
officier
2 JOURNALISTE
technicien
ingénieur
architecte
ouvrier qualifié
officier
2 TECHNICIEN
journaliste
professeur
écrivain
cadre supérieur
architecte
2 ECRIVAIN
écrivain
cadre supérieur
ingénieur
architecte
médecin
officier
artiste
journaliste
professeur
2 PROFESSEUR
cadre supérieur
architecte
officier
professeur
2 CADRE SUPERIEUR
écrivain
ingénieur
architecte
2 ARCHITECTE
ingénieur
médecin
officier
2 OFFICIER
écrivain
cadre supérieur
ingénieur
architecte
médecin
officier
professeur
2 MEDECIN
cadre supérieur
ingénieur
architecte
médecin
officier
écrivain
2 INGENIEUR

La deuxième question demandait un choix plus réaliste si le premier choix s'avérait impossible :

"Dans le cas contraire, avez-vous pensé à une autre profession ? laquelle ? "

Le croisement de ces deux questions pour ceux qui ont répondu aux deux, donne, après élimination des lignes ou colonnes à effectifs trop faibles, le tableau croisé de la figure 1. Ce tableau à un Khi-deux de 2447,8 : si on le soumet à l'analyse des correspondances, l'approximation des deux premiers facteurs en prend en compte 64% dont 41% pour le seul premier facteur. Le premier vecteur propre donne pour chaque ligne ou colonne une coordonnée qui sert habituellement pour la construction d'un graphique de représentation simultanée : nous ne présenterons pas ce graphique. En effet, et bien qu'il soit parfaitement interprétable, il s'agit d'une approximation des données puisqu'on ne prend en compte que 64% du Khi-deux. Nous préférons donc examiner l'intégralité des écarts à l'indépendance et nous servir de la coordonnée factorielle de chaque ligne ou colonne pour ordonner l'ensemble.

Dans le graphique de la figure 2 on a ordonné lignes et colonnes en fonction de leur coordonnées factorielles pour le premier facteur et on a fait la représentation en surface des écarts à l'indépendance. Les lignes (premier choix) sont précédées de "1", les colonnes (deuxième choix) de "2".

Dans une représentation en surface, nous représentons pour chaque case du tableau son écart à l'indépendance :

- au-dessus de la ligne de référence (en grisé) si celui-ci est positif. Cas d'attraction entre une ligne et une colonne.
- au-dessous de la ligne de référence (en blanc) si celui-ci est négatif. Cas de déficit par rapport- à l'indépendance : opposition entre ligne et colonne.
- sans écart à la ligne de référence s'il n'y a pas d'écart à l'indépendance.

Chaque surface est proportionnelle au nombre d'individus en écart à l'indépendance : cette surface est obtenue par le produit de deux proportions. En hauteur, l'écart à la proportion moyenne dans la colonne considérée ; en largeur la proportion de la ligne considérée.

Soit par exemple le croisement de la première ligne et de la première colonne selon l'ordre du premier facteur. Il s'agit de premier choix "employé" et deuxième choix "employé"'. Soit les effectifs de case de de marges correspondants ainsi que les pourcentages en ligne.

	2 employé	total
1 employé	118	258
	45,7%	100%
Total	610	3767
	16,2%	100%

S'il y avait indépendance entre les lignes et les colonnes, dans la ligne 1 employé, 16,2% seulement se trouveraient choisir 2 employé ; or ce n'est pas le cas puisque 45,7% parmi les 1 employé font ce choix. On a donc un écart au pourcentage moyen de 45,7 - 16,2 soit +29,5 : c'est ce nombre qui sert de hauteur dans la première case du graphique. Quant à la base c'est simplement la proportion de la ligne 1 Employé dans la population, c'est à dire 258/3767 soit 6,8%.

On vérifiera que le produit de l'écart par la base en proportion est : 0,45736 - 0,16193 = 0,29543

0,29543 x 0,06849 = 0,02023

écart base prop. des individus en écart à l'indépendance.

Cette proportion multipliée par l'effectif total 3767 nous donne l'effectif de cette case en écart à l'indépendance soit +76,22. C'est ce même effectif que l'on aurait obtenu par produit des marges divisé par le total (1)

Interprétation

Les 4 derniers profils indiquent d'où viennent d'une manière privilégiée ceux qui choisissent comme deuxième choix les professions d'ingénieur, de médecin, d'officier et d'architecte. Or on constate que ce sont surtout ceux qui ont choisi ces 4 professions en premier choix qui alimentent ce groupe en deuxième

---

(1) on trouvera en annexe 2 un procédé de construction à la main de graphiques de ce type.

choix. Ceci se traduit visuellement par la ressemblance des 4 profils qui ont des écarts positifs pour ces catégories et des déficits ailleurs. Le point commun de ces professions est de se situer à un niveau élevé de la hiérarchie sociale, au moins dans l'imaginaire de jeunes arrivant en fin de 3e en 1965 date de l'enquête. Alors que les questions de l'enquête étaient posées en terme de choix idéal et de choix réaliste, elles ont été comprises ici en terme de substitution : "si je n'arrive pas à être ingénieur, je serai médecin". On accepte éventuellement de changer de domaine à condition que l'on reste au même niveau social.

On retrouve des traits analogues dans les trois profils suivants de 2 Cadre supérieur, 2 Professeur et 2 Ecrivain avec simplement l'apparition de professions plus littéraires (professeur, journaliste, artiste, écrivain) perçues toujours à un haut niveau de prestige. Dans les 7 premiers profils de 2 Ingénieur à 2 Ecrivain, les substitutions de deuxième choix sont toujours au même niveau de la hiérarchie sociale.

Il n'en est plus de même avec le profil suivant de 2 Technicien où deux blocs de premiers choix apparaissent : celui de technicien et celui d'ingénieur. Ceux qui ont choisi technicien en premier n'ont fait que confirmer leur décision mais par contre ceux qui ont choisi d'abord ingénieur ont ici introduit une nouvelle logique de comportement. Elle consiste à accepter en choix réaliste une baisse de ses espérances sociales avec pour consolation le fait de ne pas changer de secteur d'activité. Ce raisonnement de substitution "régressive" se retrouve dans le profil voisin de 2 Instituteur dont la masse des écarts positifs est constituée par ceux qui avaient choisi professeur en premier choix. Le raisonnement est le même : "Si je ne peux pas être professeur, je serai instituteur". On retrouvera encore le même raisonnement mais à un niveau plus bas de la hiérarchie sociale dans le profil 2 Ouvrier Qualifié dont les écarts positifs ne sont faits que par des premiers choix de technicien.

Par contre les profils restants sont des substitutions à égalité de niveau hiérarchique au niveau le plus bas. Les secrétaires ou comptables en 2e choix se recrutent surtout dans

les catégories instituteur, secrétaire, infirmière ou technicien en premier choix. Les infirmières ou assistantes sociales de même, avec cependant quelques substitutions régressives de premier choix de médecin ou de professeur. Enfin tout en bas de la hiérarchie on voit les profils de 2 Employé et 2 Chauffeur à recrutement de même niveau.

Quelques profils, tout en s'inscrivant dans la logique générale ont quelques particularités.

- 2 Agriculteur : la plupart de ceux qui avaient choisi agriculteur en premier choix persistent dans leur choix. Ils sont peu nombreux mais exclusifs, ce qui explique la forme allongée de leur écart dans ce profil. Dans les autres, le premier choix d'agriculteur est toujours en écart négatif sauf pour le deuxième choix d'officier (1).

- 2 Artiste, 2 Journaliste et 2 Ecrivain : ces 3 profils ont des écarts très faibles. On doit remarquer que les premiers choix d'artiste ou de journaliste sont toujours inférieurs à la moyenne pour les catégories les plus élevées (de 2 Cadre supérieur à 2 Ingénieur) mais se rencontrent plus souvent dans les profils de niveau inférieur. Ces premiers choix expriment un désir de conquête de la hiérarchie sociale qui se fixe sur des professions quelque peu mythiques pour des jeunes. On ne s'étonnera pas que les deuxièmes choix soit très réalistes. Cette ascension sociale mythique s'oppose aux deux voies beaucoup plus réelles que sont l'enseignement ou la filière technique (ingénieur ou technicien).

En résumé on voit que la représentation graphique met en relief trois logiques de sustitution des choix :

1) Substitution en restant au niveau élevé de la hiérarchie sociale (ingénieur, médecin, officier, architecte, cadre supérieur et professeur).

---

(1) Avec pour cette case "1 Agriculteur - 2 Officier" un effectif théorique de 0,9 individu et un effectif observé de 2.

2) Substitution régressive : de professeur à instituteur, d'ingénieur à technicien, de technicien à ouvrier qualifié, de médecin à infirmière.

