MÉTHODES D'ANALYSE MULTIVARIEE DANS LA GÉOGRAPHIE ANGLO-SAXONNE, ÉVALUATION DES TECHNIQUES ET DES APPLICATIONS

Marie-France Ciceri-Marchand

Marie-France CICERI

MÉTHODES D'ANALYSE MULTIVARIEE DANS LA GÉOGRAPHIE ANGLO-SAXONNE

ÉVALUATION DES TECHNIQUES ET DES APPLICATIONS

N° d'ordre 73225 89 K

THESE

Présentée

A L'UNIVERSITÉ DE PARIS 1 PANTHÉON-SORBONNE

pour obtenir un

DOCTORAT DE 3e CYCLE

Par

Marie-France CICERI

MÉTHODES D'ANALYSE MULTIVARIEE DANS LA GÉOGRAPHIE ANGLO-SAXONNE

ÉVALUATION DES TECHNIQUES ET DES APPLICATIONS

Soutenue le 25 Novembre 1974 devant la commission d'examen

MM. Ph. PINCHEMEL, Président
M. BARBUT
J.-P. TRYSTRAM

Mention Très Honorable avec les félicitations du jury.

Marie-France Cicéri-Marchand, 1942-2009
géographe (Master of Science, Penn State University ; thèse de III° cycle, Paris-1), a travaillé en aménagement au Vénézuéla, comme urbaniste à Toronto, puis au laboratoire de Mathématiques du CEA, enfin à France-Telecom (modèles de tarification). Elle a ensuite suivi, jusqu'à sa mort, des études de philosophie et préparé un Master sur Spinoza.

Autre publication : Cicéri M-F, B Marchand & S Rimbert (2012) *Introduction à l'analyse de l'espace*, Armand Colin, Coll U,

Table des Matières

TROISIEME PARTIE : LES DIMENSIONS SOCIO-ECONOMIQUES DE L'ETAT DE PENNSYLVANIE : UN EXEMPLE DE REGIONALISATION

Chapitre

Liste des Tables

PREMIERE PARTIE

DEUXIEME PARTIE

TROISIEME PARTIE

Liste des Figures

PREMIERE PARTIE

DEUXIEME PARTIE

TROISIEME PARTIE

PRÉFACE

Ce travail a été effectué à l'Unité d'Enseignement et de Recherche de Géographie, sous la direction de Monsieur le Professeur PINCHEMEL. Je tiens à lui exprimer toute ma reconnaissance pour m'avoir guidée dans le choix et l'orientation de cette étude.

Qu'il me soit permis également, d'exprimer à Monsieur le Professeur TRYSTRAM ma gratitude pour la bienveillante attention qu'il m'a témoignée au cours de mes recherches.

Je prie Monsieur le Professeur BARBUT d'agréer mes remerciements pour l'honneur qu'il m'a fait en acceptant de faire partie du jury et d'examiner mon travail.

Mes remerciements s'adressent également :

- à Peter GOULD, Professeur à l'Université d'État de Pennsylvanie, dont les cours excellents ont fourni les bases statistiques nécessaires à l'élaboration de cet ouvrage.

- à Claude DENIAU, Maître-Assistant à l'Université René Descartes, pour les conseils qu'il a aimablement accepté de me dispenser sur le plan des mathématiques.

- à Bernard MARCHAND, Maître-Assistant à l'Université de Paris I, pour m'avoir incitée à suivre la voie passionnante de la géographie quantitative.

INTRODUCTION

Un grand nombre de géographes anglo-saxons ont eu recours depuis une vingtaine d'années à la formulation mathématique pour inventorier, classer, décrire, comparer et "modéliser" divers aspects de leur champ d'études. Le but scientifique de cette investigation est de repérer et de comprendre les forces responsables de "la différenciation et de l'organisation de la surface de la terre", pour reprendre la définition que donne Philippe PINCHEMEL (1968) de l'objet d'étude du géographe.

Deux ouvrages, l'un brillant, provocateur, *Theoretical Geography* (BUNGE, 1962), l'autre, clair et convaincant *Locational Analysis* (HAGGETT, 1965), ont représenté de façon remarquable l'esprit et la lettre de cette "nouvelle géographie". Nouvelle effectivement dans sa philosophie faisant une part essentielle à. la formulation logique de théories générales et de modèles; nouvelle dans ses techniques empruntant leur langage aux statistiques et aux mathématiques. Depuis, en dehors d'ouvrages plus spécifiques sur telle ou telle branche, d'autres ouvrages généraux tels que *Quantitative Ceography* (GARRISON and MARBLE, 1967) et *Models in Ceography* (CHORLEY and HAGGETT, 1967), ont montré le développement florissant de cette « révolution quantitative » Dernièrement un manuel pédagogique, *Spatial Organization* (ABLER et al., 1971),destiné aux étudiants du ler cycle prouve que la "révolution" est achevée puisque les idées et les méthodes qui la caractérisent sont devenues parties intégrantes de l'enseignement conventionnel.

Le lecteur pourra se reporter à ces ouvrages pour avoir une indication de l'étendue et de la variété de ces méthodes, que 1a géographie quantitative emprunte à la physique, à la géométrie, à l'économétrie, au calcul des probabilités, aux statistiques, etc. Parmi celles-ci, l'analyse multivariée occupe une place importante. Elle fait partie de l'ensemble des méthodes statistiques descriptives et inférentielles, traitant des relations entre les phénomènes et, particulièrement, des relations spatiales pour ce qui concerne la géographie.

Les méthodes de régression et de corrélation linéaire simple ont depuis longtemps prouvé qu'elles pouvaient être fort utiles aux géographes, par des travaux comme ceux de STRAHLER (1954) en géomorphologie. Décrivant en termes d'un modèle linéaire les associations entre deux variables, elles peuvent montrer les principales divergences spatiales par rapport à ce modèle grâce à la cartographie des résidus. Lorsqu'on s'interroge sur l'origine de telles anomalies, cela conduit souvent à émettre l'hypothèse qu'elles pourraient venir de relations entre d'autres variables**.** L'analyse de régression multiple sert précisément à déterminer dans quelle mesure plusieurs variables indépendantes sont responsables de la variation d'une variable dépendante. De la même manière, l'analyse de corrélation multiple mesure le degré d'interrelation entre une variable et plusieurs autres. Cependant il est possible d'examiner la relation entre deux variables quand les autres variables sont dans le problème, en maintenant leur effet constant par la corrélation partielle. L'analyse de corrélation canonique est utilisée pour établir des relations entre des ensembles de variables. Les thèmes d'application de ces méthodes sont nombreux[1] : estimation de paramètres pour des modèles de localisation, de

[1] Il serait trop long de citer les études elles-mêmes dont on peut trouver un excellent échantillon dans *Methodological Developments since the Fifties* où GOULD (1969 c) fait le point de progrès méthodologiques de la géographie depuis les années 1950

migration, ou de diffusion ; estimation de surfaces de prix, de développement des réseaux de transport, de croissance de population, de valeur de la terre, etc.

Les analyses de dérive spatiale ou *trend surface analysis* qui font appel à une technique semblable, sont peut être encore plus géographiques dans la mesure où elles incorporent les coordonnées spatiales dans l'équation de régression. Ce type d'analyse a surtout eu la faveur des spécialistes de géographie physique, dans leurs tentatives de "filtrer" les principales composantes régionales et locales de la distribution d'un phénomène physique (faciès des sols, pénéplanation, distribution de la végétation, etc ...). En géographie humaine, elle a paru particulièrement indiquée pour étudier la progression des vagues de colonisation des terres.

Les problèmes de classification et de régionalisation ont été traités par les méthodes multivariées de la taxonomie numérique, "linkage analysis" ou "cluster analysis", faisant intervenir des critères de similarité comme la corrélation entre les observations destinées à être regroupées.

Bien qu'une utilisation étendue de ces méthodes continue à se poursuivre, elles comportent des dangers et des limitations qui, très tôt, leur ont fait préférer l'utilisation d'analyses multivariées du type de l'analyse factorielle. Par exemple, il existe bien peu de situations concrètes dans lesquelles on trouve des relations fonctionnelles claires entre le phénomène dont on recherche la cause et les variables indépendantes qui seraient, dans une plus ou moins grande mesure, responsables de ce phénomène. La présence d'intercorrélation entre les variables dites "indépendantes" et d'autocorrélation entre les observations, affectent les résultats sans qu'on puisse savoir exactement de quelle manière, et jusqu'à quel point. Bien plus, même si des relations fonctionnelles existent et si l'hypothèse d'indépendance entre les variables tient, il est presque impossible d'identifier toutes les variables qui peuvent servir à rendre compte d'un phénomène particulier. C'est pour les mêmes raisons - besoin d'une information plus étendue avec un plus grand nombre de critères, et d'une information plus claire avec l'indépendance de ces critères - que les méthodes de taxonomie numériques se sont adjoint aussi des méthodes factorielles et parfois se sont vues remplacées par elles.

Parmi toutes les méthodes d'analyse multivariée que nous venons sommairement d'énumérer, les méthodes d'analyse factorielle ont sans conteste joué un rôle primordial dans la recherche géographique anglo-saxonne depuis une quinzaine d'années. Et cela essentiellement pour trois raisons :

1 – par le nombre même des recherches et donc par la quantité d'information traitée,

2 – par la richesse des résultats,

3 – et enfin par l'abondance et la force des critiques qu'elles ont suscitées.

En effet, la violence des critiques que les méthodes factorielles subissent depuis quatre ou cinq ans est à la mesure de l'engouement qu'elles ont provoqué. Ces critiques sont souvent contradictoires. D'une part on reproche aux analyses factorielles d'aboutir à des truismes, à des généralités évidentes depuis longtemps reconnues par le bon sens commun ; et d'autre part on les accuse de proposer, sans une apparence d'exactitude conférée par le langage mathématique, des explications fausses, parce que ne tenant pas compte de l'"inquantifiable" (psychologie, mentalités, traditions).

De telles critiques viennent non seulement de chercheurs qui se sont toujours opposés à la géographie quantitative dans son ensemble et à toute mise en formule, quelque soit la

méthode utilisée :

> "I1 n'est que trop évident que l'infinité des données aux volumes continuellement changeants constituant une réalité régionale ne saurait se mettre en formules. La formule initiale est fausse par définition parce qu'elle exprime d'une façon erronée et incomplète une réalité complexe et continuellement changeante" .(P. GEORGES, 1968, p. 204)

mais elles viennent aussi de géographes qui, touchés et conquis par la "révolution quantitative", ont été rapidement déçus par les méthodes d'analyse multivariée en général et en particulier par les méthodes d'analyse factorielle.

La raison de cette déception est simple. Pour être utilisées, les méthodes factorielles ne requièrent pas de la part du chercheur la connaissance des mécanismes de leur fonctionnement, ni celle de la logique mathématique qui les met en place. Cela est vrai pour d'autres méthodes d'analyse multivariée, dans la mesure où il existe des programmes d'ordinateur prêts à l'emploi. Cependant, des méthodes progressives, moins puissantes, qui ne peuvent traiter à la fois qu'un petit nombre de variables sélectionnées avec soin, font davantage appel à la réflexion ; alors que, les méthodes d'analyse factorielle, du fait de leur capacité à traiter simultanément un grand nombre de données, encouragent d'une certaine manière l'ignorance.

Que ce soit par paresse ou par curiosité, de nombreux chercheurs ont utilisé de façon indiscriminée des quantités de variables directement disponibles à partir du recensement. Cette masse de données engloutie par la machine ressort automatiquement, sous une forme différente plus ou moins interprétable. Que se passe-t-il dans l'intervalle ? Que faut-il introduire au départ ? Quelles sont les précautions à prendre, les contraintes à observer ? Beaucoup l'ignorent et, convaincus après des essais infructueux de l'inutilité de la méthode, ne cherchent pas à le savoir. C'est de cette ignorance dont souffrent l'analyse factorielle et les méthodes équivalentes, et à travers elles toute la géographie quantitative en général.

Pour éviter que le discrédit, largement immérité, des méthodes factorielles n'encourage ceux qui regardent déjà avec méfiance l'analyse quantitative en géographie à lui tourner définitivement le dos, pour éviter que ceux qui abordent aujourd'hui résolument l'analyse quantitative ne subissent comme leurs prédécesseurs le même engouement aveugle, suivi de la même réaction de rejet, nous avons voulu que ce mémoire soit une contribution à la réhabilitation et à une meilleure compréhension des méthodes d'analyse factorielle. Le plan découle de ces considérations et comprend trois parties:

1 - La première partie est une présentation des techniques factorielles. La construction d'un modèle hypothétique, cadre conceptuel "idéal", à partir de théories bien établies, permet d'exposer la méthode dans son fonctionnement le plus simple, dépouillé de toutes, les contraintes qu'impose d'ordinaire la réalité. Il est montré ensuite comment l'adaptation à cette réalité se traduit par trois grandes alternatives, incorporant des options annexes,

2 – La deuxième partie est une évaluation critique des applications des méthodes d'analyse factorielle dans la géographie anglo-saxonne. Ces applications sont regroupées en trois grandes catégories. Les études thématiques se caractérisent par l'examen d'un phénomène particulier dont on cherche à déterminer les causes ; les études de régionalisation par la description et le regroupement des portions d'espace en zones intérieurement homogènes et hétérogènes entre elles ; et les études urbaines par la classification des villes et la recherche de structures urbaines fondamentales. I1 est inévitable qu'une telle partition contienne une part d'arbitraire, car ces trois catégories d'études font tour à tour appel, dans une plus ou moins

grande mesure, aux capacités descriptives classificatrices et explicatives d'un même ensemble de méthodes. Puisqu'il n'est pas possible de dresser un inventaire général des applications et de ménager un développement à chaque point, une sélection s'impose qui est délicate. Les exemples sont choisis en fonction de leur contribution au champ d'étude, de leur originalité et des problèmes de contraintes conceptuelles ou inhérentes à la méthode, qu'elles permettent de soulever.

3 – La troisième partie est un exemple d'exploitation des méthodes factorielles dans une étude de régionalisation de l'Etat de Pennsylvanie. Cet Etat recouvre en partie une région programme établie essentiellement d'après un indice de niveau moyen de revenu. Il nous a paru intéressant de rechercher, à un niveau plus fin d'analyse, des indices plus discriminants. L'évaluation de l'invariance de l'analyse sous différentes échelles de mesure est finalement examinée.

Première partie

LES TECHNIQUES FACTORIELLES

CHAPITRE I

Présentation de la méthode idéale de base

Le modèle d'analyse factorielle essaie de réduire un grand nombre de variables en un nombre plus petit de dimensions sous-jacentes indépendantes, appelées facteurs, qui sont responsables de la variation entre les variables. Ainsi, pour chaque étude empirique, le point de départ est une matrice de données de m variables par n observations qui, en recherches géographiques, sont en général des unités spatiales allant du *"census tract*[2]*" à* l'échelle nationale. Un excellent exemple pédagogique de prédiction spatiale, développé par KLOVAN (1968), illustre certaines relations apparemment obscures, entre la matrice des données brutes et les *"causes"* ou facteurs cachés. Le modèle hypothétique qui suit est une adaptation du modèle géologique de KLOVAN dans le domaine de la géographie urbaine.

L'une des techniques multivariées les plus couramment exploitées par les géographes anglo-saxons est une forme de modèle factoriel appelé écologie factorielle. Pour illustrer la méthode générale, nous allons procéder *à* l'écologie factorielle *"idéale"* d'une ville hypothétique ; mais avant de développer la partie analytique du modèle, nous devons présenter le contexte théorique duquel sont tirées certaines hypothèses. Les hypothèses que nous aurons à tester sont au nombre de trois, provenant chacune d'une des trois théories classiques sur l'usage du sol urbain.

[2] Les *census tracts* sont des découpages administratifs urbains constitués par le regroupement de plusieurs îlots d'après des critères d'homogénéité.

LES ORIGINES DE L'ECOLOGIE FACTORIELLE

L'Ecole d'Ecologie Urbaine de Chicago[3] culmine avec une théorie de la croissance et de la structure urbaines, connue sous le nom de « *théorie des zones concentriques »* (BURGESS, 1925). Cette théorie prétend que la ville, au fur et à mesure de sa croissance, s'organise selon une structure en zones concentriques dont voici, du centre vers la périphérie, les caractères distinctifs

- la zone 1, au centre, constitue le quartier des affaires, le C.B.D (*Central Business District)*,
- la zone 2 est une zone de transition où l'on trouve les taudis et les résidences des familles aux revenus les plus modestes,
- la zone 3 est occupée par les résidences des travailleurs
- la zone 4 est formée par la plupart des quartiers résidentiels de la classe moyenne,
- la zone 5 est une zone de commutation en dehors des limite de la ville.

Le processus dynamique qu'implique la théorie vient de ce que, lorsque la situation d'un individu s'améliore, celui-ci a tendance à s'éloigner du centre de la ville dans le but de trouver un logement meilleur et plus récent. La théorie de BURGESS reflète de façon très satisfaisante la structure de Chicago ; mais lorsqu'on veut généraliser son application, la structure simple en anneaux concentriques apparaît le plus souvent déformée par les caractéristiques physiques (collines, vallées, rivières, etc ...) et par le réseau routier.

La *"théorie des secteurs"* (HOYT, 1939), construite d'après un travail empirique sur des données relatives au logement, affirme que l'usage du sol urbain ne suit pas un schéma concentrique mais sectoriel. « La répartition spatiale des résidences urbaines est déterminée, selon HOYT, par le choix de ceux qui ont les moyens de payer les loyers les plus élevés, (BERRY, 1970, p. 307) » On montre que ce choix s'appuie sur la topographie et sur la structure radiale des routes. Les détracteurs du modèle d'HOYT l'accusent de ne pas parvenir à expliquer l'aspect dynamique du phénomène.

Enfin, la *"théorie des noyaux multiples"-,* développée par HARRIS et ULLMAN (1945), prétend que l'usage du sol s'organise autour de plusieurs noyaux, parce que certaines activités commerciales ou industrielles ont tendance à s'agglomérer pour tirer profit d'une économie d'échelle.

Une vaste controverse, touchant non seulement les géographes urbains mais aussi les planificateurs et les économistes, a tenté d'établir les mérites relatifs de ces trois théories devenues classiques. On considère généralement aujourd'hui que les structures spatiales apparemment incompatibles que ces théories ont dégagées, sont en fait complémentaires. Chacune d'elle peut être considérée comme une composante indépendante et additive de la structure socio-économique globale d'une ville.

---

[3] L'écologie des plantes a inspiré le travail des sociologues et des géographes de l'Université de Chicago qui, en appliquant des concepts et des techniques similaires à l'étude de la ville, ont créé l'écologie urbaine. Bien que la première étude d'écologie urbaine ait été publiée il y a presque 60 ans (PARK, 1916), c'est seulement en 1936, dans une étude où il est question « de divers groupes d'intérêt au sein d'une population urbaine ; de la domination d'un groupe dans une zone dite "fonctionnelle" ; d'invasion d'une zone par un groupe ; de compétition conduisant à la "succession" et à la domination de la zone par un nouveau groupe », que les analogies avec le monde biologique ont été établis d'une façon claire (PARK, 1936).

C'est une méthode de classification des census tracts, *"l'analyse des zones sociales"* (BELL et SHEVKY, 1955), proposée par des sociologues urbains, qui a conduit à l'intégration de ces trois théories. La méthode comprend l'élaboration de trois indices composites mesurant le statut socio-économique, le statut familial et le statut ethnique d'une population urbaine. La construction de ces indices a été mise au point au cours d'études à Los Angeles et à San Francisco (SHEVKY, WILIAM et BELL, 1949 ; BELL, 1953), à partir de six variables de recensement ayant trait à l'éducation, la profession, le loyer, la fertilité ou le travail féminin, le type de logement et l'origine ethnique.

En effectuant une analyse factorielle sur ces mêmes six caractéristiques, BELL a pu vérifier la validité de la structure typologique de *1"analyse des zones sociales",* à Los Angeles et à San Francisco (BELL, 1955). D'autres sociologues (VAN ARSDOL, CAMILLERI et SCHMID, 1958), ont étendu ce test à dix villes américaines et ont trouvé, sauf pour quatre d'entre elles, que les variables entrant dans la composition des indices pouvaient être effectivement réduites, par l'analyse factorielle, aux trois dimensions distinctes de statut économique, statut familial et statut ethnique. Les quatre exceptions vont conduire, par la suite, à la modification du modèle de SHEVKY et BELL, dans le sens de l'incorporation d'un ensemble de variables plus important.

Cependant, il restait à examine la structure spatiale associée à *1"analyse des zones sociales".* Une analyse de variance effectuée par ANDERSON et EGELAND (1961) dans quatre villes américaines de topographie uniforme, Akron, Dayton, Indianapolis et Syracuse, a montré que la répartition des trois indices de SHEVKY et BELL avait tendance à s'organiser de la façon suivante :

- les indices de statut socio-économique varient principalement selon un schéma sectoriel;
- les indices qui mesurent les caractéristiques familiales et l'âge de la population s'organisent en anneaux concentriques ;
- les indices qui isolent les minorités ethniques ont tendance à se regrouper en noyaux distincts dans certaines parties de la ville.

C'est ainsi que s'est faite, grâce à *1"analyse des zones sociales",* la réconciliation complète des trois théories de l'usage du sol urbain, soutenues respectivement par HOYT, BURGESS, ULLMAN et HARRIS.

Depuis, avec le concours d'un nombre plus grand de variables socio-économiques, de nombreuses analyses factorielles ont isolé, dans plusieurs villes différentes, des structures typologiques et des structures spatiales qui ont pu être comparées à celles de *1"'analyse des zones sociales"* ; et qui ont contribué à la généralisation, ou à la modification des théories classiques d'écologie urbaine. Le terme d'écologie factorielle est communément réservé à cette application étendue de l'analyse factoriel à l'écologie urbaine.

Le nombre des écologies factorielles, commencées en 1961 avec celle de Toledo, Ohio, par ANDERSON et BEAN (1961) dépasse maintenant la quarantaine, dont presque la moitié a été publiée . Issus récemment de *"L'Ecole de Chicago"* deux manuels, *Geographic Perspectives on Urban Systems* (BERRY and HORTON, 1970, chap. 10), et *City Classification Handbook (*BERRY and SMITH, 1972, chap. 10), fournissent une classification exhaustive de ces études. Certaines d'entre elles sont examinées avec plus d'attention, comme l'Ecologie factorielle de Chicago dans le premier manuel (REES, 1970); ou bien elles servent de base à une analyse comparative comme les écologies factorielles de Toledo, Ohio (ANDERSON and BEAN,

1961), d'Helsinki (SWEETSER, 1965), du Caire (ABU-LUGHOD, 1966), de Cardiff et Swansea (HERBERT, 1968), de Toronto (MURDIE, 1969) et de Calcutta (BERRY and REES, 1969), dans le second manuel.

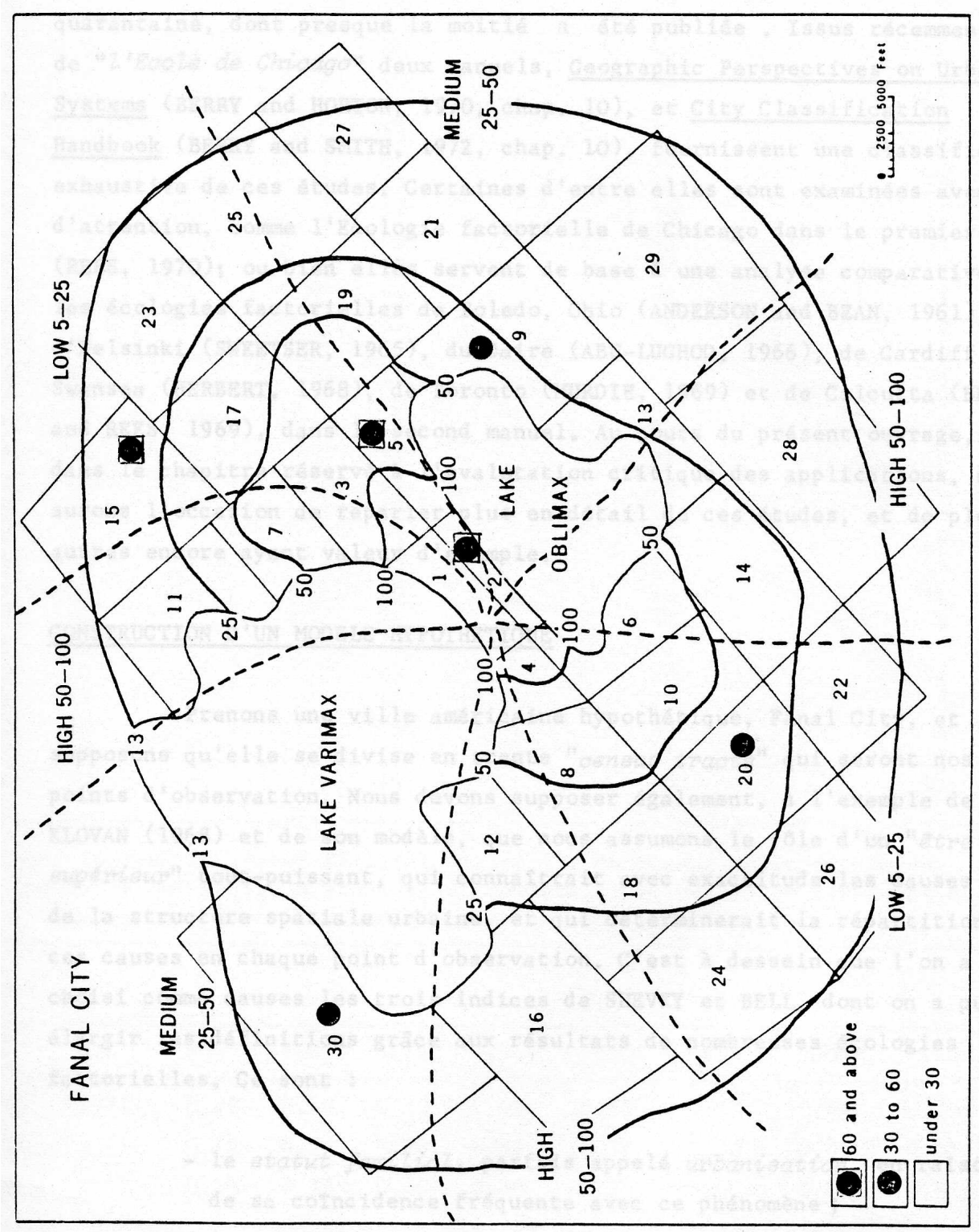


Figure 1 – Distribution du statut familial (contours), du statut socio-économique (secteurs délimités par des pointillés) et du statut ethnique et commercial (points).

Au cours du présent ouvrage, dans le chapitre réservé à l'évaluation critique des applications, nous aurons l'occasion de reparler plus en détail de ces études et de plusieurs autres encore ayant valeur d'exemple.

CONSTRUCTION D'UN MODELE HYPOTHETIQUE

Prenons une ville américaine hypothétique, Fanal City, et supposons qu'elle se divise en trente *"census tracts"* qui seront nos points d'observation. Nous devons supposer également, à l'exemple de KLOVAN (1968) et de son modèle, que nous assumons le rôle d'un *"être supérieur"* tout puissant, qui connaîtrait avec exactitude les causes de la structure spatiale urbaine, et qui déterminerait la répartition de ces causes en chaque point d'observation. C'est à dessein que l'on a choisi comme causes les trois indices de SHEVKY et BELL, dont on a pu élargir les définitions grâce aux résultats de nombreuses écologies factorielles. Ce sont

- le *statut familial,* parfois appelé *urbanisation,* en raison de sa coïncidence fréquente avec ce phénomène;
- 1e statut *socio-économique ou rang social;*
- le *statut ethnique et d'activité commerciale.*

Et pour rester dans la même ligne théorique, nous avons décidé que la répartition spatiale de ces indices devait suivre respectivement les structures en anneaux concentriques, en secteurs et en noyaux multiples (fig. 1 et table 1).

Si nous suivons à nouveau KLOVAN dans son rôle omnipotent, « nous devons préciser dans quelle mesure des variables socio-économiques communément utilisées reflètent ces trois causes » (KLOVAN, 1968, p. 46). Par exemple, les trois premières colonnes de la Table 2 indiquent que la variation dans : la structure par âge, la taille de la famille, la proportion de femmes dans la population active, est fortement liée, pour plus de 70 °/o, à la cause *urbanisation,* bien qu'une certaine partie de la variation soit due aux deux autres causes. Les colonnes 4, 5 et 6, illustrent le fait que les variables représentant respectivement le revenu, l'emploi et l'éducation dépendent du *rang social.* De la même manière, les quatre dernières variables : proportion de noirs, proportion de personnes d'origine étrangère, unités commerciales et surfaces en parkings, sont caractérisées principalement par la dernière cause.

Puisqu'il a été décidé que les trois causes devaient expliquer parfaitement toute la variation contenue dans les dix variables, le total de chaque colonne, dans la Table 2, doit être égal à 100 %. En effectuant simplement le produit de deux matrices.(Table 1 x Table 2) on peut créer une troisième matrice de 30 x 10 qui donne la valeur de chaque variable à chacun des trente points d'observation (Table 3). Cette matrice résultante représente les données non transformées, telles qu'elles sont utilisées par le géographe lorsqu'il commence une étude empirique. Par exemple, la première colonne représente les pourcentages de personnes âgées de moins de dix huit ans, tels qu'on aurait pu les obtenir à partir du recensement.

CAUSES	1	2	3
LOCALITY 1	142	53	6
LOCALITY 2	120	8	65
LOCALITY 3	80	54	5
LOCALITY 4	110	11	15
LOCALITY 5	77	9	70
LOCALITY 6	58	72	3
LOCALITY 7	55	75	7
LOCALITY 8	85	31	13
LOCALITY 9	27	38	43
LOCALITY 10	88	12	15
LOCALITY 11	18	78	2
LOCALITY 12	54	83	3
LOCALITY 13	22	70	5
LOCALITY 14	25	74	4
LOCALITY 15	16	9	73
LOCALITY 16	20	80	3
LOCALITY 17	49	10	17
LOCALITY 18	51	32	10
LOCALITY 19	53	35	13
LOCALITY 20	48	14	41
LOCALITY 21	21	39	8
LOCALITY 22	45	18	18
LOCALITY 23	20	16	12
LOCALITY 24	22	30	8
LOCALITY 25	17	33	7
LOCALITY 26	18	20	14
LOCALITY 27	14	37	8
LOCALITY 28	19	29	4
LOCALITY 29	17	40	9
LOCALITY 30	16	50	45

Table 1 - Distribution du statut familial (cause 1), du statut socio-économique (cause 2), et du statut ethnique et commercial (cause 3) en trente points d'observations dans la ville hypothétique de Fanalcity.

VARIABLES	AGE	FAM	WOW	INC	STAT	EDUC	RACE	FOR	COMM	PARK
CAUSE 1	0,85	0,75	0,80	0,10	0,10	0,05	0,15	0,10	0,45	0,10
CAUSE 2	0,15	0,15	0,10	0,75	0,70	0,65	0,30	0,30	0,05	0,05
CAUSE 3	0,00	0,10	0,10	0,15	0,20	0,30	0,55	0,60	0,50	0,85

Table 2 - Influence relative des causes sur la variation de dix variables socio-économique

Variables		AGE	FAM	WOW	INC	STAT	EDUC	RACE	FOR	COMM	PARK
LOCALITY	1	64,32	57,52	59,75	27,42	26,25	21,67	20,25	16,85	34,77	10,97
LOCALITY	2	51,60	48,85	51,65	13,87	15,30	15,35	28,07	26,70	43,45	33,82
LOCALITY	3	38,05	34,30	34,95	24,62	23,40	20,30	15,47	13,60	20,60	7,48
LOCALITY	4	47,57	42,82	45,30	10,75	10,85	8,57	14,02	11,65	28,77	12,15
LOCALITY	5	33,40	33,05	34,75	12,47	14,00	15,35	26,37	26,20	35,05	33,82
LOCALITY	6	30,05	27,30	26,95	30,12	28,40	25,30	15,97	14,60	15,60	5,97
LOCALITY	7	29,00	26,60	26,10	31,40	29,70	26,80	17,30	16,10	16,00	7,60
LOCALITY	8	38,45	34,85	36,20	16,85	16,40	14,15	14,60	12,80	23,15	10,55
LOCALITY	9	14,32	15,12	14,85	18,82	18,95	19,47	19,55	19,95	17,77	20,57
LOCALITY	10	38,30	34,65	36,55	10,02	10,10	8,35	12,52	10,70	23,85	11,07
LOCALITY	11	13,50	12,70	11,20	30,30	28,40	26,10	13,60	13,20	6,50	3,70
LOCALITY	12	29,17	26,62	25,90	34,05	32,05	28,77	17,32	16,05	14,97	6,05
LOCALITY	13	14,60	13,75	12,55	27,72	26,10	24,05	13,52	13,10	7,95	4,97
LOCALITY	14	16,17	15,12	13,90	29,30	27,55	25,27	14,07	13,55	8,47	4,80
LOCALITY	15	7,47	10,32	10,50	9,65	11,25	14,27	22,62	24,05	22,07	32,05
LOCALITY	16	14,50	13,65	12,15	31,22	29,30	26,95	14,32	13,90	7,25	4,27
LOCALITY	17	21,57	19,97	20,95	7,47	7,65	7,02	9,85	9,05	15,52	9,92
LOCALITY	18	24,07	22,02	22,50	15,30	14,75	13,17	11,37	10,35	14,77	7,60
LOCALITY	19	25,15	23,15	23,60	16,75	16,20	14,65	12,80	11,80	16,05	9,05
LOCALITY	20	21,45	21,10	21,95	10,72	11,40	11,90	16,97	16,80	21,40	20,17
LOCALITY	21	11,85	11,20	10,75	16,27	15,50	14,40	9,62	9,30	7,70	5,43
LOCALITY	22	20,47	19,12	19,80	10,35	10,35	9,67	11,02	10,35	15,07	10,35
LOCALITY	23	9,70	9,30	9,40	7,90	7,80	7,50	7,20	7,00	7,90	6,50
LOCALITY	24	11,60	10,90	10,70	12,95	12,40	11,50	8,35	8,00	7,70	5,25
LOCALITY	25	9,70	9,20	8,80	13,75	13,10	12,20	8,15	7,90	6,40	4,65
LOCALITY	26	9,15	8,95	8,90	9,45	9,30	9,05	8,20	8,10	8,05	7,35
LOCALITY	27	8,72	8,42	7,85	15,17	14,45	13,57	8,80	8,65	6,07	5,02
LOCALITY	28	10,25	9,50	9,25	12,12	11,50	10,50	6,87	6,50	6,00	3,37
LOCALITY	29	10,22	9,82	9,25	16,52	15,75	14,77	9,75	9,55	7,07	5,67
LOCALITY	30	10,55	12,00	11,15	22,92	22,80	23,40	21,07	21,80	16,10	2L,l

Table 3 – Matrice des données brutes

INTERPRETATION DES VARIABLES

AGE = structure par âge

FAM = taille de la famille

WOW = femmes dans la population active

INC = revenu

STAT = statut professionnel

EDUC = éducation

RACE = population non-blanche

FOR = étrangers

COMM = unités commerciales

PARK = parkings

En commençant cette étude, nous avons supposé que la structure sous-jacente de la matrice des données brutes était parfaitement connue, puisque c'est à partir de cette structure qu'ont été créées les données elles-mêmes. Cependant, nous devons maintenant demander au lecteur d'oublier comment les données ont été obtenues. Les géographes, aux prises avec une étude réelle, ne sont pas censés connaître les propriétés structurelles de l'ensemble des données. C'est précisément dans l'espoir de découvrir ces relations qu'ils utilisent des techniques multivariées du type de l'analyse factorielle.

		1 AGE	2 FAM	3 WOW	4 INC	5 STAT	6 EDUC	7 RACE	8 FOR	9 COMM	10 PARK
2	FAM	0,99									
3	WOW	0,99	0,99								
4	INC	0,14	0,12	0,07							
5	STAT	0,16	0,15	0,10	0,99						
6	EDUC	0,07	0,07	0,02	0,97	0,98					
7	RACE	0,50	0,55	0,55	0,24	0,32	0,41				
6	FOR	0,35	0,41	0,40	0,20	0,28	0,39	0,98			
9	COMM	0,83	0,87	0,88	- 0,13	- 0,07	- 0,06	0,78	0,70		
10	PARK	0,26	0,33	0,35	- 0,32	- 0,24	- 0,11	0,82	0,85	0,75	0

Table 4 – Matrice des corrélations entre variables

		FACTOR I	FACTOR II	FACTOR III
AGE	1	0,84	0 09	0,52
FAM	2	0,87	0,11	0,46
WOW	3	0,87	0,16	0,46
INC	4	0,20	- 0,97	0,11
STAT	5	0,27	- 0,96	0,06
EDUC	6	0,27	- 0,95	- 0,12
RACE	7	0,88	- 0,11	- 0,44
FOR	8	0,80	- 0,10	- 0,58
COMM	9	0,93	0,34	0,01
PARK	10	0,66	0,40	- 0 62

Table 5 - Matrice initiale des saturations

	Eigenvalue	Percent of variance explained
FACTOR I	5,192	51,92
FACTOR II	3,128	31,28
FFACTOR III	1,680	16,80

TOTAL	10,000	100,00

Table 6 - Valeurs propres de la matrice des corrélations

		Factor I	Factor II	Factor III
AGE	1	0,98	- 0,08	- 0,13
FAM	2	0,97	- 0,075	- 0,20
WOW	3	0,97	0,025	- 0,21
INC	4	0,06	- 0,99	0,05
STAT	5	0,07	- 0,99	- 0,02
EDUC	6	0,03	- 0,98	- 0,17
RACE	7	0,36	0,27	- 0,89
FOR	8	0,20	0,24	- 0,94
COMM	9	0,77	0,14	0,61
PARK	10	0,16	0,27	0,94

Table 7 – Matrice des saturations après rotation

LES ETAPES DE L'ANALYSE EN COMPOSANTES PRINCIPALES

Le modèle factoriel utilisé dans l'étude présente, connu sous le nom d'analyse en composantes principales, comprend les étapes suivantes :

- 1° : Organisation du tableau des données

L'analyse commence avec une matrice de données de *m* mesures sur *n* observations (Table 3).

- 2° : Calcul de la matrice des corrélations entre variables

A partir de la matrice des données, on calcule une matrice de m par m qui contient les coefficients de corrélation simple de chaque variable avec chacune des autres variables (Table 4). Nous n'examinerons pas en détail le calcul algébrique du coefficient de corrélation, puisqu'il existe à ce sujet de nombreux ouvrages en français comme en anglais, dont nous ne citerons, pour ceux de nos lecteurs dont la formation mathématique est élémentaire, que les plus accessibles :

- *Mathématique des sciences humaines* (BARBUT, 1968)
- *Statistical Analysis in Geography* (KING, 1969)
- *Spatial Organisation, the Geographer's View of the World* (ABLER et alii, 1971)

Géométriquement, on peut essayer d'imaginer nos dix variables comme autant de vecteurs dans un espace à trente dimensions. Après que les variables aient été standardisées, les vecteurs correspondants sont de longueur unité et ont pour origine la moyenne des variables. Sous cette forme, la projection de chaque vecteur sur chacun des autres, c'est-à-dire le cosinus de l'angle qu'ils forment entre eux, sont les coefficients de corrélation. Intuitivement, cela représente bien ce que chaque variable a en commun avec chacune des dix autres.