3) Substitution au niveau inférieur de la hiérarchie (employé, chauffeur, agriculteur, infirmière, secrétaire).

Retour aux données

Nous voyons maintenant que le seul tableau de données ne nous permettra pas de savoir pourquoi nous sommes en présence de plusieurs logiques de substitution. Il est évident en effet que ces différents projets doivent être mis en rapport avec la position sociale d'origine. La substitution au niveau élevé est plus le fait d'originaire de haut niveau de la hiérarchie sociale qui désirent maintenir leur position sociale. Au contraire, la substitution régressive correspond à une trajectoire d'ascension sociale à partir des milieux intermédiaires. Enfin la substitution au niveau inférieur est déjà le signe d'un départ très bas (1).

II Le Pourcentage du Khi-deux Maximum (PKM)

Le fait que l'analyse factorielle permette de trouver un ordre sur les lignes et les colonnes va également être utilisé pour juger si le Khi-deux observé dans un tableau de contingence représente une part notable du Khi-deux qu'il y aurait s'il y avait une liaison maximum entre les lignes et les colonnes.

La pratique habituelle pour répondre à cette question est celle du V de Kramer : cet indice a pour numérateur le Phi-deux observé et pour dénominateur n-1 où n est le plus petit nombre de lignes ou de colonnes. Cette valeur du dénominateur se trouve bien être le Phi-deux maximum dans un cas tout à fait spécifique, celui où il existe une dépendance fonctionelle entre les lignes et les

---

(1) Cf. notre rapport "Demande individuelle d'éducation. Etude de cas :France" doc.SME/ET/76.16, OCDE 1976,pp.56-61.

colonnes. Si par exemple c'est le nombre de colonnes qui est inférieur au nombre de ligne cela suppose que pour une ligne donnée, les effectifs de toute la ligne se trouvent concentrés sur une seule colonne. On concoit bien que ce maximum ne soit atteint sur des tableaux empiriques que dans des cas asez rares et que de ce fait, pour un tableau donné, le V de Kramer soit assez pessimiste.

Le problème de la recherche du Khi-deux maximum (ou ce qui revient au même du Phi-deux maximum), se ramène à l'exploration des extrémités d'un cléroèdre (1) associé à des marges données : on sait que cette étude est complexe du fait du grand nombre de sommets possibles aussitôt que le tableau a un nombre suffisamment grand de degrés de liberté. Par contre il se simplifie en se ramenant à deux possibilités si l'on dispose d'un ordre sur les lignes et sur les colonnes. En effet on peut alors trouver les effectifs qui associent autant qu'il est possible une ligne à une colonne, soit en chargeant la première diagonale, soit la seconde. Cette procédure, que G.Th.Guilbaud appelle procédure du "Quart Nord-Ouest" consiste à :

1- partir d'une des extrémités quelconque de la diagonale

2- y mettre la plus faible des deux marges

3- mettre le résidu de la marge restante au plus près de la diagonale

4- calculer le résidu pour l'autre marge

5- recommencer en 3 l'alternance ligne/colonne jusqu'à la fin.

On commencera par la case du tableau située en haut et à gauche si l'on veut charger la première diagonale ou en haut et à droite si l'on veut charger la seconde.

---

(1) Cf. G.Th.Guilbaud,"Exercice de calcul pour préparer à l'usage raisonnable du Khi-deux", Mathématiques et Sciences Humaines, n°14, 1966,pp.31-40. et Ph.Cibois,"La représentation factorielle des tableaux croisés et des données d'enquête : étude de méthodologie sociologique", Paris, LISH, 1980.

Puisque pour un ordre donné on a mis autant qu'il était possible les lignes et les colonnes en liaison et que l'on a que deux solutions, il est possible maintenant de calculer le Khi-deux lié à chacune d'elles. On dispose donc maintenant d'un dénominateur beaucoup plus plausible que celui employé dans le V de Kramer. En effet on dispose de la liaison maximum pour des marges données, ce que ne prenait pas en compte l'indice V .

En ce qui concerne l'ordre à choisir pour les lignes et les colonnes, il s'impose parfois de lui-même quand les intitulés sont ordonnés. Si ce n'est pas le cas, on peut toujours en trouver une approximation en utilisant celui donné par le premier facteur d'une analyse des correspondances. Enfin en ce qui concerne la diagonale à choisir, c'est l'interprétation qui la donne facilement puisqu'il s'agit de choisir entre deux solutions opposées : il est évident qu'une seule est en rapport avec ce que l'on observe.

L'implémentation de cet algorithme de calcul du PKM est en cours à l'intérieur du programme REPFAC.

Exemple d'utilisation

Soit par exemple deux questions issues de l'enquête sur L'Ouvrier Français en 1970 (1), et qui sont les suivantes:

1) "Est-ce que l'attitude de la CGT pendant le mois de Mai 1968 vous à tout à fait satisfait, plutôt satisfait, plutôt déçu, très déçu ?"

2) Même question mais vis-à-vis de la CFDT.

En n'utilisant que la population des répondants aux deux questions, on observe les effectifs suivants :

---

(1) G.Adam et al., L'ouvrier français en 1970,Paris, Presses de la FNSP,1970.

Satisfaction vis-à-vis de la CGT	Satisfaction vis-à-vis de la CFDT très satisfait	plutôt satisfait	plutôt déçu	très déçu	TOTAL
très satisfait	36	33	20	20	109
plutôt satisfait	7	144	45	8	204
plutôt déçu	10	47	154	17	228
très déçu	3	11	19	117	150
TOTAL	56	235	238	162	691

Une lecture rapide du tableau montre que tous les écarts à l'indépendance positifs sont sur la diagonale du tableau et tous les écarts négatifs en-dehors. Ceci signifie que quand on a un certain degré de satisfaction pour une centrale, on a plutôt tendance à avoir le même degré de satisfaction pour l'autre.

Si cette hypothèse était pleinement satisfaite, quels effectifs aurions-nous ? Pour répondre à cette question, il faut essayer de charger au maximum la diagonale du tableau qui associe les niveaux de satisfaction.Par exemple pour la case "très satisfait" par CGT et CFDT, le maximum possible nous est donné par la plus petite des deux marges, ici la marge colonne. En effet comme la colonne des "très satisfaits" de la CFDT n'est que de 56,c'est le plus que l'on puisse mettre dans la case de la diagonale avec un effectif nul pour les autres cases de la colonne.

La colonne des "très satisfaits" CFDT est terminée mais il reste 109 - 56 = 53 "très satisfaits" de la CGT à placer. Comme on fait l'hypothèse que la liaison est la plus forte possible, si on ne peut pas les placer au même niveau, on les place au niveau le plus proche, c'est à dire comme plutôt satisfaits de la CFDT.

Parmi les 235 de cette colonne, 53 sont déjà placés et il nous en reste 235 - 53 = 182 à placer. Tous ceux-ci peuvent être classés au niveau de satisfaction équivalent pour la CGT puisque le total de cette ligne est de 204. Il en reste donc 204 - 182 = 22 à placer que nous plaçons au niveau immédiatement inférieur.

Une fois l'algorithme terminé on a les résultats suivants:

Satisfaction vis-à-vis de la CGT	Satisfaction vis-à-vis de la CFDT très satisfait	plutôt satisfait	plutôt déçu	très déçu	TOTAL
très satisfait	56	53	0	0	109
plutôt satisfait	0	182	22	0	204
plutôt déçu	0	0	216	12	228
très déçu	0	0	0	150	150
TOTAL	56	235	238	162	691

Sur ce tableau artificiel mais unique pour les règles données, on constate que c'est la faiblesse de la marge" très satisfaits" de la CFDT qui déplace des effectifs au-dessus de la diagonale.

On peut maintenant calculer le Khi-deux tant sur le tableau observé que sur le tableau artificiel correspondant au maximum de liaison entre les deux questions. On a les résultats suivants :

- Khi-deux observé = 584,1
- Khi-deux maximum = 1460,8

La part que le Khi-deux observé prend du Khi-deux maximum est de 584,1 / 1460,8 = 0,400 soit 40,0%. Ce "pourcentage du Khi-deux maximum" doit-il être considéré comme élevé ? Pour cette réponse on ne se satisfera pas de la connaissance de son minimum 0 et de son maximum 100 théoriques. En effet pour utiliser validement ce genre d'indicateur, il faut en acquérir une expérience qui peut varier selon les populations et les circonstances étudiées. Chacun devra donc en quelque sorte l'étalonner pour son propre compte. On peut par exemple dire que pour l'enquête citée, un pourcentage du maximum de 40% peut-être considéré comme l'indice d'une forte liaison.

On remarquera que la technique du V de Kramer est très imparfaite car elle exagère le maximum du Khi-deux : si on l'avait appliquée au cas de la satisfaction vis-à-vis des syndicats on aurait trouvé un Khi-deux maximum de (4 - 1) x 691 = 2073 et donc un pourcentage du maximum de 584,1 / 2073 = 0,282 soit 28,2% beaucoup plus pessimiste que celui noté plus haut de 40%.

On voit tout l'intérêt qu'il y a à trouver un ordre, même imparfait sur les modalités des questions. Pour y arriver et pour la même raison que pour la représentation en surface d'un tableau, on peut utiliser l'ordre des lignes et des colonnes donnée par le premier facteur d'une analyse des correspondances.

III La méthode Tri-deux

L'analyse factorielle présente un certain nombre de difficultés quand on l'applique au dépouillement d'enquête. Au coeur de celles-ci se trouve le fait que le statut de ce qui est mis en avant par l'analyse est délicat à établir : s'il s'agit d'un type idéal, s'applique-t-il à une population importante ou marginale ? Ce type comprend-il beaucoup d'individus "purs" ? S'agit-il d'un artefact du à un facteur de correction ? Les proximités entrevues sont-elles dues à un effet de chaînage ? A certaines de ces questions on peut trouver des éléments de réponse dans l'examen des contributions des modalités mais il s'agit là d'éléments d'information qui sont extérieurs au graphique. Si l'on se satisfait de cette technique, on renonce à l'ambition qui est à la source de l'analyse des données, c'est à dire le fait de représenter les éléments pertinents des données sur un graphique. L'analyse factorielle ne présente que certains de ces éléments pertinents et de plus leur interprétation à partir de données numériques est délicate et suppose une bonne connaissance de la méthode.

La méthode Tri-deux au contraire va essayer de représenter graphiquement l'élément que le sociologue juge pertinent, c'est à dire l'attraction entre deux modalités de réponse.

Dans le cas d'un simple tableau croisé on s'intéresse à l'attraction entre une ligne et une colonne ; dans le cas du dépouillement d'enquête il peut exister pour une ligne ou une colonne donnée de multiples attractions simultanées qui n'ont pas toutes la même importance. Le principe de la méthode Tri-deux consiste à prendre les plus importantes de ces attractions et à les représenter graphiquement sur un fonds de carte d'analyse des correspondance.

D'une manière plus formelle, on suppose que l'on traite une enquête où les réponses aux questions soient sous forme disjonctive complète. Il est d'ailleurs toujours possible de se ramener à cette situation en discrétisant les variables continues et en ne tenant pas compte des ordres éventuels sur les modalités de réponse aux questions.

On considère l'ensemble de tous les croisements deux à deux entre toutes les questions de l'enquête (tableau de Burt). Pour chaque croisement on calcule pour chaque case l'écart à l'indépendance. On retient tous les écarts positifs qui indiquent de ce fait une certaine "attraction" entre la ligne et la colonne. On range en ordre décroissant l'ensemble des écarts issus de tous les tableaux.

On a ainsi un graphe valué dont les sommets sont constitués par l'ensemble de toutes les modalités de réponse issues de toutes les questions. En ce qui concerne ses arêtes elles sont définies de la manière suivante :

- elles ne peuvent exister qu'entre modalités issues de questions différentes.
- elles existent quand, dans le croisement des deux questions dont elles sont respectivement issues, l'écart à l'indépendance est positif.
- elles ne sont pas orientées.
- elles sont valuées par l'intensité de l'écart à l'indépendance. Ce dernier point est justifié par le fait que tous les tableaux ont même effectif total. Si l'on est géné par l'hétérogenéité des marges, on peut en option prendre la contribution au Khi-deux de la case considérée plutôt que l'écart brut à l'indépendance.

Un tel graphe est connexe à la condition que toute modalité de réponse soit prise au moins par un individu. On suppose que l'enquête a été recodée pour arriver à ce résultat. Par contre si l'on ne s'intéresse qu'au sous-ensemble des arètes dont la valuation est supérieure à un certain niveau on a alors plusieurs parties disjointes.

S'il existe des cliques dans le graphe, cela signifie que des modalités issues de plusieurs questions sont en attractions simultanées. Du point de vue de l'interprétation on voit apparaitre dans ce cas un univers de modalités qui manifeste un type de répondant. On retrouve un univers sous une forme affaiblie quand, pour un certain niveau de la valuation, il y a simplement connexité entre des modalités.

Il serait possible de faire une recherche automatique des cliques ou des simples parties connexes mais on a préféré prendre l'option de la représentation graphique. En effet comme un univers de modalités peut être plus ou moins bien constitué et sa représentation sur le graphe passer de la clique maximale à la simple connexité sans solution de continuité, on a préféré laisser cette discrimination à la charge de l'utilisateur. Il suffit pour cela de lui représenter graphiquement un certain nombre d'outils qui sont:

- l'ensemble du graphe.
- son résumé sous la forme de son arbre minimum.
- le sous-ensemble des parties disjointes pour un certain niveau de la valuation (ou encore en remplaçant chaque partie par son arbre minimum).

Chacune de ces options est utilisable soit en considérant la valuation apportée par l'écart brut à l'indépendance soit en considérant la forme pondérée de cet écart, c'est à dire la contribution au Khi-deux qui en est issue.

Pour la représentation graphique, on estime que la longueur des arêtes sera minimisée si on utilise pour représenter les sommets le premier plan d'une analyse factorielle quelconque des données d'origine (composantes principales ou correspondances). C'est en ce sens que la méthode tri-deux peut être considérée comme une méthode "post-factorielle". Elle utilise l'analyse factorielle d'une manière pratique pour optimiser la représentation d'un graphe qui seul est pertinent pour le chercheur.

La méthode tri-deux permet de contourner le délicat problème des "contributions" en analyse factorielle des correspondances. En effet, sur un plan factoriel une proximité entre modalités peut être fallacieuse car les contributions n'y sont pas

représentées. Avec la méthode tri-deux la question ne se pose plus puisque ce qui est pertinent n'est plus la proximité des modalités mais la présence ou l'absence d'une arête qui de plus peut être valuée graphiquement par l'épaisseur du trait. On tend mieux ainsi vers cet idéal de la représentation graphique qui est de manifester tout ce qui est pertinent.

Implémentation

Dans sa version actuelle la méthode tri-deux enchaîne les divers programmes nécessaires en cumulant les résultats intermédiaires sur un fichier auxiliaire. On a les programmes suivants :

- CODLOG : codage logique des données.
- YAGOL : programme d'analyse factorielle (version UER MLFI de l'Université Paris V).
- ECARTS : calculs et ordonnancement des écarts.
- TRACES : représentation graphique.

Tous les programmes sont en Fortran, le dernier utilise le système GPGS pour piloter une unité graphique. Tous ces programmes sont implantés au CIRCE ainsi qu'au CNUSC. Ils sont communicables (programme source) sur simple demande ainsi que les notices d'utilisation.

Exemple d'utilisation

Cette recherche a pour origine les pertes de voix communistes à l'élection présidentielle de 1981 et aux diverses élections qui l'ont suivie. Certains en effet on affirmé que les électeurs qui auparavant votaient communistes étaient en quelque sorte "captifs" de la gauche. S'ils ne votaient plus communistes, ils avaient donc toutes chances de voter socialiste.

Cette manière de voir repose implicitement sur la notion d'une opposition gauche/droite linéaire, la gauche communiste se situant plus à gauche que les socialistes. Or, cette hypothèse n'a pas été confirmée par diverses analyses des affiliations faite sur l'enquête de l'ouvrier français en 1970 (1). On y a vu en effet qu'entre les trois pôles de la gauche communiste,

---

(1) Cf. Ph.Cibois, "La représentation...", op.cit.

de la gauche non-communiste et de la droite, il y avait une disposition en triangle et non pas en ligne. Comme il existait une population intermédiaire entre la gauche-communiste et la droite, il n'était donc pas impossible qu'il y ait un passage direct de l'un à l'autre au point de vue électoral. De ce fait il devenait intéressant de se servir de la population de l'ouvrier français de 1970 comme une source historique pouvant renseigner sur des évolutions ultérieures.

Pour arriver à ce but on s'est intéressé à la sous-population de ceux qui avaient voté communiste à l'élection présidentielle de 1969. Dans la suite c'est donc sur la sous-population des 234 votants pour Duclos que nous ferons porter nos analyses, dans le but de faire l'inventaire de leurs opinions et de leurs caractéristiques. On fait l'hypothèse que si certains cessent de voter communiste, le choix d'un autre vote se fera en fonction des types d'opinions qui étaient les leurs en 1970.

Questions prises en compte

On a retenu 20 questions comportant 92 modalités pour faire une typologie des opinions de la sous-population étudiée. Ces questions sont soit d'opinion, soit d'affiliation, soit de statut.