– 3° : Extraction des vecteurs propres et des valeurs propres

Les composantes principales sont calculées à partir de la matrice des corrélations. La première composante peut être représentée géométriquement comme l'axe principal d'un groupe de vecteurs. Un tel axe est appelé vecteur propre. En prenant un exemple simple, avec deux variables, on, peut montrer commente les vecteurs propres et les valeurs propres qui leur sont associées, se calculent. Soit une matrice de corrélation de 2 x 2 :

$$R = \begin{vmatrix} 1 & 0.5 \\ 0.5 & 1 \end{vmatrix}$$

et la représentation de ses vecteurs V_1 et V_2 (Fig. 2a). La résultante des deux vecteurs V_1 et V_2 est formée par un troisième vecteur C_1. La surface comprise entre 0, V_1, C_1 et V_2 est égale au déterminant de la matrice R, que l'on peut aussi calculer en soustrayant le produit des éléments de sa diagonale secondaire du produit des éléments de sa diagonale principale :

$$\text{Det. } R = (1.1) - (0{,}5\text{-}0{,}5) = 0{,}75$$

que l'on écrit aussi :

$$\text{Det. } R = \begin{vmatrix} 1 & 0.5 \\ 0.5 & 1 \end{vmatrix} = 0.75$$

Les valeurs propres λ doivent être telles que, retirées de la diagonale principale de R, elles réduisent à zéro le déterminant de R

$$\text{Det. } R = \begin{vmatrix} 1-\lambda & 0.5 \\ 0.5 & 1-\lambda \end{vmatrix} = 0$$

Intuitivement on voit bien que si la surface 0, V_1, C_1 et V_2 devient égale à zéro, cela signifie que les deux vecteurs V_1 et V_2 se sont fondus dans la résultante C_1. Maintenant, en termes algébriques, cela revient à résoudre les équations caractéristiques suivantes, de manière à retrouver leurs racines latentes :

$$\begin{aligned} (1-\lambda)\,(1-\lambda) - (0{,}5)\,(0{,}5) &= 0 \\ 1 - 2\lambda + \lambda^2 - 0{,}25 &= 0 \\ \lambda^2 - 2\lambda + 0{,}75 &= 0 \end{aligned}$$

La deuxième équation est de la forme $ax^2 + bx + c$, où

$$A = 1,\ b = -2,\ c = 0.75$$

Ainsi, les deux racines latentes ou *valeurs propres* sont :

$$\lambda 1_{\square} = 2 + \sqrt{(4-3)} = 3/2$$

$$\lambda 2_{\square} = 2 + \sqrt{(4-3)} = 1/2$$

De manière générale, les vecteurs propres $\underline{e}_i$ associés aux valeurs propres λ_i sont tels que :

$R\underline{e}_i = \lambda_i \underline{e}_i$ soit $(R - \lambda_i)\, e_i = 0$

où $i = 1, 2, \ldots. m$ et $\lambda_1 > \lambda_2 > \ldots. > \lambda_m > 0$

Dans notre exemple simple, les éléments e_{11} et e_{21} du vecteur propre $\underline{e}_1$ s'obtiennent donc en résolvant l'équation :

$$(R - \lambda_1 I) * \underline{e}_1 = 0$$

$$\begin{vmatrix} 1 & 0.5 \\ 0.5 & 1 \end{vmatrix} - \begin{vmatrix} 1.5 & 0 \\ 0 & 1.5 \end{vmatrix} * \begin{vmatrix} e_{11} \\ e_{21} \end{vmatrix} = 0$$

Soit

$$\begin{vmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{vmatrix} * \begin{vmatrix} e_{11} \\ e_{21} \end{vmatrix} = 0$$

et

$$\begin{vmatrix} -0.5\, e_{11} & 0.5\, e_{21} \\ 0.5\, e_{11} & -0.5\, e_{21} \end{vmatrix} = 0$$

D'où $e_{11} = 1$; $e_{21} = 1$, donc $\underline{e}_1 = (1 , 1)$

De la même manière, les éléments e_{12} et e_{22} du vecteur propre $\underline{e}_2$ se trouvent à partir de

$$(R - \lambda_2 I) * \underline{e}_2 = 0$$

$$\begin{vmatrix} 1 & 0.5 \\ 0.5 & 1 \end{vmatrix} - \begin{vmatrix} 0.5 & 0 \\ 0 & 0.5 \end{vmatrix} * \begin{vmatrix} e_{12} \\ e_{22} \end{vmatrix} = 0$$

soit

$$\begin{vmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{vmatrix} * \begin{vmatrix} e_{12} \\ e_{22} \end{vmatrix} = 0$$

et

$$\begin{vmatrix} 0.5\, e_{12} + 0.5\, e_{22} \\ 0.5\, e_{12} + 0.5\, e_{22} \end{vmatrix} = 0$$

D'où $e_{12} = 1$; $e_{22} = -1$ donc $\underline{e}_2 = (1 , -1)$

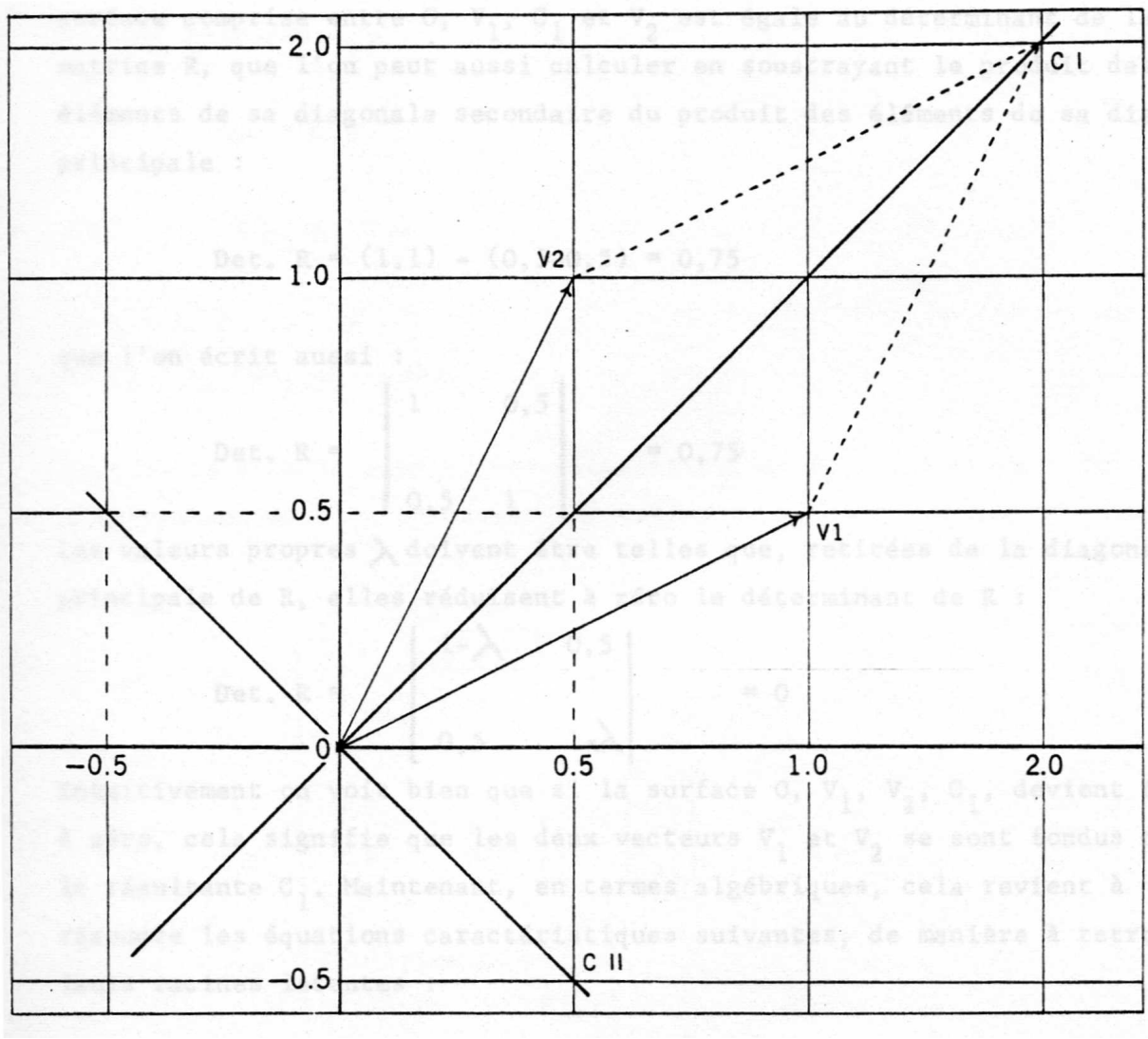


Figure 2a – Les vecteurs propres d'une matrice de corrélation 2 x 2.

Les axes engendrés par les vecteurs e_1 et e_2, vecteurs propres de R, associés aux valeurs propres λ_1 et λ_2 correspondent respectivement à la première composante principale C_I et à la deuxième composante principale C_{II} (Fig. 2a). On remarque qu'ils sont orthogonaux l'un par rapport à l'autre. Cela se passe comme si la variance des deux variables standardisées z_1 et z_2 avait été décomposée en deux parties indépendantes ; la plus grande partie est linéairement dépendante du premier vecteur propre, C_I, que l'on désigne aussi sous le nom de première composante principale. Comme nous venons de le voir, ces composantes ont la propriété intéressante d'être orthogonales, c'est-à-dire non corrélées, mais de longueur non définie.

- 4° : <u>Calcul de la matrice des saturations sur les composantes</u>

Les corrélations entre les variables et la composante sont connues sous le nom anglais de *"Loading"* ou saturations (Table 5).

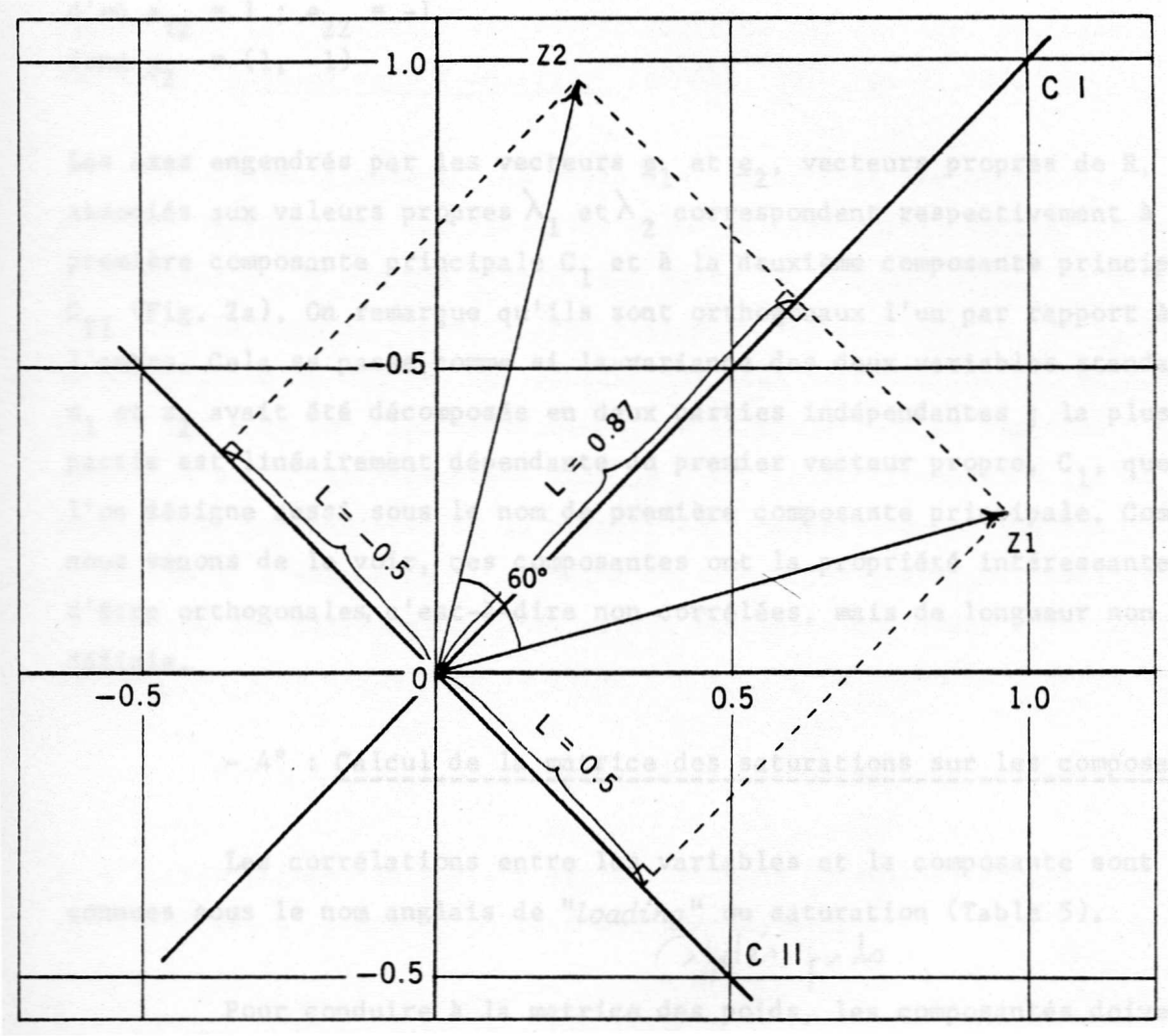


Figure 2b – Représentation géométrique des saturations

Pour conduire à la matrice des poids, les composantes doivent être normalisées et pondérées par la racine carrée des valeurs propres, ce qui s'écrit[4] :

$$L = M\sqrt{\Lambda} = M\Lambda^{1/2} \qquad (1)$$

où Λ est la matrice diagonale des valeurs propres et M une matrice composée des vecteurs propres que l'on a normalisés de manière à vérifier l'égalité

$$MM^t = I \qquad (2)$$

I étant la matrice identité.

Les lecteurs qui voudraient s'informer sur les procédures d'orthonormalisation dont la complexité dépasse le cadre de cet ouvrage, pourraient consulter *Applied Factor Analysis* (RUMMEL, 1970, p. 87, 88, 98).

[4] Dans un espace à trois dimensions, la dépendance linéaire se traduit par l'appartenance au même plan. La règle générale veut que, dans un espace à n dimensions, n vecteurs soient linéairement dépendants lorsqu'ils décrivent ensemble un espace de n-1 dimensions au plus.

Géométriquement, les poids sont égaux à la projection des vecteurs représentant les variables standardisées sur le vecteur propre ou composante. Un exemple simple peut aider la « visualisation » de cette relation. La figure 2b montre deux variables standardisées z_1 et z_2 dont la corrélation de 0,5 est représentée par deux vecteurs unités, séparés par un angle de 60°. Les saturations, ou la projection de ces deux variables sur l'axe bissecteur sont de 0,87. Leurs saturations sur le second axe, perpendiculaire au premier, sont respectivement égaux à 0,5 et -0,5. L'examen de la relation de la matrice des saturations et de la matrice des corrélations nous paraît susceptible d'aider à la compréhension de cette étape importante de l'analyse. La relation s'écrit

$$R = LL^t \qquad (3)$$

En d'autres termes, la matrice des saturations L, multipliée par sa transposée L^t donne la matricé des corrélations R :

r_{11} r_{12} ..... r_{1m}		l_{11} l_{12} ..... l_{1k}		l_{11} l_{21} ..... l_{m1}
r_{21} r_{22} ..... r_{2m}		l_{21} l_{22} ..... l_{2k}		l_{12} l_{22} ..... l_{m2}
.....	=	.....	*	.....
r_{m1} r_{m2} ..... r_{mm}		l_{m1} l_{m2} ..... l_{mk}		l_{1k} l_{2k} ..... l_{mk}

R = L * L^t

Les valeurs $l_{11}, l_{21}, \ldots, l_m$ sont les saturations ou corrélations de la variable 1, de la variable 2 et ainsi de suite jusqu'à la variable m sur la composante 1 ; et l_{12}, l_{22}, etc sont de façon identique les corrélations des variables 1 et 2 sur la composante 2.

Si l'on suppose, pour simplifier, que les variables 1 et 2 ont k composantes en commun, alors :

$$r_{12} = (r_1 C_1 \times r_2 C_1) + (r_1 C_2 \times r_2 C_2) + ... + (r_1 C_K \times r_2 C_K) \qquad (4)$$

où $r_1 C_1$ et $r_2 C_1$ sont les corrélations des variables 1 et 2 avec la composante C_1.

Ainsi, si l'on remplace dans les matrices L et L^t tous les éléments 1_{11}, 1_{12}, ... jusqu'à l_{mk}, par leurs équivalents $r_1 C_1$, $r_1 C_2$ .. .. jusqu'à $r_m c_k$, et si l'on procède à la multiplication, on remplit ainsi chaque cellule de la matrice R. Dans notre exemple de deux variables, on obtient effectivement en tenant compte des erreurs d'approximation :

$$\begin{vmatrix} 1 & 0.5 \\ 0.5 & 1 \end{vmatrix} = \begin{vmatrix} 0.87 & 0.5 \\ 0.87 & -0.5 \end{vmatrix} * \begin{vmatrix} 0.87 & 0.87 \\ 0.5 & -0.5 \end{vmatrix}$$

R = L * L^t

Nous venons de voir comment il est possible de revenir à la matrice initiale des corrélations, à partir de la matrice des saturations. Cependant, il ne faut pas oublier que l'extraction des composantes suit la démarche inverse. Ayant comme point de départ la matrice des corrélations des variables entre elles, nous cherchons à obtenir celle des corrélations des variables avec les composantes.

- 5° : Recherche de la variance expliquée

De l'équation (1), nous pouvons déduire que la somme des saturations au carré est égale, sur une composante donnée, à la valeur propre associée à cette composante :

$$_{km}L^{t}L_{mk} = {}_{k}\Lambda_{k} \qquad (5)$$

Exprimée en pourcentages de la variance totale, elle indique la proportion de variation dans les données initiales dont la composante rend compte.

Grâce à l'ordinateur, on peut calculer les valeurs propres par un procédé itératif de manière beaucoup plus simple que par le procédé manuel de résolution d'équations dont la difficulté, au-delà de trois équations, devient d'ailleurs très vite prohibitive. Les valeurs propres sont extraites dans un ordre de grandeur décroissant, si bien que la première explique la plus grande portion de la variance totale. Puis, la première composante étant retirée de la matrice des corrélations, on obtient une matrice résiduelle contenant le reste de la variance. La deuxième composante est extraite de la même façon que la première. Puis une troisième et, si nécessaire, jusqu'à m composantes sont extraites, afin que toute la variance contenue dans la matrice de corrélation soit expliquée.

Dans notre modèle hypothétique, trois composantes suffisent à rendre compte de toute la variance. On le constate du fait que la somme des trois valeurs propres, associées aux trois composantes, est égale au nombre de variables soit 10 (Table 6).

Le plus souvent, l'étape 5 constitue l'étape finale de l'analyse en composantes principales proprement dite ; car, à ce niveau, une interprétation plus poussée des résultats, c'est à dire l'identification des composantes avec des concepts précis, serait une entreprise hasardeuse. On peut dire seulement que, puisqu'elle est corrélée de façon significative avec la plupart des variables (Table 5), la première composante donne la direction positive générale autour de laquelle s'organise 1e nuage des données ; la seconde tend à être bipolaire, c'est à dire qu'environ la moitié des variables sont corrélées avec elle de façon positive et l'autre moitié de façon négative ; il en est de même pour les autres composantes.

Ainsi, toutes les composantes sont-elles très difficiles à interpréter en termes d'associations avec des groupes particuliers de variables. Cependant, la désignation de ces groupes distincts de variables peut se faire, s'ils existent, grâce à des techniques de rotation d'axes.

L' INTERPRETATION ET LA CARTOGRAPHIE DES FACTEURS

C'est donc dans le but essentiel de favoriser l'interprétation de la structure des variables que l'on fait subir une rotation aux composantes ; elles prennent alors le nom de facteurs[5]. Dans le cas présent, cette rotation suit le critère de *"simple structure",* développé par THURSTONE (1954) selon lequel les facteurs doivent être corrélés le plus étroitement possible avec un petit nombre de variables, et peu ou pas du tout avec les autres, KAISER (1958) a rendu ce critère opérationnel en créant la *"rotation varimax"* qui consiste à rechercher la variance maximum -d'où son nom- dans chacune des colonnes de saturations. Par exemple, si l'on compare la matrice des saturations sur les composantes avant la rotation, et celle des saturations sur les facteurs (Table 5 et Table 7), on voit clairement que la composante I comprend beaucoup de fortes corrélations, sorte de moyenne de la plupart des variables. Après la rotation, il n'y a plus que trois variables qui soient étroitement associées au facteur I. Ce sont : la structure par âge, la taille de la famille et la proportion de femmes dans la population active. De la même manière, le facteur II est composé principalement des variables 4, 5 et 6 alors que le facteur III est construit surtout à partir des variables 7, 8, 9 et 10. La matrice après rotation (Table 7), correspond en gros à la table 2 puisque l'une comme l'autre représente la relation entre les facteurs ou *"causes"* et les indices socio-économiques.

C'est à ce niveau que l'interprétation des facteurs intervient et cela demande certainement une compréhension profonde de la signification des variables d'origine. Les trois premières sont des mesures de ce qu'on peut appeler le statut familial. Les trois suivantes soulignent l'importance du revenu et de l'éducation que l'on considère comme faisant partie d'un facteur de statut social. Le troisième et dernier groupe forme le facteur de statut ethnique et commercial.

Cependant, la cartographie même des facteurs peut permettre une meilleure identification si, par exemple, leur organisation spatiale coïncide avec un phénomène bien connu du chercheur. Or, pour établir une carte de chacun de nos trois facteurs, il est nécessaire auparavant d'en calculer le poids local en chacun des trente points d'observation.

On reprendra le terme anglais de *"factor scores" ou "scores"* pour désigner ces *poids locaux*, dont l'avantage principal est de permettre la représentation spatiale des facteurs.

Le calcul des scores est un problème dont la complexité varie selon qu'on a affaire aux composantes principales prises dans leur totalité ou à un nombre de composantes inférieur au nombre de variables, aux facteurs après rotation, ou aux résultats d'une analyse en facteurs communs.

Etant donné que les explications à ce sujet sont souvent très peu claires (KING, 1969), en raison principalement de notations schématiques et qui varient beaucoup d'un texte à l'autre, nous allons examiner ce problème plus en détail.

---

[5] I1 existe une controverse sur l'emploi respectif des termes de *facteurs* et de *composantes.* Au sens large, facteur est plus communément employé lorsqu'on a imposé des contraintes pour limiter le nombre de facteurs considérés ; soit en jouant sur les *communautés*1 ou sur la rotation. I1 existe différentes techniques pour calculer les communautés , de même qu'il existe différents types de rotation. Le nombre et la signification des facteurs obtenus dépend étroitement de la technique utilisée et doivent donc être soutenus par une théorie. Le terme de composante sera réservé aux résultats de l'analyse en composantes principales, qui fournit une solution unique.

Le calcul des scores pour les composantes principales prises dans leur ensemble est un problème simple puisqu'on peut écrire :

$$ {}_nZ_m = {}_nS_m * {}_mL_m \qquad (6) $$

où Z est la matrice de n observations sur m variables standardisées ; L est la matrice complète des corrélations de m variables avec m composantes ; et S est la matrice des scores de chacune des n observations sur les m composantes. Si L est une matrice carrée, c'est à dire si le nombre des composantes est égal au nombre de variables m, elle peut avoir une inverse, d'où :

$$ {}_nS_m = {}_nZ_m * {}_mL_m^{-1} \qquad (7) $$

Ainsi, la matrice des scores S est la transformée, par une application linéaire, de la matrice des données d'origine.

Mais les formules (6) et (7) sont peu utiles car elles supposent que toutes les m composantes ont été calculées. Or, il est plus vraisemblable que seulement un petit nombre de composantes, $K \leq m$, ait été retenu. Dans ce cas l'équation (6) s'écrit

$$ {}_nZ_m = {}_nS_k * {}_kL^t_m \qquad (8) $$

mais comme ${}_kL^t_m$, transposée de ${}_mL_k$, matrice des corrélations des m variable sur les k composantes, n'est pas carrée, elle ne peut pas être inversée.

Cependant, en multipliant chaque terme de l'équation précédente par L, on peut écrire

$$ {}_nZ_m * {}_mL_k = {}_nS_k * {}_kL^t_m * {}_mL_k \qquad (9) $$

d'où

$$ {}_nS_k = {}_nZ_m * {}_mL_k * ({}_kL^t_m * {}_mL_k)^{-1} \qquad (10) $$

L'effet de la multiplication par la matrice $(L^t * L)^{-1}$ consiste en la normalisation de tous les vecteurs représentant les scores, ce qui revient à tous les ramener à une longueur égale.

Puisque par ailleurs, en se rapportant à l'équation (5) nous savons que $(L^t * L) = \Lambda$ où Λ est égal à la matrice diagonale de k*k des valeurs propres, si l'on substitue dans l 'équation (10) $(L^t * L)$ par sa valeur dans l'équation (5), on obtient

$$ {}_nS_k = {}_nZ_m * {}_mL_k * {}_k\Lambda^{-1}{}_k \qquad (11) $$

cela suppose, nous le voyons, une division des scores par l'inverse des valeurs propres. Ainsi les scores sur la première composante subissent une réduction puisque λ_1 est la plus grande valeur propre, tandis que les scores sur les composantes suivantes sont relativement augmentés.

D'autre part, nous rappelons que le théorème fondamental de l'analyse en composantes principales s'écrit, d'après l'équation (3) :

$$ {}_mR_m = {}_mL_k * {}_kL^t_m $$

D'où il ressort que

$$ {}_kL^t_m = ({}_mR_m * {}_mL_k)^{-1} \qquad (12) $$

et si l'on remplace L^t dans l'équation (8) par sa nouvelle valeur :

$$ {}_nZ_m = {}_nS_k * {}_mR_m * {}_mL_k^{-1} \qquad (13) $$

d'où, lorsqu'on isole S :

$$ {}_nS_k = {}_nZ_m * {}_mR_m^{-1} * {}_mL_k \qquad (14) $$

Ainsi, les formules (10), (11) et (14), dont les deux dernières généralisées par KAISER (1962), sont trois manières différentes de calculer les scores sur les composantes principales, qu'elles soient prises en nombre égal à celui des variables ou en nombre inférieur. Cependant, lorsqu'on a effectué une rotation sur la structure orthogonale des composantes, seule l'équation (10) reste valable :

$$ {}_nS_k^{\sun} = {}_nZ_m * {}_mL_k^{-t\sun} * ({}_kL_m^{t\sun} * {}_mL_k^{-\sun})^{-1} \qquad (15) $$

où $S^{\sun}$ représente les scores de n observations sur les k facteurs après rotation., et où $L^{\sun}$ contient les corrélations des m variables sur ces k facteurs, qu'il ne faut pas confondre avec L, matrice des saturations sur les composantes avant rotation. De fait, comme la rotation des axes change les saturations des variables sur les facteurs, il n'y a plus de correspondance directe entre cette nouvelle matrice de saturations $L^{\sun}$ et la matrice diagonale Λ des valeurs propres initiales d'une part, et la matrice des corrélations R, d'autre part. Ainsi les équations (11) et (14) ne peuvent pas être utilisées. A cela s'ajoute le problème que les scores sur les facteurs, après rotation, n'ont plus la propriété de non-corrélation qui caractérise les scores sur les composantes principales

Pour revenir à l'exemple étudié, un score de valeur zéro indique une intensité moyenne du facteur ; +.1 représente un écart type au dessus de cette moyenne, -1 un écart type au dessous, etc. La matrice obtenue, appelée matrice des scores permet la cartographie de la cartographie de la distribution des trois nouvelles variables.

Un carte a été construite pour chaque facteur d'après la technique automatique SYMAP. Les scores sont répartis en cinq classes auxquelles correspondent, sur l'imprimante de l'ordinateur, cinq symboles différents. Plus les valeurs sont élevées, plus le symbole est foncé. Comme on devait s'y attendre, d'après la théorie classique à parti de laquelle a été construit cet exemple, la carte qui représente le premier facteur ou statut familial montre une répartition spatiale concentrique des scores, dont les valeurs décroissent du centre de la ville vers la périphérie (Fig. 3a).

La deuxième carte, établie d'après le deuxième facteur ou statut social et économique, a une organisation nettement sectorielle (Fig. 3b) alors que la troisième carte, image du troisième facteur ou statut ethnique et commercial, comprend des noyaux distincts (Fig. 3c). De légères distorsions sont dues à l'effet des deux lacs.

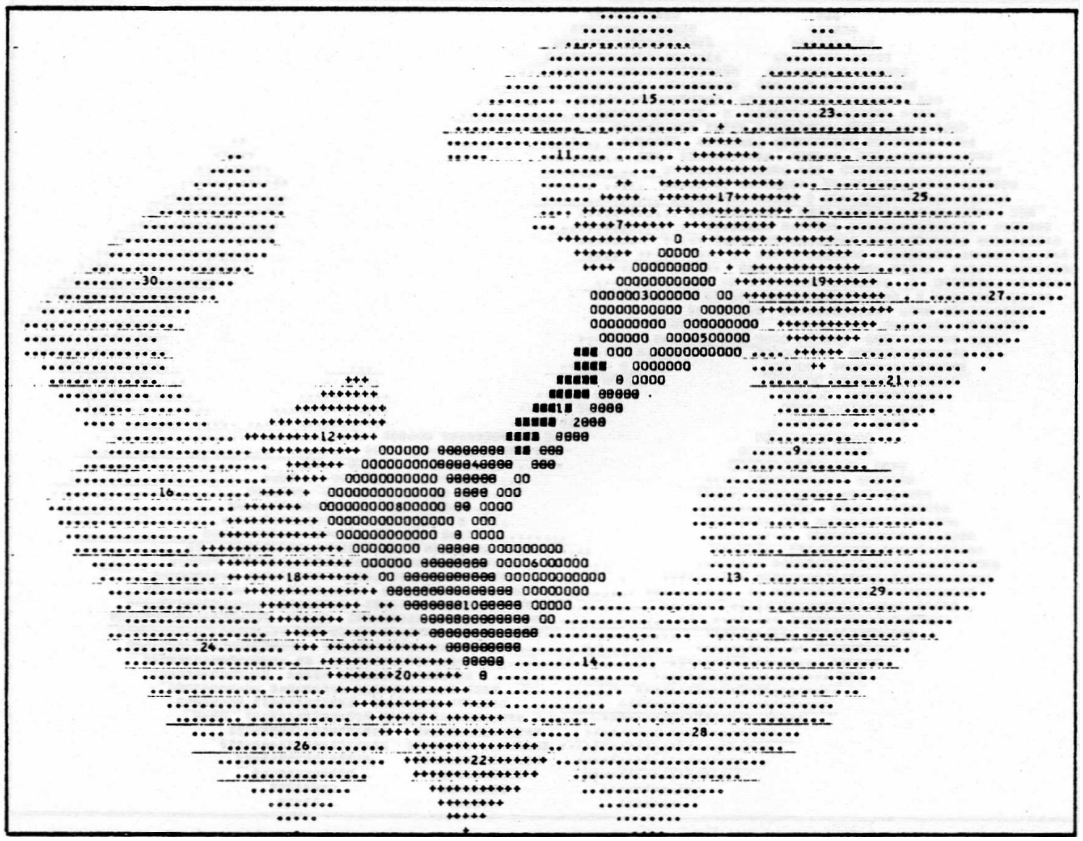


Figure 3a – Distribution spatiale du premier facteur

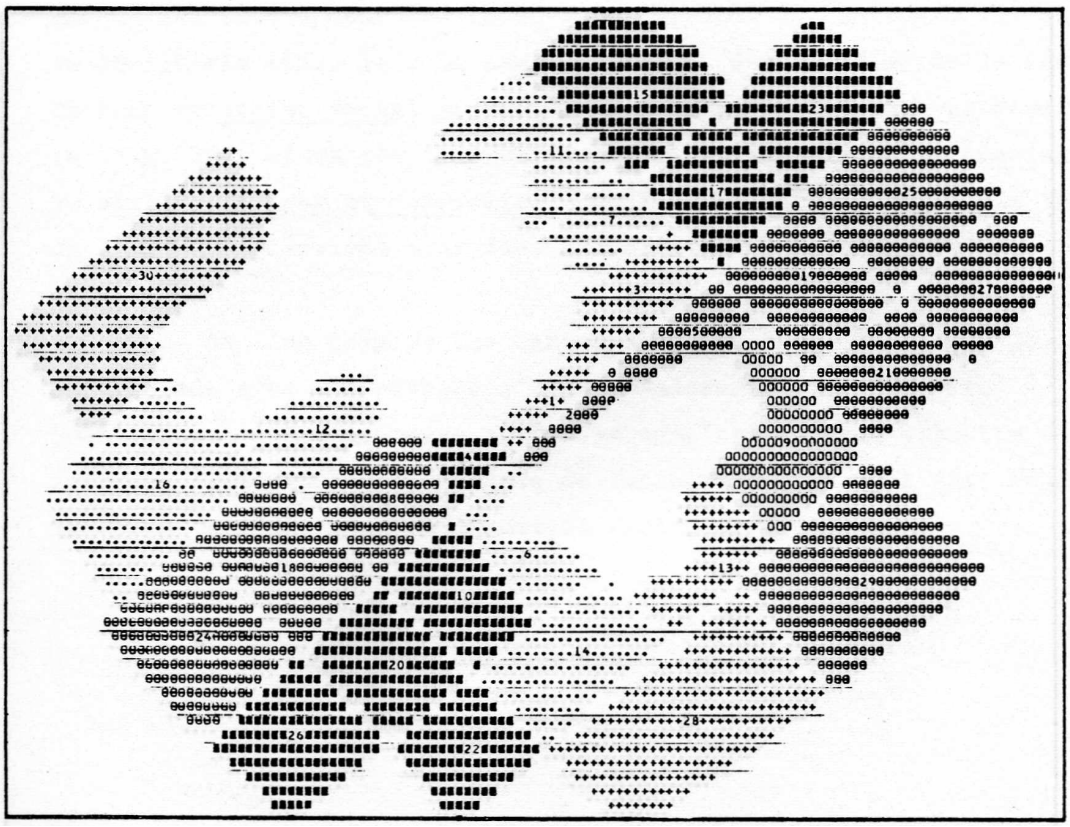


Figure 3b – Distribution spatiale du deuxième facteur

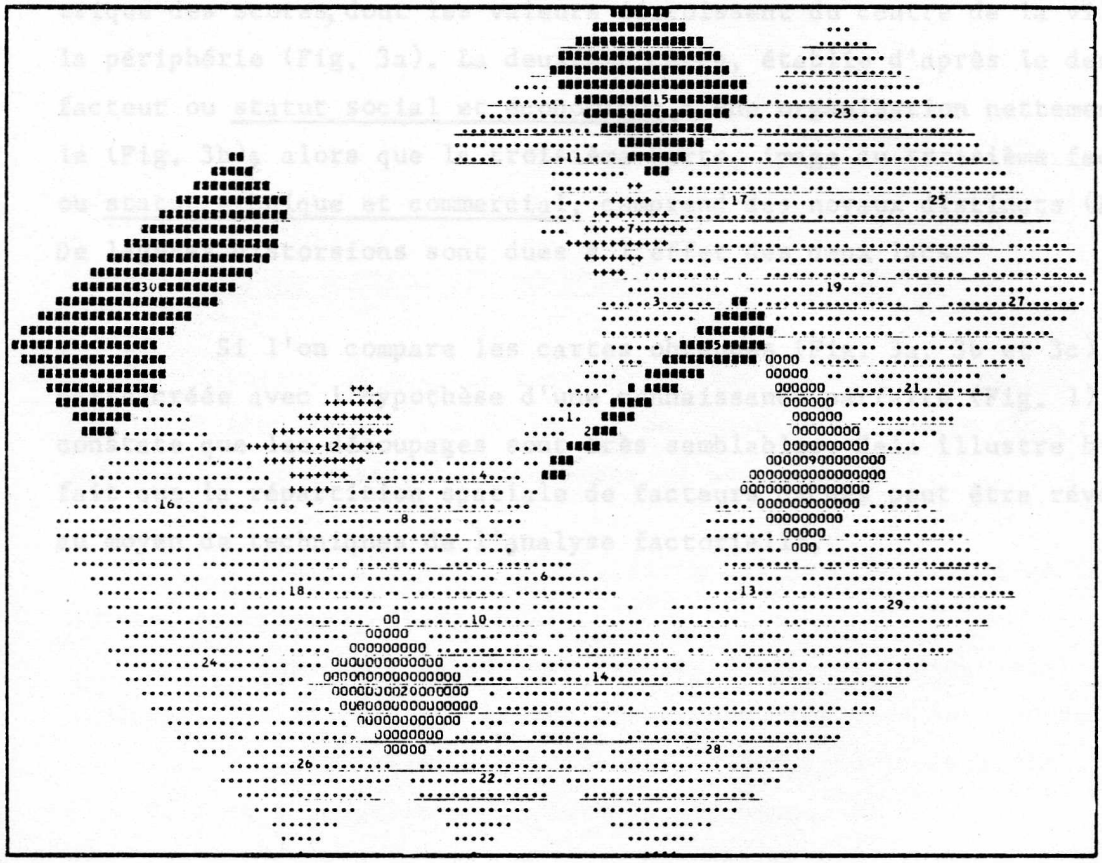


Figure 3c – Distribution spatiale du troisième facteur

Si l'on compare les cartes obtenues ((Fig. 3a, Fig. 3b et Fig. 3c) avec la carte créée avec l'hypothèse d'une connaissance parfaite (Fig. 1), on constate que les découpages sont très semblables. Cela illustre bien le fait que la répartition spatiale de facteurs cachés peut être révélée au moyen de techniques de l'analyse factorielle.

CHAPITRE II

Les grandes alternatives

L'ANALYSE DE LA MATRICE DES DONNEES ET DE SA TRANSPOSEE

Lorsqu'on cherche à énumérer tous les types de variations possibles étudiés par les sciences, on constate que trois dimensions suffisent pour décrire intégralement un phénomène :

- l'une définit les entités ou unités d'observations (individus, objets, organismes, lieux, etc ;

- une autre représente les attributs, ou indicateurs, qui caractérisent les entités et les distinguent les unes des autres;

- la troisième dimension est le temps ou l'occasion, c'est à dire la date à laquelle se produit le phénomène.

Cependant, l'analyse factorielle ne peut traiter que deux dimensions à la fois. Et, si l'on prend par paire toutes les combinaison possibles de ces trois dimensions, on obtient six manières différentes d'organiser et d'analyser un ensemble de cannées. Ce sont, d'après la terminologie proposée par CATTELL (1952), les analyses de type R et Q, P et 0, S et T.

L'analyse de type R consiste à factoriser les attributs d'après leurs mesures en différents points d'observation. C'est la technique factorielle classique, de beaucoup la plus fréquemment employée ; nous aurons l'occasion, au chapitre suivant, d'en examiner plusieurs exemples.

L'analyse de type Q, où les points d'observations sont factorisés d'après plusieurs attributs, est la transposée[6] de l'analyse de type R.

Ce type d'analyse est fourni par SCHUESSLER et DRIVER (1956) qui ont factorisé seize Sociétés primitives (entités traitées comme des variables), à travers 2500 caractéristiques. Les facteurs résultants regroupent les tribus selon leur ressemblance dans ces caractéristiques.

La technique P, où un seul objet d'observation est mesuré sur plusieurs caractéristiques à différentes dates est, après la technique R, le type d'analyse le plus couramment employé par les psychologues. Elle permet de déterminer le nombre de dimensions requises pour expliquer

---

[6] Le procédé de transposition consiste à intervertir les lignes et les colonnes d'une matrice. Dans le cas présent, l'analyse de type R travaille sur une certaine matrice de données D, et l'analyse de type Q travaille sur D^{t}., la transposée de D.

certaines modifications dans les caractéristiques d'un objet ou dans les comportements d'un individu. Sa transposée, la technique 0, n'a été, semble-t-il, utilisée que par les météorologues qui, pour établir certains modèles de prédiction, factorisent différents jours d'après des caractéristiques météorologiques relevées par une station donnée.

La technique T revient à factoriser différentes périodes, d'après leur similitude sur plusieurs entités, par une seule caractéristique. Pour reprendre l'exemple donné par RUMMEL (1970), on peut imaginer l'analyse d'un ensemble de données qui comprendrait le commerce total (caractéristique) de plusieurs nations, pour chaque année de la période 1900-1965. L'analyse de type T pourrait, durant cette période, révéler des dimensions qui représenteraient les impacts de la première et de la deuxième guerre mondiale sur le commerce.

La technique S, transposée de T, est la factorisation d'entités d'après une seule caractéristique, mesurée sur plusieurs périodes. Le résultat doit regrouper les entités qui ont évolué de façon similaire. Le seul exemple de ce type que l'on puisse citer est l'étude de JEFFREY, CASETTI et KING (1969), sur "les fluctuations économiques d'un ensemble multirégional". Les auteurs utilisent le taux de chômage mensuel de trente conurbations du "Midwest" des Etats-Unis, entre Mai 1960 et Septembre 1964, pour tester l'hypothèse que chaque série temporelle contient trois niveaux distincts qui correspondent respectivement : aux facteurs opérant à travers tout le système, aux facteurs communs aux villes d'un même groupe et au facteur spécifique de chaque ville. On peut déplorer que l'analyse de type S, qui pourrait être utile dans la recherche des mécanismes de diffusion d'un phénomène (information, innovation ou maladie), n'ait pas été exploitée dans ce domaine.

Théoriquement, on doit s'attendre à ce que les membres du couple formé par une technique et sa transposée, tendent vers la production de facteurs identiques, puisque les mesures fournies par la matrice des données sont les mêmes dans les deux cas. On peut très bien imaginer par exemple que, si des variables mesurées dans plusieurs villes se résument par la technique R en un premier facteur lié, disons, à la taille, une factorisation des mêmes villes sur les mêmes variables, dans une analyse de type Q, regroupe en un premier facteur les villes de plus grande taille. Mais il n'existe pas de réponse simple à la question de la relation entre les facteurs d'une matrice et ceux de sa transposée. Tout dépend des propriétés de la matrice des données brutes, ou du type de transformation subi par cette matrice. Voici rapidement les différents cas possibles :

1° - puisque le rang de D, la matrice des données, et de sa transposée, D^t, est le même, et que le rang détermine le nombre des facteurs, ont peut déjà affirmer que dans tous les cas le nombre de facteurs extraits à partir de D ou de D^t est le même.

2° - si D est symétrique, c'est à dire si $D = D^t$, on peut dire alors que F_D, la matrice des facteurs extraits à partir de D et F_{D^t} , la matrice des facteurs calculés à partir de D^t, sont identiques. Par exemple, on peut avoir affaire à une matrice carrée (n x n) où chaque élément représente des mesures d'interaction (migrations, mouvements de biens, interactions sociales), entre des paires d'observations (villes, régions, pays). Une telle matrice est une matrice de transaction, avec les lignes représentant les observations en tant qu'origines et les colonnes servant de destinations.

Pour KING (1969), une analyse factorielle de type R sur une telle matrice de transaction doit procéder de la manière suivante :

> "... la matrice des corrélations contient d'abord des mesures de similitude entre les vecteurs de destinations, quant à l'origine de ce qu'ils importent. L'analyse factorielle

identifie des types parmi ces destinations ; et les saturations montrent quelle destination particulière est plus typique de chaque sorte de groupe".

Dans l'approche de type Q, les lignes sont réduites à un ensemble de types d'origine ou de groupes exportateurs. Les saturations identifient les origines les plus importantes associées avec chaque groupe ; et les scores montrent à leur tour quelles destinations sont les plus importantes pour chaque groupe.

BERRY (1966) a utilisé ces deux approches dans l'analyse des flux de biens entre 36 *"blocs"* commerciaux en Inde. Parmi d'autres exemples de ce type, il faut noter l'étude de la structure commerciale au Moyen-Orient (Mc CONNELL, 1967), et l'identification des types de fermes dans l'agriculture des Barbades (HENSHALL et KING, 1966).

3° - si D n'est pas symétrique, comme c'est le cas pour la plupart des matrices de données, alors l'interprétation des facteurs de F_D sera différente de celle des facteurs de $F_D{}^t$. I1 faut noter cependant qu'il existe un moyen d'obtenir des résultats identiques avec des matrices D et D^t non symétriques. Cela consiste à effectuer un *"double centrage"* ou une *"double standardisation"*[7] des données, avant de procéder au calcul de leur corrélation. Mais cette double manipulation élimine la variation qui est associée aux différences entre les variables - ou entre les observations -, et ne laisse subsister que la variation qui existe à l'intérieur de chaque variable ou de chaque observation.

On voit donc que cette solution dépasse le cadre de notre problème, puisque les résultats obtenus sont forts différents de ceux produits par l'analyse classique[8], où ce sont précisément les variations entre les variables ou entre les observations qui sont prises en compte.

4° - que D soit symétrique ou non, on peut effectuer simultanément l'extraction des facteurs de la matrice des données et de sa transposée. Ce procédé, dit "d'analyse factorielle directe", consiste à appliquer directement l'analyse en composantes principales sur la matrice des données sans passer par l'intermédiaire d'une matrice de corrélation. Les données peuvent être exprimées en valeurs absolues ou bien en valeurs centrées ou standardisées, si l'on veut supprimer l'effet des différentes échelles de mesures. Parmi les applications, on trouve le traitement de matrices de "choix sociaux" (WRIGHT et EVITTS , 1963), et le traitement de matrices de données ordinales (BERRY, 1960-1961). Il est intéressant de souligner que BERRY et BARMUN (1962) ont utilisé l'analyse factorielle directe pour tester l'hypothèse que la "hiérarchie des places centrales peut être identifiée à celle des fonctions".

5° - enfin, que ce soit par factorisation directe de données standardisées, ou par l'intermédiaire de la matrice des corrélations, il convient de noter que l'effet de la standardisation est plus sensible dans l'analyse de type Q que dans l'analyse de type R. Dans

---

[7] Les variables centrées, x_c, s'obtiennent en soustrayant de chaque valeur de la colonne des variables x_i, la moyenne de la colonne x soit $x_c = xi - x$. Les variables standardisées z_i s'obtiennent en divisant chaque écart à la moyenne ou variable centrée x_c, par l' écart-type de sa distribution, c'est-à-dire :

$$z_i = \frac{xi - x}{\sigma_i} = \frac{Xc}{\sigma_i}$$

La variable standardisée i a une moyenne égale à zéro ($z_i = 0$), un écart-type égal à 1 ($\sigma_i = 1$). Le double centrage ou la double standardisation revient à centrer ou standardiser par colonnes (variables), puis par lignes (observations).

[8] Notons que dans l'analyse factorielle classique, le calcul de la matrice des corrélations comprend automatiquement la standardisation simple des colonnes de D ou de D^t .

l'analyse de type R, les variables sont des caractéristiques et les différences de magnitude, pour une entité à travers ces caractéristiques, n'ont en général pas grande importance. Ces différences n'ont guère de sens puisqu'elles résultent d'unités non comparables (population, âge, revenu, distances, etc ...). Dans l'analyse de type Q, les différences de mesures de caractéristiques à travers les entités sont une information importante. Les facteurs résultants peuvent réunir deux observations très différentes en magnitude (taille, développement économique, densité, etc ...) mais dont la structure de changement de magnitude, ou profil, à travers ces caractéristiques est similaire. Dans ce cas, l'information relative aux différences de magnitude peut être introduite dans l'analyse au moyen d'une mesure de magnitude, telle que le coefficient de corrélation intra-classe proposé par KENDAL et STUART (1961, vol. 2).

I1 convient ici de déplorer la méconnaissance par les géographes anglo-saxons d'une technique factorielle qui a fait depuis longtemps ses preuves en France. I1 s'agit de *L'analyse factorielle des correspondances*. On peut trouver tous les développements nécessaires à la compréhension et à l'utilisation de cette technique dans les ouvrages de C. DENIAU et L. LEBART (1969) et de J.P. BENZECRI (1973). On se contentera de souligner ici la différence essentielle qui la distingue du modèle factoriel tel que nous l'avons. présenté. Cette différence consiste dans l'utilisation, à la place de la métrique euclidienne classique donnée par

$$d^2 (x_1, x_2) = \Sigma^m_{j=1} (x_{1j} - x_{2j})^2$$

où x_1, x_2 sont des observations situées dans l'espace des variables de m dimensions ; de la métrique du Chi-deux (χ^2) telle que

$$d^2 (i, i') = \Sigma^m_{j=1} [(f_{ij} / f_i) - (f_{i'j} / f_{i'})] / f_j$$

où i et i' sont deux observations situées dans l'espace des variables de m dimensions ; f_{ij} étant la fréquence du phénomène j au point i, et f_i la fréquence agrégée de tous les phénomènes (j = 1 jusqu'à m) au point i, ce qui vaut aussi pour i'.

Cette technique est intéressante, dans le cadre de 1a discussion qui précède, du fait qu'elle autorise la représentation simultanée sur le plan des axes factoriels pris deux à deux des variables et des unités d'observations. L'examen des positions relatives concerne les relations entre les variables et les facteurs et celles des variables entre elles ; les relations entre les observations et les facteurs et celles des observations entre elles, plus les relations entre les variables et les observations. On conçoit la richesse de l'interprétation qu'on peut tirer d'un tel examen.

En résumé, nous constatons qu'il y a un nombre fini de manières différentes de traiter un ensemble de données et que, malgré ce nombre limité, la plupart des possibilités n'ont pas été exploitées.

En effet, particulièrement en Géographie, il existe fort peu d'exemples d'analyse de type Q, P, 0, T ou S ; que ce soit sur matrice symétrique ou non, par factorisation directe ou par l'intermédiaire de la matrice d'inter-corrélation. I1 est paradoxal que l'analyse de type R soit si souvent utilisée a1ors qu'on devrait s'attendre à ce que les géographes, au lieu de passer par l'intermédiaire des poids locaux, soient plus directement concernés par les différences et les similarités spatiales. Si l'analyse de type R est tellement prédominante, c'est sans doute pour des raisons de commodité ; parce que les données géographiques se présentent généralement sous la forme d'un grand nombre d'observations sur lesquelles sont mesurées un nombre inférieur de variables ; mais cela dénote aussi, chez le chercheur, un esprit routinier et un manque

d'imagination dans la méthode d'appréhender les problèmes qu'il cherche, sinon à résoudre, du moins à éclairer le mieux possible.

Ainsi, nous n'insisterons jamais assez sur le fait qu'il est souvent utile de conjuguer plusieurs méthodes, dans une même étude, pour avoir différents points de vue et pour confronter les résultats ; et que les méthodes factorielles, pléthoriques dans certains types d'études, restent sous-utilisées dans beaucoup d'autres domaines.

LE MODELE CLOS ET LE MODELE OUVERT

Dans le modèle hypothétique que l'on vient de voir, trois composantes suffisent pour expliquer la variation totale contenue dans dix variables initiales. Mais ce modèle en quelque sorte "idéal" a été construit, à partir de deux modèles différents, alternative inévitable qui se présente à tout chercheur entreprenant une étude réelle. Ce sont

- le *"modèle clos"* ou analyse en composantes principales
- le *"modèle ouvert"* ou analyse factorielle en axes principaux

d'après les termes créés par CATELL (1965, p 198).

C'est le modèle clos que l'on choisit implicitement lorsqu'on laisse des unités dans la diagonale de la matrice des corrélations. A moins qu'une variable, au moins, soit parfaitement déterminée par le reste des variables introduites dans le problème, la solution en composant principales entraîne normalement l'extraction d'un nombre de composantes égal au nombre des variables. Ainsi, aucune supposition n'est requise en ce qui concerne la structure sous-jacente des données. C'est pourquoi l'on trouve parfois les termes de *"facteurs définis",* au lieu de composantes principales, pour désigner les résultats du modèle clos, par opposition aux *"facteurs inférés"* résultats du modèle ouvert.

Nous rappelons que le modèle en composantes principales peut se résumer par l'équation suivante :

$$z_i = a_{i1}c_1 + a_{i2}c_2 + \ldots. + a_{im}c_m$$

où chacune des m variables observées est décrite linéairement en termes de m nouvelles composantes non corrélées, c_1, c_2, …, c_m, celles-ci étant à leur tour définies comme une combinaison linéaire des m variables initiales. Ainsi, une variable quelconque peut être décomposée en m composantes et peut être prédite exactement à partir de ces m composantes.

Un tel raisonnement circulaire est la contrepartie, fâcheuse, de la simplicité et de l'élégance mathématique du modèle. Cependant, ROZEBOOM en a justifié l'application en ces termes :

> « ... parfois la vue qu'on a d'un point du cercle est plus intéressante que la vue d'un autre point du cercle. L'analyse complète d'un ensemble de variables en facteurs définis est simplement une transformation linéaire de cet ensemble; et les mesures sur de tels facteurs contiennent conjointement exactement autant d'information sur cet ensemble, ni plus, ni moins, que les mesures sur les variables initiales. Mais certaines manière de dire la même chose nous éclaire davantage que d'autres ... et la transformation d'un ensemble de variables en un ensemble de facteurs peut très bien révéler des relations importantes, qui sont difficiles à discerner parmi les variables sous leur forme initiale. En particulier,

> l'extraction des facteurs exhibe de façon économique le degré de dépendance linéaire qui existe entre les variables" (ROZEBOOM, 1966, p. 213).

Nous constatons que le principal objectif du chercheur n'est pas la transformation mathématique elle-même, qui fournit autant de composantes que de variables. Le terme même *"d'économique"* employé par ROZEBOOM veut dire qu'il est surtout intéressant de connaître dans quelle mesure un plus petit nombre de composantes permet de rendre compte de la plus grande partie de l'information (la variance) contenue dans les données. De fait, puisque généralement les premières composantes expliquent une portion importante de la variance, on peut ne pas tenir compte de celles qui n'apportent qu'une contribution négligeable. Bien sûr, le nombre de composantes que l'on décidera de retenir dépend de ce qu'on entend par "négligeable".

Le critère le plus communément utilisé pour décider du nombre de composantes est celui de 1a valeur propre minimum, qui consiste à écarter comme négligeables - ou non significatives - les composantes dont la valeur propre est inférieure à 1. I1 faut rappeler, en revenant à notre exemple, que la somme des valeurs propres est égale à 1a somme des éléments diagonaux de la matrice des corrélations ; et que cette somme est égale à son tour à m, le nombre de variables. La valeur propre moyenne est donc égale à l'unité. On en déduit qu'une valeur propre inférieure à 1 est associée à une composante qui ne parvient pais à expliquer la variation contenue dans une variable au moins. C'est pour cette raison qu'en général on l'élimine. Comme les valeurs propres sont calculées en ordre de grandeur décroissant, cela pourrait impliquer que les dernières composant ne rendent compte que de la variance d'erreur. Mais comme l'a remarqué CATTEL (1965, p. 204) :

> "Normalement, la proportion de la variance d'erreur reste constante ; et par conséquent, on ne peut pas éliminer la variance d'erreur sans perdre la vraie variance, par n'importe quel procédé, si prudent soit il, d'arrêt dans l'extraction des facteurs à une phase précoce".

Ainsi, il n'existe pas de critère absolu pour définir le nombre de composantes ou de facteurs ; et nous allons voir que le "modèle ouvert" offre d'autres solutions à ce problème. A l'inverse du modèle clos, qui ne fait appel à aucune hypothèse particulière, le modèle ouvert est basé sur l'hypothèse que les corrélations observées sont principalement déterminées par quelques régularités sous-jacentes dans les données, d'où le nom de facteurs inférés que l'on donne parfois aux résultats du modèle ouvert.

Plus précisément, on suppose qu'une variable observée est influencée par différentes causes, dont quelques unes sont partagées avec les autres variables du problème ; tandis que d'autres ne sont partagées par aucune des variables présentes. Cela implique encore que la variance d'une variable ne peut pas être expliquée par toutes les autres variables présentes. La partie de la variable qui est influencée par les "causes" ou facteurs inférés - qui déterminent aussi des portions d'autres variables - est habituellement appelée "*commune*" ; et la partie qui ne peut être expliquée que par des éléments propres à chaque variable est appelée *"unique"*. Sous cette hypothèse, nous voyons que la partie unique de la variable ne contribue pas aux relations entre les variables ; par conséquent, les corrélations observées sont le résultat de l'effet de facteurs communs qui, seuls, établissent un lien entre les variables.

Dans le modèle ouvert, les éléments diagonaux de la matrice des corrélations, c'est à dire les corrélations des variables avec elles-mêmes, r_{ll}, r_{22} ... jusqu'à r_{mm}, sont connues sous le nom de "*communalities*" ou "*communautés*" et sont exprimées par l'équation suivante :

$$r_{11} = r_1f_1\ r_1f_1 + r_1f_2\ r_1f_2 + \ldots.. + r_1f_k\ r_1f_k + r_1f_u\ r_1f_u = 1$$

où les f sont les k facteurs communs et où f_u est un facteur unique que l'on peut décomposer en f_s et f_e ; dans cette décomposition, f_s est le facteur spécifique contenant cette portion de la variance qui devrait être expliquée par les variables non représentées, tandis que f_e est un facteur d'erreur qui contient la portion de la variance de chaque variable due à des erreurs d'échantillonnage ou de mesure.

De façon plus générale, en termes de la variance σ^2_{ij} d'une variable i quelconque, on peut écrire :

$$\sigma^2_i = \Sigma^k_{j=1}\ \sigma^2_{ij} + \sigma^2_{si} + \sigma^2_{ei} = 1$$

où k est le nombre de facteurs communs.

La somme $\Sigma^k_{j=1}\ \sigma^2_{ij}$, ou communauté, qui s'écrit aussi h^2_i est la variance expliquée par tous les facteurs communs. De la même manière, σ^2_{si} et σ^2_{ei} sont les variances expliquées respectivement par le facteur spécifique et par le facteur d'erreur.

Le géographe se trouve alors devant le problème de l'estimation de ces valeurs de *communauté* qu'il doit introduire dans la matrice de corrélation. Nous venons de remarquer que la communauté d'une variable due à la fois aux facteurs communs et au facteur unique constitue une limite supérieure avec une valeur de 1. Les valeurs qui fournissent la proportion de la variation expliquée dans chaque variable par toutes les autres variables comprises dans le problème, constituent une limite inférieure. Ces valeurs sont données par les coefficients de détermination multiple R^2.

Le problème de l'estimation des communautés est traité de façon extensive dans HARMAN (1968, chap. 5), et dans RUMMEL (1970, chap. 13). Parmi les nombreuses approches, l'estimation de limite inférieure que l'on vient de voir paraît être la meilleure. On connaît cependant deux autres approches qui ont conduit à des résultats satisfaisants.

L'approche de "*refactorisation*" commence par l'introduction de l'estimation de limite inférieure dans la diagonale de la matrice des corrélations. Les facteurs sont alors extraits ; de nouvelles communautés sont calculées à partir de ces facteurs ; ces communautés sont introduites dans la matrice de corrélation que l'on factorise à nouveau. On répète ce procédé jusqu'à ce que le calcul de communautés nouvelles donne un résultat très peu différent du précédent.

Une approche plus simple consiste à utiliser comme communauté d'une variable, le coefficient de corrélation le plus élevé parmi les corrélations de cette variable avec toutes les autres variables.

Aucune de ces estimations ne peut être considérée comme la seule valable ; cependant, il faut savoir que l'estimation choisie influence directement le nombre de facteurs extraits. Ainsi nous avons vu que lorsque les communautés sont égales à l'unité, le nombre de facteurs est maximum, c'est à dire égal au nombre de variables. Lorsque le coefficient de détermination multiple est utilisé comme communauté, le nombre de facteurs extrait est minimum. Or, bien qu'en général on cherche à réduire le nombre de variables en un nombre plus petit de facteurs, le but de l'analyse n'est pas de trouver le plus petit nombre possible de facteurs, mais un nombre ayant une signification théorique qui soit le plus proche possible du nombre « réel » inconnu.

Ainsi, qu'on ait affaire à des facteurs communs « inférés » ou à des composantes principales « définies », il n'existe pas de critère objectif pour décider de leur nombre optimum. Les chercheurs doivent dépendre de leur propre jugement, en tant que spécialistes dans leur domaine, pour rejeter les facteurs auxquels ils ne trouvent aucune signification. Cependant, ils doivent examiner avec un soin plus particulier les petits facteurs ; car, comme le dit RUMMEL (1970, p. 362) :

> "Les plus grands facteurs sont habituellement déjà connus par les observateurs expérimentés, à côté de la recherche systématique, tandis que les plus petits facteurs sont masqués par ces interdépendances plus grandes. En rejetant un facteur étrange, on risque par conséquent de rejeter une découverte importante".

De ce fait, le problème de l'estimation des communautés prend toute son importance ; car son influence est d'autant plus grande que les facteurs considérés sont plus petits. Mais si le chercheur, comme c'est généralement le cas en géographie, n'est intéressé que par la structure de base identifiée par les principaux facteurs, le problème des communautés est un faux problème ; car, l'introduction de facteurs inférieurs à 1 dans la diagonale de la matrice des corrélations n'a qu'a incidence imperceptible sur les résultats pour les premiers facteurs extraits. On trouvera l'illustration de cette :affirmation dans l'expérience développée dans la troisième partie de cet ouvrage, au chapitre III.

Finalement l'étape technique suivante, la rotation, peut changer l'ordre d'importance et la signification des facteurs ; et le résultat est affecté par la manière dont la rotation est faite, aussi bien que par le nombre de facteurs préalablement extraits.

LA ROTATION ORTHOGONALE ET LA ROTATION OBLIQUE

Une fois que la dimension de l'espace a été déterminée par le choix du nombre de composantes, une rotation qui conduit à une nouvelle base dans le même espace, c'est à dire une rotation des axes autour de l'origine de l'espace, n'a aucun effet sur la taille des vecteurs qui représentent les variables. Un espace vectoriel a un nombre infini de bases, chacune étant simplement une transformation linéaire de n'importe laquelle des autres. Le problème fondamental de la rotation est de déterminer la transformation linéaire qui produira des facteurs ayant certaines propriétés. Ces propriétés qui servent de critère pour la rotation dépendent pour une large part de jugements subjectifs. Par exemple, un chercheur peut décider de faire tourner les axes jusqu'à ce qu'ils correspondent à un groupe de variables particulier qui devrait, selon sa théorie, être distinct des autres. I1 y a là une tentation évidente d'essayer de contraindre les résultats à justifier la théorie. Le critère de "*simple structure*" ou de structure simple a été développé par THURSTONE (1954) pour essayer de donner au procédé de rotation une forme plus objective. C'est le critère qu'on utilise le plus fréquemment lorsqu'on procède à une rotation, qu'elle soit orthogonale ou oblique.

Pour comprendre ce que signifie *"simple structure*",nous devons auparavant rappeler quelles sont les caractéristiques de la matrice des saturations avant la rotation. La première composante est calculée de manière à rendre compte de la plus grande partie de variance possible dans les données, ce qui fait que la plupart des variables sont fortement corrélées avec elle. Les composantes suivantes rendent compte d'une quantité décroissante de la variance. Ce procédé revient souvent à localiser les composantes entre les groupes distincts que forment les variables.

Le but du critère de structure simple est obtenu lorsque, par une rotation de l'ensemble des facteurs, on parvient à maximiser la colinéarité de chaque facteur avec un groupe particulier de vecteurs, représentation géométrique des variables. La variance expliquée par la première composante avant la rotation est, après rotation, répartie à travers tous les facteurs. Chaque facteur, après la rotation vers une structure simple, s'identifie plus ou moins à un groupe particulier de variables étroitement associées. A partir d'une telle association, on devrait pouvoir établir la signification du facteur considéré comme la *cause déterminante* de cette association.

L'avantage de la structure simple, outre la possibilité qu'elle offre de donner une signification aux facteurs, est une qualité d'invariance. Alors que les saturations des facteurs initiaux dépendent du nombre de variables introduites dans l'analyse, les facteurs après rotation ont de fortes chances de désigner les mêmes groupes de variables, quelles que soient les variables qui aient été ajoutées ou retranchées par ailleurs. Cela rend possible la comparaison des résultats entre des études différentes. KAISER (1958, p. 195), par exemple, considère l'invariance factorielle comme "le critère ultime en faveur de la rotation".

Maintenant, si convaincu par les arguments de KAISER, on choisit de procéder à la rotation de la structure factorielle initiale vers une structure simple, on se trouve en présence d'une nouvelle alternative avec la technique de rotation orthogonale d'une part, et la technique de rotation oblique d'autre part.

La technique orthogonale fait subir une rotation a l'ensemble de la structure factorielle autour de l'origine de l'espace décrit par les facteurs orthogonaux, en traitant cette structure comme un cadre rigide. La rotation se poursuit jusqu'à ce que la meilleure position de ce cadre, par rapport aux variables selon le critère de structure simple, soit obtenue. Pour k facteurs et m variables, $F = F_o T_{12}$, où F est la matrice des saturations après rotation, F_o est la matrice des facteurs originels, et T_{12} est une matrice de transformation linéaire.

Dans l'exemple suivant où m = 3 et k = 2, on a

$$\begin{vmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \end{vmatrix} = \begin{vmatrix} f_{o11} & f_{o12} \\ f_{o21} & f_{o22} \\ f_{o31} & f_{o32} \end{vmatrix} * \begin{vmatrix} \text{Cos } \alpha & -\text{Sin } \alpha \\ \text{Sin } \alpha & \text{Cos } \alpha \end{vmatrix}$$

α est l'angle de rotation dans le plan des facteurs 1 et 2.

I1 existe essentiellement deux procédés algébriques qui permettent ce type de rotation orthogonale :

- ***Quartimax***, élaboré par J. NEUHAUS et C. WRIGLEY (1954), cherche à réaliser le critère de simple structure en minimisant la somme des produits de toutes les paires possibles de saturation dans chaque ligne. Cela revient à réduire la complexité des variables, c'est à dire à faire en sorte que chaque variable soit fortement corrélée avec un seul facteur, et peu ou pas corrélée avec les autres facteurs. Cependant, parce qu'il a tendance à conserver au premier facteur un caractère général en lui permettant de rendre compte encore d'une grande partie de la variance, ce procédé ne conduit pas vraiment à une structure simple.

- ***Varimax*** créé par KAISER (1958), consiste à simplifier la complexité des facteurs, en rendant maximum la variance du carré des saturations dans chaque colonne. Il en résulte que chaque facteur a de très fortes saturations avec un petit nombre de variables, et des saturations très faibles avec les autres variables. Varimax est celui des deux procédés orthogonaux qui permet de s'approcher le plus de la structure simple. I1 n'en reste pas moins que la contrainte de

l'orthogonalité des axes n'est pas vraiment compatible avec l'idéal de la structure simple, parce qu'il n'y a aucune raison pour supposer que des influences sous-jacentes soient indépendantes.

La rotation oblique permet aux facteurs d'être corrélés. On fait subir aux facteurs pris individuellement une rotation jusqu'à ce que chaque facteur corresponde à un groupe distinct de variables. I1 existe un grand nombre de procédés algébriques cherchant à respecter le critère de structure simple dans la rotation oblique. Les sources classiques de l'analyse factorielle (HARMAN, 1967 ; RUMMEL, 1970) en fournissent une description détaillée. On se contentera ici de citer les plus connues.

Avec ***Oblimax,*** PINZKA et SAUNDERS (1954) ont cherché à augmenter le nombre des saturations élevées et celui des saturations basses, en diminuant celui des saturations moyennes. Cette technique s'est avérée ne convenir qu'aux données possédant une structure simple évidente, c'est à dire aux variables qui forment des groupes bien distincts dans l'espace des observations. ***Quartimin,*** élaborée par CARROLL (1953), est en quelque sorte la version oblique de quartimax puisqu'elle s'attache à minimiser, par ligne, la somme des produits internes des saturations. Elle a l'inconvénient de conduire à de fortes corrélations entre les facteurs. A l'inverse, ***Covarimin***, une extension par KAYSER (1958) de sa technique varimax à la rotation oblique, conduit à des corrélations faibles entre les facteurs, proches finalement de celles produites par la version orthogonale. CARROLL (1957) a proposé une solution intermédiaire avec ***Biquartimin***. Il a finalement généralisé ces trois techniques sous le nom d'***Oblimin***, en jouant sur les valeurs d'un paramètre. La solution quartimin est donnée par $\delta = 0$, la solution covarimin par $\delta = 1$ et la solution biquartimin par $\delta = 0.5$. Sans compter celles qu'on vient de citer, il existe une profusion de solutions pour la rotation oblique, dont aucune ne peut être considérée comme la plus satisfaisante.

Cela explique sans doute le peu de faveur dont jouit la rotation oblique auprès des utilisateurs. Cette approche compte pourtant de nombreux avantages. En premier lieu, les groupes de variables sont mieux définis, simplement parce que les variables qui en constituent le centre sont aussi celles qui ont les plus fortes saturations sur le facteur considéré, ce qui est rarement le cas dans la version orthogonale. Deuxièmement, les résultats procurent souvent une meilleure .description de la réalité, parce que les phénomènes que désignent les facteurs sont fréquemment inter-corrélés soit individuellement soit en groupe et que les facteurs doivent refléter de telles conditions. Le troisième avantage important ressort du fait qu'on peut poursuivre l'analyse à un niveau supérieur d'abstraction. En effet, les corrélations entre les facteurs eux mêmes peuvent être introduites dans une nouvelle analyse factorielle pour rechercher les concepts plus généraux et plus abstraits qui déterminent la variation entre les facteurs.

Cependant les rotations orthogonales sont celles qu'on utilise dans la plupart des analyses, sans doute en raison de leur simplicité et de leur élégance mathématique. Il est à craindre, en outre, que leur utilisation préférentielle ne vienne aussi d'une croyance erronée. En effet, parmi ceux qui utilisent la rotation orthogonale de préférence à la rotation oblique, certains le font dans l'intention d'utiliser par la suite les scores sur les facteurs orthogonaux comme entrées dans de nouvelles manipulations qui requièrent des données indépendantes comme, par exemple, la taxonomie numérique. Ceux là oublient que seule, parmi toutes les positions possibles des axes factoriels, la solution en composantes principales fournit des scores indépendants. Après rotation, que les facteurs soient orthogonaux ou non, les scores sur les facteurs ne sont plus indépendants.

I1 nous parait intéressant, pour conclure, de signaler les articles que DAVIES et MATHER ont consacré aux intérêts respectifs de l'analyse en composantes principales, de

l'analyse factorielle et des différents procédés de rotation qui généralement les accompagnent. DAVIES (1971-1972) s'insurge contre l'utilisation indiscriminée du procédé de rotation varimax pour deux raisons. Premièrement, la simplification de la structure factorielle se fait aux dépens de la généralisation produite par l'analyse en composantes principales. Deuxièmement, on peut penser intuitivement qu'une rotation oblique devrait donner une solution plus utile, étant donnée la rigidité de la structure orthogonale. On peut reconnaître avec DAVIES que

> "de façon paradoxale, les utilisateurs de la rotation varimax ne sont pas satisfaits par les approches subjectives et descriptives et recherchent des théories générales et des modèles ; mais en utilisant varimax, ils risquent de détruire la généralité même qu'ils recherchent."

MATHER (1971-1972) réplique en citant CATTELL (1965) :

> "l'analyse en composantes principales doit être rejetée comme moyen d'investigation scientifique général, parce qu'il semble tout à fait improbable que n variables comprennent entre elles toutes les causes qui rendent compte de leur propre variance. Le modèle en composantes principales, en ce qui concerne la plus grande majorité des données réelles, doit être considéré: comme une pure fiction mathématique."

et en prétendant que la généralisation peut être atteinte plus sûrement par l'analyse factorielle, dont l'objectif est d'identifier des ensembles de variations communes dans les données. I1 recommande de compléter cette recherche d'influences sous-jacentes par une rotation en accord avec le critère de structure simple.

Le fait qu'il y ait plusieurs procédés de rotations qui soient tous aussi valables mathématiquement ne veut pas dire que l'on peut prendre n'importe quelle solution. Il n'y a qu'une solution qui soit géographiquement bonne, au moins dans un cadre théorique donné. Par exemple, le problème de la spécificité contre la généralité dépend de l'idée conceptuelle que se fait le chercheur du système qu'il examine. La généralité d'une solution peut être tout simplement fonction de l'échelle sélectionnée au départ, ou bien elle peut être obtenue par l'intermédiaire de rotations obliques telles qu'oblimax ou *oblimin*, suivies d'une analyse factorielle d'ordre supérieur, puisqu'un facteur de second degré peut exprimer de façon plus générale et plus abstraite la relation qui existe entre les facteurs de premier ordre.

En résumé, les grandes alternatives qu'offrent les modèles factoriels, et qui en font toute la richesse, ont été peu ou mal exploitées. En ce qui concerne la construction de la matrice des corrélations, on a vu que la plupart des analyses sont restreintes aux relations de type R alors que beaucoup de combinaisons possibles restent à explorer. Quant aux choix du modèle clos ou du modèle ouvert, de la rotation orthogonale ou de la rotation oblique, ils n'ont d'intérêt que si le chercheur propose au départ une théorie qui permette de les justifier. Dans le cas contraire, malheureusement très fréquent, d'analyse factorielle empirique de caractère exploratoire, on voit le chercheur suivre aveuglément la même route (communautés = 1, rotation orthogonale, nombre de facteurs limité) sans se soucier de savoir si c'est celle qui convient le mieux au phénomène qu'il étudie.

Dans la partie suivante , une revue des principales applications, regroupées en trois grandes catégories, va nous permettre de mesurer l'importance des modèles factoriels dans la géographie anglo-saxonne et d'en évaluer les apports et les dangers pour la géographie en général.

DEUXIEME PARTIE

EVALUATION CRITIQUE DES APPLICATIONS

Chapitre I

LES ETUDES THEMATIQUES

Les études thématiques se caractérisent par l'examen d'une variable particulière dont on cherche à expliquer la répartition spatiale, en fonction d'un petit nombre de variables spécifiques indépendantes. Les méthodes de régression et de corrélation multiple ont été couramment appliquées à ce genre d'analyse. I1 n'est pas dans notre intention de présenter ici les innombrables applications dont on trouve une liste sélectionnée dans *Statistical Analysis in Geography*, à la suite d'un exposé détaillé des procédés de calcul (KING, 1969 ; p. 135-152). D'ailleurs, les méthodes de régression et de corrélation n'ont pas l'exclusivité des études thématiques. Très tôt, les modèles factoriels ont apporté leur contribution dans ce vaste champ d'opération. Puisque nous essayons de dégager les mérites relatifs du modèle factoriel, il serait intéressant de faire une mise à jour des applications par régression et corrélation, en insistant sur celles où les problèmes apparaissent le plus clairement, avant de présenter les études thématiques où l'analyse en composantes principales, et l'analyse factorielle, ont été de préférence utilisées.

LES ANALYSES DE REGRESSION ET DE CORRELATION MULTIPLE

BROWN et LONGBRAKE (1970) ont voulu établir une fonction *"d'utilité du lieu"* fondée sur des caractéristiques de migration à l'intérieur d'une ville, en l'occurrence Cedar Rapid, Iowa. Une méthode de régression estime les paramètres de cette fonction. Les résultats montrent que les migrations entre des voisinages de type différent sont plus fluides qu'on ne s'y attendait. Les variables indépendantes les plus déterminantes se révèlent être les aspirations de la famille touchant le logement et les caractéristiques spatiales du marché. On peut critiquer 1e fait que la proportion de la variation expliquée n'est pas assez forte pour que le modèle soit utilisable dans la planification. Aussi, l'utilisation des unités d'observation disponibles, d'un degré d'agrégation assez élevé, rend suspect l'authenticité des corrélations observées, corrélations dites *"écologiques"* qui, comme nous le verrons plus loin, ont peu de chance d'être valables au niveau individuel.

Exploitant le même thème de migration intra-urbaine, CLARK (1973) a cherché à définir les types de pressions qui sont liés à la mobilité individuelle. Les résultats montrent que cinq types de pression sont significativement corrélés avec le désir de déménager. Ce sont dans l'ordre la taille et les aménagements du logement; les gens du voisinage, la proximité des amis, la pollution de l'air, la distance du lieu de travail. Il est évident que pour un thème aussi complexe, l'essentiel des variables sélectionnées n'est pas assez étendu, et que certaines variables comme le bruit ou la proximité de certains services devraient être considérées.

HARRIES (1973) a étudié la répartition spatiale de la violence en relation avec les populations métropolitaines. Pour cela, il a examiné la variation des activités criminelles avec la taille des SMSAs (Standard Metropolitan Statistical Areas) au moyen de simples corrélations.

Les résultats montrent que quatre variables ont une corrélation significative avec la taille de la population. Ce sont par ordre d'importance, le vol, le viol, le meurtre et l'attaque à main armée. Parmi les variations régionales, on remarque la prédominance des villes du Sud pour les taux de meurtre et d'attaques à main armée, celle des SMSAs supérieurs à deux millions d'habitants en ce qui concerne les vols, et la concentration des viols en Californie. Si ces résultats améliorent notre compréhension des processus générateurs de violence, ils sont loin d'être suffisamment explicites. Par exemple, on peut se demander si les taux élevés de viols en Californie sont plus particulièrement liés au climat, à la croissance urbaine, ou tout simplement à la bonne qualité des statistiques.

Deux études, à dix années d'intervalle, vont nous servir plus spécifiquement d'exemples. Très complètes et scrupuleusement menées, elles devraient nous permettre de dégager l'originalité des méthodes de régression et de corrélation multiple par rapport aux modèles factoriels, les qualités ou les défauts que ces méthodes ont en commun, et finalement l'apport des modèles factoriels.

En 1961, KING s'est intéressé à l'espacement des agglomérations en relation avec différents indices économiques physiques et sociaux sur un échantillon de 200 villes américaines (KING, 1961). Ces indices au nombre de six sont : la taille des villes, la taille de l'exploitation agricole, la densité rurale, le taux d'industrialisation, la densité de la population totale, et la valeur des terres et des bâtiments par unité de surface. Bien qu'ils aient été sélectionnée, en tant que variables indépendantes, pour tester par l'analyse de régression l'hypothèse logique de leur association avec l'espacement des agglomérations, deux d'entre eux seulement contribuent à l'explication de l'espacement des agglomérations de façon significative. Ce résultat décevant vient de ce qu'il existe une très forte inter-corrélation des variables dites *"indépendantes",* en particulier avec la densité de la population et la taille de l'exploitation.

Pour surmonter cette difficulté inhérente au modèle de régression, deux possibilités se présentent au chercheur. L'une est de jouer sur le nombre et la qualité des variables indépendantes. Pour cela il faut étudier la représentation cartographique des résidus[9], formuler de nouvelles hypothèses, supprimer certaines variables, en ajouter d'autres et poursuivre ce procédé par tâtonnements, qui risque d'être fort long, jusqu'à ce qu'on obtienne un bon niveau d'explication. La seconde solution, celle qu'à choisie KING, consiste à évaluer, par l'intermédiaire de l'analyse de covariance[10], l'importance du rôle joué par certains regroupements des unités d'observation dans la variation de la variable dépendante. Les regroupements en *places centrales* définies comme servant de centre de services pour la zone rurale environnante, et en *"places non centrales",* en régions au relief plus ou moins accentué, et en zones d'agriculture de type différent, parviennent dans l'ensemble à rendre les relations entre l'espacement des villes et les variables indépendantes désignées plus haut, davantage significatives.

Par exemple, deux variables expliquent une proportion importante de la variance dans l'espacement des places centrales. Ce sont la taille de la ville et l'emploi dans l'industrie. Pour lesplaces non-centrales, c'est la densité de population totale qui offre la meilleure explication. Dans la zone d'agriculture extensive, 40 % de la variation dans la variable dépendante est attribuée à la relation étroite entre l'espacement des villes et leur taille. Dans la zone

---

[9] Les résidus sont les différences entre les valeurs calculées d'après l'équation de régression et les valeurs observées.

[10] L'analyse de covariance, d'après KING (1969), est une combinaison de l'analyse de régression et de l'analyse de variance. Etant donnée la régression de X_0 sur X_1, on divise les observations en groupes d'après l'effet régional que l'on désire introduire. Puis, on établit un test pour savoir si les moyennes x des groupes diffèrent de manière significative.

d'agriculture spécialisée, le niveau d'explication est nettement plus bas avec 0,9 % seulement de la variance expliquée par le pourcentage de la population employée dans l'industrie.

On voit ainsi que certaines relations sont significatives dans certaines régions et pas dans d'autres. Lorsque l'analyse est faite au niveau individuel, sur l'ensemble des villes sans distinction d'appartenance à une catégorie ou à une région particulière, on comprend que des relations puissantes mais contradictoires parviennent à s'annuler. Dans l'étude précédente c'est par des manipulations contrôlées, fondées sur une meilleure définition des hypothèses de départ, que l'auteur a pu obtenir finalement des résultats satisfaisants.