1) Questions d'opinion

- portant sur la marche de l'entreprise : comment l'entreprise doit-elle être dirigée (par l'état, par les syndicats, par tout le personnel, comme actuellement) ; sur la discipline (inutile, génante mais utile,indispensable) ; sur l'interlocuteur privilégié en cas de difficulté (le patron, le syndicat, quelqu'un d'autre).

- portant sur l'opinion vis-à-vis de la politique : participation à des discussions en cas de distribution de tracts (souvent, parfois, jamais) ; intérêt pour la politique (beaucoup, assez, un peu, pas du tout) ; si la politique est une activité honorable (très honorable, honorable, a peu près, pas du tout) ; si l'on parle ou non politique au travail, en famille, dans le quartier ; si l'on juge son entourage professionnel de droite, de gauche ou du centre ; comment on se juge soi-même orienté politiquement.

- portant sur le racisme : y a-t-il ou non trop de nord-africains ou de juifs en France ?

2) questions d'affiliation

Syndicat auquel on est affilié ; parti politique dont on se sent le plus proche ; syndicat pour lequel on vote aux élections professionnelles.

3) Questions de statut

Sexe, age, revenu, catégorie de commune.

Analyse

Pour faire l'analyse de ces 20 questions on utilise la méthode Tri-deux : sur un fonds de carte d'une analyse en composantes principales des données en codage logique (1), on représente les 60 attractions les plus importantes entre modalités issues de deux questions. Dans l'ensemble des 190 tableaux deux à deux possibles entre 20 questions, les 60 attractions les plus importantes correspondent à des écarts à l'indépendance supérieurs ou égaux à 9,8 individus.

Sur la figure 3 qui représente ces 60 attractions (2), les 92 modalités ne sont pas représentées : certaines en effet ne sont reliées au graphe que par des attractions inférieures au seuil retenu. On y distingue 5 graphes distincts dont deux seulement regroupent plus de deux modalités. Ce sont ces deux graphes que nous allons maintenant examiner : chacun d'entre eux constitue un type particulier.

---

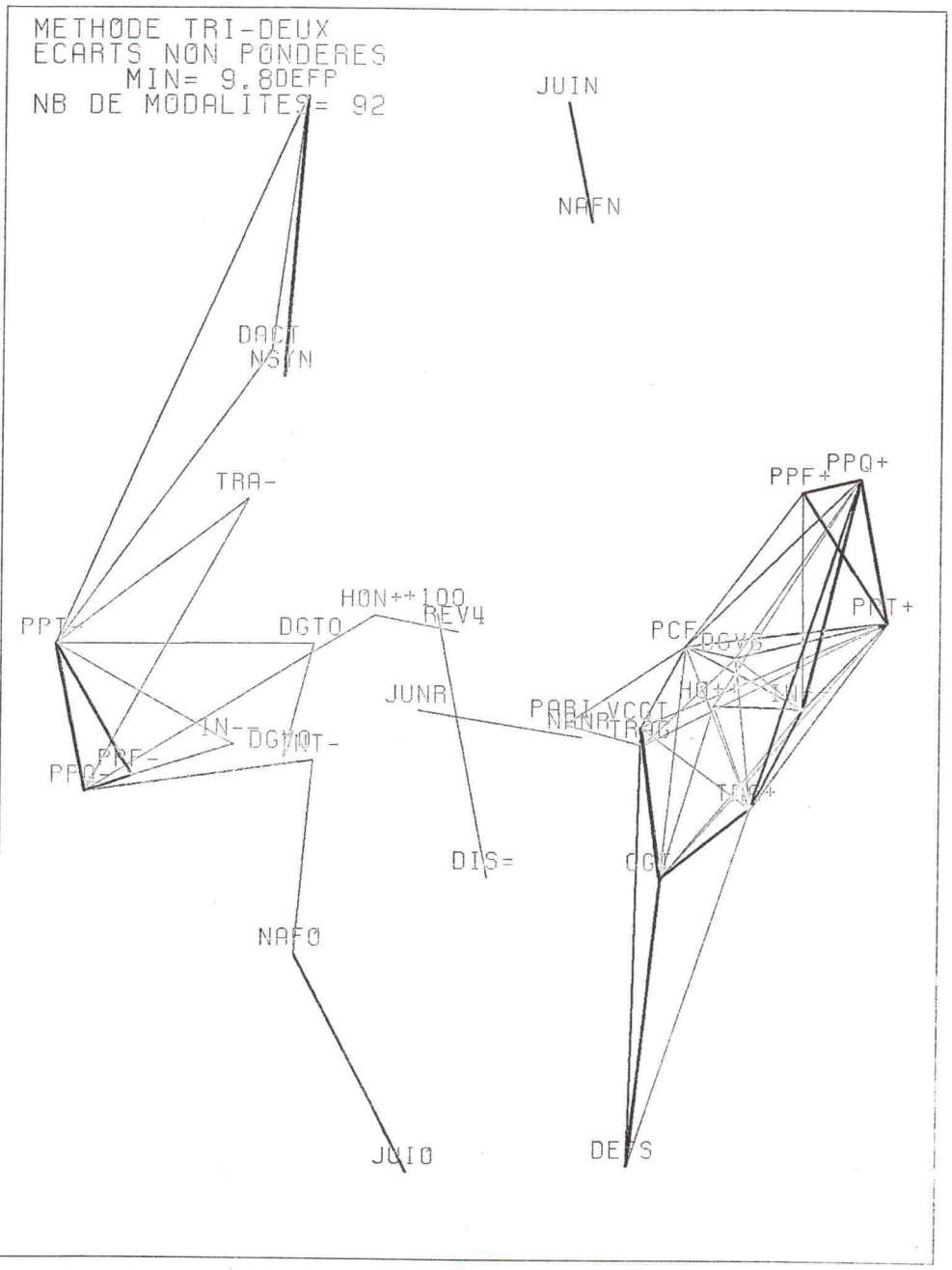
(1) On a retenu l'analyse en composantes principales (sur des données centrées et normées) plutôt que l'analyse des correspondances pour des raisons de lisibilité. Dans le cas présent, ce type d'analyse rend le graphique des écarts plus lisible : ce n'est pas une loi générale. Il convient pour chaque cas de prendre le fonds de carte le plus adapté.

(2) Graphique imprimé directement par programme sur imprimante électrostatique.

FIGURE 3

OUVRIER FRANCAIS 1970 : électeurs de Duclos

Graphe des 60 attractions les plus fortes



1) On voit sur la partie droite un graphe en forme de "diamant" : cette structure particulière est réalisée quand toutes les modalités d'un sous-ensemble sont en relation deux à deux. Dans ce cas le graphe est aussi compact que possible : il ne l'est ici que d'une façon approchée mais cependant assez nette. En commençant par le haut, les différentes modalités en sont : PPQ+, parle politique dans son quartier ; PPF+, en famille ; PPT+, au travail ; PCF, se sent proche du parti communiste ; DGVG, dans l'opposition droite/gauche se situe à gauche ; HO++, pense que la politique est quelque chose de très honorable ; IN++ a beaucoup d'intérêt pour la politique ; PARI, réside dans l'agglomération parisienne ; VCGT, vote CGT ; TRAG, considère que son entourage de travail est de gauche ; TRA+, participe toujours aux discussions quand un tract est distribué ; CGT, appartient à la CGT ; DEFS, en cas de difficulté dans l'entreprise, fait appel à un membre d'un syndicat.

Ce premiers univers de modalités n'est pas surprenant pour une population d'électeurs communistes : il s'agit du type-idéal du militant PC-CGT confiant dans l'action politique et syndicale, il est motivé et actif. On le trouve plus que la moyenne en région parisienne.

2) Le deuxième ensemble de modalités, qui se trouve sur la gauche du graphique ne possède pas cette structure forte en diamant. Autour de quelques triangles, on trouve un simple enchainement de modalités ce qui dénote une plus grande hétérogénéité du type-idéal. On trouve de haut en bas les modalités suivantes : DEFP, en cas de difficulté essaye de régler directement l'affaire avec le patron ; DACT, pense que l'entreprise doit être dirigée comme actuellement (et non par l'état, les syndicats ou le personnel) ; NSYN, n'est pas syndiqué ; TRA-, en cas de distribution de tracts ne discute que rarement ; PPT-, PPQ-, PPF-, ne parle politique ni au travail, ni dans le quartier, ni en famille ; DGTO et DGVO, ne répond pas à la question de savoir si son entourage de travail ou lui-même est de gauche ou de droite ; HON+, pense que la politique est honorable ; REV4, appartient à la tranche médiane des revenus ouvriers ; INT- ou IN--, n'a que peu ou pas du tout d'intérêt pour la politique ; NAFO ou JUIO, pense qu'il y a trop de nords-africains ou de juifs.