Cet exemple nous permet d'introduire une mise en garde contre les effets des regroupements non contrôlés. Ce phénomène, étudié par ROBINSON (1956) sous le nom *d'"erreurs écologiques"*[11], et dont les géographes se sont fort peu occupés depuis, mérite qu'on s'y attarde.

ROBINSON a établi les équations écologiques suivantes :

$$C_{xy} = WC_{xy} + EC_{xy}$$

ce qui signifie que la covariance des variables X et Y pour n individus est égale à la somme de la covariance à l'intérieur de la région WC_{xy} et de la covariance écologique EC_{xy} entre les régions. Maintenant, la covariance de X et de Y, divisée par la racine carrée de la variance de X et de Y, donne par définition la corrélation entre X et Y :

$$R_{XY} = \frac{WC_{XY} + EC_{XY}}{\sqrt{C_{XX} * C_{YY}}}$$

Les effets de covariations entre les régions et à l'intérieur des régions interfèrent dans la relation des corrélations individuelles. L'exemple qu'en donne ALKER dans l'ouvrage de DOGAN et al. (1969) est frappant. La corrélation ente le nombre de noirs et l'analphabétisme aux Etats-Unis est de 0,95 à l'échelle des régions alors qu'elle n'est plus que de 0,20 à l'échelle des individus. Cette *"erreur écologique"* vient du fait que les régions où il y a beaucoup de noirs se trouvent être aussi celles ou les blancs illettrés sont nombreux. Mais, à l'intérieur de chaque région, la proportion de noirs analphabètes est faible. On voit comment cela interdit les inférences entre deux niveaux différents d'agrégation des unités d'observation. On voit aussi quel peut être l'intérêt, pour un géographe, de développer des analyses à différents niveaux, pour étudier justement l'effet régional ou national sur tel phénomène comme, pour citer un autre exemple d'ALKER, le comportement électoral. Ainsi, sachant qu'au niveau individuel les ouvriers ont tendance a être radicaux, peut-on en conclure que dans les états où il y a une plus grande proportion d'ouvriers, il y aura une plus forte tendance radicale ? Pas nécessairement, car les plus fortes concentrations d'ouvriers se trouvent dans les états les plus avancés économiquement et, sous de telles conditions, les ouvriers ont tendance à devenir conservateurs.

On pourrait multiplier les exemples de ce type. Cela signifie-t-il qu'il faut, dans la mesure du possible lorsqu'on veut étudier tous les aspects d'un phénomène, établir un programme d'analyses faites à des échelles différentes ? Cela est recommandé en effet puisqu'on ignore généralement les implications écologiques du choix qu'on aura fait de telle ou telle échelle. Cependant on comprend que le temps requis et la non disponibilité des mêmes variables à différents niveaux interdisent souvent la répétition des analyses. C'est pourquoi THOMAS et ANDERSON (1965) ont cherché à développer une méthode dont le but est

[11] *Ecologique* a gardé ici son sens étymologique de rapport des phénomènes avec leur environnement immédiat.

d'éliminer l'effet de 1a taille de l'unité spatiale d'observation. Avant eux, ROBINSON (1956) a proposé une solution qui consiste à pondérer les équations qui, permettent de calculer les paramètres de régression et les coefficients de corrélation par la surface des unités spatiales considérées.

L'exemple élémentaire qui suit montre trois cas avec leurs paramètres et coefficients respectifs :

A = 2 X = 2 Y = 4	A = 2 X = 2 Y = 4		A = 2 X = 2 Y = 4				
				A = 4 X = 2 Y = 4		A = 8 X = 2 Y = 4	
A = 2 X = 2 Y = 4	A = 2 X = 2 Y = 4		A = 2 X = 2 Y = 4				
A = 2 X = 4 Y = 6	A = 2 X = 3 Y = 8		A = 2 X = 4 Y = 6	A = 2 X = 3 Y = 8		A = 2 X = 4 Y = 6	A = 2 X = 3 Y = 8
Cas I			Cas II			Cas III	
A = 1,429 b = 1,429 r = 0,715			A = 1,625 b = 1,375 r = 0,687			A = 3,00 b = 1,00 r = 0,50	

L'équation générale qui permet de trouver b, la pente de régression, s'écrit :

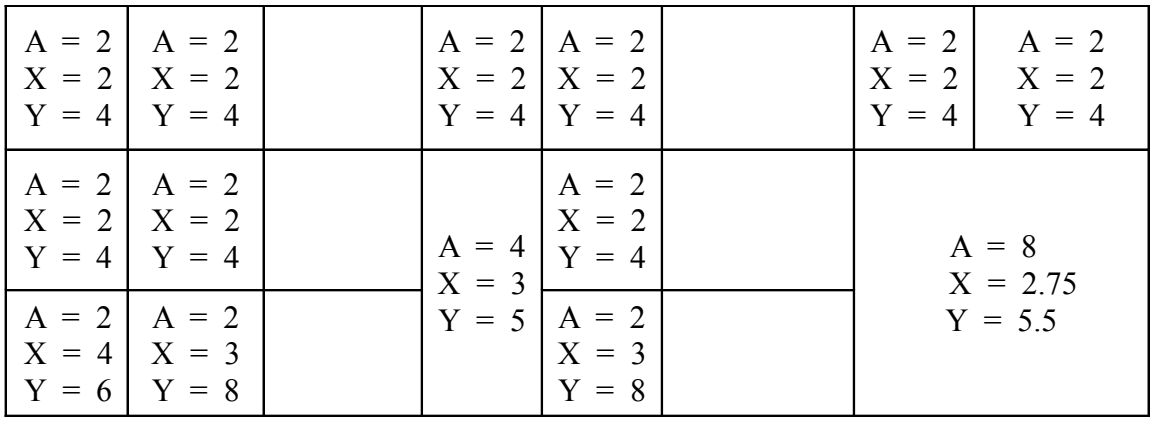
$$b = (N \sum xy - \sum x \sum y) / (N \sum x^2 - N \sum y^2)$$

L'équation, une fois pondérée, devient :

$$b = (\sum A \sum A_{xy} - (\sum A_x \sum A_y) / (\sum A \sum A_x^2 - (\sum A_x)^2)$$

où A est la surface de l'unité spatiale d'observation. Après pondération par A, on trouve, dans les cas II et III, les mêmes paramètres que dans le cas I.

Cependant, la pondération par la surface ne convient que lorsque les zones combinées en zones plus vastes ont, à l'origine, la même distribution pour les X et les Y. Trois nouveaux cas très simples servent à illustrer ce problème :



et après pondération par la surface, on voit qu'on ne peut plus retrouve les paramètres du cas A initial :

Cas A	Cas B	Cas C	Cas B pondéré	Cas C pondéré
a = 1,429	a = -1,00	a = 0,00	a = 0,00	a = 0,00
b = 1,429	b = 2,50	b = 2,00	b = 2,00	b = 2,00
r = 0,715	r = 0,829	r = 1,00	r = 0,707	r = 1,00

Puisque la pondération par la surface n'est pas une solution générale au problème, THOMAS et ANDERSON ont imaginé d'utiliser des notions de statistique inférentielle. Etant donné qu'on peut découper une même zone d'étude en unités d'observations différentes et qu'on obtient des paramètres différents selon la taille et le nombre de ces unités, comment doit-on interpréter ces différences ? Sont-elles significatives d'une réelle différence géographique ou sont-elles simplement dues au hasard ? Le test utilisé pour confirmer l'hypothèse d'une différence aléatoire entre les coefficients de corrélation est connu sous le nom de z-test, et consiste à évaluer le rapport suivant :

$$(z_m - z_p) / (1/N_m - 3) + 1/N_p - 3)^{1/2}$$

où N est le nombre d'unités spatiales dans une étude donnée, m et p sont les études que l'on cherche à comparer, et où $z = 1/2 \ln ((1+r) / (1-r))$

Pour tester a et b, les paramètres de régression, le modèle inférentiel choisi par THOMAS et ANDERSON fait appel à une analyse de covariance légèrement modifiée. La différence entre les coefficients de régression pour deux études différentes, m et p, est testée par :

$$(b_m - b_p) / \sigma (1 / \Sigma x^2_m + 1 / \Sigma x^2_p)$$

où σ est l'erreur type de l'estimation pour l'ensemble.

Cependant, un autre problème de caractère spatial reste à résoudre, car les effets de l'agrégation des distributions spatiales diffèrent selon que les valeurs de X et de Y, dans des

unités spatiales données, sont systématiquement très semblables aux valeurs des unités voisines ou au contraire très différentes. Ce phénomène de continuité, connu sous le nom *"d'auto-corrélation spatiale",* est souvent mentionné dans la littérature anglo-saxonne (CHORLEY et HAGGETT, 1967 ; KING 1969) pour dire qu'il est très difficile à éliminer et aucune solution satisfaisante n'a été proposée jusqu'à ce jour. Pourtant dès 1965, MATHERON a publié un ouvrage, où, sous le nom de *Les variables régionalisées et leur estimation*, il expose "une application de la théorie des fonctions aléatoires aux sciences de la nature".

Les variables régionalisées ont été substituées aux variables aléatoires, objets de la statistique habituelle, afin de pouvoir tenir compte précisément de la structure spatiale des phénomènes naturels. Les caractères structuraux que la statistique ordinaire est incapable d'exprimer et qui sont pris en charge par la théorie des variables régionalisées, sont essentiellement la localisation, la continuité et l'anisotropie. Pour tenir compte de la localisation, il faut définir de manière précise les dimensions, la forme, et l'orientation de l'unité spatiale, ou *"support géométrique",* sur lequel on effectue les mesures, à l'intérieur de la région bien déterminée que l'on étudie ou *"champ géométrique".* Ainsi, pour reprendre les propres termes de MATHERON (1965, p. 7) :

> « L'une des tâches de la théorie des variables régionalisées que l'on appelle parfois aussi géostatistique quand elle s'applique à des problèmes géologiques ou miniers, consiste à prévoir les caractéristiques de la variable définie sur un support V connaissant, par exemple, celles de la variable ponctuelle dans un champ différent V' »

Deuxièmement, il est important de savoir si un phénomène est extrêmement continu ou s'il présente de nombreuses irrégularités ou des discontinuités dans ses manifestations. Et finalement, une régionalisation peut être anisotrope, c'est à dire qu'elle peut connaître :

> « Une direction privilégiée le long de laquelle les valeurs se modifient lentement tandis qu'elles varient beaucoup plus vite dans la direction perpendiculaire » (MATHERON 1965, p. 8).

L'essentiel de la méthode consiste à exprimer ces caractéristique structurales majeures sous la forme synthétique d'une fonction g(h), appelé *covariogramne transitif***,** ou de son équivalent probabilistique, la fonction d'autocorrélation K(h). Un exposé plus complet de cette méthode, infiniment plus complexe, subtile et puissante que ne le laisse entrevoir ce bref résumé, dépasse le cadre de notre étude sur les modèles factoriels. On peut cependant regretter que les géographes d'Outre-Manche et d'Outre-Atlantique n'aient, semble-t-il, pas eu connaissance de la théorie des variables régionalisées, déjà mise en pratique maintes fois avec succès par les géologues de l'Ecole des Mines de Fontainebleau.

L'étude de KING nous a conduit à examiner les difficultés que présente l'analyse simultanée en général, au niveau des "*corrélations écologiques*" et de "*l'autocorrélation spatiale*". Ces difficultés essentiellement liées à la définition de l'unité spatiale d'observation, nous pourrions les appeler les "*pièges spatiaux*". L'étude qui suit va nous servir à mettre en évidence des problèmes d'ordre conceptuel ou "*pièges conceptuels*" ayant trait à la définition des variables et aux relations causales qu'on leur attribue.

MORRIL et WOHLEMBERG (1971) ont consacré tout un ouvrage à *La géographie de la pauvreté*. Le but de leur étude est une meilleure connaissance de la variation régionale de la pauvreté, et de la persistance de ce phénomène aux Etats-Unis. Comme variable dépendante, et bien qu'ils aient eux-mêmes critiqué son caractère simpliste, les auteurs, emploient la proportion de familles dont le revenu est inférieur à 3 000 dollars en 1959, parce que c'est la

seule information accessible au niveau d'observation auquel ils travaillent. Les unités d'observation sont les SEAs (State Economic Areas), intermédiaires entre les comtés et les états. Ici, on peut reprocher aux auteurs de ne pas avoir cherché à explorer dans quelle mesure des changements dans la définition de l'unité d'observation utilisée pouvait affecter la distribution spatiale des résultats. Quant au nombre de variables utilisées (51) pour essayer de démêler l'enchevêtrement des relations qui concourent à engendrer et à maintenir la pauvreté, il n'est pas sûr qu'il soit une solution à la complexité du problème ; témoin en est la grande redondance qu'on trouve parmi ces variables, dont quelques unes seulement contribuent à l'explication. Encore faut-il savoir exactement lesquelles ; or il ne semble pas que le procédé analytique utilisé soit vraiment approprié.

La méthode de régression multiple, dans sa version *"stepwise"* ou progressive, introduit théoriquement les variables indépendantes dans l'ordre de leur capacité à réduire la variance non expliquée. Ainsi, MORRILL.et WOLHEMBERG trouvent que sept variables, qu'ils considèrent comme vraiment indépendantes, expliquent 93 % de la variance dans le pourcentage des familles pauvres pour 146 SEAs ; et ils en concluent que les 44 autres variables sont redondantes, ou pas assez pertinentes, pour expliquer la variation spatiale de la pauvreté. Les sept variables retenues sont des caractéristiques sociales (médiane des années de scolarité, etc), des caractéristiques professionnelles et économiques (pourcentages d'artisans, proportion d'inactifs, etc) et des caractéristiques spatiales (localisations relatives).

Pour deux raisons, il faut être très prudent dans l'interprétation de ces résultats apparemment très satisfaisants :

La première raison vient du fait que les résultats d'une régression par étape sont extrêmement sensibles à l'ordre dans lequel les variables *"explicatives"* sont entrées dans l'analyse, et que cet ordre est arbitraire. La deuxième raison est qu'il ne faut pas attacher trop d'importance aux relations de cause à effet entre ces variables et la variable dépendante. Par exemple, quand bien même une forte corrélation négative existe entre le niveau d'éducation et l'incidence de la pauvreté, il est impossible de dire laquelle de ces variables est la cause, laquelle est l'effet. Est-ce le bas niveau d'éducation qui entraîne une mauvaise préparation aux emplois bien rémunérés, et par conséquent une grande proportion de pauvres dans la population ? Ou bien est-ce l'environnement physique et psychique créé par les conditions de pauvreté qui rend difficile la réussite scolaire ? 0u bien encore, le faible niveau de revenu de la communauté ne peut-il pas fournir les fonds nécessaires à une éducation de qualité ? On conçoit qu'il y a en réalité une interaction des deux phénomènes, éducation et pauvreté. Ainsi, lorsque l'on considère que l'un d'entre eux seulement est déterminant (l'éducation en l'occurrence), on gonfle artificiellement son pouvoir explicatif. C'est ce que fait le chercheur lorsqu'il choisit de couper la chaîne des inter-corrélations à un point donné pour le prendre comme point de départ.

Ainsi, en intervenant d'une part dans l'ordre d'entrée des variables, et en donnant une direction préférentielle aux relations résultantes, le chercheur a encore par deux fois (sans compter le choix initial de l'ensemble des variables et des unités d'observations), la possibilité d'orienter l'analyse dans le sens qu'il désire, c'est à dire dans le sens qui correspond le mieux à l'idée qu'il se fait du phénomène.

On peut observer alors trois attitudes. L'attitude *"machiavélique"* de celui qui, bien que tout à fait conscient de ses choix, les tait pour profiter de l'apparence d'objectivité que confère l'appareil mathématique à ses résultats et pour prouver ainsi aisément le bien fondé de sa théorie; ce cas est heureusement assez rare. Plus courante est l'exposition explicite des choix et la mise en garde sur les conséquences que ces choix peuvent avoir sur les résultats. Cependant il est rare qu'on s'attarde à examiner les conséquences elles mêmes ; et l'invitation à la prudence

devient une routine sans effet. C'est le plus souvent à cette troisième attitude qu'on a affaire. D'ailleurs, fréquemment, le choix des chercheurs leur est dicté par les études précédentes et par leur adhésion plus ou moins consciente à l'idéologie qui les entoure. L'exemple de MORRIL et WOLHEMBERG est significatif à cet égard. Les deux variables introduites en premier dans l'analyse, soit la médiane des année de scolarité et le pourcentage de population *"non blanche",* expliquent déjà 67 % de la variance de la pauvreté. Mais les auteurs sont assez gênés par ce résultat. Pour en atténuer l'effet, ils soulignent les fortes inter-corrélations entre ces caractéristiques sociales et les variables économiques et spatiales qui, ensemble, parviennent à expliquer 93 °' de la variance, et graduellement ils en viennent à la conclusion suivante :

> « Seul un déplacement vers une autre région, où les conditions du marché du travail sont meilleures, devrait donner à la majorité des pauvres l'espoir d'améliorer leur statut ».

Même si cela est vrai, on peut dire assurément que ce ne sont pas les méthodes analytiques utilisées qui ont conduit, logiquement et irrémédiablement, à une telle conclusion.

Enfin, en considérant que la pauvreté est associée à des caractéristiques différentes selon les régions, les auteurs ont procédé à une régionalisation. Cette classification spatiale a été effectuée en regroupant les SEAs les plus semblables, d'après leurs mesures sur les sept caractéristiques les plus étroitement associées à la pauvreté. Il est bien évident qu'on aurait pu utiliser, au lieu des simples variables, les scores sur les facteurs dérivés à partir de toutes les variables ; ce qui aurait permis d'utiliser davantage d'information. Mais les auteurs ont pensé que l'utilisation de variables simples rendrait plus facile pour le lecteur la compréhension de la nature des régions.

Cela nous conduit à examiner ce qu'aurait apporté une méthode comme l'analyse factorielle, par rapport aux méthodes de régression et de corrélation multiple, dans des études thématiques comme celles que nous venons d'examiner. D'abord, nous devons reconnaître que beaucoup de problèmes sont communs aux deux types d'analyses. Par exemple, le choix de la taille de l'unité d'observation est très important dans les deux cas, puisqu'il intervient au niveau des corrélations entre les variables, corrélations qu'on retrouve à la base des deux méthodes. Le choix de la taille dépend, bien sûr, du niveau de généralité qu'on recherche ou de la fréquence spatiale du phénomène qu'on étudie. Dans tous les cas, la prudence veut qu'on ne tire de conclusion qu'au niveau où l'on a entrepris l'analyse, et qu'en aucun cas on n'essaie d'inférer des résultats pour des niveaux inférieurs ou supérieurs. Le choix des variables est aussi une étape cruciale, encore que pour l'analyse factorielle où les variables sont traitées simultanément en grand nombre, ce choix soit moins déterminant que pour la régression où quelques variables sont introduites progressivement et sont considérées séparément dans l'interprétation des résultats.

Au niveau des procédés de calcul et de l'interprétation, chaque méthode a ses inconvénients. Dans l'analyse de régression. le principal danger vient de ce qu'on croit mieux comprendre ce qui se passe. Mais les associations simples qu'on voit, et qu'on traduit trop souvent par des relations de cause à effet, cachent en réalité tout un réseau complexe d'interrelations et de *"feed bock"* dont il est impossible de suivre exactement le sens. L'analyse factorielle ne présente pas un tel danger. Théoriquement elle devrait fournir davantage d'information en proposant, avec les facteurs, des concepts où la contribution de chaque variable est clairement définie. Mais cette abondance d'information est rarement exploitée. Le plus souvent, le chercheur se contente d'identifier les deux ou trois variables qui possèdent les plus fortes saturations, et de donner un nom à leur association. Ce nom même est bien souvent inspiré directement de celui de la variable qui est liée le plus étroitement au facteur, et l'on retombe ainsi dans l'interprétation grossière. L'idéal serait une combinaison des deux méthodes.

Nous allons avoir l'occasion de présenter quelques exemples de ce type en examinant maintenant les applications de l'analyse factorielle aux études thématiques.

LA CONTRIBUTION DES MODELES FACTORIELS

Longtemps avant HARRIES, que nous avons cité plus haut, SCHMID (1960) s'était penché sur le problème de la criminalité dans les villes. Cependant, le problème est posé de façon très différente. I1 ne s'agit plus de chercher à connaître l'influence que peut avoir la taille d'une ville ou sa situation géographique sur l'incidence de 1a criminalité ; il s'agit de déterminer de la façon la plus significative possible les zones de crime dans une grande communauté urbaine, en termes économiques, démographiques et sociaux.

Pour servir ce but, SCHMID a choisi la méthode d'analyse factorielle dont l'utilisation commençait à se développer dans les sciences sociales. Cela était justifié en raison de l'abondance des données dont il disposait : 65 000 cas d'après les fichiers de la police, regroupés en vingt indices de crime[12] et 18 attributs économiques démographiques et sociaux, d'après le recensement de 1950, soit trente huit variables pour 93 census tracts.

La démarche de l'analyse est classique. Une analyse factorielle en axes principaux est faite à partir de la matrice des corrélations des 38 variables. Les 8 premiers facteurs sont retenus parce qu'ils rendent compte de la presque totalité de la variation commune contenue dans les données. Huit nouveaux facteurs sont obtenus après une rotation orthogonale des huit axes principaux. Il est intéressant de noter que ce n'est pas avec les deux premiers facteurs, qui représentent respectivement un *"faible statut familial"* et un *"faible statut professionnel",* que la plupart des variables de criminalité sont le plus étroitement liées, mais avec le troisième facteur, où les plus bas niveaux de statut familial et de statut économique se combinent.

Les autres facteurs, à part le septième qui est une réplique du premier et parait donc très ambigu, sont satisfaisants, dans la mesure où ils définissent bien des groupes particuliers de variables. Le quatrième facteur, appelé *"mobilité de la population"* est assez étroitement lié aux activités de suicide, fraude bancaire, vols avec effraction … Le cinquième facteur regroupe des *"crimes non-typiques"* : vols de vélos, attentats à la pudeur, qui ne sont pas associés à des variables socio-économiques particulières. Le sixième facteur intitulé *"faible mobilité de la population",* n'inclut aucune variable de crime. Le huitième facteur associe la population noire et les vols non-résidentiels.

Aucun de ces facteurs n'est contre notre attente, et dans ce sens, on ne peut pas dire que l'analyse ait vraiment apporté quelque chose de nouveau à l'étude de la criminalité dans les villes. Pourtant elle permet de réfléchir sur ce problème et d'aller plus loin dans sa compréhension. L'étape suivante, que propose SCHMID guidé par ses premiers résultats, est d'établir une relation entre le crime et la typologie urbaine d'après SHEVKY (1955). Cette typologie, nous le rappelons, classe les zones intra-urbaines d'après leur *"statut familial",* leur *"statut économique",* et leur *"ségrégation".* Ainsi, SCHMID trouve qu'il y a une corrélation, d'ordre zéro, inverse entre la Ségrégation et le Suicide à Seattle et à San-Francisco. Peut-on en conclure que la ségrégation a tendance à défavoriser le suicide ? SCHMID montre que non ; car,

[12] Ce terme est employé par SCHMID au sens large d'activité répréhensible car il inclut, outre les vols, les meurtres et les attaques à main armée, les suicides, les fraudes fiscales, etc ....

lorsqu'on corrèle la ségrégation et le suicide en maintenant constant le statut familial (corrélation partielle d'ordre 1), on constate alors que la corrélation reste négative, quoique moins forte à Seattle, mais qu'elle devient positive à San-Francisco. Finalement, lorsque SCHMID effectue une corrélation d'ordre 2 entre la ségrégation et le suicide, en maintenant constants le statut familial et le statut économique, l'inégalité entre les deux villes apparaît plus nettement, avec une corrélation négative à Seattle et une corrélation positive encore plus forte à San-Francisco. On a ici un aperçu de la richesse qu'on peut tirer d'une telle analyse, lorsqu'on ne se contente pas de donner une appellation concrète plus ou moins exacte aux êtres mathématiques abstraits que sont les facteurs mais lorsqu'on se sert des associations qu'on découvre pour affiner les résultats au moyen d'autres méthodes.

BERRY, dont on a du mal aujourd'hui a suivre le *"compte des analyses factorielles* ", ainsi que 1e compte de celles de ses disciples du Département de Géographie de Chicago a, il faut le reconnaître, donné l'exemple d'études originales qui, en raison de leur succès même, ont été beaucoup imitées par la suite, souvent malheureusement de façon incomplète ou peu pertinente. En 1961, BERRY a effectué une étude sur *Les structures de base du développement économique*, pour accompagner un atlas de la répartition du développement économique dans le monde (BERRY, 1961), Bien qu'on puisse la considérer aussi comme une classification de régions, il nous paraît intéressant de la présenter dans le cadre des études thématiques. En fait, le problème est bien thématique puisqu'il s'agit de savoir, en comparant plusieurs pays sur quarante trois variables économiques, démographiques et spatiales (accessibilité), quelles sont les conditions qui conduisent au développement économique, ou le favorisent.

L'une des originalités de l'étude tient dans l'utilisation des rangs de chacun des 95 pays sur chacun des 43 indices, plutôt que des valeurs précises des indices pour chaque pays. Le fait de ne tenir compte que des positions relatives indique déjà un souci de généralisation que BERRY omet d'exposer de façon explicite. La seconde originalité vient de ce qu'une analyse factorielle directe est effectuée sur ces données ordinales. BERRY utilise ce procédé essentiellement dans le but d'éliminer la redondance qui existe parmi les quarante trois variables. Ainsi, il considère que les quatre premiers facteurs qu'il retient résument l'essentiel de ce que toutes ces variables ont en commun (Table 1).

Indices	Basic Pattern Values			
	1	2	3	4
1. Kms. railroads p.u.a.*	148	— 4	25	—34
2. Kms. railroads p.c.†	146	—14	—10	24
3. R.r. freight ton-km. p.y.p.c.‡	149	—10	28	21
4. Ton-km. freight p. km. r.r.	141	12	42	15
5. Km. roads p.u.a.	146	—10	20	—36
6. Km. roads p.c.	141	—16	—13	25
7. Motor vehicles p.c.	147	—31	—35	5
8. Motor vehicles p.u. road	142	— 3	—35	—16
9. Motor vehicles p.u.a.	149	—23	— 3	—36
10. Value for. Trade	150	— 4	24	11
11. For. trade p.c.	145	—28	—44	5
12. Exports p.c.	149	0	24	17
13. Imports p.c.	151	— 5	26	13
14. % Exports to N. Atlantic	132	38	21	—17
15. % Exports raw materials	147	2	3	—14
16. Kw-h. electricity p.c.	152	—30	— 8	7
17. Energy cons. in kw-h.	147	14	45	17
18. Energy cons. p.c.	151	—27	08	11
19. Comml. energy p.c.	153	—28	— 6	6
20. % Energy cons. comml.	157	—14	—10	2
21. Energy res. in kw-h.	142	25	29	47
22. Energy res. in kw-h.p.c.	140	13	— 6	50
23. % Hydro pot. dev.	148	— 9	17	—12
24. Hydro pot. dev. p.c.	147	— 8	1	8
25. Fiber cons. p.c.	152	—25	—13	1
26. Petrol. ref. capc. p.c.	143	8	8	10
27. Pop. density	138	26	30	—61
28. Crude birth rates	123	81	—46	— 1
29. Crude death rates	129	74	—12	12
30. Pop. growth rates	127	64	—57	8
31. Infant mort. rates	118	94	—17	6
32. % Pop. in cities >20.000	151	—20	—14	— 6
33. Physicians p.c.	150	—28	—10	— 4
34. % Land area cultiv.	137	25	28	—40
35. Wheat yields	147	— 6	25	10
36. Rice yields	129	54	—11	— 5
37. Pop. p.u. cultiv. land	134	28	— 6	—40
38. Newspaper circ. p.c.	151	—30	—14	— 5
39. Telephones p.c.	151	—35	—18	— 2
40. Mail flows p.c. domestic	151	—27	— 3	— 1
41. Mail flows p.c. internat.	144	—26	—33	—12
42. National product	148	11	40	12
43. National product p.c.	151	—30	—19	1

* p.u.a. = per unit area.
† p.c. = per capita.
‡ p.y. = per year.

Table 1 – Valeur de 43 indices sur 4 structures de base. (Source : BERRY, 1961)

Le premier facteur ou *"première structure de base"*, dégage le plus fort effet commun aux 43 indices. Il représente l'association, sur une seule dimension, des indices d'accessibilité, de transport, de commerce, de relations extérieures, de technologie, d'urbanisation, de produit national. En raison de la relative importance des indices de technologie, cette dimension est appelée *"échelle technologique »* (Fig. 1) :

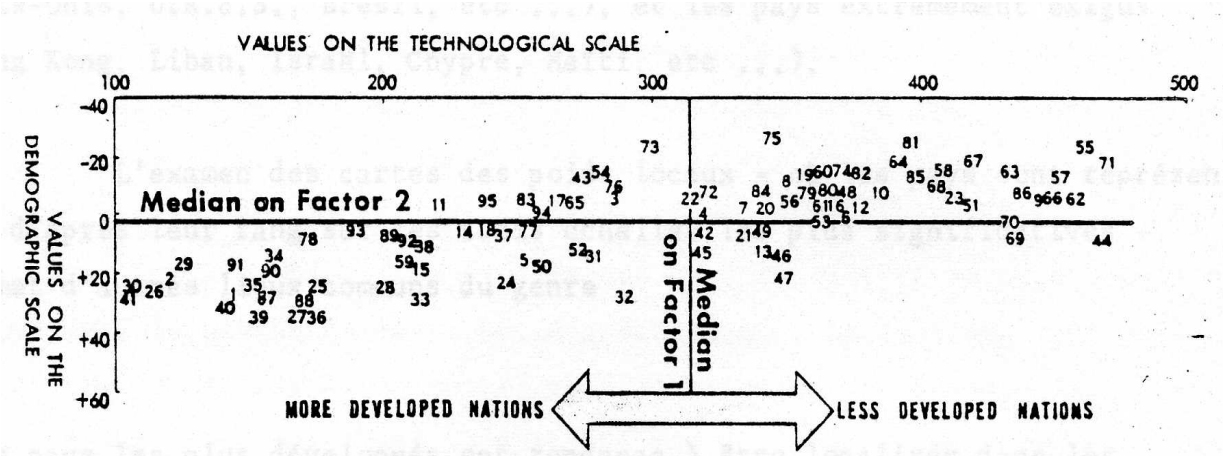


Figure 1 – Echelle de développement économique et démographique
(Source : BERRY, 1961)

Le calcul des poids locaux permet de voir que les pays apparaissent distribués de façon régulière le long de cette échelle. On ne peut distinguer aucun regroupement particulier de pays ayant le même rang. BERRY en conclut qu'il n'existe pas de *« groupes naturels »* de pays développés ou sous-développés ; mais il oublie de rappeler que cet effet de continuité est dû essentiellement à l'utilisation de données ordinales.

Le deuxième facteur comprend un groupe d'indices qui sont peu représentés dans l'échelle technologique ; ce sont ceux qui touchent la population des pays - taux de natalité et de mortalité, mortalité infantile, taux d'accroissement - ainsi que des indices associés négativement au commerce extérieur par habitant, à la consommation d'énergie par habitant, etc ... Cette structure révèle donc essentiellement les effets de la pression exercée par la population ; et on lui a attribué pour cette raison le nom d'*échelle démographique* (Fig. 1). De la même manière que pour l'échelle technologique, et pour la même raison, les pays se répartissent régulièrement sur cette échelle.

On passera rapidement sur la description des troisième et quatrième facteurs, car notre but n'est pas ici de reprendre en détail l'étude de BERRY ; il nous suffit de savoir pour notre propos que le troisième facteur représente *"le contraste"* entre des pays au produit national élevé et aux relations extérieures très actives, comme certains pays des Caraïbes, et des pays où c'est l'inverse que l'on rencontre, comme les pays d'Asie et le bloc soviétique. Quant au quatrième facteur, il ne fait que souligner les différences évidentes qui opposent les pays de grande étendue (Canada, Etats-Unis, U.R.S.S., Brésil, etc ...*),* et les pays extrêmement exigus (Hong Kong, Liban, Israël, Chypre, Haïti, etc).

L'examen des cartes des poids locaux - où les pays sont représentés d'après leur rang sur les trois échelles les plus significatives - permet d'autres lieux communs du genre : « les pays les plus développés ont tendance à être localisés dans les latitudes moyennes » ; ou encore : « les pays les moins développés ont une économie commerciale spécialisée (industrielle pour les plus développés d'entre eux) », ce qui ne fait que reprendre la définition même du développement économique et n'apporte rien à notre connaissance initiale.

Ces généralisations paraissent bien élémentaires et l'auteur le reconnaît en ces termes :

> « Ces schémas ne sont pas des nouveautés, d'autres personnes les ont décrits en d'autres temps et d'autres lieux, quoique peut être pas d'une manière aussi explicite et générale que nous l'avons fait ».

Pourtant, en passant par de telles généralisations, on peut finalement dégager les exceptions et proposer une explication plus nuancée de la répartition du développement économique, et peut-être de façon plus claire, qu'en commençant par examiner toutes les nuances et les cas particuliers. C'est là qu'intervient pour le chercheur l'utilisation d'autres méthodes que l'analyse factorielle. BERRY utilise une méthode de régression pour tenter de tester la validité des généralisations qu'il vient de faire. Ainsi, en prenant comme variables dépendantes, chacune à leur tour, l'échelle technologique, l'échelle démographique et l'échelle mesurant le rapport (produit national / activité extérieure), il essaie de voir jusqu'à quel point on peut prédire la position d'un pays sur ces trois échelles en utilisant de simples informations binaires (0 ou 1) sur leurs caractéristiques économiques, leur statut politique, leur localisation, leur climat, etc .. soit en tout seize variables

La première régression, qui concerne la position d'un pays sur l'échelle technologique, montre que cette position varie avec le type d`économie et la région ; mais pas avec le statut politique, et que si la localisation dans Ies régions de latitude moyenne est significative, celle dans les pays tropicaux ou équatoriaux ne l'est pas et ne peut déterminer à coup sûr le niveau de développement économique. La deuxième régression sur l'échelle démographique montre que les colonies, ou ex-colonies, sont relativement mieux partagées au point de vue de la pression démographique que les pays indépendants, de structure par ailleurs très voisine. La troisième régression n'apparaît pas assez significative pour qu'on en mentionne ici les résultats.

Ainsi, on voit comment, dans le but de déterminer les conditions du développement économique dans le monde, BERRY a recours en fin de compte - comme MORRILL et WOLHEMBERG, dans le cas parallèle de la pauvreté dans une nation - à l'analyse par régression. La différence essentielle entre les deux approches vient de ce que BERRY utilise auparavant l'analyse factorielle. Cela lui permet, d'une part, d'incorporer davantage d'information (43 variables résumées en trois *"structures de base"),* au lieu de restreindre le concept de développement économique à une seule variable - comme le revenu familial dans le cas du concept de la pauvreté - et, d'autre part, par l'intermédiaire de la cartographie des poids locaux, il peut établir ou plutôt réaffirmer quelques grandes généralités qu'il cherche à nuancer ensuite par la méthode de régression.

COX (1968) a examiné le comportement politique électoral dans la ville métropolitaine de Londres. C'est un problème géographique dans la mesure où on a remarqué, depuis longtemps, qu'il existe une opposition entre le centre de la ville et la banlieue concernant les préférences politiques et l'activité électorale. A quoi tient exactement cette opposition ? Peut-on évaluer le rôle relatif que jouent la *"banlieurisation"* et d'autres phénomènes socio-économiques dans l'explication de l'attachement à une politique et de la participation électorale ? Pour répondre à ces questions, COX construit un modèle causal sur la base des résultats d'une analyse factorielle en axes principaux, ayant subi une rotation *"varimax"*.

Etant donnée la qualité théorique des variables introduites dans l'analyse, ces résultats montrent que quatre facteurs suffisent (avec près de 90 % de la variance expliquée), pour définir les aspects sociaux et géographiques de la ville de Londres qui devaient être liés au comportement électoral de la population et à la notion de *"banlieue"*. D'après les fortes

saturations de certaines variables (classes sociales, éducation, profession) sur le premier facteur, celui-ci est considéré comme un facteur de rang social. Le second facteur est une dimension qui oppose le centre-ville et la banlieue. D'après ce facteur, la banlieue type est une zone localisée à une grande distance du centre des affaires, avec une densité relativement faible, mais un fort accroissement de population, une forte proportion d'hommes et de femmes mariées, une faible proportion de femmes qui travaillent, etc ... Le troisième facteur regroupe les variables qui caractérisent l'importance de la commutation ville-banlieue ; et le quatrième facteur décrit la structure par âge, en opposant la population de jeunes adultes (15 à 44 ans) et la population au-delà de 65 ans.

L'étape suivante consiste à établir des corrélations simples (d'ordre zéro) entre les quatre facteurs et quatre variables de comportement électoral, qui sont : les pourcentages de votes conservateurs en 1950 et en 1951, et les pourcentages de participation en 1950 et en 1951. Le résultat suggère que les deux aspects de l'activité politique (votes conservateurs et faible participation) sont directement liés à la localisation en banlieue mais qu'il existe aussi une influence indirecte des autres dimensions opérant à travers l'opposition banlieue/centre-ville. Cela permet à COX de proposer une première hypothèse qui voudrait que les différences de comportement politique soient dues aux différences sociales (revenu, éducation, occupation, etc *...)* résumées dans la dimension de rang social. Or, lorsque cette dimension est contrôlée dans une analyse de corrélations partielles, la tendance conservatrice de la banlieue est maintenue ; d'où la nécessité d'émettre une nouvelle hypothèse. La localisation en banlieue pourrait influencer le comportement électoral de deux manières : 1° - par la conversion, du fait de leur nouvel environnement, d'ouvriers anciens partisans du parti travailliste, ayant habité le centre-ville ; 2° - par la migration sélective des conservateurs du centre-ville vers la banlieue, les uns comme les autres ayant gardé les caractéristiques sociales généralement attachées à leur ancienne localisation. Ainsi, cela se passerait bien indépendamment de toute amélioration du statut social.

I1 est certain que ces dernières hypothèses restent encore à être testées. Mais on voit qu'en un petit nombre d'étapes logiques, où l'analyse factorielle joue un rôle non négligeable, on arrive à mieux connaître les mécanismes écologiques qui conduisent à un certain comportement humain.

La prochaine étude que nous avons choisi de présenter est une application originale de l'analyse en composantes principales à la recherche des structures de préférence spatiale. GOULD (1965) est l'instigateur de ce type d'études. I1 a récemment regroupé les résultats de ses recherches et de celles de ses collaborateurs dans un ouvrage "sur les Cartes mentales" : *On Mental Maps* (GOULD and WHITE, 1974). Cet ouvrage, adressé aux non spécialistes comme aux spécialistes, réussit le pari de pouvoir réellement intéresser les deux. Clarté, simplicité et humour le rendent de lecture facile et agréable pour les non spécialistes. La conjonction d'un fourmillement d'idées, d'un grand éventail de méthodes allant des plus *"qualitatives"* aux plus *"quantitatives",* et d'exemples pris dans le monde entier, ouvre un vaste champ d'études qui promet d'être extrêmement fertile pour les spécialistes.

La démarche commune à toutes ces études est la construction d'une*"carte mentale",* à partir de l'image que les gens ont d'une certaine portion de territoire. Bien sûr, il est reconnu que la carte mentale de chaque personne est unique ; mais il peut y avoir une concordance importante entre les cartes mentales de plusieurs personnes. La *"cartographie homomorphique"* permet de construire une seule carte représentative de ce qu'il y a de commun dans tous les points de vue individuels. Plus le groupe est homogène en termes d'âge, d'expérience, de lieu de résidence, plus on s'attend à ce que la concordance entre les différentes images mentales soit grande. Dans *The mental maps of British School Leavers* (1968), l'âge et l'expérience sont

considérés comme identiques et les cartes mentales représentent l'image préférentielle des comtés britanniques vue à partir de différentes localisations. La même expérience a été répétée pour les différents états des Etats-Unis. *The Perception of Residential Desirability in the Western Region of Nigeria* (GOULD, 1970) montre comment les préférences changent avec l'âge et l'expérience. Dans l'exemple suédois *The Black Boxes of Jönköping* (GOULD, 1972), les deux situations précédentes sont combinées. Il serait trop long, malheureusement, de citer tous les exemples. En ce qui concerne le côté méthodologique qui, essentiellement, nous préoccupe ici, il nous a paru intéressant de présenter "*La structure préférentielle spatiale de la Tanzanie*" (GOULD, 1969-a). Cette étude offre un intérêt méthodologique particulier car on peut y évaluer les effets de l'usage de différentes échelles de mesure dans une analyse en composantes principales.

A des étudiants placés dans l'éventualité de l'obtention d'un poste dans un des 60 districts du pays, on a demandé de classer ces districts par ordre de préférence, en leur assignant un nombre de 1 à 60. A partir de ces mesures ordinales, une matrice de corrélation est construite entre les participants, pour être soumise à une analyse en composantes principales. La première composante principale constitue une dimension générale où se regroupent tous les participants qui sont le plus d'accord entre eux. Les scores de poids locaux de chaque district sur cette dimension permettent, par l'interpolation linéaire, de construire une carte d'isolignes (Fig. 2).

La *"surface de préférence"* ainsi obtenue est assez gondolée, présentant des pics aux points les plus désirables - en général les plus grandes ville des plaines-, d'indifférence, et des vallées de répulsions dans les campagnes les plus éloignées ou isolées. Une telle carte pourrait permettre au Gouvernement de développer une politique de salaires compensatoires pour attirer les gens dans les zones les moins prisées. Cela se fait déjà bien entendu, mais de manière plus ou moins logique, ou d'après le point de vue de quelques personnes - gouvernants ou planificateurs - et non pas en tenant compte de l'avis des intéressés eux-mêmes qui, en fin de compte, sont les seuls qui pourraient rendre cette politique efficace ; car, comme le dit justement GOULD :

> « L'isolement est une construction mentale qui ne s'exprime pas obligatoirement en termes simples, comme la côte arctique ou le désert australien »

Dans une seconde expérience, par souci d'approfondir l'étude, il a été demandé aux étudiants de Tanzanie de placer le chiffre correspondant à un salaire annuel de base dans le district préféré et, dans les autres districts, celui de salaires aussi élevés qu'ils le jugent nécessaire pour compenser le manque d'attrait. En conséquence, l'intensité du choix, absent dans la première étude, peut être nuancée.

A partir de ces mesures cardinales, une matrice de corrélation, que l'on soumet ensuite à une analyse en composantes principales, est établie entre les participants. Or, les résultats dérivés de ces mesures cardinales, théoriquement plus subtiles et devant contenir davantage d'informations, sont décevants. La portion de la variation totale expliquée par la première composante décroît considérablement par rapport à la première expérience.

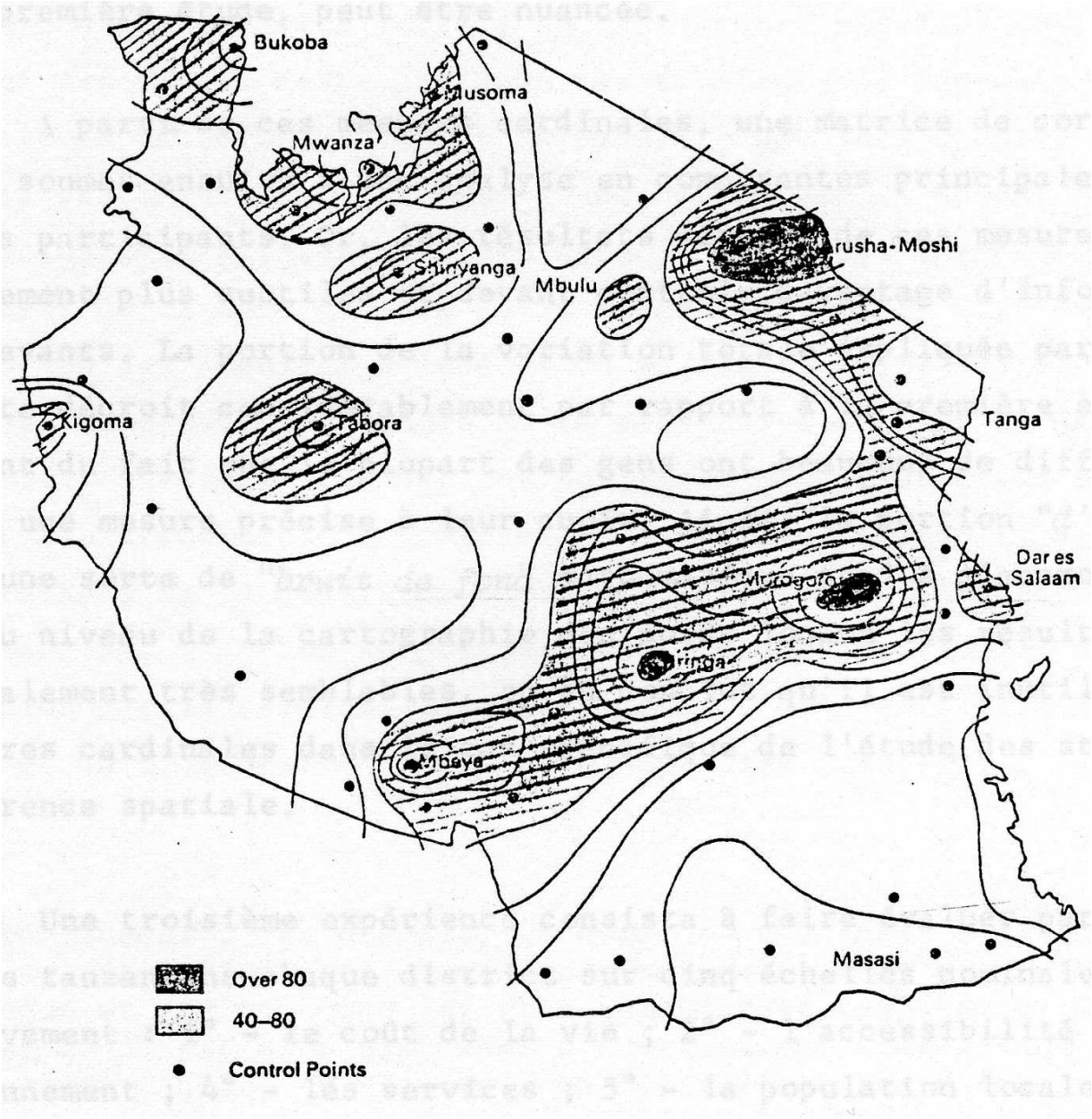


Figure 2 – Carte mentale des étudiants de Tanzanie.
(Source : GOULD, 1974)

Cela vient du fait que la plupart des gens ont beaucoup de difficultés à donner une mesure précise à leur choix. Ainsi, la portion *"d'indifférence"* formant une sorte de *"bruit de fond"* mathématique tend à s'accroître. Et puisqu'au niveau de la cartographie des poids locaux, les résultats apparaissent finalement très semblables, on en conclut qu'il est inutile d'employer des mesures cardinales dans le cas spécifique de l'étude des structures de préférence spatiale.

Une troisième expérience consista à faire évaluer par les étudiants tanzaniens chaque district sur cinq échelles nominales mesurant respectivement : 1° - le coût de la vie ; 2° - l'accessibilité ; 3° - l'environnement ; 4° - les services ; 5° - la population locale. Pour ne prendre qu'un exemple, soit le coût de la vie, voici comment se présente cette échelle de mesure nominale :

Coût de la vie très élevé : il est très difficile d'économiser sur la salaire	+3 +2 +1 0 -1 -2 -3	Coût de la vie très bas : il est facile d'économiser à partir du salaire

Les résultats de l'analyse en composantes principales, dans ce dernier cas, confirment les préférences spatiales établies par les deux premières expériences. Dans l'ensemble, les cinq échelles sont très étroitement liées et elles ont de fortes saturations avec la première composante. Cependant, la deuxième composante représente uniquement la population locale, qui est moins fortement associée aux quatre autres échelles. Sur cette échelle, la cartographie des poids locaux représente à quel point les étudiants se sentent en accord avec la population locale de tel ou tel district. Il apparaît qu'en moyenne, les scores sont très faibles, et que les districts de scores négatifs se trouvent le long de la côte, là où l'influence arabe est très forte, ou bien dans l'extrême-ouest du pays, dominé par les Masaï.

Cette attitude des étudiants de Tanzanie vis-à-vis de populations différentes risque de paraître peu surprenante. Cependant elle n'est pas inévitable. La même étude, effectuée au Ghana, montre que la population locale intervient peu dans le choix des Ghanéens. On a montré aussi de la même manière qu'il existe une grande différence dans la perception des population locales de Malaisie, entre les étudiants malais et les étudiants chinois de la même université de ce pays. Cela permet d'envisager comment une politique gouvernementale pourrait s'adapter aux structures de préférences spatiales. Lorsqu'il apparaît que les raisons de la répulsion ou du choix sont déterminées par les populations locales, l'action du gouvernement ne peut se limiter à une simple compensation par le salaire. Il est évident qu'elle doit être alors plus subtile, et qu'elle doit répondre à de nouvelles questions :

> "Lorsqu'en Tanzanie, par exemple, un homme ou une femme - disons un agronome ou une infirmière - veut avoir un poste dans sa région d'origine où la langue et les coutumes lui sont commodes et familières, doit-on considérer qu'il est plus efficace à court terme, travaillant et servant dans sa propre région, ou bien doit-on fixer l'horizon d'une planification à long terme vers une:nation forte et unifiée ?" (GOULD and WHITE, 1974, p. 170).

Nous n'insisterons pas davantage sur les intérêts multiples que peut présenter l'étude des structures de préférence spatiale. Quant à la méthode elle-même, mise au point pour ce type d'étude, on a vu que l'analyse en composantes principales y jouait un rôle fondamental et pourrait très bien s'accommoder d'un domaine essentiellement qualitatif. Effectivement, il s'avère que les mesures cardinales les plus rigoureuses n'ajoutent rien aux résultats obtenus grâce aux mesures les plus qualitatives ordinales et nominales.

D'autres méthodes ont été employées pour étudier les phénomènes de préférence spatiale et de perception de l'environnement. Nous avons vu, plus haut, comment BROWN et LONGBRAKE (1970) ont essayé d'établir une fonction *"d'utilité du lieu",* à l'intérieur d'une ville, par l'intermédiaire de la régression multiple ; et comment, dans le même cadre intra-urbain et avec la même méthode, CLARK (1973) a cherché à définir les types de pression qui sont liée à la mobilité individuelle. Nous en avions conclu que l'utilisation d'unités administratives trop vastes et d'un nombre de variables assez restreint, introduites dans un ordre arbitraire, ne permettait que des résultats de portée limitée. Là encore, une combinaison de l'analyse de régression multiple et d'un modèle factoriel pourrait éclairer plus complètement le problème.

GOULD (1967) a effectué une intéressante expérience dans ce sens. Des scores de préférence résidentiel pour chacun des états des Etats-Unis, établis selon le procédé que nous

venons de décrire, ont été introduits dans une analyse de régression avec comme variable dépendante un indice d'assistance sociale. Ses conclusions ($R^2 = 0,61$) indiquent que les cartes mentales, représentatives du groupe étudié, reflètent assez bien les mesures d'assistance. En outre, les principales valeurs résiduelles permettent d'intéressants commentaires sur la surestimation des "images" de certains états.

Une autre méthode, très voisine de l'analyse factorielle, spécialement adaptée par les psychologues TORGERSON (1952 et 1958), SHEPARD (1962) et KRUSKAL (1964), aux études du comportement, jouit depuis quatre ou cinq ans d'une certaine faveur auprès des géographes anglo-saxons intéressés par la perception de l'environnement. I1 s'agit du *"multidimensional Scaling " ou "analyse multidimensionnelle des proximités".* Cette technique, dont nous ne voulons pas exposer ici tous les mécanismes, traite les données obtenues à partir de la ressemblance ou *"proximité"* de certains objets, plutôt que les données classiques composées de magnitudes variables d'attributs spécifiques. Ainsi, les objets sont localisés par les personnes interrogées en termes ci-après : "plus proche que", "plus grand que", "plus semblable à", "préféré à", dans un espace *"psychologique"* qui n'est pas forcément euclidien, c'est à dire, par exemple, qui ne vérifie pas l'inégalité triangulaire.

Le but de l'analyse multidimensionnelle est de retrouver, à partir de ces données non métriques, une configuration spatiale des points dans un espace identifiable, de dimension minimum. Elle est donc particulièrement adaptée aux données de choix, de préférence spatiale, et plus généralement de perception de l'environnement. Parmi les applications dans ce domaine, on peut citer : les préférences spatiales vues à travers les migrations entre régions (SCHWIND, 1971) et vues à travers l'attrait résidentiel des différents états des Etats-Unis (GOULD, 1969-b) ; la comparaison des distances perçues et réelles dans un espace intra-urbain (GOLLEDGE, 1969), (MARCHAND, 1974).

La différence essentielle avec les modèles factoriels, qui fait toute l'originalité et l'intérêt de l'analyse multidimensionnelle des proximités, concerne l'adoption d'une hypothèse plus faible (l'hypothèse de *"monotonité"*[13]), dans la transformation des mesures psychologiques en mesures réelles. Cependant, tout comme les autres méthodes métriques d'analyse multivariée, l'utilisation de l'analyse multidimensionnelle pose le problème fondamental de l'interprétation des dimensions.

La dernière étude dont nous parlerons dans le cadre des études thématiques : "*Economic fluctuation in a Multiregional Setting*" (JEFFREY, CASETTI and KING, 1969) est intéressante à plusieurs titres :

- 1° : c'est le seul exemple que l'on connaisse d'application de l'analyse factorielle à des séries temporelles (analyse de type S)
- 2° : elle utilise une approche originale de l'analyse factorielle connue sous le nom de *"bi-factor analysis"* ou analyse bi-factorielle.
- 3° : son but consiste explicitement à vérifier une hypothèse bien définie, ce qui est rare en géographie
- 4° : et enfin, elle ne s'appuie pas uniquement sur l'analyse factorielle et comprend l'utilisation de méthodes annexes.

L'approche bi-factorielle, mise au point par HOLZINGER (1941) il y a plus de trente ans, suppose une théorie selon laquelle il y aurait un facteur commun à l'origine de la variation

[13] Le premier objectif de la contrainte de monotonité est d'assurer que l'ordre des distances entre les points, dont la configuration est obtenue à partir des mesures initiales, est identique à l'ordre des mesures initiales de dissemblance (ou à l'ordre inverse des mesures de proximité).

contenue dans les données, tout le reste de la variation étant imputé à des facteurs spécifiques correspondant à des groupes de variables qui ne se recouvrent pas. De ce fait, la complexité de chaque variable peut être décomposée en deux parties (d'où le nom de la méthode) : l'une "expliquée" par le facteur général ; l'autre "expliquée" par le facteur du groupe particulier auquel appartient la variable.

Pour JEFFREY et ses co-auteurs, il s'agit de vérifier l'hypothèse qu'un ensemble de séries temporelles est fait d'influences provenant d'échelles spatiales différentes - nationales, régionales, sous-régionales et locales - en utilisant comme données les taux de chômage mensuels de trente villes du *"Mid Wes*t" américain entre Mai 1960 et Septembre 1964. La méthode choisie implique d'abord que soient retirées les fluctuations attribuées aux facteurs nationaux. Il suffit pour cela de régresser les séries temporelles de chaque ville sur les séries temporelles de chômage national. Les séries résiduelles sont ensuite corrélées et, dans une troisième étape, la matrice des corrélations est soumise à un algorithme de classification (*linkage analysis*)[14], pour que des groupes bien distincts de villes, dont les séries sont similaires, soient formés. Les cinq groupes obtenus correspondent à un découpage en cinq régions, dont les centres sont Chicago, Détroit, Indianapolis, Cleveland et Pittsburg.

L'hypothèse de départ est donc que chaque série temporelle résiduelle contient un facteur général, un facteur correspondant à chacun des cinq groupes régionaux, et un facteur unique spécifique de chaque ville. Les saturations des villes sur ces différents facteurs sont calculées (Table 2), à partir desquelles on construit une nouvelle matrice de corrélation. Les différences entre la matrice de corrélation initiale et la matrice reconstruite sont obtenues et un test montre qu'elles sont très proches de zéro. Ainsi, l'hypothèse de départ est-elle vérifiée.

Nous ne reviendrons pas sur les mérites de cette étude que nous avons énumérés avant de la présenter ; mais nous nous placerons à un niveau plus général pour conclure avec JEFFREY et ses collaborateurs :

> « l'analyse bi-factorielle diffère des études factorielles que l'on trouve en géographie jusqu'à présent, en ce qu'elle vise à tester la consistance de certaines hypothèses avec un ensemble de données, plutôt que d'identifier des dimensions inconnues, souvent confuses ».

[14] Dans le chapitre suivant sur les études de régionalisation, nous aurons l'occasion de parler plus en détail de quelques méthodes de classification

City	General Trend	FACTORS					
		Cleveland group	Pittsburgh group	Detroit group	Indiana group	Chicago group	Unique
Akron	.78	. 36					. 51
Cincinnati	.44	.58					.68
Dayton	.84	.30					.43
Hamilton	.44	.64					.62
Cleveland	.82	.42					.39
Columbus	.70	. 40					. 59
Lorain	.72	.41					.56
Youngstown	.64		.66				.40
Steubenville	.44		.70				.57
Wheeling	.68		.56				.46
Pittsburgh	.69		.64				.33
Gary	.68		. 66				. 31
Toledo	.57			.31			. 76
Detroit	.77			.46			.44
Flint	.36			.74			.56
Lansing	.39			.84			.36
Saginaw	.71			.65			.25
Evansville	.79				.53		.29
Indianapolis	.87				.43		.23
Terre Haute	.82				.44		.36
Louisville	.75				.52		.39
Chicago	.63					.48	. 60
Davenport	.62					. 46	. 64
Peoria	.78					.37	.50
Rockford	.85					.32	.54
Grand Rapids	.61					.48	.63
Kalamaz oo	.66					. 40	. 64
Muskegon	.64					.45	.62
Milwaukee	.51					. 43	.74
Racine	.72					.53	.46
Contribution to total Variance	48.32%	4.96%	6.93%	6.69%	3.10%	5.80%	24.20%

Table 2 – La structure bi-factorielle
(Source : JEFFREY, CASETTI & KING, 1969)

CONCLUSION

On pense généralement que le rôle de l'analyse factorielle se limite à substituer à un grand ensemble de variables un ensemble plus petit de facteurs. Par exemple, certains chercheurs ayant à tester des hypothèses concernant la relation de variables spécifiques (pauvreté, crime, espacement des agglomérations) et qui de ce fait n'ont pas besoin de traiter simultanément de grands ensembles

de variables, pensent qu'ils n'ont pas de raison d'employer l'analyse factorielle et que certaines techniques comme la régression et la corrélation multiple sont plus appropriées.

Nous avons vu quelques exemples d'utilisation de ces méthodes dans des situations inférentielles, et nous avons remarqué que l'interprétation causale de corrélations écologiques en termes de certaines variables spécifiques n'est pas toujours convaincante. Cela vient de ce qu'on s'attend implicitement à certaines relations causales entre des variables théoriques dont les variables observées ne sont que des mesures imparfaites. Ainsi bien souvent, les corrélations de variables spécifiques simplifient arbitrairement une situation complexe.

Qui plus est, il existe un problème de multi-colinéarité entre les variables qui empêche de connaître la part exacte de telle ou telle variable dans la relation. Or, puisque 1e facteur est un agrégat pondéré de corrélations partielles entre des variables primaires, les inter-corrélations entre ces variables ont déjà été prises en compte en extrayant les facteurs. Cela suffit a justifier l'utilisation de facteurs dans des analyses inférentielles soit comme "causes" de la cohésion de certaines variables primaires, soit comme variables composites indépendantes qui peuvent être l'objet d'études inférentielles ultérieures. Par exemple, BERRY extrait une échelle technologique et une échelle démographique, à la fois mesures et "causes" du développement économique de divers pays du monde. GOULD extrait une vue commune de la désirabilité résidentielle qui est à la fois la manifestation et l'explication synthétique de ce phénomène. KING a pu tester l'hypothèse que le chômage est fonction de facteurs nationaux, régionaux et locaux relativement indépendants. Enfin, les mêmes auteurs, ainsi que par exemple SCHMID et COX, ont cherché à tester des hypothèses engendrées par les premiers résultats d'une analyse factorielle en utilisant les facteurs dans des analyses de corrélations partielles ou de régression multiple.

Ces applications soulèvent plusieurs problèmes techniques comme celui de l'utilisation de différentes échelles de mesure. Nous avons consacré une place importante au problème des corrélations écologiques fallacieuses car à notre sens, ce problème a été généralement sous-estimé par les géographes anglo-saxons. Inhérent à l'utilisation de mesures ayant un support spatial de taille modifiable, il est présent dans les applications des méthodes d'analyse multivariée qui ont des buts explicatifs ; donc il touche particulièrement les analyses thématiques inférentielles comme celles que nous venons d'examiner. Il intervient aussi, nous le verrons, dans les études régionales et urbaines dans la mesure où ces études ne s'arrêtent pas à la description ou à la classification d'objets mais proposent des explications.

CHAPITRE II

LES ETUDES DE REGIONALISATION

En géographie, toute étude se réfère au moins implicitement à l'espace. En ce sens elle peut être qualifiée de régionale ou « d'écologique », pour reprendre le terme qu'utilisent de préférence les anglo-saxons.

Dans notre esprit, régionalisation évoque plus précisément l'idée de classification, c'est à dire de partition de l'espace en types définis. Or, d'après le groupe CHADULE (1974), dont on reprendra les définitions, il existe trois catégories de partition de l'espace.

- 1° . la "délimitation d'aires d'extension", qui suppose une classification monothétique c'est à dire fondée sur un seul critère ;

- 2° : le "zonage" qui se fait par des procédés de classification polythétique et qui correspond au découpage de l'espace en portions homogènes selon plusieurs critères ;

- 3° : la "régionalisation" proprement dite, dont l'objectif difficile à atteindre est la région, "conçue comme un système cohérent, structuré, multiforme" (CHADULE, 1974).

A notre sens, l'utilisation du modèle factoriel peut intervenir à divers degrés dans les trois cas. En effet, l'analyse en composantes principales et l'analyse factorielle qui permettent de transformer des données complexes linéairement interdépendantes, ou au moins distinctes, peuvent être considérées en elles-même comme des méthodes efficaces de classification polythétique. L'analyse de type n regroupe directement les observations en zones possédant une combinaison commune de caractéristiques. L'analyse de type R regroupe les variables en facteurs ; et la cartographie des poids locaux montre des zones où, sous l'appellation commune d'un facteur, sont regroupés plusieurs des critères initiaux. Cependant, dans ces deux cas, la partition de l'espace ne s'exerce que sur les facteurs pris un à un : c'est pourquoi à un .autre niveau on peut considérer que chaque facteur est lui-même un critère ; et que l'analyse factorielle ne sert qu'à mieux définir ces critères.

Les critères factoriels, s'ils sont pris un à un, peuvent alors servir à une classification monothétique (cartographie des scores) ou bien, pris dans leur ensemble, ils peuvent servir de base à des classifications polythétiques ultérieures, faisant appel à des algorithmes plus rigoureux.

Les méthodes de classification, qui forment à elles seules un sujet complet d'étude, ne seront pas examinées ici avec beaucoup de détail. Cependant les principales opérations, nécessaires à la compréhension des applications que nous allons présenter, vont être maintenant rapidement esquissées.

De manière générale, les techniques de classification regroupent les observations d'après

une mesure quelconque des relations entre ces observations, prises deux à deux. Par exemple, si l'on considère le modèle factoriel comme une méthode de classification, cette mesure de similarité est fournie par le coefficient de corrélation - pour l'analyse en composantes principales - ou par le "Chi-deux" -pour l'analyse des correspondances. Parmi les techniques de classification proprement dites, "*linkage analysis*" (l'analyse des liens), recherche dans la matrice des corrélations les paires à plus forte corrélation réciproque, puis, alloue progressivement à ces paires les variables restantes les mieux corrélées. Cette démarche, qui se fait par ressemblance de proche en proche, a l'inconvénient de conduire souvent à des classes "étirées", peu homogènes.

"*Cluster Analysis*", ou méthode des emboîtements, a été utilisée de préférence dans les nombreuses études anglo-saxonnes de régionalisation et de classification. Comme, en outre, cette technique suppose la participation indirecte du modèle factoriel, nous allons en présenter les différentes phases.

La première phase consiste à déterminer le degré de similarité de n observations prises paire par paire. Pour cela, "Cluster Analysis" utilise la distance entre deux observations telle qu'elle est définie par le théorème de Pythagore. Soit n observations mesurées sur m caractéristiques ; la distance entre deux d'entre elles, X et Y, de coordonnées respective (x_1, x_2, ..., x_m) et (y_1, y_2, ..., y_m) est donnée par

$$d_{xy} = (\Sigma^{m}_{i=1} (x_i - y_i)^2)^{1/2}$$

qui se développe comme suit

$$d_{xy} = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_m - y_m)^2]^{1/2}$$

Lorsque des groupes d'observations sont en jeu, la distance entre les paires de groupes est définie par l'*indice de distance généralisée* de MAHALANOBIS (1936), qui s'écrit

$$D^2_{xy} = \frac{(X_1 - Y_1)^2}{S_1} + \frac{(X_2 - Y_2)^2}{S_2} + ... + \frac{(X_m - Y_m)^2}{S_m}$$

où, pour chacune des caractéristiques de 1 à m, X et Y sont les moyennes respectives des groupes X et Y, et S représente la dispersion commune.

Cependant, ces deux définitions de la distance obéissent à certaines contraintes qu'il convient de rappeler, Premièrement, les axes servant de système de référence pour les observations doivent être mesurées sur la même échelle. Cela peut s'obtenir par une simple standardisation des m caractéristiques. Deuxièmement, ces axes doivent être mutuellement orthogonaux. Cette dernière condition, qui suppose la non corrélation des variables, n'est généralement jamais réalisée par l'ensemble des données aux multiples relations qu'utilise le

géographe. La transformation des m variables en axes orthogonaux non corrélés peut se faire alors au moyen de 1'analyse en composantes principales. Les scores des observations sur ces m axes sont les nouvelles coordonnées, dont on se sert pour calculer la distance d ou D^2 entre les observations ou les groupes d'observations.

Si le chercheur utilise l'analyse en composantes principales, non seulement comme un moyen d'obtenir m axes orthogonaux, mais aussi comme un moyen de condenser ses données, comme cela arrive fréquemment, il peut ne retenir que les k premières composantes ($k < m$) pour définir l'espace dans lequel seront mesurées les distances. En outre, s'il juge nécessaire de pondérer ces k composantes selon leur importance respective, par exemple d'après la valeur propre qui leur est associée, la distance pondérée dans un espace à K composantes devient

$$d_{xy} = (\Sigma^{k}_{i=1} (x_i - y_i)^2 * W_i)^{1/2}$$

où W_i est le poids donné à la composante i.

Chaque distance ainsi définie est calculée par paire de points, pour produire une "matrice de similarités" de n x n entre les observations.

Au cours de la deuxième phase, des procédés de groupement sont appliqués à la matrice des similarités. L'objectif est de réduire progressivement les n points en un nombre plus petit de groupes, jusqu'à ce qu'un seul groupe soit formé, de telle manière qu'à chaque étape la perte d'information soit minimum. Pour atteindre cet objectif, il faut :

1° - identifier la paire d'observations pour laquelle la distance d^2 est minimum ;
2° - combiner les vecteurs lignes et les vecteurs colonnes de ces deux observations en un seul vecteur ligne et un seul vecteur colonne, représentant le nouveau groupe. Les éléments de ce nouveau vecteur sont les distances carrées du centroïde du groupe à tous les autres points. La matrice de similarité est maintenant d'ordre (n-1) (n-1).
3° - répéter ce procédé pour aller de n-1 à n-2 groupes, et ainsi de suite, jusqu'à l'obtention d'un seul groupe contenant toutes les observations de départ.

Le résultat est un arbre complet de connections ou "dendrogramme", qui montre la hiérarchie entière des groupes d'observation. Cependant, n'importe quel groupement suppose une certaine généralisation et une perte d'information de détail. On peut obtenir, à chaque niveau de groupement dans la hiérarchie, une mesure de la distance séparant les points à l'intérieur d'un groupe qui équivaut à la variance totale à l'intérieur du groupe.

A chaque étape successive, l'accroissement de la variance à l'intérieur du groupe mesure la quantité d'information perdue. Le rapport de cette augmentation à la distance totale entre les points décroît, lors du processus de groupement, vers une valeur minimale, et croît à nouveau. Théoriquement, l'étape pour laquelle cette valeur est minimale offre un nombre optimal de groupes. Cependant, tout en sachant qu'il subira proportionnellement une plus grande perte d'information, le chercheur peut sélectionner n'importe quel niveau hiérarchique ; en particulier, si, pour les besoins de son étude, il lui faut un nombre de groupes plus petit.

Au cours d'une troisième et dernière phase, quelque soit le niveau hiérarchique sélectionné, il reste à s'assurer de l'homogénéité des groupes. On doit à E. CASETTI la procédure discriminante itérative[15] qui converge vers la solution optimale, c'est à dire vers la minimisation de la variation intra-groupe, par rapport à la variation totale ; et vers la maximisation de la variation entre les groupes par rapport à la variation totale.

Mis à part le modèle factoriel lui-même, la méthode des emboîtements est le procédé de classification le plus largement utilisé par les géographes anglo-saxons, dans les études de régionalisation dont nous allons présenter quelques exemples, mais aussi dans les études de classification des villes qui feront l'objet du chapitre suivant.

Parmi les études géographiques anglo-saxonnes qui font appel au modèle factoriel, les études de régionalisation occupent une place plus importante que les études thématiques. La plupart d'entre elles se trouvent citées, classées et parfois commentées dans certains ouvrages généraux sur l'analyse factorielle (RUMMEL, 1970), ou sur l'analyse statistique (KING, 1969), et dans un récent ouvrage de synthèse sur l'écologie factorielle (BERRY, 1971). En consultant ces trois ouvrages, le lecteur pourra assurément obtenir une liste exhaustive de ces études ; c'est pourquoi nous avons décidé de ne pas présenter ici une telle liste, qui ne pourrait être qu'une longue et fastidieuse répétition. Notre objectif reste essentiellement de montrer la place du modèle factoriel parmi les diverses méthodes utilisées, et de souligner sa contribution à des résultats originaux ou triviaux.

Notre commentaire sera limité à quelques études exemplaires, les unes en raison de leur caractère typique qui leur a valu d'être beaucoup imitées, d'autres en raison de leur caractère original permettant d'ouvrir l'éventail des possibilités.

DERIVATION DE RÉGIONS UNIFORMES A FACTEURS MULTIPLES

BERRY ne fut pas le premier à se pencher sur le problème d'établir des régions uniformes à partir de critères multiples. Bien avant lui, HAGOOD (1943) et WEAVER (1954) ont réussi à délimiter des régions agricoles de façon très satisfaisante ; mais BERRY fut l'initiateur d'une méthode s'appuyant à la fois sur l'analyse factorielle et sur l'algorithme des emboîtements. "*A Method for Deriving Multi-Factor Uniform Regions*" (BERRY, 1961), qui contient la description de la méthode, illustrée d'un exemple simple, fut à l'origine de nombreuses applications dont nous avons choisi de présenter l'une des plus représentatives, maintes fois reproduite par la suite, plus ou moins complètement, et dans des contextes légèrement différents.

Cinq cent cinquante cinq municipalités, prises comme unités spatiales d'observation, et trente et une variables à partir desquelles sont crées, par des pourcentages et des taux, seize variables additionnelles, servent de point de départ à l'"*Étude de la pauvreté rurale en Ontario*" (BERRY, 1965). Les variables portent sur les revenus des exploitants agricoles et des familles rurales non agricoles, sur l'éducation, les aménagements du logement, la taille des exploitations,

[15] *Statistical analysis in Ccogrsphy* (KING, 1969, p. 204-215) fournit certaines informations utiles sur les fonctions discriminantes.

la mortalité infantile, etc ...

Inévitablement, certaines de ces variables ont un comportement très similaire sur l'ensemble des 555 unités d'observations. Afin d'éliminer la redondance qui existe entre les variables, et de ne faire ressortir que les quelques facteurs qui sont responsables de la concomitance de certaines variations, deux analyses factorielles en axes principaux sont effectuées ; l'une sur les trente et un indices bruts, l'autre sur les seize indices composés.

Certains auteurs déconseillent l'utilisation de taux et de pourcentages pour deux raisons ; la première, parce qu'un tel système "clos" risque de renforcer de façon artificielle la redondance des variables ; la seconde parce que, selon la façon dont ils sont construits, ils risquent d'introduire entre les variables des relations non linéaires que le coefficient de corrélation linéaire utilisé ne pourra pas faire apparaître. Par exemple, il est recommandé d'utiliser comme mesure de la proportion de femmes F, par rapport à celle des hommes H, 1e rapport F/(H+F) plutôt que le rapport F/H dont la distribution a tendance à être dissymétrique.

L'expérience de BERRY montre qu'au niveau général d'interprétation où il se place, ne retenant que les tout premiers facteurs, les résultats ne sont influencés que de façon négligeable par l'utilisation d'indices construits à la place des valeurs absolues. En effet, dans 1es deux cas, les quatre mêmes facteurs ont pu être identifiés, rendant compte d'environ 70% de la variance. Pour prévenir une autre critique, celle qui voit dans la nécessité et la difficulté de réaliser l'hypothèse de normalité[16] un obstacle sérieux à l'utilisation de l'analyse factorielle, BERRY entreprend une seconde expérience.

Une nouvelle analyse factorielle est effectuée sur chacun des deux groupes de variables, après que les distributions de ces variables aient été transformées[17]. Là encore, bien que des différence sensibles apparaissent dans les résultats après transformation, celles-ci nous semblent négligeables en regard de la généralité des conclusions de BERRY ; et en particulier, elles ne changent en rien l'interprétation des facteurs.

Après une rotation varimax qui devait rendre leur interprétation plus facile, les quatre facteurs retenus ont reçu la signification suivante :

- le premier facteur est une dimension de "pauvreté rurale agricole",
- le second est un facteur de "pauvreté rurale non agricole",
- le troisième, mélange de pauvreté rurale et de densité rurale, ne correspond à aucune définition simple ;
- le quatrième, associé aux plus bas niveaux d'éducation et à une mortalité infantile élevée, est intitulé facteur de "désavantage social".

---

[16] Nous rappelons que la normalité de la distribution des variables assure la linéarité de leur relation ; en particulier, elle permet d'affirmer l'indépendance statistique de deux variables dont le coefficient de corrélation est égal à zéro.

[17] La transformation logarithmique est celle qu'on applique le plus fréquemment, car elle permet de redresser les distributions étirées vers la droite, formes de distribution qu'on rencontre souvent dans les sciences sociales. D'autres types de transformation sont signalées par RUMMEL (1970, p. 280-286).

Les scores des 555 municipalités sur ces quatre facteurs sont ensuite calculés. L'examen des scores, et plus encore leur cartographie, contribue à mieux les définir. En particulier, il devient alors visible qu'une seule municipalité ressort sur le troisième facteur. Celle-ci a une distribution bimodale où se détachent à la fois les classes les plus basses et les plus élevées de revenu rural non agricole ; sans doute en raison de la récente installation dans cette municipalité de commutants de la ville proche. On voit par là que lorsqu'un facteur parait obscur, il est toujours possible - et il est recommandé - d'avoir recours aux indices et aux observations initiales pour en éclairer la signification.

Dans le contexte de son étude, BERRY décida d'ignorer ce facteur "exceptionnel". Restent trois facteurs. Le premier oppose l'est de l'Ontario où les sols minces du bouclier canadien n'offrent que de maigres ressources agricoles, et le sud, nettement plus riche. La cartographie du second offre des "taches" de pauvreté rurale non agricole, localisées le long des séparations des zones d'influences urbaines. Enfin le troisième (facteur 4). identifie les zones de désavantage social qui s'avèrent correspondre aux zones de peuplement français.

Ainsi l'analyse factorielle produit-elle des types de région et peut-elle être utilisée par elle-même comme un procédé de classification. Cependant la délimitation de régions fondées sur un "continuum" de scores reste assez vague et arbitraire. On peut se contenter d'un tel résultat si l'on considère que toute classification, au sens étroit du terme, est impossible parce que les phénomènes étudiés sont continus et n'apparaissent pas "naturellement" en unités discrètes.

Mais si l'on accepte la notion de classification en groupes ou en régions exclusives, alors des procédés plus rigoureux doivent être utilisés . C'est en fonction de cette dernière considération que BERRY poursuit son analyse, par l'application de la méthode des emboîtements décrite plus haut. Pour l'alléger, il ne garde que les 158 municipalités de l'Ontario oriental, qui s'est révélé être la région où se manifeste le plus gravement le syndrome de pauvreté.

1° - Les 158 municipalités sont localisées dans un espace à trois dimensions engendré par les facteurs identifiés, 1, 2 et 4 ; les distances entre les municipalités prises deux à deux sont calculées produisant une matrice de 158 x 158.

2° - La plus petite distance entre deux observations permet de former le premier groupe et de réduire la matrice à 157 x 157. La combinaison des deux unités d'observation séparées par la plus petite distance conduit à une matrice de 156 x 156, et ainsi de suite jusqu'à l'obtention d'un seul groupe.

3° - Enfin, l'application des fonctions discriminantes itératives de CASETTI permet d'obtenir une classification optimale minimisant la distance à l'intérieur des groupes.

L'arbre des connections montre les différents niveaux hiérarchiques du regroupement. Une telle classification n'assure pas forcément la réunion de zones contiguës, et produit des types régionaux plutôt que des régions contiguës uniformes.

Pour perfectionner son modèle, BERRY décide, au cours d'une seconde analyse,

d'introduire une contrainte de contiguïté. Cela consiste à coder par le signe - toutes les distances entre des observations non contiguës et à ne considérer, lors du processus de groupement, que les distances positives. I1 suffit qu'une observation soit contiguë à un membre quelconque d'un groupe, avant que le groupe soit formé, pour qu'elle soit considérée comme contiguë à ce groupe tout entier.

Les résultats de l'analyse typologique et de l'analyse régionale (fig. 3) s'accordent à désigner les mêmes groupes de municipalités les plus désavantagés - celui de la "pauvreté rurale agricole", situé sur le bouclier (C), et celui du "désavantage social" dans l'extrémité orientale (B) ; et ils concordent assez bien sur le troisième groupe relativement prospère situé au sud (A). Cependant, l'analyse typologique montre que quelques municipalités, de type (A) et (B), se retrouvent de façon dispersée en dehors de leur propre région ; alors que l'analyse régionale fait apparaître au nord-est une quatrième région (D) de pauvreté rurale agricole, où l'on rencontre par contraste une certaine prospérité rurale non agricole.

Le choix d'un niveau plus fin de régionalisation, d'après le dendrogramme, devrait conduire à la subdivision de ces principales régions en sous-régions, chacune ayant d'après les propres termes de BERRY :

> "une combinaison très distincte de prospérité et de désavantage relatifs, et la présence d'un désavantage d'une sorte ne conduit pas nécessairement à un désavantage d'une autre sorte, parce que différents processus interviennent […] de ce qui précède, il est clair que des politiques publiques différentes seront nécessaires pour attaquer les divers et distincts syndromes de pauvreté".

L'originalité de l'étude "pilote" de BERRY tient dans son plan en trois phases :

I - Organiser les données
II - Etablir des mesures de similarité
III - Regrouper les observations

accordant une place importante à la première phase d'organisation des données par rapport aux méthodes classiques de taxonomie (SOKAL and SNEATH, 1963), qui se concentrent principalement sur les deux dernières phases.