Ce deuxième type-idéal est caractérisé par un certain nombre de réactions comme l'apolitisme, le respect de la hiérarchie et le racisme que l'on ne s'attend pas à voir habituellement partagé par des électeurs communistes. Si depuis cette enquête l'électorat du parti communiste s'est réduit, on peut raisonnablement penser que ceux qu'il va perdre se situent plus dans cette frange idéologique que du coté du noyau examiné précédemment.

Alors qu'on comprendrait assez bien que des militants du noyau "dur", s'ils devenaient contestataires de la ligne du Parti, aille encore dans un parti de gauche, on voit mal pourquoi cette frange d'électorat communiste irait rejoindre la gauche non-communiste alors que tous ces thèmes semblent plutôt appartenir à la droite. Si le parti communiste pert des voix, celles-ci ont de fortes chances soit d'être perdues pour tout le monde (abstentionnisme), soit d'être gagnées par un parti de droite. Par exemple, le RPR, du fait de son implantation populaire en récupère peut-être.

Il est d'ailleurs possible que ce soit cette frange idéologique que le parti communiste ait visé à travers diverses campagnes ayant eu lieu avant l'élection présidentielles de 1981. Des thèmes comme la drogue ou les travailleurs immigrés ont pu avoir là un certain impact.

En conclusion on notera donc qu'il y a tout lieu de penser qu'une partie de la population ouvrière d'électeurs communistes n'est pas électoralement captive de la gauche non-communiste si elle abandonne le vote communiste. Il est très possible qu'elle rejoigne directement la droite gaulliste ou reste dans l'abstention (1).

---

(1) Il reste à noter dans le graphique la présence de trois groupes de deux modalités qui forment des ensembles isolés dont on ne peut rien dire de ce fait. Il s'agit de ceux qui pensent qu'il n'y a pas trop de juifs ou de nords-africains (JUIN,NAFN). De ceux qui ne répondent pas à ces questions (JUNR,NANR) et enfin d'une sous-population qui habite des communes de plus de 100.000 habitants autres que la région parisienne (+100) et qui ont une attitude moyenne vis-à-vis de la discipline (DIS=, pensent que la discipline et le réglement sont génants, mais que ça accroit le rendement).

IV Les variables idéales-typiques

Une question se pose cependant : quelle est l'importance numérique des deux groupes dont nous avons repéré la présence. Pour y répondre nous allons utiliser la technique des variables idéales-typiques que nous allons exposer maintenant.

En effet les groupes de modalités que nous avons repérés (et il s'agit là d'un cas tout à fait général) ne correspondent pas a des effectifs importants de répondants si on exige la présence simultanée des traits qui les constituent. Par exemple dans le cas présent alors que le groupe "militant" est constitué par 12 modalités, il n'y a aucun individu qui les possèdent toutes et seulement 7 qui en possèdent 11. De même du côté "abstentionniste" formé par les réponses à 13 questions, personne n'en a 13 ni 12 et un seul en a 11.

Cette constatation serait inquiétante si nous n'utilisions pas ici le concept wéberien de type-idéal : l'ensemble des modalités que nous avons observé ne correspond à des effectifs restreints que si l'on exige le type à l'état pur. Par contre si on le prend d'une manière approchée, on a des effectifs notables si l'on décide que le critère d'appartenance au type sera non le fait d'avoir la totalité des traits qui le constituent mais seulement un certain nombre.

Cependant pour un même nombre de traits du type, il y a plusieurs manières d'y appartenir : si certaines étaient prépondérentes, les analyses factorielles préalables les auraient manifestées. Comme ce n'est pas le cas on peut donc tenir pour équivalents les divers indicateurs constitutifs du type.

Dans le cas présent le type-idéal du "militant" est constitué par 12 modalités tandis que celui de l'"abstentionniste" l'est par 14 modalités issues de 13 questions (car on peut recoder ensemble l'intérêt faible et l'intérêt très faible). On va donc faire l'hypothèse que pour appartenir à un type, il suffit d'en posséder un nombre suffisant de modalités.

Nous ne déterminerons par a priori quelle doit être le nombre "suffisant" de modalités pour appartenir au type. Cette détermination sera faite au vu du croisement des types : on suppose en effet qu'il est contradictoire d'appartenir aux deux en même temps. Il suffira donc de voir à partir de combien de modalités les deux types s'excluent mutuellement.

En premier lieu il faut construire par programme une variable auxiliaire qui pour chaque individu fait le compte du nombre de modalités possédées de chaque type. Ceci est réalisable facilement dans la plupart des logiciels de dépouillement d'enquête. Il suffit pour chaque type de tester successivement si l'individu en cours a choisi les différentes modalités du type. Si c'est le cas on augmente d'une unité un compteur de modalités. Après passage dans les 12 tests du type "militant" et dans les 13 tests du type "abstentionniste", on dispose pour chaque individu de deux variables auxiliaires qui sont le nombre de modalités de chaque type (1).

On peut maintenant faire un tableau qui croise ces variables idéal-typiques brutes que sont le nombre de traits de chaque type. C'est ce qui est fait à la figure 4 où l'on a mis en ligne le nombre de modalités du type "militant" et en colonne le nombre de modalités du type "abstentionniste". On constate que :

- comme nous l'avons signalé, ni pour les militants, ni pour les abstentionnistes, le type n'existe à l'état pur, le maximum est atteint pour 7 individus avec 11 modalités sur 12 du type militant et pour 1 individu qui possède 11 modalités sur les 13 du type abstentionniste. Dans de telles situations on constate toujours que les types-idéaux n'apparaissent pas à l'état pur mais beaucoup plus souvent d'une manière approchée.

- si l'on regarde la forme générale de la distribution des individus, on constate qu'elle s'ordonne selon l'une des diagonales du tableau, celle qui va du type militant maximum au type abstentionniste maximum. Si le nombre de modalités d'un type diminue, le nombre de modalités de l'autre augmente : il n'y a personne qui n'ait aucun type ou les deux types à la fois (2).

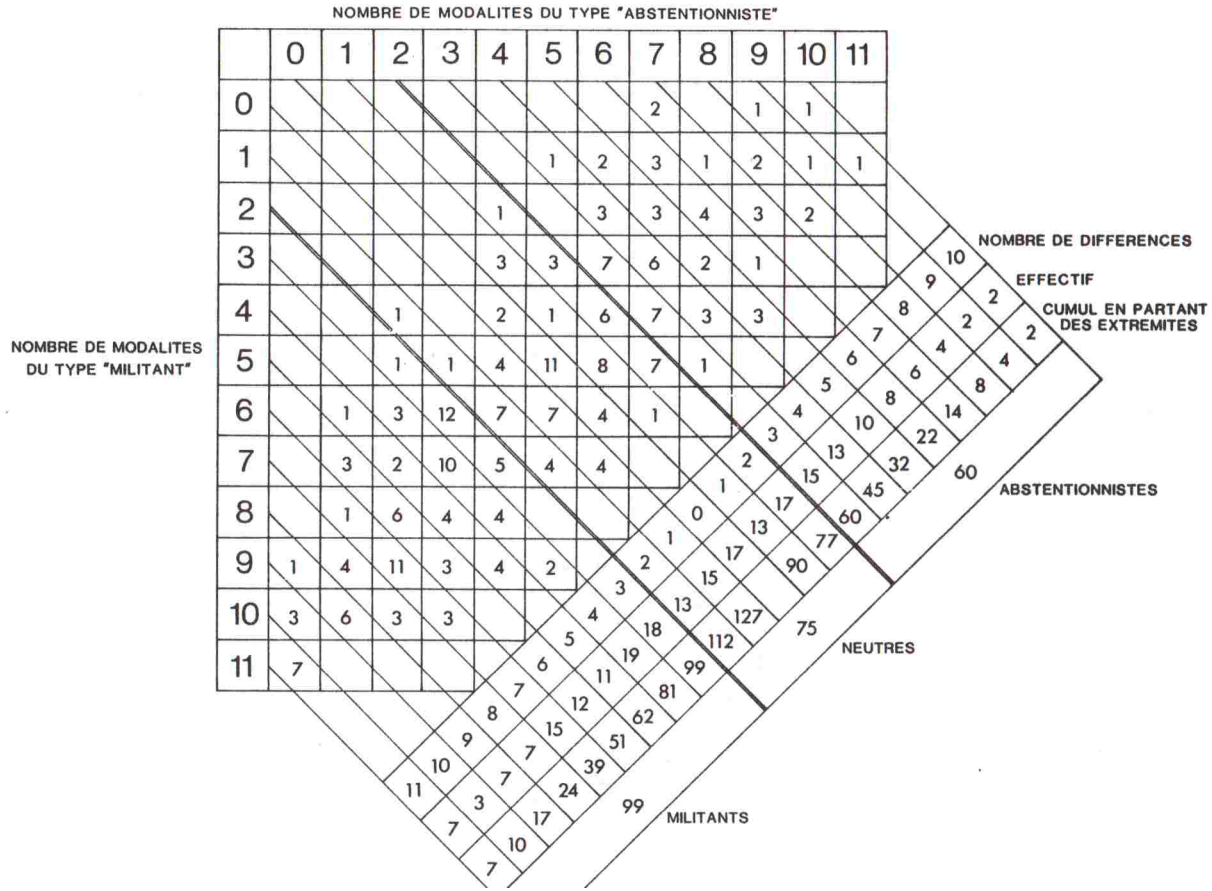
---

(1) Le nombre de modalités n'a pas besoin d'être rigoureusement identique pour chaque type, il suffit qu'il soit du même ordre de grandeur.

(2) Comme les deux types possèdent en commun 10 questions, on ne peut pour des raisons formelles posséder en même temps les deux types, cependant rien n'empêcherait que des individus n'appartiennent à aucun des deux types en choisissant des réponses qui ne sont ni dans l'un ni dans l'autre.

FIGURE 4

OUVRIER FRANCAIS 1970 : électeurs de Duclos
Variables Idéal–Typiques



Comme on voit bien sur la figure 4 qu'il n'y a pas de coupure nette dans la distribution, on doit donc en déduire que la population passe insensiblement de l'un à l'autre ce qui nécessite la prise en compte d'une sous-population intermédiaire que l'on peut dire "neutre" ou "non-classée" quant à l'opposition. Appartiennent à coup sûr à celle-ci, ceux qui ont autant de modalités de chaque type, c'est à dire :

- les 2 individus qui en ont 4 traits de chaque
- " 11 " " " " 5 " " "
- " 4 " " " " 6 " " "

Total : 17

De même ceux qui ont une modalité de plus dans un type, donc qui se situent autour de la diagonale où les nombres de traits sont identiques, sont à classer dans le type neutre. Pour savoir où l'on doit mettre le seuil d'appartenance, c'est à dire pour pouvoir juger à quel niveau de différence il faut mettre la frontière entre le type neutre et les "vrais" types, on a reporté sur la figure les effectifs de chacune des diagonales du tableau. Dans chaque diagonale se trouvent les effectifs de ceux qui ont la même différence entre le nombre de modalités d'un type et le nombre de modalités de l'autre. Par exemple dans la partie supérieure du graphique, 2 individus ont 10 modalités du type abstentionniste de plus que de modalités de type militant : le premier parce qu'il en à 10 et zéro, le deuxième en ayant 11 et 1. La première série de nombre indique le nombre de différences, la deuxième l'effectif correspondant, somme des effectifs de la diagonale considérée, la troisième étant un cumul des effectifs de chaque type en partant des extrémités, c'est à dire des types purs.

Le choix de la frontière ne s'impose pas de lui-même, il faut simplement trouver un compromis entre le désir d'avoir des types aussi purs que possibles et cependant assez nombreux. On peut par exemple dire que pour appartenir à un type, il faut avoir au moins 3 modalités de plus du type. Avec cette convention on a les effectifs suivants :

- Type-idéal "militant" : 99
- population neutre : 75
- Type-idéal "abstentionniste" : 60

De toute façon, quelle que soit la frontière, on constate que le type "militant" est plus important en effectif que le type "abstentionniste". Par exemple en étant plus exigeant d'une différence pour établir les frontières, on aurait fait passer 15 individus du type abstentionniste au centre ainsi que 18 du type militant. On aurait les résultats suivants :

- Type-idéal "militant" : 81
- population neutre :108
- Type-idéal "abstentionniste" : 45

On a toujours le même effet de prééminence numérique du type militant mais la population non classée d'un tiers passe maintenant à près de la moitié ce que l'on peut juger excessif.

Ayant ainsi apprécié numériquement l'importance des deux types-idéaux, il est possible d'utiliser cette variable idéale-typique dans des tris croisés ou dans d'autres analyses selon les orientations que l'on désire privilégier. Cette technique des variables idéales-typiques, simple à mettre en oeuvre et a utiliser permet de revenir aux données d'une manière compréhensible par tous. Elle a, par rapport à d'autres techniques de typologies automatiques, l'avantage d'être construite de manière raisonnée par le chercheur : avec elle on sait toujours pourquoi des individus sont rassemblés dans un même type. Ici c'est le type conceptuel qui guide le chercheur et non le type empirique obtenu par un regroupement automatique des individus que l'on cherche à comprendre ensuite. Par contre, les types-idéaux qui ont permis le classement de la population n'ont pas été construit a priori par le chercheur mais au vu des données ce qui est plus satisfaisant.

A chaque moment de la recherche, le chercheur doit faire alterner observation et interprétation : les variables idéal-typiques sont l'opérationalisation de la phase active de même que le graphe des attractions de la méthode Tri-deux lui permettait une observation aisée.

V Extensions diverses

La représentation post-factorielle peut aussi s'appliquer aux matrices de corrélations et plus généralement à toute matrice symétrique. Se pose alors cependant le problème du statut à donner à ce qui se trouve sur la diagonale : comme un général il ne s'agit pas du résultat d'une observation, il devient possible de manipuler celle-ci afin de réduire le rang de la matrice. On retrouve là la procédure de l'analyse factorielle classique (dite "des psychologues") qui tend à minimiser le rang de la matrice étudiée.

Pour arriver à ce résultat on renonce complètement aux artifices de calcul de l'analyse factorielle classique pour utiliser l'algorithme de l'analyse en composantes principales de la manière suivante :

1- fixer un seuil de précision pour l'arrêt qui donne la part de la Somme des Carrés que l'on désire retrouver à un rang donné (99% ou 99,9% par exemple)

2- faire l'analyse en composantes principales sur le nombre désiré de facteurs (en commençant à un)

3- reconstituer la diagonale pour le nombre de facteurs en cours

4- refaire l'analyse avec la nouvelle diagonale reconstituée mais en utilisant les données d'origine pour ce qui est en dehors de la diagonale ; puis retourner en 2

5- arrêter la boucle 2 à 4 à stabilité de la diagonale pour une précision donnée

6- tester si l'on obtient avec le nombre de facteurs en cours la part de la Somme des Carrés prévue par le seuil

7- si oui, arrêt ; sinon augmenter le nombre de facteurs désirés de un et retourner en 2

En abandonnant le facteur tout en nombres positifs (premier facteur dit facteur "de taille") et en se plaçant dans le plan des facteurs 2 et 3 on obtient ainsi un fonds de carte possible pour une représentation des valeurs de la matrice. Comme dans la méthode tri-deux, on considère les intitulés de la matrice comme des sommets d'un graphe et les valeurs comme des valuations des

arêtes. Ici le graphe est complet : on le présente par ordre d'importance décroissante des valuations jusqu'à un seuil donné et on fait apparaitre graphiquement cette décroissance par des traits d'épaisseur différente.

Cette technique est particulièrement utile dans le cas de l'Analyse de Similitude qui ne disposait pas d'un système de représentation graphique : plutôt que de tenter par la constitution d'un filtrant (1) de sélectionner les cliques maximales les plus intéressantes, on laisse ce soin à l'utilisateur à partir d'un graphique où celles-ci apparaissent facilement.

Cette technique peut d'ailleurs s'étendre au cas de l'exploration générale d'un graphe non valué dont on désire simplement explorer les connexités, par exemple pour une analyse de réseaux. Dans ce cas la matrice sera simplement une matrice de 0 et de 1 pour exprimer l'absence ou la présence d'une arête entre deux sommets.

Du point de vue de l'implémentation de la méthode, seul le programme d'analyse factorielle classique a été mis au point sous le nom de PSYFAC (analyse FACtorielle des PSYchologues)

Exemples d'utilisation

1) Analyse de similitude

Nous utilisons l'exemple de Degenne et Verges qui donne un indice de similitude entre 9 variables décrivant des techniques agricoles:

	1	2	3	4	5	6	7	8	9
1 Outillage agricole	-	10	9	4	15	15	16	17	10
2 Artisanat courant		-	2	3	9	8	7	15	19
3 Assolement			-	3	4	14	11	11	5
4 Fumure organique				-	19	2	6	11	6
5 Fumure minérale					-	6	6	18	11
6 Sélection semence						-	17	13	9
7 Sélection arbustive							-	17	6
8 Sélection animaux								-	19
9 Mécanicien									-

---

(1) Pour un exposé d'ensemble cf. Alain DEGENNE et Pierre VERGES, "Introduction à l'analyse de similitude", Revue Française de Sociologie (4) XIV, 1973, pp.471-512.