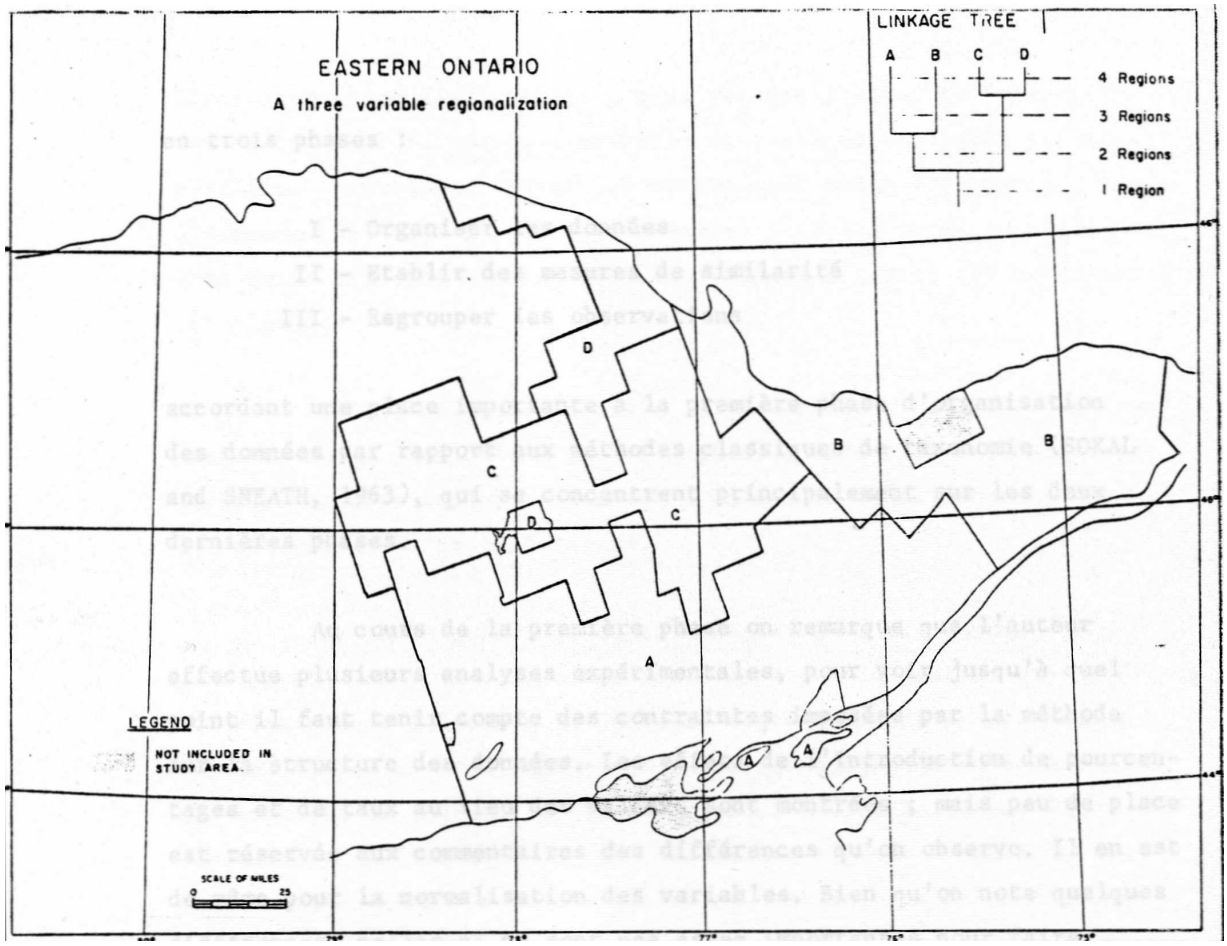


Figure 3 - La régionalisation de la partie orientale de la province d'Ontario d'après trois critères factoriels
Source : BERRY (1965)

Au cours de la première phase, on remarque que l'auteur effectue plusieurs analyses expérimentales, pour voir jusqu'à quel point il faut tenir compte des contraintes imposées par la méthode sur la structure des données. Les effets de l'introduction de pourcentages et de taux au lieu des valeurs sont montrés, mais peu de place est réservée aux commentaires des différences qu'on observe. Il en est de même pour la normalisation des variables. Bien qu'on note quelques différences, celles-ci ne sont pas assez importantes pour faire changer la définition donnée aux facteurs ; et l'auteur poursuit son étude sans tenir compte de ces différences.

I1 est vrai qu'étant donné le niveau très général où se place BERRY, on peut les considérer comme négligeables. Cependant, il faut admettre que les expériences peu concluantes de BERRY ont eu, indirectement, un effet néfaste. A cause d'elles, de nombreux auteurs se sont crus par la suite dispensés de la vérification des hypothèses en question ; vérification qui, dans un autre contexte et surtout à un niveau plus fin d'analyse, aurait pu être fort utile.

Au cours de la seconde phase, il faut signaler l'utilisation des scores comme mesures de similarités sur les facteurs après rotation. La rotation des axes à partir de la solution en

composantes principales présente certainement des avantages, en ce qui concerne l'interprétation des facteurs ; mais puisque l'effet de "varimax" est de répartir la grande proportion de variance expliquée par les premières composantes parmi les composantes moins importantes, les distances calculées d'après les premiers facteurs après rotation sont moins précises que celles qu'on obtient à partir des composantes principales.

Il est toujours possible et recommandé d'utiliser "varimax" comme un moyen parallèle d'investigation ; mais cette technique ne devrait pas être employée avec les procédés de classification dont les définitions sont rigoureuses.

Un autre aspect discutable concerne la standardisation des scores. BERRY la recommande pour supprimer l'effet de prépondérance des plus grands facteurs. Il justifie cette attitude en prétendant que les variables employées ne sont que des approximations de celles qui existent en théorie. La sélection d'un nombre différent de "succédanés" pour chaque variable théorique conduit à une solution tendancieuse. I1 n'y a aucune raison de donner un poids plus grand à une dimension, pour laquelle il existe dans le Recensement un grand nombre de succédanés, alors que pour une autre dimension, peut-être théoriquement aussi importante, il n'existe qu'un ou deux indices.

On sait le rôle primordial que BERRY attribue à l'analyse factorielle dans l'élimination de la redondance ; or, dans cette perspective il est contradictoire de ne considérer que les premiers facteurs comme le fait BERRY, alors qu'il prétend accorder à chacun autant d'importance. Il est possible, en effet, que parmi les plus petits facteurs se cache un concept important mais peu identifiable, parce que mal représenté par un trop petit nombre d'indices partiels.

Cela nous conduit à formuler envers BERRY une dernière critique : celle de ne pas introduire dans son analyse des caractéristiques fondamentales pour les régions rurales et qui, bien que pouvant être codifiées de façon à être traitées mathématiquement, ne sont pas aussi directement utilisables que les indices du recensement. Ce sont par exemple le relief, le drainage, les conditions climatiques, les sols, la couverture forestière, etc ... ; en d'autres termes, toutes les caractéristiques physiques d'une région. La même remarque est valable pour les caractéristiques humaines telles que le genre de vie des habitants, leur conception du niveau de vie, leur histoire locale, leur folklore, etc ... Cette critique est d'autant plus justifiée qu'on est gêné de voir l'auteur se référer aux contrastes français-anglais, bouclier-basses terres, et de tenter d'expliquer les syndromes de pauvreté en s'appuyant uniquement sur les systèmes d'associations très généraux dérivés de son analyse.

I1 est clair que l'analyse factorielle est un outil très puissant pour l'identification des caractéristiques locales et pour décrire de vastes systèmes ; mais ce n'est qu'une opération préalable qui ne peut pas faire partie de l'explication en soi, surtout quand on sait qu'une grande partie des données n'est pas dans le problème.

RAY (1965), associé à BERRY, a utilisé I'ensemble des procédés analytiques décrits plus haut, dans une étude des 120 comtés du Québec et de l'Ontario, sur 88 variables culturelles, démographiques, agricoles, urbaines et industrielles. Des analyses séparées ont été encore effectuées pour des groupes de variables et pour l'ensemble des variables. Chaque fois, les résultats ont donné trois dimensions de base :

- les différences entre le Canada français et le Canada anglais,
- une opposition des comtés urbains et des comtés ruraux,
- les contrastes entre le bouclier du nord et les terres basses du sud.

Des cartes montrent la distribution spatiale de chacun des facteurs et, finalement, des procédés de groupement avec et sans contrainte de contiguïté permettent une régionalisation sur la base de ces facteurs.

SPENCE (1968), au cours d'une régionalisation des comtés britanniques, s'appuie directement sur les travaux de BERRY. Les données qu'il utilise sont 152 variables d'emploi, exprimées en pourcentage de l'emploi total de chaque comté, pour tenir compte des différentes tailles de population. Une analyse en composantes principales effectuée dans le but d'éliminer la redondance des variables, produit huit composantes principales qui expliquent un peu plus de 70 % de la variance et se prêtent difficilement à l'interprétation. Une rotation varimax permet d'identifier quatre facteurs :

- le premier regroupe les emplois concernant la production et la distribution des produits alimentaires;
- le second est celui des industries nouvelles (chimie, métallurgie);
- le troisième représente les industries traditionnelles (charbon, textile);
- le quatrième est un facteur des services (commerce, assurances, banques, communications).

Les quatre autres facteurs, plus complexes, n'expliquant qu'une part infime de la variance, ne sont pas identifiés.

La cartographie des poids locaux fait correspondre le facteur 1 à la zone de production agricole de l'est ; le facteur 2 à la mégalopolis, le facteur 3 aux zones d'industrie lourde du nord ; et le facteur 4 au centre commercial et financier de Londres.

On voit comment des groupes d'activités liées, isolés par la structure factorielle, peuvent prendre une expression régionale. Mais il s'agit là plutôt d'une typologie régionale établie sur la bas d'un seul facteur. Pour obtenir une régionalisation sur plusieurs facteurs à la fois, SPENCE fait appel à des techniques de groupement, en suivant l'exemple de BERRY. Chaque facteur ayant un poids spécifique pour chaque observation, les comtés sont exprimés comme des points dans l'espace des huit facteurs considérés comme orthogonaux ; et une matrice de 42 x 42 des distances entre les points (D^2) est calculée. D'un premier groupement découlent des zones uniformes d'offres d'emploi. Après que la contrainte de contiguïté ait été introduite, les groupes

de comtés contigus forment des régions d'offres d'emploi uniformes et exclusives.

On peut citer encore l'exemple de BROWN (1968) sur la régionalisation de la Pologne, d'après soixante indices économiques et vingt deux "voivodships", pour les années 1958 et 1964. Pour chacune des deux années, quatre dimensions expliquent presque les 3/4 de la variance totale. Elles comprennent

1° - un facteur agricole,
2° - un facteur d'industrie lourde,
3° - un facteur d'économie socialisée,
4° - un facteur de croissance industrielle.

Ces dimensions sont utilisées pour regrouper les voïvodships en un système de types régionaux basés sur les caractéristiques agraires, le degré d'industrialisation et le degré de socialisation de l'agriculture. Les changements qui portent sur une période de six .ans sont notés.

On pourrait, à l'occasion de cette étude, calquée sur celle de BERRY, reprendre les critiques de cette dernière, et discuter à nouveau de la pertinence des variables utilisées et de la valeur des distances mesurées sur des axes ayant subi une rotation. Cependant, nous nous contenterons d'attirer ici l'attention sur un nouveau problème : celui des comparaisons de deux analyses factorielles[18]. En effet, l'auteur se croit autorisé à noter les changements qui sont survenus entre les années 1958 et 1964, sur la base des résultats de ses deux régionalisations. Or, même lorsque des variables identiques sont utilisées d'une période à l'autre, les facteurs seront différents - ce qui est normal - car il n'est pas possible de penser que l'ensemble des variables apportent à nouveau exactement le même ensemble de contributions relatives. I1 devient alors difficile de parler de changement ou de stabilité relative de certaines régions, sur des facteurs dont la définition n'est plus exactement la même. A plus forte raison, il est impossible de comparer cette régionalisation économique d'un pays non capitaliste avec celle d'un pays capitaliste, comme on pourra être tenté de le faire, étant donné la similitude des indices et des techniques utilisées.

Les études précédentes montrent que l'association de l'analyse factorielle, qui transforme des données complexes linéairement interdépendantes en un ensemble de critères bien définis et de la méthode des emboîtements qui classe au mieux les observations d'après ces critères, conduit à des régionalisations satisfaisantes. Les aspects discutables que nous avons notés plus haut concernent surtout des problèmes théoriques de choix de variables au départ, et de tentative d'explications en fin d'analyse. Ces problèmes, les méthodes classiques de régionalisation les ont toujours connus et affrontés ouvertement. Cependant, du fait qu'ils sont moins visibles dans les régionalisations faisant appel à un appareil mathématique complexe, ils doivent être davantage soulignés.

En outre, les critiques sur le respect de certaines contraintes mathématiques n'apparaissent valables que dans le cas d'études dont les résultats sont suffisamment détaillés pour être influencés par la non observation de ces contraintes. Enfin, pour résumer l'apport de la méthode de BERRY aux études de régionalisation, nous ferons appel à l'expérience que KING a

[18] La comparaison des structures factorielles est un problème délicat que LAWLEY et MAXWELL abordent dans leur ouvrage : *Factor Analysis as a Statistical Method* (1963, p. 6, 92-93)

effectuée d'après l'étude de WEAVER.

L'étude de WEAVER (1954) sur les régions agricoles du Midwest des Etats-Unis est un excellent exemple de régionalisation traditionnelle. En reprenant les données utilisées par WEAVER (sept types de récoltes dans quatre vingt huit comtés), KING (1969) a montré qu'avec l'appareil technique plus complexe que propose BERRY, on parvenait à des résultats très semblables. En conséquence, il ne semble pas que l'on puisse remettre en question l'utilité de ces techniques dites objectives ; car outre des résultats qui ne contredisent pas le jugement du spécialiste, elles offrent un triple avantage :

- celui d'isoler les éléments séparés (linéairement indépendants) constitutifs du phénomène étudié,
- celui du choix de différents niveaux de découpage régional.,
- et, bien sûr, celui de traiter plus rapidement un ensemble plus vaste de données.

Exemples de typologies régionales

Un grand nombre d'applications de l'analyse factorielle, qu'on peut classer avec les études de régionalisation, ne fait pas intervenir d'algorithme de groupement. L'analyse en composantes principales ou l'analyse factorielle y sont considérées elles-mêmes comme des méthodes efficaces - sinon de classification au sens étroit du terme - du moins, d'ordonnance des données.

La dernière phase de classification, en classes rigoureusement exclusives et homogènes, n'est pas jugée nécessaire et même, pour certains chercheurs qui pensent qu'elle est incompatible avec le concept de zones de transition si répandu en géographie, elle n'est pas souhaitable. Cependant, comme les études qui appliquent ce point de vue ne sont que des reproductions incomplètes des exemples que nous venons d'examiner, nous ne ferons que les citer rapidement et nous concentrerons notre attention sur les plus originales.

Il est à noter, en premier lieu, que parmi les études thématiques examinées dans le chapitre précédent, certaines, comme celle de BERRY (1961) sur les structures de base du développement économique, peuvent être considérées en partie comme des études de régionalisation ; tant il est vrai que la détermination des causes d'un phénomène, problème essentiellement thématique, passe bien souvent par le classement de portions d'espace sur quelques facteurs fondamentaux.

Ce classement, cette régionalisation au sens large, sont le plus souvent le seul but d'un grand nombre d'applications de l'analyse factorielle. Par exemple, THOMPSON, SUFRIN, GOULD et BUCK (1962), sont concernés par l'identification de facteurs qui expriment au mieux la variation spatiale de la santé économique dans l'état de New-York. Ces facteurs furent interprétés comme :

1° - un facteur général de santé économique,
2° - un facteur rural-urbain
3° - un facteur de croissance et d'emploi.

Le même type d'analyse se retrouve dans l'étude que Mary MEGEE (1965) a entreprise à l'échelle nationale américaine pour tester les hypothèses traditionnellement émises sur l'importance relative de certaines régions au point de vue du développement économique. Dans une autre étude semblable, utilisant des procédés identiques, OLSEN (1965) a analysé quinze états de la région sud-est des Etats-Unis sur quarante caractéristiques de développement économique.

La même approche dans un domaine différent a été utilisée par HENSHALL et KING (1966) pour tenter de découvrir l'organisation structurelle et régionale de l'agriculture de la Barbade. Une analyse de type R parvient à briser la complexité des associations de culture et d'élevage, en produisant quatre types de cultures. L'analyse de type Q extrait les quatre types de fermes correspondants. Leur répartition régionale bien définie incite les auteurs à suggérer une politique d'intervention, pour encourager certaines productions dans les zones les plus défavorisées.

E. SOJA (1968), dans son ouvrage sur la *Géographie de la modernisation au Kenya*, consacre toute une partie à la recherche des dimensions de modernisation. Une analyse en composantes principales fait ressortir une première dimension générale de développement et une seconde dimension opposant deux sous-systèmes : l'un africain, l'autre européen. Après une rotation varimax, un premier facteur d'urbanisation associé au commerce de type asiatique apparaît, suivi par un second facteur indiquant le niveau d'éducation des africains ; et un troisième marquant l'impact du peuplement européen. Des aires d'extension correspondant à chacun de ces trois aspects de la "modernisation" peuvent ainsi être délimités.

On pourrait multiplier les exemples de ces typologies régionales, mais il nous a paru plus intéressant de consacrer le reste de ce chapitre à une catégorie d'études régionales peu représentées dans la, littérature géographique anglo-saxonne et pourtant fondamentale : il s'agit des études consacrées à l'établissement de régions fonctionnelles.

DELIMITATION DE RÉGIONS FONCTIONNELLES

Là encore, nous devons la principale contribution à BERRY. Au cours d'une double analyse sur le commerce de l'Inde, BERRY a établi un ensemble de régions fonctionnelles ; puis il a tenté de définir des systèmes d'interdépendances entre des types de comportement spatial d'une part, et la structure spatiale d'autre part.

La première analyse (BERRY, 1966) comprend au départ une matrice de 36 x 36 qui contient le total des flux de soixante trois biens de consommations, entre trente six "blocs" commerciaux représentés par les Etats de l'Inde et les villes les plus importantes. Les flux sont notés entre chacune des trente six paires de blocs - ou "*dyads*" - allant des lignes vers les colonnes.

Une analyse factorielle de type R montre la ressemblance des lieux de destination quant à l'origine de leurs biens de consommation. Les facteurs, après rotation, font ressortir des groupes de régions destinataires ou "consommatrices" ; et les scores sur ces facteurs indiquent

les sources principales d'envoi pour chaque groupe.

Une analyse de type Q regroupe les régions productrices d'après leur ressemblance sur les lieux de destination de leurs produits. Les scores indiquent les destinations principales pour chaque groupe. Enfin, un algorithme de classification - dont on a vu plus haut les mécanismes - permet de délimiter des régions fonctionnelles (Fig. 4).

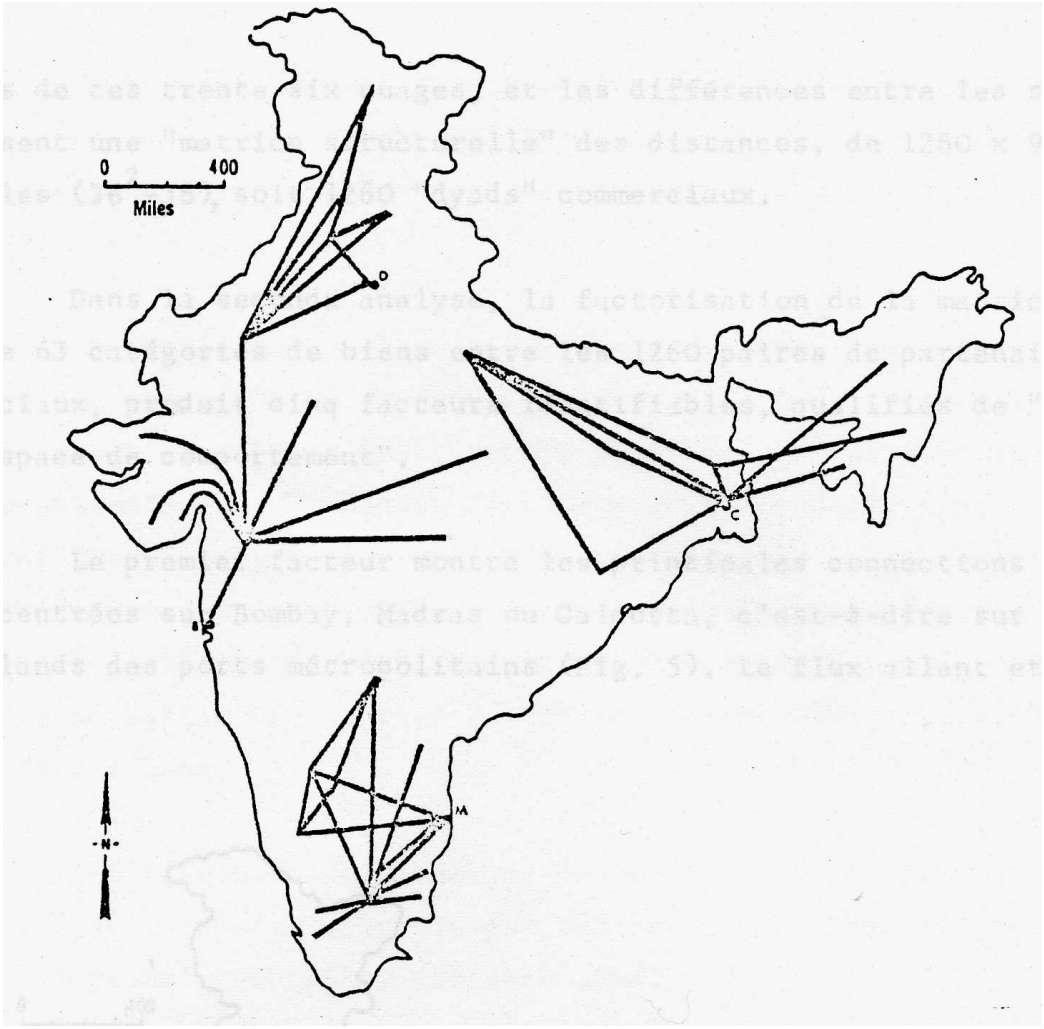


Figure 4 - Régions fonctionnelles fondées sur l'expédition de biens de consommation (Source : BERRY, 1966)

Dans la seconde étude, BERRY (1968) entreprend à nouveau deux séries d'analyses factorielles. L'une vise à trouver une "*structure spatiale*" sous-jacente dans un ensemble de quatre vingt douze attributs socio-économiques, agricoles et d'emploi, mesurés sur les 325 districts de l'Inde. L'autre cherche à établir une "*structure de comportement*", d'après les échanges de soixante trois catégories de biens de consommation, entre trente six "blocs commerciaux" du même pays.

Au cours de la première analyse, neuf facteurs sont retenus dont les scores définissent la

structure spatiale suivante :

- le premier facteur identifie les régions urbaines et industrielles,
- le second correspond aux régions de cultures intensives irriguées,
- le troisième fait ressortir les contrastes agricoles entre l'est et l'ouest,
- le quatrième désigne la vallée du Gange,
- le cinquième l'agriculture sèche du nord-ouest.

Les quatre derniers facteurs correspondent tous à des spécialités régionales de production minière. Puisqu'il n'existe que trente six blocs commerciaux et 325 districts, les districts appartenant à un même bloc sont traités comme un nuage de points dans un espace à neuf dimensions. Les poids locaux des neuf dimensions sont calculés à partir des centroïdes de ces trente six nuages, et les différences entre les scores produisent une "matrice structurelle" des distances, de 1260 x 9, entre les (36^2-36), soit 1260 "dyads" commerciaux.

Dans la seconde analyse, la factorisation de la matrice des flux de 63 catégories de biens entre les 1260 paires de partenaires commerciaux, produit cinq facteurs identifiables, qualifiés de "base de l'espace de comportement".

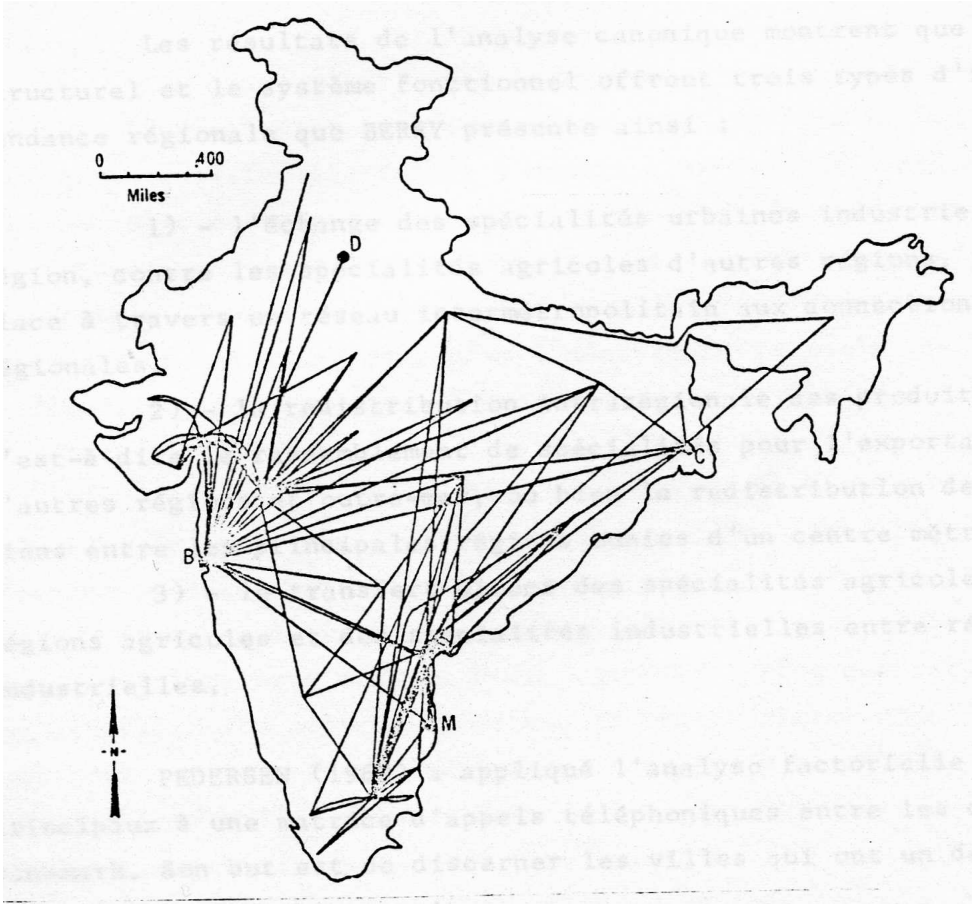


Figure 5_- Scores supérieurs à 2,0 sur le premier facteur de comportement commercial (Source : BERRY, 1968)

Le premier facteur montre les principales connections régionales centrées sur Bombay, Madras ou Calcutta, c'est à dire sur les hinterlands des ports métropolitains (Fig. 5). Le flux allant et venant du nord-ouest forment le second facteur, alors que le facteur 3 réunit les flux qui viennent de l'est, etc ... Les scores sur ces cinq facteurs donnent une "matrice de comportement" spatial de 1260 x 5.

La dernière phase, objectif final de BERRY, est consacrée à examiner les relations qui existent entre le système structurel - représenté par la matrice de structure spatiale - et le système fonctionnel - représenté par la matrice de comportement spatial-. Dans ce but, BERRY utilise l'analyse de corrélation canonique[19]. Cette technique multivariée peu connue est très proche de l'analyse de corrélation classique, à la différence près qu'elle mesure la relation de deux ensembles de variables, au lieu de mesurer la relation de variables prises deux à deux.

Les résultats de l'analyse canonique montrent que le système structurel et le système fonctionnel offrent trois types d'interdépendance régionale que BERRY présente ainsi :

1)- l'échange des spécialités urbaines industrielles d'une région, contre les spécialités agricoles d'autres régions, prenant place à travers un réseau inter-métropolitain aux connections interrégionales,

2)- la redistribution intra-régionale des produits régionaux, c'est à dire le rassemblement de spécialités pour l'exportation dans d'autres régions et outre-mer, ou bien la redistribution des importations entre les principales régions munies d'un centre métropolitain,

3)- le transfert direct des spécialités agricoles entre régions agricoles et des spécialités industrielles entre régions industrielles.

PEDERSEN (1968) a appliqué l'analyse factorielle en axes principaux à une matrice d'appels téléphoniques entre les districts, au Danemark. Son but est de discerner les villes qui ont un degré élevé de "centralité" en ce qui concerne leur trafic téléphonique, et de délimiter les zones d'influence de ces places centrales.

Les données sont le nombre d'appels enregistrés entre chacun des 62 districts. Un district comprend en général une ville et son hinterland. Dix facteurs communs ont été extraits, qui expliquent 70 % de la variance totale contenue dans la matrice. Lorsqu'on cartographie des poids locaux, on voit nettement que le facteur n°1 représente les appels de Copenhague au reste du pays. Les facteurs 2, 3 et 4 regroupent les communications des districts respectifs de Århus, Odense et Ålborg ; les facteurs 5, 6 et 7 indiquent que les districts de Hadersley, Holstebro et Esbjerg respectivement, sont aussi des centres régionaux pour le trafic téléphonique, quoique d'influence plus faible. Quant aux facteurs suivants, qui n'expliquent qu'une très petite partie de la variance, on ne les détaillera pas ici.

Les zones d'influence de chacun de ces centres sont données par les districts dont les saturations sont les plus fortes avec le facteur en question (la figure 6 montre l'exemple de

[19] On n'entrera pas dans les détails de la formulation mathématique qu'on peut trouver dans tous les ouvrages généraux sur l'analyse multivariée (ANDERSON, 1968) et (KING, 1969).

HADERSLEY). Ainsi on voit nettement ressortir les centres régionaux importants, entourés d'une zone d'influence nette, et les centres d'influence plus faible. Devant ces résultats, on peut s'interroger : quel est l'intérêt de cette approche pour l'analyse factorielle des trafics téléphoniques par rapport à d'autres mesures de "centralité" ?

En général, on utilise comme mesure de "centralité" la fréquence de certaines fonctions définies comme centrales (banques, commerce de luxe, hôpital, etc ...) ou bien, par exemple, l'emploi dans le commerce de gros. Au Danemark, l'utilisation de ces mesures confirme la position dans la hiérarchie des quatre centres les plus importants ; mais elle ne donne pas la même position pour les villes plus moyennes qui suivent. Elles ont tendance à attribuer une grande influence aux villes situées dans des régions densément peuplées, proches des principaux centres métropolitains. Au contraire l'analyse factorielle fait ressortir des centres plus indépendants, éloignés des principales zones urbaines. Cela vient d'une qualité inhérente au modèle factoriel. En effet, puisque

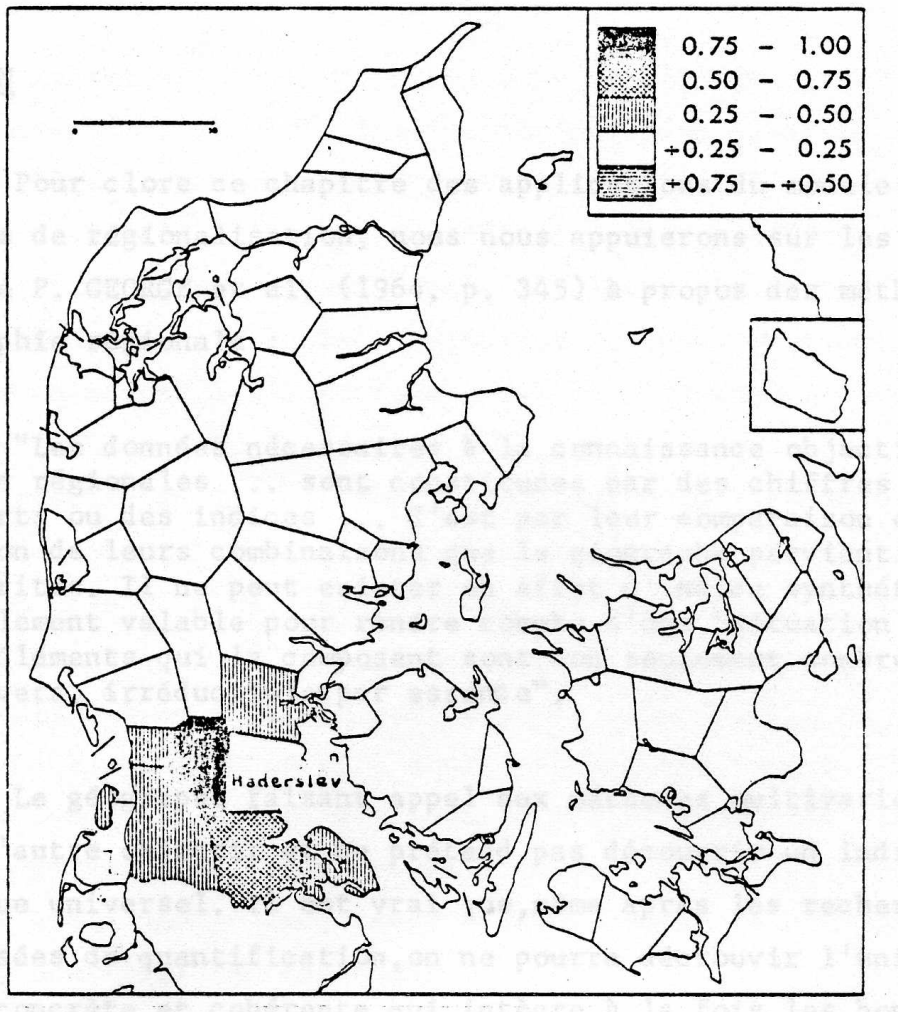


Figure 6 - La zone d'influence du district de HADERSLEV (saturations sur le facteur 5)
Source : PEDERSEN (1968)

le premier facteur rend compte de la plus grande proportion possible de la variance contenue dans les observations, il regroupe, en même temps que les partenaires de la plus grande ville, la plupart des partenaires des districts qui entourent cette ville et qui d'ailleurs ont tendance à être les mêmes. Puisque le second facteur explique le plus possible de la variance qui n'a pas déjà été prise en compte par le premier, il doit représenter une ville importante localisée à quelque distance de la plus grande, l'influence des villes proches du plus grand centre ayant déjà été incorporée dans le premier facteur. Et il en va de même pour les facteurs suivants.

Ainsi les centres obtenus par une analyse factorielle ont tendance à être espacés, chacun dominant son propre hinterland et localisé en dehors de l'influence des autres centres. En conclusion, on remarquera avec Pedersen que

> "des centres régionaux définis selon de tels principes seront souvent plus satisfaisants pour la localisation des activités régionales, comme par exemple un gouvernement local".

CONCLUSION

Pour clore ce chapitre des applications du modèle factoriel aux études de régionalisation, nous nous appuierons sur les considérations de P. GEORGE et al. (1964, p. 345) à propos des méthodes de la géographie régionale

> "Les données nécessaires à la connaissance objective des disparités régionales ... sont constituées par des chiffres bruts, des rapports ou des indices ... C'est par leur comparaison et par la comparaison de leurs combinaisons que le géographe parvient à définir des disparités. I1 ne peut exister en effet d'indice synthétique universellement valable pour rendre compte d'une "situation régionale [...] tous les éléments qui la composent sont non seulement nombreux mais encore divers, irréductibles par essence".

Le géographe faisant appel aux méthodes multivariées n'a pas d'autre objectif et ne prétend pas découvrir un indice synthétique universel. I1 est vrai que, même après les recherches les plus poussées de quantification, on ne pourra découvrir l'unité d'espace concrète et cohérente qui intègre à la fois les homogénéités et les polarisations, les stades de développement et les comportements psychologiques, les servitudes administratives et les liaisons extérieures. Il n'y a pas une unité d'espace mais plusieurs, selon le contenu qu'on veut donner à celui-ci.

Cependant, du point de vue opérationnel, on pourra décider de travailler dans l'un ou l'autre de ces espaces selon le but recherché. C'est l'ensemble des données sélectionnées par le chercheur qui fournit le contenu conceptuel de l'espace à "régionaliser" : caractéristiques touchant par exemple le développement économique, les types de culture, l'emploi, la modernisation, les échanges, les flux, etc ... Etant donné ce contexte, il est possible, nous l'avons vu, de découvrir par les méthodes factorielles des types de région - plus synthétiques donc plus généralement utilisables que d'autres - qui, sans prétendre à l'intégration de tous les critères, prennent en compte les plus importants d'entre eux.

Composantes ou facteurs peuvent être considérés comme les critères les plus importants,

dans la mesure où ils regroupent parmi toutes les données initiales celles qui varient ensemble dans un système de relations causales. Et plus ce système intègre de variables initiales et plus il est cohérent, plus il apparaît comme important. I1 faut ici cependant apporter une restriction car, comme nous l'avons signalé, l'importance d'un facteur peut être artificielle si ce dernier intègre un grand nombre d'indices redondants, c'est à dire linéairement dépendants du fait technique de leur construction, et non du fait de relations causales conceptuelles. (Un exemple évident est celui de variables comme "proportion de femmes actives dans la population totale" et "population active dans la population totale", mais il existe un grand nombre de liaisons artificielles qui ne sont pas aussi évidentes). Au niveau de la classification, ce problème n'apparaît plus si l'on a pris soin d'effectuer une standardisation des scores. Car l'étude régionale ne s'arrête pas toujours à la définition des disparités les plus importantes et à leur cartographie successive. Elle se poursuit souvent par une classification des observations sur ces disparités, considérées simultanément. Cela conduit à toute une hiérarchie de régions homogènes et exclusives rendant compte le plus synthétiquement possible d'une situation régionale.

Ainsi, avec la cartographie des scores sur les facteurs d'une part, et la classification hiérarchique d'autre part, on obtient bien de façon optimale les 2 formes de découpage régional reconnues par Pierre GEORGE et ses collaborateurs (1964, p. 27) en ces termes :

> "Il est exceptionnel que l'on puisse proposer une seule et unique forme de parcellement ... L'une montre une hiérarchie des régions, l'autre les chevauchements".

La compétence du modèle factoriel s'étend à la délimitation de régions fonctionnelles. L'exemple de l'analyse factorielle des flux téléphoniques de PEDERSEN montre qu'elle permet, mieux qu'une autre méthode, de faire apparaître des régions fonctionnelles secondaires après que l'influence des plus grands centres ait été prise en compte. La seconde étude de BERRY en Inde est une tentative d'intégration de la région formelle et de la région fonctionnelle. Elle aboutit à montrer les 3 combinaisons essentielles d'interdépendance entre ces 2 systèmes, alors qu'il existe un beaucoup plus grand nombre de combinaisons possibles inessentielles, dans la considération successive desquelles une observateur, même bien entraîné, risque de se perdre ou du moins de perdre un temps précieux.

CHAPITRE III

LES ÉTUDES URBAINES

Les études urbaines ont toujours occupé une place importante dans la Géographie anglo-saxonne ; et particulièrement, depuis que le modèle factoriel leur a apporté son concours, elles n'ont cessé de se multiplier. La tentative 1a plus récente de leur synthèse peut être trouvée dans *City classification HandBook* (BERRY and SMITH, 1972). A l'aide de cet ouvrage, et des principaux travaux auxquels il se réfère, nous essaierons pour notre part de mettre en évidence le rôle du modèle factoriel :

1° - dans les applications au niveau interurbain (classifications, comparaisons), et,

2° - au niveau intra-urbain (recherche des structures socio-économiques et de leur organisation spatiale).

LES ÉTUDES INTERURBAINES

"*Factor Analysis in the Study of Metropolitan Center*" (PRICE, 1942) est, à notre connaissance, l'une des premières études urbaines qui fasse appel au modèle factoriel. Les données rassemblées - caractéristiques d'emploi et attributs socio-économiques, soit en tout quinze variables pour quatre vingt treize villes de plus de 100 000 habitants - portent sur l'année 1930.

Les dimensions relativement modestes de la matrice des données répondent à des contraintes de temps, dues à l'absence d'ordinateur. Une analyse factorielle en axes principaux permet de retenir quatre facteurs, sur lesquels est effectuée une rotation graphique. Cette méthode consiste à faire tourner visuellement, sur un graphique, les axes orthogonaux pris deux à deux, autour de leur origine, jusqu'à l'obtention pour chaque paire de facteurs d'un nombre maximum de saturations proches de zéro. C'est une bonne approximation de la méthode varimax dont la mise au point n'est venue que plus tardivement.