Le graphe valué des arêtes les plus importanes entre les 9 sommets est représenté sur la figure 5A. Le fonds de carte a été obtenu en faisant l'analyse factorielle classique sur 3 facteurs de la matrice de similitude. On retient le plan des facteurs 2 et 3 (en éliminant le facteur 1 qui, tout en termes positifs, est un facteur de taille) et on relie les points ainsi obtenus par des traits d'épaisseur décroissante selon l'importance de l'arête : on a choisi le seuil de 10 comme un bon compromis entre la richesse de l'information et la clarté nécessaire.

A titre de comparaison, on a porté sur la figure 5B la représentation de l'arbre maximum du graphe donnée par les auteurs. Si on compare les deux graphiques on y retrouve bien la position centrale de la variable 8 et les 3 groupes 4-5, 8-9, 3-6-7 mais on y voit mieux la place de 1 qui doit plutôt se trouver du coté du groupe 3-6-7. Même si l'on ne retenait que les arêtes formant l'arbre maximum on aurait intéret à faire une représentation graphique qui tienne compte de l'intégralité de l'information afin que la représentation spatiale n'induise pas en erreur, ce qui est le cas de la figure 5B

2) Exploration d'un graphe

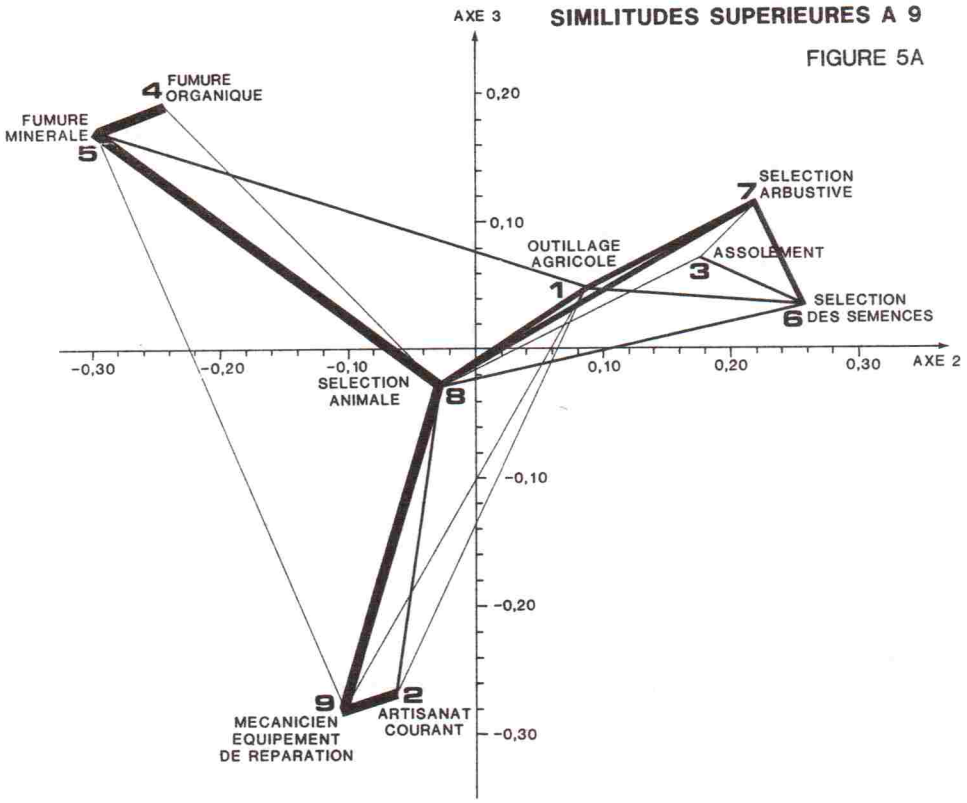
Des procédures existent qui permettent de trouver automatiquement les cliques maximales d'un graphe (1) : leur inconvénient est que l'on passe de la simple connexité à la clique maximale sans solution de continuité et que de ce fait le chercheur est souvent autant intéressé par une clique presque complète que par une clique maximale. Pour lui faciliter le choix on représente le graphe en utilisant en fonds de carte la représentation du plan des facteurs 2 et 3 de la matrice d'incidence dont on a fait au préalable l'analyse factorielle.

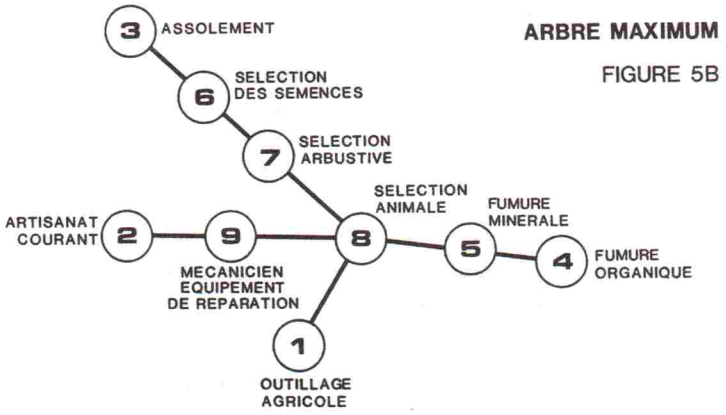
L'exemple traité ici est le jeu d'essai de Giraud : on a représenté à la figure 6A la représentation post-factorielle du graphe et à la figure 6B le graphe donné par l'auteur.

---

(1) Christian Giraud, "Cliques maximales d'un graphe", Informatique et Sciences Humaines, n°55, décembre 1982, pp.45-59.

ANALYSE DE SIMILITUDE





La représentation factorielle choisie est une analyse en composantes principales car l'analyse factorielle classique modifie la diagonale ce qui est ici dommageable. En effet dans une matrice d'incidence le "un" d'une diagonale, même s'il est toujours présent a le même sens qu'en dehors de celle-ci : il signifie une relation. De ce fait on constate que la modification de la diagonale dans un tel cas, perturbe plus qu'elle n'éclaire la représentation graphique.

Dans le graphique 6A on a utilisé le plan des facteurs 2 et 4 et non des facteurs 2 et 3. En effet le facteur 3 est constitué par les sommets 1 à 4 (2 et 3 étant d'ailleurs confondus parce qu'étant dans des situations strictement identiques). De ce fait le reste du graphe est écrasé sur l'axe 1 ce qu'on évite en prenant l'axe 4 plutôt que l'axe 3 en coordonnées verticales.

En examinant la figure la figure 6A on voit bien que, en plus des sommets 1 à 4, deux ensembles s'opposent : à gauche les sommets 6 à 12 et à droite les sommets 13 à 22. Ce dernier groupe de sommet présente la particularité d'un allongement qui peut s'interpréter comme la juxtaposition de 8 cliques maximales à 4 sommets qui différent d'une seul sommet de l'une à l'autre. On a en remontant du bas vers le haut du graphique :

6 14 15 21	13 14 15 21	13 15 21 22
13 15 18 21	15 17 18 21	17 18 20 21
17 19 20 21	19 20 21 22	

On vérifie que de la première à la seconde les sommets 14 15 21 sont communs,13 15 21 de la 2e à la 3e etc...

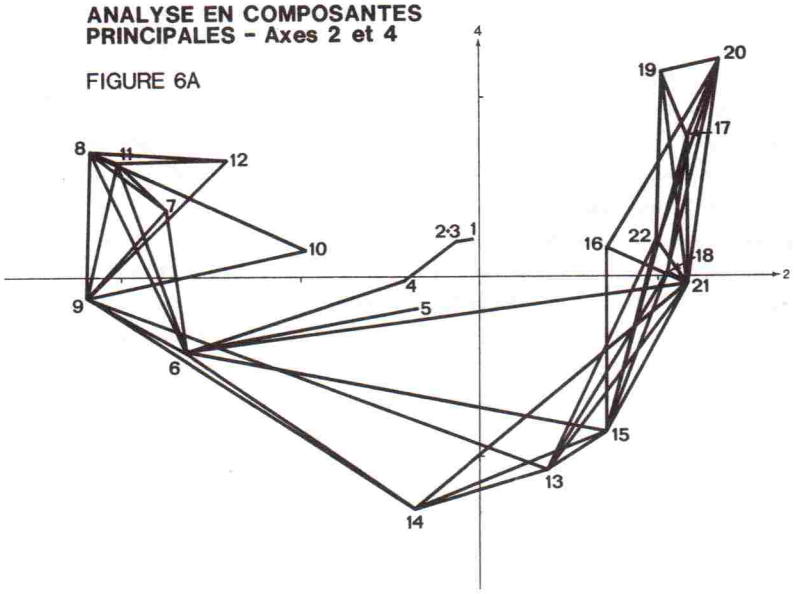
Par contre dans le groupe des sommets de gauche il existe une clique maximale à 5 sommets : 6 7 8 9 11. On voit donc que le graphique visualise bien l'opposition entre une clique plus ample mais unique à gauche et une "tresse" de cliques nombreuses mais de moins grand nombre de sommets. Ceci n'apparait pas dans le graphique 6B dont ce n'était d'ailleurs pas le but puisqu'il n'est donné que comme référence.