Avec l'aide du rang des villes d'après leurs scores sur les facteurs, PRICE a donné à ces facteurs l'interprétation suivante :

Facteur 1 : grands centres métropolitains anciens;
Facteur 2 : spécialisation dans les services, opposée à la spécialisation dans l'industrie;
Facteur 3 : niveau de vie général;
Facteur 4 : volume commercial par habitant.

PRICE affirme lui-même qu'il ne peut répondre de l'utilité des concepts qu'il a découverts avec un seul exemple d'analyse. Il recommande que d'autres études soient faites avec la même méthode à des fins de comparaisons. Mais i1 a fallu attendre une vingtaine d'années avant que, grâce à la vulgarisation d'ordinateurs rapides à grande capacité, ce type d'analyse soit repris et étendu à un grand nombre d'applications.

L'objectif de MOSER et SCOTT (1961) est encore exploratoire et descriptif : classer 157 villes anglaises et galloises de plus de 50 000 habitants en quelques catégories homogènes, d'après la considération simultanée d'un grand nombre de caractéristiques. Les soixante variables du recensement de 1951, utilisées pour ce classement, peuvent se répartir en huit rubriques : taille et structure de la population (7 indices), changement démographique (8 indices), caractéristiques du logement et de la famille (15 indices), caractéristiques économiques (10 indices), classes sociales (4 indices), comportement électoral (7 indices), santé (7 indices), éducation (2 indices). Le déroulement de l'analyse s'effectue en deux temps. En un premier temps, "parce que les nombreuses séries qui décrivent les villes ne sont pas indépendantes, elles se répètent dans l'histoire qu'elles racontent …" .

une analyse en composantes principales est utilisée pour éliminer la redondance dans les caractéristiques initiales, afin de retrouver les structures de base d'après lesquelles sont organisées les 157 villes. Les résultats montrent que quatre composantes seulement peuvent rendre compte de 60 % de la variance contenue dans les soixante variables. Ce sont les dimensions désignant :

1° - les classes sociales,
2° - la croissance démographique de 1931 à 1951,
3° - la croissance démographique récente,
4° - les conditions du logement.

Malgré un ensemble de données différent, et surtout l'absence de rotation dans le second cas qui interdit toute comparaison scrupuleuse, on peut constater que la dimension de niveau de vie général, identifiée par PRICE, se retrouve en partie avec la dimension de MOSER et SCOTT désignant les classes sociales.

En un second temps, les 157 villes, considérées comme un nuage de points dans un espace à quatre dimensions, reçoivent une valeur (score) sur chacune de ces dimensions. Les villes sont classées par une méthode visuelle. Les points adjacents sont regroupés dans un espace graphique à deux dimensions ; puis les distances sont mesurées entre les membres des groupes ainsi définis et la moyenne des groupes dans l'espace à quatre dimensions. Enfin, les villes en bordure des groupes sont redistribuées de façon à avoir au moins dix villes par groupe. Cette méthode en partie arbitraire conduit à la formation de quatorze groupes, auxquels s'ajoutent deux cas uniques : Londres et un faubourg champignon de Liverpool, qui n'ont pas pu être intégrés.

Les auteurs se heurtent à la difficulté de donner des noms descriptifs aux groupes. En fait, au lieu de les décrire d'après les quatre composantes - ce qui serait logique - ils les décrivent d'après les attributs des villes membres, que l'on aperçoit peu ou pas du tout dans les

composantes. Par exemple si l'on s'en tient, pour une lecture plus simple, à l'agrégation de ces quatorze groupes en trois catégories :

1) - les villes touristiques et les villes administratives,
2) - les villes industrielles,
3) - les villes de banlieue,

on voit que cette classification contredit d'une certaine. manière les dimensions de structure urbaine révélées par l'analyse en composantes principales ; en particulier la première dimension qui a trait aux variations entre classes sociales.

D'après la même analyse, reproduite par ANDREWS (1971) en appliquant l'algorithme de classification optima que nous avons décrit au chapitre précédent, on n'obtient que deux catégories de groupes. Il existe bien une différence entre les villes principalement administratives et principalement industrielles, en termes des quatre dimensions ; mais les villes de banlieue ont plus de traits communs avec les villes administratives ou commerciales, dont elles dépendent, qu'entre elles. ANDREWS en conclut que la division en trois groupes de MOSER et SCOTT n'est pas une représentation précise et objective des données utilisées dans leur analyse, et que ces derniers font intervenir leur jugement de façon implicite. Cependant, si la classification optimale contredit le jugement des spécialistes, cela ne veut pas forcément dire que les spécialistes se trompent. Il se peut simplement que les données identiques, sur lesquelles sont fondées les deux classifications, ne contiennent pas les indices qui pourraient faire se détacher les banlieues comme distinctes des autres types de villes. La confrontation des deux points de vue ne peut être que fructueuse, si elle conduit à un réexamen des données.

AHMAD (1965) a suivi l'exemple de MOSER et SCOTT pour procéder à la classification de 102 villes indiennes au-delà de 100 000 habitants, d'après soixante-deux variables. Une analyse factorielle en axes principaux fournit vingt vecteurs propres dont dix, ayant une valeur propre supérieure à un, subissent une rotation vers une position varimax. Ces dix facteurs rendent compte de 72 % de la variation entre les villes contenues dans les soixante deux variables. Voici la signification attribuée aux principaux d'entre eux :

1 -proportion de la population masculine par rapport à la population féminine et proportion de femmes dans la population active,
2 - accessibilité généralisée,
3 -compacité,
4 - opposition commerces-industries,
5 - orientation rurale,
6 -taille de la ville.

En comparant ces dimensions avec celles des villes américaines et anglaises, AHMAD constate que la taille de la ville sous diverses formes - plus grande accessibilité, migrations plus importantes, structure industrielle plus diversifiée - paraît jouer un rôle significatif dans les différents systèmes urbains. Cependant, il ne semble pas qu'en Inde la concentration des .activités industrielles se fasse en fonction de l'accessibilité au marché, comme c'est le cas dans les pays plus industrialisés. Les plus grands centres métropolitains de l'Inde n'ont pas, proportionnellement à leur concentration de main d'oeuvre, la plus forte concentration

d'activités industrielles. AHMAD pense que cela reflète le rôle dominant de planification et du contrôle gouvernemental, en ce qui concerne 1a localisation des industries.

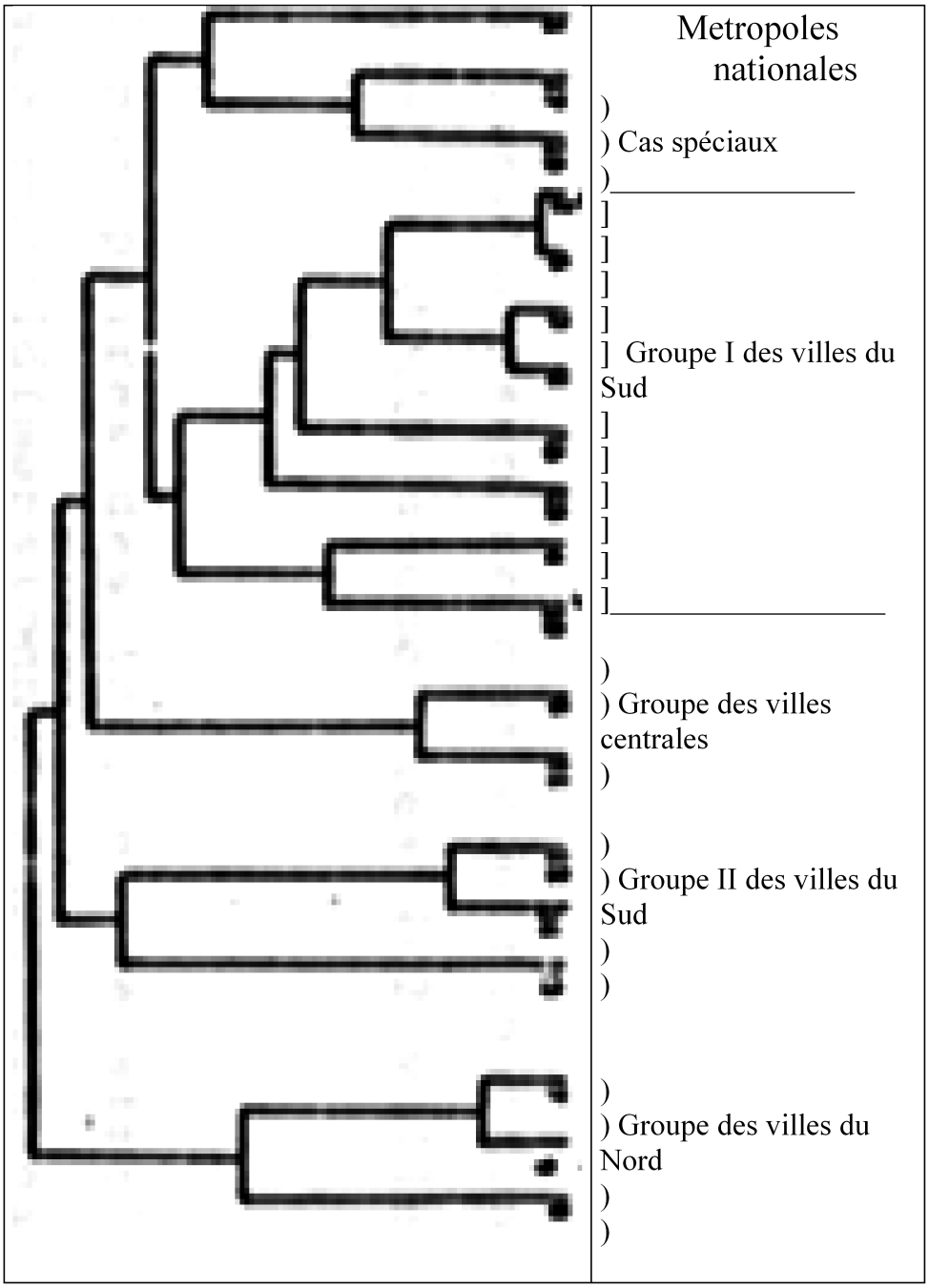


Figure7 – Arbre de classification des villes indiennes.
(Source : AHMAD, 1965)

Quant au premier facteur, soulignant une forte proportion d'hommes dans les villes du Nord, par rapport aux villes du Sud, AHMAD l'explique par la migration sélective de la population masculine dans les centres industriels du Nord, et :par une migration de la campagne

vers la ville, incluant davantage de femmes, dans le Sud.

Cette attitude pourrait avoir des causes culturelles et historiques, les femmes de la population chrétienne du Sud ayant tendance à participer davantage à la population active que celles de la population musulmane du Nord. AHMAD reconnaît que ses hypothèses s'appuient sur une base étroite de données et qu'il faudrait un ensemble de caractéristiques plus élaborées, par exemple en termes culturels et historiques, pour pouvoir les confirmer.

Pour notre part, sans nier l'intérêt de ce premier facteur, il nous semble en effet qu'un choix plus large mais surtout plus varié des indices serait une meilleure mesure de son importance ; car un grand nombre de variables démographiques redondantes (pourcentage de femmes dans la population active, pourcentage de femmes dans le commerce, pourcentage de femmes dans tel ou tel type d'industrie, etc ...) pourraient être la raison essentielle de sa prépondérance.

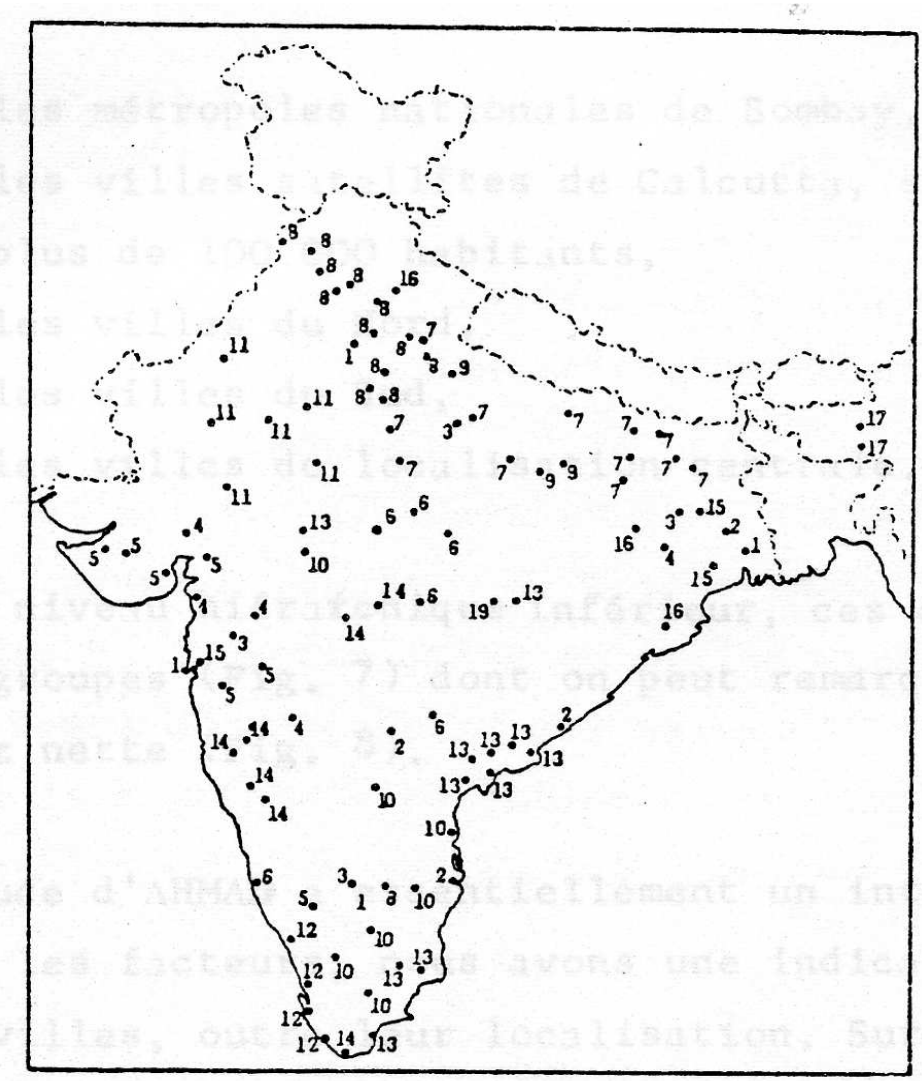


Figure 8 – Sous-groupes régionaux des villes indiennes
(Source : AHMAD, 1965)

Par l'examen des scores, les villes sont analysées d'après la contribution de chaque facteur à leur caractère ; puis elles sont classées d'après la combinaison de ces contributions. La classification optimale à laquelle AHMAD soumet les 102 villes suit la méthode des emboîtements. Le calcul par paires des indices de similarité entre les villes, dans l'espace multidimensionnel engendré par les facteurs, permet leur groupement progressif de telle manière qu'à chaque étape, l'homogénéité interne des groupes soit maximum. L'itération de fonctions discriminantes vérifie l'optimalité du groupement.

Cinq grandes classes apparaissent qui regroupent :

1 - les métropoles nationales de Bombay, Delhi et Calcutta,
2 - les villes satellites de Calcutta, soit huit villes de
plus de 100 000 habitants,
3 - les villes du Nord,
4 - les villes du Sud,
5 - les villes de localisation centrale.

A un niveau hiérarchique inférieur, ces classes se subdivisent en dix neuf sous-groupes (Fig. 7) dont on peut remarquer l'organisation régionale assez nette (Fig. 8).

L'étude d'AHMAD a essentiellement un intérêt descriptif. Cependant, par les facteurs, nous avons une indication de ce qui rapproche ces villes, outre leur localisation. Sur cette nouvelle base, un réexamen des indices bruts initiaux devrait apporter davantage d'information de détail et suggérer l'introduction de nouveaux indices pour des analyses ultérieures.

L'étude de HADDEN et BORGOTTA (1965), sur "*Les caractéristiques sociales des villes américaines"*, apporte sa contribution à une meilleure connaissance des villes, sous la forme de deux expériences méthodologiques d'une part, et d'une réflexion théorique touchant la validité de la classification des villes d'après leur fonction, d'autre part. Au cours d'une première expérience, HADDEN et BORGOTTA ont cherché à montrer les conséquences de l'utilisation, comme unités d'observations dans une analyse factorielle, des différentes définitions de la zone urbaine. Le recensement compte trois définitions :

1 - la ville centrale légale,
2 - la zone urbanisée définie d'après les critères de densité pouvant recouper diverses frontières administratives,
3 - la S.M.S.A. (Standard Metropolitan Statistical Area), formée de la somme de tous les comtés possédant une zone urbanisée, et pouvant comprendre de larges espaces non construits.

Les résultats des trois analyses factorielles séparées, effectuées sur 644 villes de plus de 25 000 habitants, d'après ces trois définitions, montrent une remarquable invariance. Lorsque les trois séries de scores obtenus sur les facteurs sont corrélés, on constate que deux facteurs seulement, la densité de population et la croissance démographique, ont une corrélation inférieure à 0.85 entre n'importe quelle paire de définition. Il n'existe donc pas une unité d'observation qui, sur le plan technique, apporte de meilleurs résultats qu'une autre. La valeur

attendue de la corrélation entre les variables tend à s'accroître avec le degré d'homogénéité des unités et avec la taille de l'unité. Or dans la pratique, les influences de taille et d'homogénéité s'équilibrent car les petites unités ont tendance à être plus homogènes que les grandes.

La question de l'utilisation de l'unité d'observation la plus appropriée est finalement davantage du domaine théorique. Une unité doit être sélectionnée essentiellement en fonction de l'échelle du phénomène que l'on étudie - par exemple les îlots sont trop petits pour étudier les taux de délinquance à l'intérieur d'une ville -. Aussi il est bien évident que lorsqu'on veut établir des comparaisons entre diverses analyses, il vaut mieux utiliser la même définition d'unité d'observation.

Dans une seconde expérience, HADDEN et BORGOTTA divisent l'ensemble des 644 villes en quatre groupes différents où les villes sont réparties en fonction de leur taille. Les résultats des quatre analyses factorielles suivies d'une rotation varimax montrent une structure similaire. Les facteurs :

1 - statut socio-économique,
2 - structure d'âge,
3 - mobilité résidentielle,
4 - éducation,
5 -densité de population et
6 - concentration de commerce et d'industrie,

sont partout présents. A l'intérieur de cette tendance générale commune, on peut noter quelques différences intéressantes comme, par exemple, l'apparition d'un facteur de "population non-blanche" dans les ville inférieures à 150 000 habitants, facteur qui est confondu avec celui de statut socio-économique dans les plus grandes villes.

Une telle structure invariante est une base solide pour la réflexion théorique. Elle conduit HADDEN et BORGOTTA à critiquer les classifications urbaines fondées uniquement sur les fonctions des villes, c'est à dire ne comportant que des indices économiques et des caractéristiques d'emploi. Constatant que la proportion de personnes employées dans le commerce a une forte corrélation (0,90 ou plus) ave la taille de la population, ils s'interrogent :

> "Est-ce que cela a un sens de parler de villes spécialisées dans le commerce de gros, de détail, dans certaines industries, etc ..., si la quantité de ces industries est directement proportionnelle à la taille ? "

En effet, un nombre croissant d'activités économiques sont maintenant orientées vers le marché. Leur localisation tend donc à se différencier selon l'accessibilité aux marchés nationaux, et selon leur position dans la hiérarchie urbaine plutôt que d'après les facteurs classiques de localisation. Bien plus, à part certaines vastes différences régionales - les communautés de plus bas statut se trouvant dans les régions rurales pauvres, et les communautés de retraités dans l'Ouest et le Sud, en général - les différences socio-économiques les plus importantes se font sentir de plus en plus à l'intérieur des métropoles. Les centres métropolitains sont multifonctionnels et la plupart de leur croissance est auto-engendrée. I1 en

résulte que l'approche taxonomique traditionnelle fondée sur les fonctions économiques ne sépare plus que des communautés relativement petites, selon deux types d'activités

- les activités économiques pour lesquelles les facteurs de localisation non métropolitaine sont encore prédominants, telles que les activités orientées vers la matière première (mine et agriculture),

- les activités localisées sans but économique telles que celles qui caractérisent les collèges, les universités, les bases militaires.

Ces réflexions fondées sur l'exemple des villes américaines ne peuvent pas s'étendre sans nuance aux villes d'Europe, où le poids considérable du "stock historique" justifie davantage la classification des villes d'après leur localisation fonctionnelle et leurs fonctions présentes, souvent héritées du passé.

KING reproche aux trois dernières études de ne rien apporter de nouveau par rapport à l'analyse empirique développée par PRICE en 1942. En introduisant un plus grand nombre de variables et d'observations, elles ne font que mettre en évidence des dimensions plus générales. Ces informations sont valables pour les différents systèmes urbains étudiés ; mais le fait que non seulement les observations, mais aussi l'ensemble des variables utilisées ne soit pas les mêmes, rend toute comparaison difficile, et en conséquence interdit toute inférence vers une théorie.

Puisqu'une comparaison systématique ne peut pas être faite, il est préférable de chercher à tester les théories existantes. C'est dans ce but que KING entreprend "*l'analyse des dimensions urbaine canadiennes*" (KING, 1966).

HADDEN et BORGOTTA ont montré que les dimensions urbaines étaient stables lorsqu'on utilisait différentes classes de ville. KING pense que cela contredit ce qu'on connaît de la "dynamique urbaine' et de la tendance vers une "métropolisation" des villes, caractérisée par des formes d'organisation complexes, des moyens de communication compliqués, un emploi tertiaire pléthorique, et une importante stratification sociale et économique. L'hypothèse veut que des variables mesurant ces caractéristiques identifient des dimensions de métropolisation ; dimensions qui devraient s'accentuer au cours du temps. Par exemple, on devrait voir apparaître plus clairement une dimension suburbaine au fur et à mesure que les villes vieillissent. En outre, les villes qui dépendent d'une activité économique particulière, comme les villes minières, pourraient subir d'importants changements individuels. Elles pourraient être plus ou moins fortement associées à des dimensions de base qui, pour leur part, n'auraient pas variées.

Pour permettre la comparaison temporelle, deux analyses en composantes principales, l'une portant sur 106 villes de plus de 10 000 habitants en 1951, l'autre sur le même ensemble de villes en 1961, sont effectuées à l'aide de cinquante deux variables économiques, démographiques, sociales et de localisation spatiale. Dans les deux analyses, environ 83 % de la variance originale de l'ensemble des cinquante deux variables est prise en compte par douze composantes, dont six seulement peuvent être identifiées, sans qu'on puisse leur donner une définition simple. Les scores indiquent à quel type de villes elles sont le plus étroitement associées. Ces dimensions n'offrent pas une grande évidence de stabilité. On retrouve dans la

première composante de 1961 des caractéristiques de la seconde composante de 1951, et vice versa.

On peut le voir par la position respective des villes dans l'espace défini parles composantes I et II pour chacune des deux périodes (Fig. 9). Mais la raison de cette inversion n'est pas claire. Est-elle due à un renforcement des relations de certaines caractéristiques dans les mêmes villes ? Ou bien y-a-t-il eu une augmentation du nombre de villes exhibant ces mêmes relations ? Cela pourrait être dû encore à un recul relatif d'autres relations.

La quatrième composante de 1961 ressemble beaucoup à la cinquième de 1951. Quant aux autres composantes de 1961, elles n'apparaissent pas ou peu en tant que telles en 1951.

Une telle comparaison n'a pas grand sens. Pour tenter de saisir les différences associées à la structure factorielle, il faudrait revenir à la matrice des corrélations et examiner soigneusement le degré et la direction des changements dans les coefficients. Et l'on sait qu'à ce niveau aussi il faut être prudent, car beaucoup de corrélations risquent d'être artificielles, du fait de la multi-colinéarité de certaines variables ou de la présence de relations non linéaires.

Cependant, le changement des dimensions urbaines au cours des dix années est indéniable ; et en poursuivant l'analyse, il est possible d'en faire apparaître quelques uns des mécanismes. Par exemple, ce changement s'est -il opéré de façon parallèle sur l'ensemble des villes ? Ou bien a-t-il touché de façon sélective certaines villes individuelles ou certains groupes de villes ?

Pour répondre à cette question, KING applique la méthode des emboîtements aux 106 villes placées dans l'espace orthogonal des douze composantes. Les distances entre les villes dans cet espace sont considérées comme des mesures de similarités ; et les villes sont regroupées progressivement d'après des mesures suivant le principe de minimiser la distance à l'intérieur des groupes. Le processus est arrêté lorsqu'il ne reste plus qu'une seule ville n'ayant pas été attribuée à un groupe. De cette manière, d'après les scores de 1951,onze groupes sont formés qui concordent avec la division régionale généralement reconnue du pays. Après une procédure identique sur les scores de 1961, le même nombre de groupes apparaît.

D'une classification à l'autre, les premier et deuxième groupes contenant les villes industrielles de la frontière ont peu changé.

> "Le fait que ces deux groupes soient restés à part au cours des dix années d'évolution reflète l'immaturité relative du système urbain et le manque d'intégration à l'ensemble du système canadien" (KING, 1966).

Ces villes dépendent encore fortement du développement des ressources naturelles locales et des fonctions industrielles correspondantes.

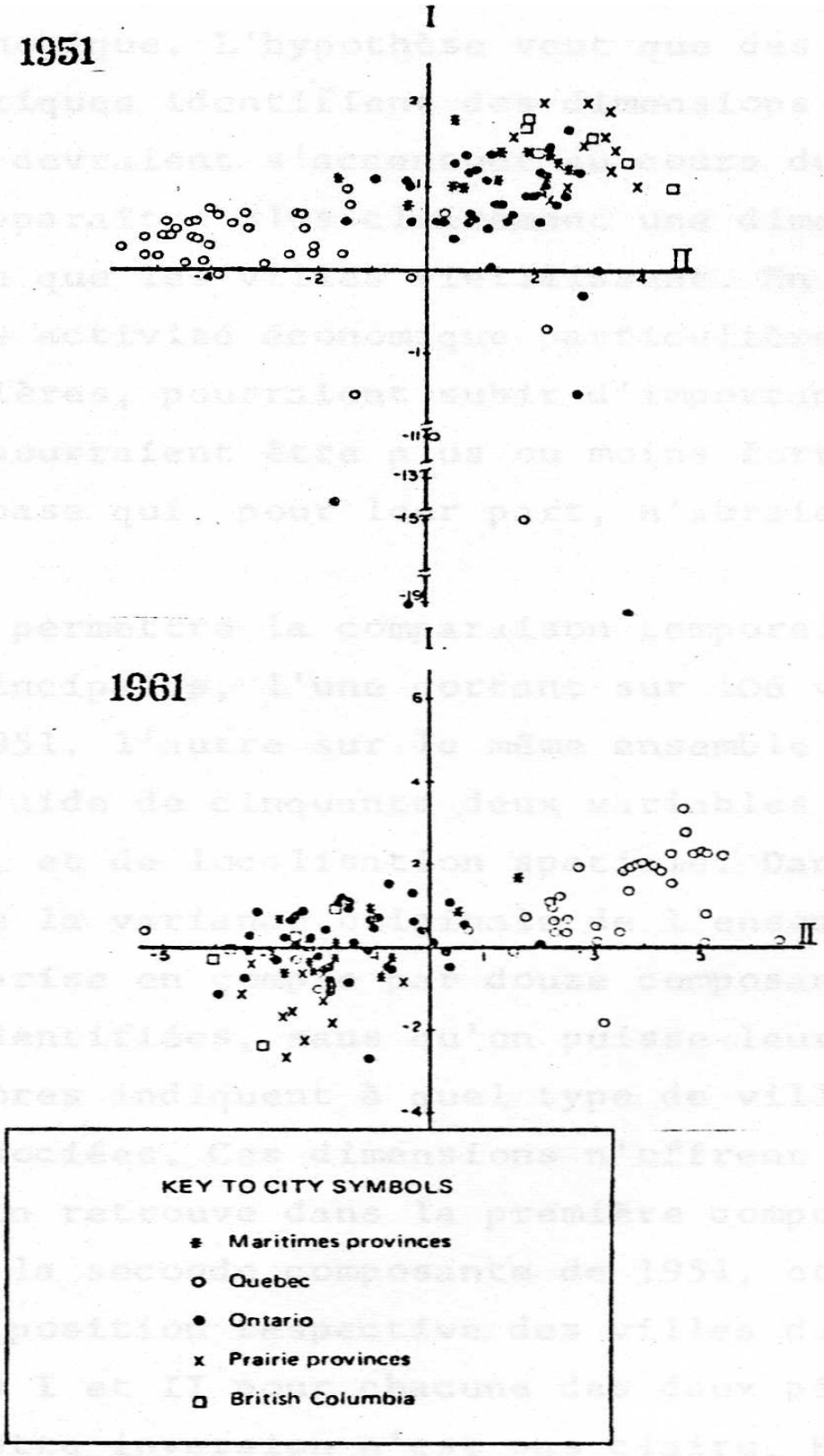


Figure 9 – Poids locaux sur les composantes I et II des villes canadiennes en 1951 et 1961. (Source : KING, 1966)

Le troisième groupe des villes d'Ontario de 1951 n'est que partiellement reproduit en 1961. On n'y retrouve plus certaines villes marginales. Celles-ci font partie en 1961 du quatrième groupe, avec les villes des Provinces maritimes et de la Prairie. Le cinquième groupe, qui inclut les petites villes du Québec, et le sixième groupe formé des riches communautés suburbaines voient également leur homogénéité renforcée. Montréal et Toronto sont à part dans un septième groupe. Les autres groupes n'incluent pas de villes régionalement intégrées.

Finalement, malgré les changements importants survenus dans les dimensions urbaines, 1a stabilité relative du classement des villes d'après ces dimensions apporte une information précieuse. Cela est indicatif d'une tendance générale vers une définition plus étroite et plus homogène des mêmes groupes de villes. KING invoque la tendance vers un ordre telle qu'on l'a étudiée depuis longtemps dans les systèmes de "places centrales" et conclut :

> "Il. se peut fort bien, au niveau plus étendu d'un système urbain national dans toute sa complexité, qu'une tendance similaire vers un ordre existe ; et que les contrastes croissants entre les groupements régionaux des villes puissent être considérés comme autant d'exemples supplémentaires de processus amplificateurs de déviations, à l'intérieur du système." (KING, 1966).

Nous venons d'examiner les études qui ont le plus contribué à la connaissance des structures urbaines et de leur évolution. Une multitude d'autres études ont été entreprises selon le même principe, avec des ensembles de données et d'observations différentes, et dans diverses parties du monde. BERRY a tenté de synthétiser leurs résultats (BERRY, 1972, Part. I). Se référant à l'étude de PRICE, à celle de HADDEN et BORGOTTA, à la sienne propre effectuée sur 1762 villes et 97 variables (BERRY, 1972, p. 15-49), et à deux ou trois autres études non publiées, BERRY met en évidence la généralité des structures latentes parmi les villes américaines.

Reprenant les définitions de HODGE (1967), il affirme que pour la période de 1930 à 1960 :

> "La structure urbaine peut être définie en termes d'un ensemble de dimensions "indépendantes", concernant au moins :
>
> a - la taille de la population,
> b - la qualité du développement physique,
> c - la structure d'âge de la population,
> d - la base économique,
> e - le niveau d'éducation de 1a population,
> f - l'orientation ethnique ou religieuse,
> g - le degré d'assistance sociale,
> h - la situation géographique
>
> La base économique des centres urbains tend à agir indépendamment des autres traits structuraux urbains […] avec l'exception de l'organisation hiérarchique des activités orientées vers le marché".

Puis BERRY établit des comparaisons d'après les études faites dans d'autres pays comme celle de MOSER et SCOTT en Angleterre, celle de KING au Canada, de AHMAD en Inde, de MABOGUNGE (1965) et de Mc NULTY (1972) au Nigéria et au Ghana, de FISHER (1966) en Yougoslavie et de lui-même au Chili (BERRY, 1965).

I1 montre que, dans chaque société, les principales dimensions socio-économiques qui peuvent servir de façon stable à distinguer les villes entre elles, sont celles de statut social et de structure par âge, bien que ces facteurs n'apparaissent de façon indépendante que dans les pays "de plus haut niveau de développement". A un moindre niveau, comme en Angleterre, par exemple, le revenu et la structure familiale ne sont pas séparés. A un niveau encore inférieur, les différences de rang social et de structure d'âge se retrouvent à travers des différences régionales, opposant le coeur du pays aux tendances modernistes et la périphérie encore marquée de traditionalisme. En Inde, par exemple, les quatre grandes métropoles correspondant à la structure de plus grande accessibilité servent de foyer de modernisme. Au Canada, les différences de statut économique et de structure d'âge coïncident avec les différences culturelles opposant les anglais aux français et, de ce fait, les dimensions socio-économiques et ethniques se trouvent confondues ; alors qu'aux Etats-Unis, la dimension ethnique et raciale apparaît de façon indépendante.

LES ETUDES D'ECOLOGIE FACTORIELLE INTRA-URBAINE

Le chapitre I de la première partie de cet ouvrage retrace brièvement l'histoire de l'écologie factorielle urbaine qui est née et s'est développée grâce aux chercheurs de l'Ecole de Chicago. Nous rappelons que de nombreuses écologies factorielles ont été entreprises à la suite des travaux des sociologues SHEVKY et BELL (1955), sur "l'analyse des zones sociales". Les résultats de ces analyses confrontés aux modèles des économistes BURGESS (1925), HOYT (1939) et HARRIS et ULLMAN (1945) ont conduit à la construction d'un cadre théorique présentant trois hypothèses :

1 - le statut économique tend à être associé avec des indice de revenu, d'emploi et d'éducation, et tend à être organisé en secteurs.
2 - le statut familial tend à être associé avec l'âge, la taille de la famille active, et tend à être organisé en anneaux concentriques.
3 - le statut ethnique est associé à la race, la nationalité d'origine, la religion, et tend à former des noyaux qui se superposent à la structure cellulaire créée par la combinaison des schémas sectoriels et concentriques.

Parmi toutes les études qui ont contribué à la formulation et à la vérification de ces hypothèses, nous examinerons celle qui, par la précision de ses analyses et par la richesse de ses résultats, nous parait l'une des plus éminentes. Puis, nous essaierons de faire la synthèse des résultats qu'offrent l'ensemble des applications.

Le premier objectif de l'"écologie factorielle de Chicago" (REES, 1970) est de tester les hypothèses ci-dessus à travers la recherche des dimensions de différenciation qui sont responsables de la configuration socio-économique de la ville de Chicago. Un objectif

secondaire -non mentionné explicitement par l'auteur - semble être aussi celui de tester la validité d'une méthode, applicable de manière identique à d'autres villes ; ce qui est la condition nécessaire à toute entreprise comparative. On pourrait contester la valeur de ces tests mutuels s'ils ne s'appuyaient pas précisément sur Chicago, sans doute une des villes les plus étudiées et les plus connues au monde.

La vérification de la théorie guide REES dans la sélection de 57 caractéristiques ayant trait à la population d'une part (âge, statut ethnique et religieux, revenu, emploi, éducation), et au logement d'autre part (valeur, loyer, taille, vétusté). La question de savoir dans quelle mesure la théorie correspond mieux à certains champs d'étude conduit l'auteur à entreprendre 3 analyses concernant : 1) la ville centrale ; 2) les faubourgs proches et la banlieue, et 3) la ville métropolitaine qui est formée de la somme des 2 précédentes définition. En raison d'une plus grande maniabilité des observations et d'un certain souci de généralisation, les unités d'observations choisies sont les "communities areas" de la ville centrale et les municipalités de banlieue, soit 222 unités en tout pour la ville métropolitaine. Dans un second temps, une analyse au niveau des 1 324 census tracts, mais n'incluant plus que 12 variables sélectionnées d'après leur "performance" dans les analyses précédentes, permet d'apporter des modifications de détail dans les résultats.

Les 4 analyses sont construites sur le même modèle factoriel - extraction des facteurs par une analyse en composantes principales, suivie d'une rotation varimax vers une structure simple. Nous ne considérerons ici que les résultats de l'analyse sur les 222 unités d'observation de la ville métropolitaine parce que ce sont les plus conformes à la théorie ; et aussi parce qu'il serait trop long d'examiner en détail les divergences d'ailleurs peu importantes et aisément explicables notées dans les autres analyses.

Dix principaux facteurs ressortent. Leur interprétation à l'aide des caractéristiques dont les saturations sont les plus élevées sur les facteurs en question, permet de tester la validité des indices de SHEVKY et BELL. Les poids locaux, c'est à dire les scores des unités d'observation sur ces facteurs répartis en 4 classes de taille à peu près égale, sont cartographiés à deux fins : 1) faciliter l'interprétation des facteurs, et 2) voir pour quel fait leur distribution spatiale corres-pond aux modèles de BURGESS, HOYT et ULLMAN et HARRIS.

Les fortes saturations positives de certaines variables : scolarité élevée, cols blancs, professions libérales, commerces, bureaux, revenus élevés, loyers élevés, logements de bonne qualité, et les fortes saturations négatives des variables opposées, désignent le premier facteur comme un facteur de statut socio-économique. Le deuxième facteur associé aux variables de taille de 1a famille et de structure d'âge fait ressortir les différences entre les communautés de population d'après leur cycle de vie (célibataires, jeunes couples, gens âgés, familles nombreuses). I1 est intitulé "stage dans le cycle de vie" par "statut familial". Les variables étroitement liées au 3ème facteur, appelé "race et ressource", sont des indices ethniques (population noire) et aussi certains indices de statut économique (emploi dans les services, revenus bas, logement de qualité médiocre, absence de voiture). Parmi les sept autres facteurs identifiés par REES, 5 sont des facteurs ethniques associés aux immigrants catholiques, à la population juive, aux populations irlandaises et suédoises, et aux autres populations non blanches. La taille et la densité de la population d'une part, sa mobilité d'autre part apparaissent encore avec 2 autres facteurs.

Bien que ces 7 facteurs soient intéressants dans la mesure où ils complètent une description de la structure particulière de 1a ville de Chicago, nous laissons le soin au lecteur qui le désire d'en voir le commentaire dans l'étude de REES (REES, 1970). Puisque nous nous attachons à montrer l'intérêt de l'écologie factorielle de Chicago, en tant que modèle général pouvant servir de base de comparaison pour d'autres études, nous concentrerons notre attention sur les 3 premiers facteurs, dont on constate qu'ils correspondent de façon remarquable aux 3 indices proposés par SHEVKY et BELL. Reste à savoir si la distribution spatiale de ces facteurs correspond à un schéma clair et défini, comme celui que proposent BURGESS, HOYT et ULLMAN et HARRIS.

L'inspection des poids locaux à travers leur représentation cartographique renseigne déjà. Trois cartes montrent respectivement que :

1) la structure spatiale du statut socio-économique a tendance à être sectorielle ;
2) le statut familial tend à s'organiser en zones concentriques et
3) le statut ethnique se regroupe en secteurs et en noyaux distincts.

Mais il faut se méfier de cette tendance ordonnée qu'offre chacune des cartes sur les 3 facteurs ; car elle n'est peut-être, en grande partie, qu'une impression visuelle dictée par les modèles que nous avons à l'esprit. Pour s'assurer de l'objectivité de cet ordre, REES a recours à 2 techniques annexes.

La première consiste à diviser la carte de la métropole de Chicago en zones concentriques et en secteurs réguliers (Fig. 10) et à examiner les profils qu'offrent ces différents secteurs et ces différents anneaux sur les 3 facteurs. Sur le premier facteur, on note des différences .assez nettes entre les profils des divers secteurs, mais avec des variations internes des profils des secteurs en fonction de la distance au centre (Fig. 11). On remarque en effet que près du centre de la ville, un profil peu élevé correspond aux jeunes célibataires et aux gens âgés à revenus moyens dans l'ensemble ; ou aux familles nombreuses pauvres ne pouvant se permettre de loger dans les maisons individuelles de la périphérie. Le pic du profil indique que les résidents suburbains sont en général au sommet de leur capacité de revenu. Quand on s'éloigne encore davantage de 1a ville, l'apparition de résidents ruraux à revenus plus modestes fait s'infléchir à nouveau la courbe.

Sur le 2ème facteur, des différences marquées se voient entre les profils des anneaux, mais avec des variations à l'intérieur de chaque anneau selon la direction du secteur où l'on se trouve.