Philippe Cibois
Laboratoire d'Informatique
pour les Sciences de l'Homme
CNRS Paris

ETUDE D'UN GRAPHE

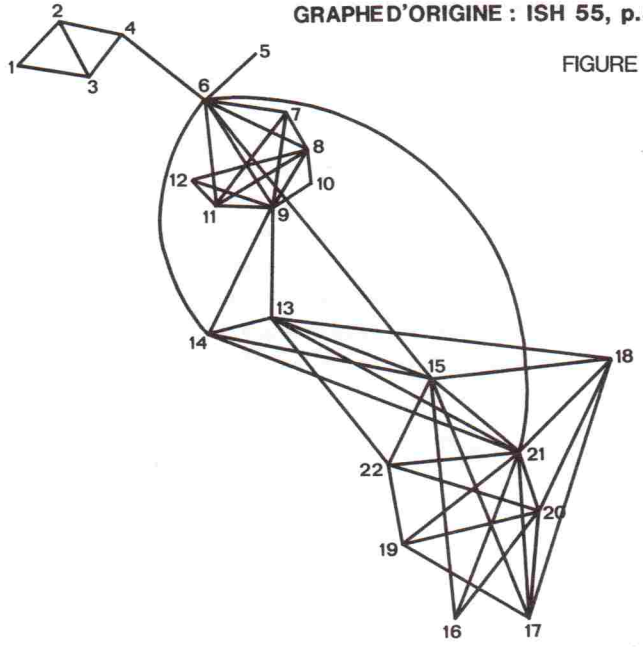
ANALYSE EN COMPOSANTES PRINCIPALES - Axes 2 et 4

FIGURE 6A



GRAPHE D'ORIGINE : ISH 55, p.56

FIGURE 6B



Annexe 1 : formalisation de la représentation graphique d'un tableau croisé

Soit un tableau de contingence dont les lignes sont indicées par i et les colonnes par j : on adoptera les notations suivantes :

n_{ij} : effectif d'une case quelconque

n_i : total marginal d'une ligne $n_i = \Sigma_j n_{ij}$

n_j : total marginal d'une colonne $n_j = \Sigma_i n_{ij}$

n : total général $n = \Sigma_i \Sigma_j n_{ij}$

A ces effectifs on associe les fréquences correspondantes

$$f_{ij} = \frac{n_{ij}}{n} \quad \text{avec} \quad \Sigma_i \Sigma_j f_{ij} = 1$$

$$f_i = \frac{n_i}{n} \quad \text{avec} \quad \Sigma_i f_i = 1$$

$$f_j = \frac{n_j}{n} \quad \text{avec} \quad \Sigma_j f_j = 1$$

On définit les fréquences conditionnelles suivantes :

$$f_j^i = \frac{f_{ij}}{f_i} = \frac{n_{ij}}{n_i}$$

$$f_i^j = \frac{f_{ij}}{f_j} = \frac{n_{ij}}{n_j}$$

avec les relations $\Sigma_j f_j^i = 1$ et $\Sigma_i f_i^j = 1$

Ces fréquences conditionnelles sont à un facteur 100 près les pourcentages en ligne (f_j^i c'est à dire f_j si i) et en colonne (f_i^j).

On définit l'écart à la moyenne (en ligne ou en colonne) comme la différence entre la fréquence conditionnelle pour une ligne donnée (resp. une colonne) et la fréquence marginale toutes lignes confondues (resp. colonnes). Cette différence est conditionnelle pour une ligne donnée (resp. une colonne). On pose :

$$e_j^i = f_j^i - f_j \quad \text{et} \quad e_i^j = f_i^j - f_i$$

On vérifie que $\Sigma_j e_j^i = \Sigma_i e_i^j = 0$ car $\Sigma_j e_j^i = \Sigma_j \frac{f_{ij}}{f_i} - \Sigma_j f_j = 1 - 1$

et il en est de même pour $\Sigma_i e_i^j$

Pour construire une représentation en surface on utilise la différence entre la proportion en ligne et la proportion toutes lignes confondues : c'est donc e_j^i qui correspond à la hauteur de chaque rectangle.

En ce qui concerne la base du rectangle, elle est prise proportionnelle à la fréquence de chaque ligne c'est à dire f_i.

La surface du rectangle est donc proportionnelle au produit $f_i\ e_j^i$. On peut vérifier que ce produit représente la proportion d'individus qui pour une case donnée sont en écart à l'indépendance, en effet :

$$f_i\ e_j^i = f_i\ (f_j^i - f_j) = f_i\ f_j^i - f_i\ f_j$$

or $$f_i\ f_j^i = f_i \frac{f_{ij}}{f_i} = f_{ij}$$

donc $f_i\ e_j^i = f_{ij} - f_i\ f_j$: cette dernière proportion d'individus en écart à l'indépendance sera notée e_{ij}

Pour un profil donné qui représente une colonne on a $\sum_i e_{ij} = 0$

en effet $$\sum_i e_{ij} = \sum_i f_{ij} - \sum_i f_i\ f_j = f_j - f_j$$

il en est de même pour $\sum_j e_{ij}$ qui représente cette fois la superposition des rectangles de même largeur mais qui appartiennent à différents profils.

Si l'on veut représenter en surface non plus les écarts à l'indépendance mais la contribution au Khi-deux de chacune des cases il suffit de modifier la base de chaque rectangle.

En effet l'écart à l'indépendance e_{ij} a été exprimé en construisant des profils de colonnes : il est possible de les construire en ligne et on aurait dans ce cas :

$e_{ij} = f_j\ e_i^j$: on peut donc exprimer les différences à la moyenne en fonction de l'écart à l'indépendance. On a :

$e_j^i = \frac{e_{ij}}{f_i}$ et $e_i^j = \frac{e_{ij}}{f_j}$; le produit des écarts à la moyenne est égal à la contribution au Phi-deux de chaque case (et est donc proportionnel à la contribution au Khi-deux). En effet $e_j^i\ e_i^j = \frac{(e_{ij})^2}{f_i\ f_j}$ c'est à dire le carré des écarts à l'indépendance divisé par le produit des fréquences marginales. Il suffit donc de remplacer f_i par e_i^j.

Dans une représentation en surface des écarts à l'indépendance les bases de chaque rectangle sont identiques d'un profil à l'autre. Dans la représentation en surface des contributions au Khi-deux ce n'est plus le cas : si le rapport de l'écart à l'indépendance à l'effectif théorique est plus petit que 1, la base du rectangle subit une réduction et la contribution au Khi-deux est plus faible que l'effectif en écart à l'indépendance. Inversement si le rapport est plus grand que 1, c'est à dire si l'écart est plus grand que le théorique, la base du rectangle subit une amplification et la contribution au Khi-deux est plus forte que l'effectif en écart à l'indépendance. Il n'y a pas de différence entre les deux représentations quand l'écart est est égal à l'effectif théorique (cas fréquent quand l'effectif observé est nul).

Annexe 2 : pratique de la représentation graphique d'un tableau croisé

- On met en ligne l'origine, la variable de tri ou dite "explicative" ; en colonne la destination, l'institution, la variable dite "à expliquer".
- On calcule les pourcentages pour chaque ligne ainsi que les pourcentages des deux marges par rapport au total général.
- Pour chaque case on calcule la différence entre le pourcentage de la ligne dans une colonne donnée et le pourcentage de cette même colonne toutes lignes confondues . On vérifiera que pour une ligne donnée la somme des écarts positifs et négatifs est égale à zéro.
- Pour chaque profil on construit une ligne de 10 cm divisée selon les pourcentages de chaque ligne (arrondir au demi-milimètre).
- Affecter une ligne à chaque colonne . Sur la base de chaque ligne construire un rectange au-dessus ou en-dessous de la ligne de référence selon le signe de l'écart et de hauteur égale à cet écart (exprimer les écarts en mm et arrondir au demi mm).
- Faire l'opération pour chaque destination : on obtient autant de profils que de colonnes du tableau original.
- Pour compléter l'opération on peut exprimer l'échelle par une surface. (Sachant qu'un carré de 10 cm sur 10 cm correspond par construction à l'ensemble de la population, une règle de trois donnera la surface d'un carré correspondant à un nombre rond d'individus et une extraction de racine carrée donnera le côté correspondant).
- On peut également représenter l'importance de chaque colonne par un graphique en "tranches de gâteau"(utiliser un rapporteur en grades où 200 gr correspondent à 100%).