Sur 1e 3ème facteur, les profils montrent des "bonds" remarquables dans 2 directions et à une distance bien définie du centre de la ville.

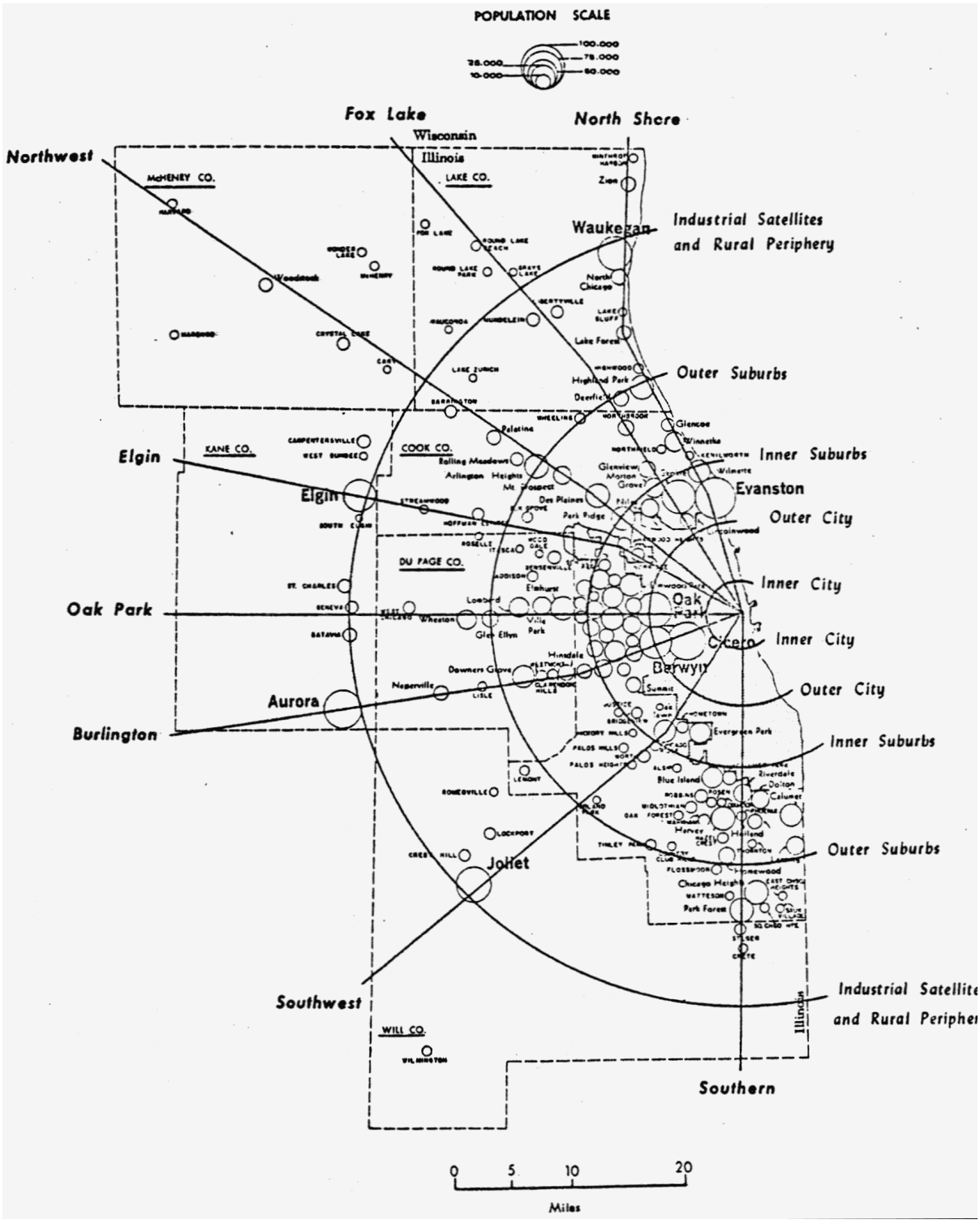
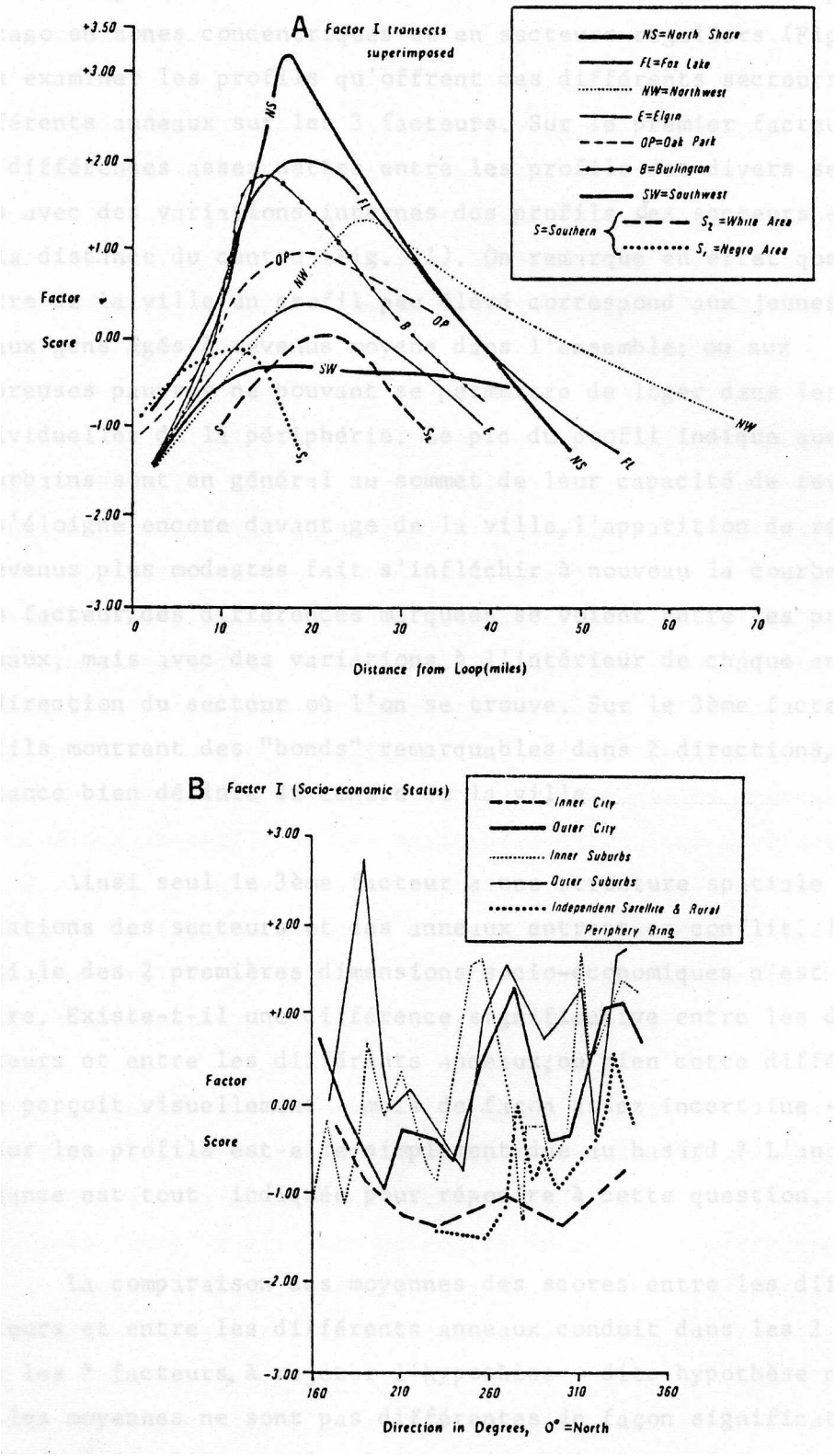


Figure 10 - Division de la ville métropolitaine de Chicago en secteurs et en zones concentriques. (Source : REES, 1970)



Distance from Loop

Figure 11 – Profil des secteurs et des anneaux sur le facteur de statut socio-économique. (Source : REES, 1970)

Ainsi, seul le 3ème facteur a une structure spatiale ambiguë. Les variations des secteurs et des anneaux entrant en conflit, l'organisation spatiale des 2 premières dimensions socio-économiques n'est pas aussi claire. Existe-t-il une différence significative entre les différents secteurs et entre les différents anneaux ? Ou bien cette différence que l'on perçoit visuellement - mais de façon assez incertaine - sur la carte et sur les profils est-elle simplement due au hasard ? L'analyse de variance est tout indiquée pour répondre à cette question.

La comparaison des moyennes des scores entre les différents secteurs et entre les différents anneaux conduit dans les 2 cas, et pour les 2 facteurs, à rejeter l'hypothèse - dite hypothèse nulle - que les moyennes ne sont pas différentes de façon significative, au seuil de 0,01, Bien que les 2 types de variation soient significatives pour chacun des 2 facteurs, il apparaît que pour le 1er facteur, la variation par secteur est davantage significative, alors que pour le 2ème facteur, la variation zonale est plus forte. Par contre, si l'on compare les moyennes des secteurs avec celles des anneaux, on a 5 % de chance pour que leurs différences soient dues au hasard, ce qui conduit à accepter l'hypothèse nulle dans ce cas.

Les résultats de l'analyse de variance permettent donc à REES d'affirmer que le statut socio-économique et que le statut familial varient essentiellement et respectivement par secteurs et par zones (alors que la variation de zones à secteurs n'est pas significative) ; et que les modèles de BURGESS et de HOYT sont bien adaptés à la ville de Chicago.

Si, à la lumière des théories qui ont conduit BURGESS, HOYT, ULLMAN et HARRIS à établir leurs modèles classiques de localisation des résidences urbaines, on examine à nouveau les 3 dimensions de base de la structure urbaine, on peut faire la part des forces générales qui les mettent en place et des distorsions dues à l'histoire et à la situation géographique particulière de la ville. La théorie économique d'enchères sur les prix des terrains les mieux situés par leur accessibilité au centre, ou par l'agrément de leur environnement est à la base du modèle de HOYT. On voit en effet qu'à Chicago, les communautés de statut élevé monopolisent les meilleures situations géographiques : le long des principaux axes de transport vers le nord-ouest et vers l'est, le long de la façade du lac au nord et, dans une moindre mesure, au sud. Les communautés de bas statut, écartées par des loyers et des prix trop élevés, s'insèrent entre les secteurs privilégiés.

Cependant, à l'intérieur de chaque secteur, il y a interférence de la structure en anneaux, dérivée de la théorie "d'invasion-succession" de BURGESS, sur la croissance urbaine. Le statut socio-économique varie directement avec la distance du centre de la ville parce que les immigrants, en général pauvres, s'établissent dans les résidences anciennes et délabrées du centre puis s'éloignent dans des quartiers plus récents au fur et à mesure que leur statut s'améliore. La présence dans ce schéma régulier de noyaux secondaires d'activités - industriels pour la plupart - autour desquels sont concentrées les communautés de travailleurs de statut plus élevé, manifeste l'interférence du modèle de ULLMAN et HARRIS.

La répartition des communautés par stage dans le cycle de vie suivant un schéma concentrique dépend pour une grande part des types de logements offerts, qui sont eux-mêmes fonction de la distance du centre de la ville. Les grands immeubles à étages au centre reflètent le coût du terrain et attirent de façon sélective, en raison de la plus grande accessibilité aux lieux d'activité, une population plus âgée que la moyenne, composée de célibataires, de jeunes

couples ou de couples âgés sans enfants. Dans les anneaux de la périphérie, se trouvent des résidences individuelles entourées de vastes pelouses qui attirent les familles ayant des jeunes enfants. Entre ces deux extrêmes, des maisons urbaines de 2 ou 3 étages abritent des familles d'âge intermédiaire.

Les exceptions que constituent des différences significatives de statut familial entre certains secteurs sont dues à la superposition de la structure spatiale du facteur ethnique et de la géographie physique locale. L'interférence du statut ethnique se manifeste par la présence près du centre de 1a ville de 2 noyaux de populations jeunes, ghettos noirs où sont confinées des familles chargées d'enfants. Quant à la géographie physique locale, elle entraîne une révision de la notion de centre-ville en termes de valeur du terrain. En effet, la présence du lac fait que le "centre" de 1a ville s'étire le long de sa façade en un étroit ruban de hauts immeubles résidentiels, formant un secteur, plutôt qu'un cercle particulier de cycle de vie.

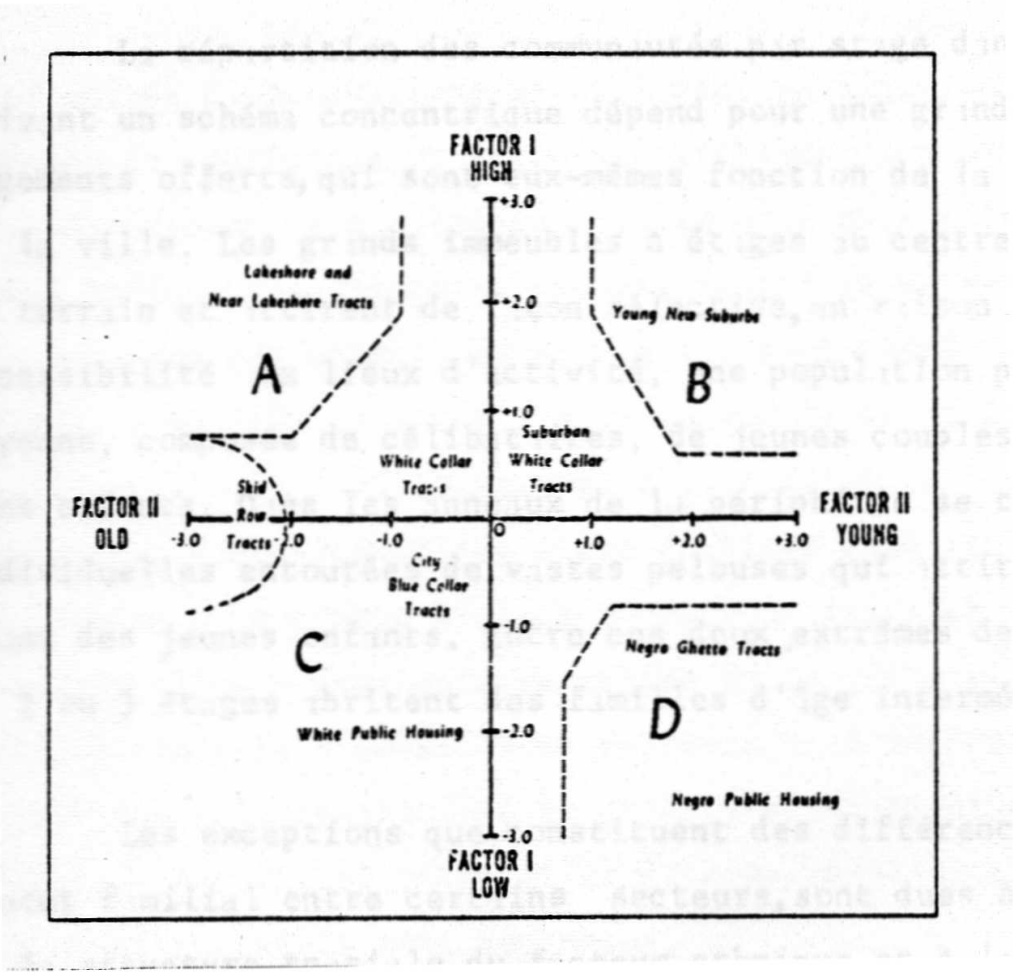


Figure 12 – L'espace social de Chicago
Facteur I : statut socio-économique ; Facteur II : Stage du cycle de vie
(Source : REES, 1970)

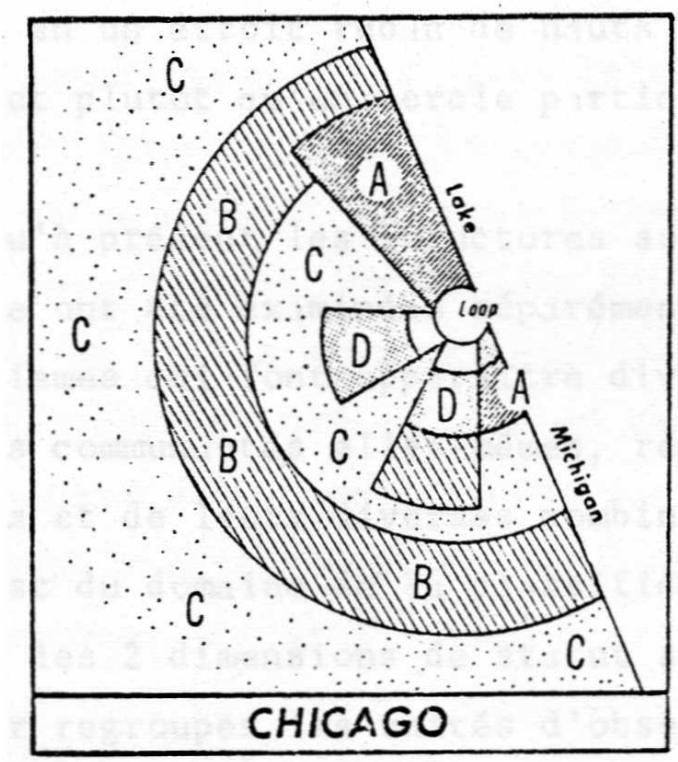


Figure 13 – Les zones sociales de la métropole de Chicago
(Source : REES, 1970)

Jusqu'à présent, les structures socio-économiques et leur organisation spatiale ont été examinées séparément pour mieux saisir les forces ou les mécanismes qui font apparaître divers types de communautés. Reste à définir les communautés elles-mêmes, résultats de l'action simultanée de ces forces et de leurs diverses combinaisons possibles. Une telle définition est du domaine de la classification. REES utilise l'espace engendré par les 2 dimensions de statut socio-économique et de statut familial pour regrouper les unités d'observation qui se ressemblent. Une première répartition grossière peut se faire à vue, en considérant les quadrants délimités par les dimensions (Fig. 12). Ces 4 classes sont alors

utilisées pour construire un croquis schématique qui fournit une image résumée de la répartition des communautés (Fig. 13).

Une classification plus précise, considérant les distances entre les unités d'observations et utilisant un algorithme de groupement, permet de construire un arbre hiérarchique dans lequel il est possible de sélectionner à différents niveaux un nombre différent de groupes. RÉES choisit d'examiner 22 groupes de taille moyenne. Parmi ceux-ci, on remarque par exemple le groupe 1 : communauté à statut social élevé, plus âgée et moins familiale que le reste de la population, résidant à 5 ou 8 miles du centre ; le groupe 3 : communautés familiales de statut socio-économique élevé établies dans des zones plus éloignées du centre ; le groupe 16 : communauté de familles moyennes, de statut bas, établies à proximité du centre de la ville et dans la périphérie rurale. On aura l'occasion de reparler de l'intérêt d'une telle classification synthétique et objective, qui rejoint celui de la classification des villes entre elles et même au-delà, celui des régionalisations utilisant la même technique.

Cependant, notre objectif principal. en présentant l'étude de REES était de montrer la

participation du modèle factoriel à la vérification d'une théorie d'écologie urbaine ; et avant de conclure sur les grandes capacités de la méthode dans ce cas, il nous reste à examiner à travers d'autres écologies factorielles dans quelle mesure méthode et théories, devenues indissociables, s'appliquent à d'autres villes.

La comparaison d'écologies factorielles individuelles qui diffèrent par les variables, le champ d'étude, les unités d'observation, la période et le modèle factoriel sélectionné est délicat. REES (1972) l'a abordé en préparant des résumés de format standard pour 35 écologies factorielles de villes américaines, et les quelques écologies factorielles de villes non américaines qu'il a pu rassembler . Ces résumés établissent des équivalences entre les diverses variables utilisées et les regroupent en ensembles, correspondant aux 3 dimensions théoriques de statut socio-économique, statut familial et statut ethnique. Jusqu'à quel point les diverses analyses ont effectivement fait apparaître les 3 facteurs correspondant à ces 3 ensembles de variables théoriques ? En s'inspirant d'un schéma proposé par REES (1970), on peut essayer de répondre à cette question. Voici quelques unes des différentes combinaisons qu'offrent les correspondances entre les 3 ensembles de variables théoriques et les facteurs effectivement découverts.

Combinaison n° 1

Ensemble de variables	Correspondance	Facteurs
SSE = statut socio-économique	---------->	A
SF = statut familial	---------->	B
SR = statut racial ou ethnique	---------->	C

Chaque ensemble de variables donne un facteur indépendant. C'est la combinaison théorique "idéale" qui n'apparaît jamais exactement comme telle dans la réalité. Dans les diverses études, on trouve sous des noms différents le facteur identifié par les variables socio-économiques: "*statut socio-économique*", "*rang social*" ou "*prestige du quartier*" . De même, le facteur identifié par les variables démographiques et les caractéristiques d'habitat prend les noms de "*statut familial*", "*stage dans le cycle de vie*", "*urbanisation*", "*âge de l'habitat*". Enfin, le facteur identifié par les caractéristiques ethniques peut s'intituler "*statut ethnique*", ou "*race*", ou encore "*ségrégation*".Certains géographes, voulant se distinguer par là des sociologues, préfèrent employer des termes invoquant une structure spatiale comme "*prestige du quartier*", "*urbanisation*", ou "*ségrégation*", plutôt que des termes sociologiques. L'organisation spatiale des trois facteurs suit respectivement des schémas sectoriels, concentriques, et polynucléaires.

Combinaison n° 2

Ensemble de variables	Correspondance	Facteurs
SSE = statut socio-économique	----------->)	A
SR = statut racial ou ethnique	----------->)	C
SF = statut familial	---------->	B

Les trois facteurs sont séparés, mais il existe un certain degré d'association du facteur ethnique et du facteur socio-économique, sous la forme de saturations secondaires de certaines caractéristiques du statut social avec le facteur ethnique. Cette combinaison est typique de la plupart des villes américaines de l'est, du nord, et de l'ouest, comme Manhattan (CAREY, 1966), Boston (SWEETSEF, 1965), Chicago (BERRY et TENNANT, 1965; et REES, 1971), Los Angeles et San Francisco (BELL, 1955). On la retrouve au Canada (MURDIE, 1969) avec la population italienne comme principal indicateur du statut ethnique.

En cherchant à vérifier, au Canada, le postulat que les petites villes devaient montrer moins d'ordre et moins de différenciation interne, BOURNE et BARBER ont conclu que toutes les villes étudiées montrent quelque forme de différenciation spatiale des groupes socio-économiques, et en particulier, ils ont noté une zonation concentrique marquée du statut familial.

Combinaison-n° 3

Ensemble de variables	Correspondance	Facteurs
SSE = statut socio-économique	---------- /---------->	A
SR = statut racial ou ethnique	----------/	C
SF = statut familial	---------->	B

Seuls les 2 premiers facteurs A et B apparaissent séparés. Les variables mesurant le statut racial ne font pas ressortir de facteur distinct, car elles possèdent de fortes saturations avec le facteur de statut socioéconomique dans lequel elles se fondent. Cette combinaison caractérise les villes du sud des Etats-Unis telles que Birmingham, Miami, Shreveport (C.U.S., 1968), où les noirs sont désavantagés socialement de façon plus systématique qu'ailleurs. Pour une raison similaire, quoique dans un sens opposé, cette structure se retrouve dans une ville comme Helsinki (SWEETSER, 1960) où la minorité ethnique représentée en l'occurrence par les Suédois occupe systématiquement le haut de la hiérarchie sociale et économique. Le cas de la ville de Copenhage, où le 3ème facteur est aussi absent, est particulier, puisque l'auteur (PEDERSEN, 1967), ayant perçu la ville comme éthniquement homogène, n'a introduit aucune caractéristique de groupe minoritaire.

Il en va de même pour les grandes villes suédoises étudiées par JANSON. Dans ces dernières, en outre, on constate que le statut familial est représenté par 3 dimensions qui correspondent aux 3 stages du cycle de vie : "familles jeunes" associées avec les banlieues nouvelles ; "familles établies" dans les zones développées vers les années 1940 ; et population "post familiale" dans les zones centrales les plus anciennes. Cette apparition de 3 facteurs au lieu d'un seul reflète plusieurs sous-groupes résidentiels nés de l'immobilité relative de la population suédoise par rapport à la population américaine.

Combinaison n° 4

Ensemble des variables	Correspondance	Facteurs
SSE = statut socio-économique	----------/---------->	A
SF = statut familial	----------/	(B)
SR = statut racial ou ethnique	---------->	C

Le statut socio-économique et le statut familial sont confondus, si bien que 2 facteurs seulement sont mis en évidence. C'est le cas de la ville du Caire (ABU LUGHOD, 1969) où, de façon générale, plus les familles sont de rang socio-économique élevé, moins elles ont tendance à avoir de nombreux enfants. La même structure apparaît à Calcutta (BERRY et REES, 1969) et dans d'autres villes indiennes (BERRY, 1971). A ce propos il convient de signaler que le facteur reconnu comme facteur de statut familial par BERRY, dans son zèle à faire correspondre la théorie à la réalité, représente en fait une concentration d'hommes dans les zones centrales de la ville. Le même facteur apparaît aussi au Caire où J. ABU LUGHOD lui donne le nom de facteur de *dominance masculine*. Le fait qu'un grand nombre d'hommes, célibataires ou non, migrent pour travailler dans l'industrie urbaine laissant leurs familles ou leurs femmes au village, explique l'importance de ce facteur dans les villes non occidentales. Le facteur ethnique (castes, religions) se manifeste de façon distincte bien qu'il soit assez fortement lié au facteur socio-économique,

D'autres études non publiées, comme celle de PYLE et MORRIS à Rio de Janeiro ou celle de FOREST PITTS à Séoul, montrent le même schéma structurel. Quant au schéma spatial correspondant, il est aussi typique des villes pré-industrielles. Les plus hauts statuts se trouvent au centre et les classes inférieures occupent la périphérie (favelas). Cependant, parmi les villes les plus développées, on voit se dessiner une nouvelle tendance due à l'impact de l'industrialisation. Les basses classes s'introduisent à l'intérieur de la ville et l'élite investit de plus en plus les zones les plus agréables de la périphérie (plages, collines verdoyantes, etc ...).

Combinaison n° 5

Ensemble des variables	Correspondance	Facteur
SSE = statut socio-économique	----------/--------->	A
	/	
SF = statut familial	----------/---------->	B
SR = statut racial ou ethnique	---------->	C

Les 3 facteurs apparaissent mais il ne sont pas en position dominante et viennent après d'autres facteurs prépondérants. Cette structure peu. marquée caractérise les villes britanniques comme Merseyside (GITTUS, 1964), Sunderland (ROBSON, 1969) et Cardiff et Swansen (HERBERT, 1970). Dans ces villes, les facteurs les plus évidents sont liés à l'habitat, opposant d'une part les logements privés avec les logements appartenant à la municipalité, et d'autre part les habitations anciennes de qualité médiocre et les habitations nouvelles de qualité conforme

ou supérieure aux normes. A travers ces dimensions d'habitat, on retrouve bien sûr des caractéristiques du statut socio-économique; - les classes élevées occupant les logements de meilleure qualité - , et du statut familial - les jeunes familles étant plus généralement logées par la municipalité - ; mais les associations sont trop faibles et varient trop d'une ville à l'autre pour qu'on puisse considérer ces dimensions comme de bons indices du statut socio-économique et du statut familial. En outre, aucune structure spatiale définie ne semble associée à ces dimensions. On pense que cela est dû à l'intervention fréquente des autorités publiques dans le domaine de l'habitat (rénovation des vieux quartiers du centre ou constructions nouvelles en banlieue).

On constate, avec les exemples des combinaisons 4 et 5 et surtout avec la dernière combinaison, que la théorie de l'écologie factorielle a atteint les limites de sa validité. Il faudrait construire un autre cadre théorique, adapté au contexte particulier des villes non occidentales ou des villes britanniques et, pour diriger les investigations futures dans ce domaine, les divergences notées avec la première théorie sont des indications précieuses.

Une des critiques les plus vivement formulées envers les études intra-urbaines concerne l'orthogonalité des 3 dimensions socio-économiques qu'elles font apparaître. En effet, toutes les écologies factorielles utilisent une technique similaire ; analyse en composantes principales ou analyse factorielle, suivie d'une rotation varimax vers une structure simple. Or, l'orthogonalité des axes ainsi maintenue signifie que théoriquement les groupes de variables désignés par ces axes sont linéairement non corrélés et donc statistiquement indépendants. Cette notion d'indépendance, si elle est prise à la lettre, peut conduire par exemple à l'absurdité de considérer 1e rang social et le statut racial comme étant parfaitement indépendants. La non coïncidence du langage conceptuel et du langage mathématique peut avoir trois origines :

1°) elle peut venir du fait que les variables ne vérifiant pas l'hypothèse de la multi-normalité de leur distribution, sont sans doute linéairement non corrélées mais ne sont pas indépendantes, et peuvent être liées par des relations courbes ou autres.

2°) elle peut être due à l'impossibilité de réaliser parfaitement le critère de structure simple. En conséquence les variables qui donnent leur nom au concept de statut social ou de race, par exemple, sont fortement mais incomplètement corrélées avec l'axe qui les désigne, et de ce fait, elles ont des corrélations secondaires avec les autres axes.

3°) et enfin, même si les contraintes de multi-normalité et de structure simple sont respectées, il reste le fait que les concepts par lesquels on décrit ces variables ne représentent qu'imparfaitement les "êtres mathématiques" indépendants qu'elles forment ; êtres abstraits, difficiles à saisir, qui ne correspondent pas à l'ensemble du statut social ou du statut racial, mais plutôt aux portions de ces concepts qui sont mutuellement indépendants.

En raison de ces difficultés sémantiques, et en particulier pour respecter plus parfaitement le critère de structure simple, certains chercheurs préconisent la rotation oblique qui devrait donner une image plus juste de la réalité: Par exemple BELL (1955) dans ses travaux sur Los Angeles et San Francisco, a fait tourner graphiquement à la main les facteurs économiques et ethniques, de façon à ce qu'ils correspondent au mieux aux 2 groupes de caractéristiques qui les décrivent respectivement. La relation de ces 2 groupes apparaît sous la

forme de la corrélation modérée (-0,62) des 2 facteurs. Par contre HAYNES, dans une étude non publiée sur la ville de Montréal, n'a trouvé, en effectuant des rotations obliques, que des corrélations extrêmement faibles entre les facteurs et en conclut que la structure de cette ville est naturellement orthogonale.

Ainsi, il peut y avoir divers degrés d'association entre les facteurs, depuis l'indépendance presque totale jusqu'à la dépendance complète, comme nous en avons vu plusieurs exemples dans les quelques combinaisons présentées plus haut. I1 est certain que le degré d'association des facteurs est un moyen de comparaison précieuse, et que cela devrait encourager l'utilisation des rotations obliques.

Une autre critique concerne le manque de rigueur de certaines conclusions, inférant sur le comportement des individus à partir de données agrégées spatialement. Nous avons discuté longuement du problème des corrélations écologiques fallacieuses, dans le 1er chapitre de cette 2ème partie. Nous rappellerons simplement que les "census tracts", unités les plus couramment employées dans les écologies factorielles, peuvent manquer d'homogénéité. Une valeur moyenne pour un tract peut cacher une description de la zone considérée ayant un sens. Par exemple, il suffit de longer les bords du lac à Chicago pour voir que le secteur décrit par REES comme riche et de moyenne d'age élevée n'est souvent pas plus large qu'un ilôt ou même qu'un immeuble. Très vite, parfois simplement en regardant de l'autre côté de la rue, on voit apparaître des maisons à un ou 2 étages délabrées, abritant des familles nombreuses pauvres.

Au niveau de l'interprétation, il faut regretter aussi que trop souvent les dimensions soient nommées sur la base d'un trop petit nombre de variables clefs. Les auteurs semblent satisfaits avec la simple vérification d'une théorie et ne cherchent pas à voir s'il existe d'autres régularités significatives. Celles-ci pourraient être dévoilées non seulement en examinant les saturations secondaires sur les premières dimensions et les dimensions mineures , mais aussi en introduisant en début d'analyse des caractéristiques plus variées. Puisqu'il s'agit d'écologie, si l'on prend le terme dans son sens propre, cela ne suppose-t-il pas une meilleure connaissance du comportement des gens vis-à-vis de leur milieu ? Par exemple, des données sur une utilisation du sol autre que résidentielle - industries, commerces, centres d'achat, administrations, parcs, distractions - seraient fort utiles. Enfin, des caractéristiques du paysage urbain avec la structure de ses constructions, de ses rues, ses arbres, ses monuments, ses symboles, pourrait aider à saisir la notion d'appartenance à un quartier ou à un voisinage aussi bien que l'âge, la richesse de la profession de ses résidents, ou tout au moins de façon complémentaire.

CONCLUSION

En résumé de ce chapitre sur l'application des méthodes factorielles aux études urbaines, nous noterons qu'elles consistent essentiellement en la recherche de structures fondamentales qui différencient les villes entre elles ou différencient les quartiers d'une même ville. Alors que les études de la géographie traditionnelle s'attachent à une description "idiographique" des villes, expliquant leur caractère propre par leur histoire ou leur situation géographique unique, les analyses factorielles devraient permettre de faire la part de ce qui est général dans un cas apparemment unique en tous points. Nous avons vu qu'elles y réussissent, en particulier celles

qui, comme les écologies factorielles intra-urbaines, sont fondées sur une théorie. Les études interurbaines y parviennent dans une moindre mesure. Par exemple, elles montrent que certaines fonctions économiques dépendent de manière générale de la taille de la ville, et qu'il ne sert à rien de différencier les villes d'après ces fonctions, comme le font certaines classifications traditionnelles ; mais qu'il vaut mieux concentrer son attention sur d'autres dimensions plus discriminantes, comme le statut social et la structure par âge, dimensions qui de façon constante servent à distinguer les villes dans des systèmes urbains variés.

Cependant, on atteint très vite les limites de la généralisation, non seulement parce qu'en passant d'un type de société à l'autre -sociétés capitalistes à libre concurrence, sociétés planifiées, sociétés pré-industrielles - les conditions changent de façon brutale, mais aussi parce que les différentes études sont trop disparates dans leurs définitions du champ d'application, des données et dans les méthodes utilisées. Tant qu'on peut trouver des dimensions latentes qui ont des significations similaires, les cas peuvent être comparés ; et, dans ce but, il apparaît que le recours à la rotation conduisant à une plus grande invariance de la structure factorielle est indiqué. Par exemple, HADDEN et BORGOTTA qui ont utilisé les mêmes données et la même méthode -faisant appel à la rotation varimax -, pour analyser différentes classes de villes, ont retrouvé sensiblement la même structure factorielle dans ces différentes classes. KING par contre, comparant les résultats de 2 analyses en composantes principales identiques, faites sur les mêmes données et dans le même champ d'étude à deux périodes différentes, n'a pu rendre compte de façon satisfaisante des grandes différences survenues dans les composantes.

En ce qui concerne les classifications effectuées sur les résultats d'analyses factorielles, nous rappellerons que les distances entre les observations dans l'espace des composantes orthogonales sont évaluées de façon plus satisfaisantes que les distances calculées dans l'espace des facteurs après rotation. Mise à part cette restriction qui contredit celle que suscite le besoin de comparaison, les classifications ont fourni des résultats descriptifs intéressants, montrant parfois comme en Inde des régionalisations marquées. Elles permettent aussi, à KING par exemple, en montrant l'évolution de certains groupements dans le temps, de tirer un meilleur parti des résultats d'analyse en composantes principales qu'une comparaison hasardeuse des dimensions.

CONCLUSION

Renonçant à fournir une liste exhaustive de toutes les applications des méthodes d'analyse multivariée de la géographie anglo-saxonne, nous avons examiné avec soin un éventail assez large d'exemples qui nous a permis, nous l'espérons, de saisir l'essentiel des qualités et des défauts de ces applications. Qu' apportent-elles de plus que l'analyse géographique définie par exemple par P. GEORGE, GUGLIEIMO, KAYSER et LACOSTE (1964, p. 339) en ces termes :

> "Dans la recherche, elle s'ingénie à utiliser toutes les sources et les méthodes possibles et, dans la présentation, elle s'attache avant tout aux convergences c'est à dire aux situations" ?

Les différents stades de recherche que proposent ces géographes sont :

1 - la recherche documentaire des éléments qui. se veut compréhensive, couvrant tous les aspects, ne laissant aucun "blanc".

2 - l'utilisation systématique de l'expression cartographique ; la superposition de cartes simples permettant l'examen de "corrélations".

3 - "l'examen des problèmes synthétiques (qui se sont) dégagés progressivement de l'étude analytique".

Or, il nous semble que l'utilisation des méthodes d'analyse multivariée n'entrave en rien cette démarche scientifique ; bien au contraire, elles y apportent davantage de systématisme et d'objectivité. Car, il est clair que lorsqu'on fait intervenir 1e plus grand nombre de critères possibles comme le recommande P. GEORGE et ses collaborateurs, l'examen de toutes les combinaisons possibles de ces critères, la comparaison d'une grande quantité de cartes, est un art extrêmement délicat que seuls les plus éminents spécialistes, dont l'expérience assure l'objectivité, parviennent réellement à maîtriser.

Le premier intérêt de l'analyse factorielle consiste à ramener les nombreuses caractéristiques de départ à un nombre plus petit de critères de différenciation qui, parmi toutes les combinaisons possibles de ces caractéristiques, regroupent celles qui varient ensemble et qui forment une première "convergence". Les unités d'observation peuvent être alors classées et cartographiées d'après ces critères objectifs synthétiques, pris un à un. Cette base descriptive ne peut constituer une fin en soi pour le chercheur utilisateur de méthodes multivariées. S'ouvrent alors devant lui plusieurs voies possibles vers une analyse explicative.

1. - La voie la plus couramment suivie est aussi 1a plus courte. C'est celle qui consiste à rechercher, dans la définition des variables elles-mêmes, les forces qui lient ensemble les caractéristiques d'un critère factoriel. Cette recherche inférentielle se fait d'ailleurs la plupart du temps à l'aide de la distribution spatiale des critères. Si elle a conduit parfois à des généralisations peu satisfaisantes telle que : "une technologie avancée et une situation dans un pays tempéré déterminent le développement économique", cela n'est pas la conséquence

inévitable de l'utilisation des mathématiques qui seraient déterministes par essence comme certains l'affirment. Cela est dû à la philosophie particulière d'un auteur. D'ailleurs, un tel point de vue déterministe de la géographie n`a pas attendu, avec HUNTINGTON, l'utilisation des mathématiques pour se manifester. Au contraire, on pourrait dire que l'attitude possibiliste de VIDALde la BLACHE a été influencée par le calcul des probabilités.

2 - Une seconde voie, nous l'avons vu au niveau des études thématiques, consiste à utiliser les critères généraux de l'analyse factorielle comme variables indépendantes dans des analyses de corrélation ou de régression, afin de voir par exemple dans quelle mesure ces critères sont responsables, séparément, d'un comportement politique, criminel, etc .... ou d'un phénomène comme le développement économique ou le chômage. Cette voie a été jusqu 'à présent assez peu empruntée par les utilisateurs de l'analyse factorielle et elle mériterait de l'être davantage ; à moins qu'on la considère plutôt du ressort de disciplines plus spécialisées que la géographie, dont l'investigation devrait se poursuivre au niveau "horizontal".

3 - Une troisième possibilité concerne la classification polythétique des observations, en considérant simultanément leur ressemblance sur plusieurs critères factoriels. Le contenu de ces critères déterminé par les variables introduites au départ dépend en partie du jugement du chercheur. Etant donnée cette part d'arbitraire, l'indépendance entre les critères est assurée de façon objective par l'orthogonalité des dimensions, et le regroupement des observations s'obtient de façon optimale. L'arbitraire ne réapparaît que dans le choix du nombre de classes retenues pour la représentation cartographique et le commentaire. Ce type de classification a l'avantage, sur la classification traditionnelle, de minimiser la part de la décision humaine. L'intérêt d'établir une classification objective, à partir de dimensions orthogonales dérivées d'une grande variété de caractéristiques, est évident. Outre ses vertus de clarté descriptive ou pédagogique, elle permet au planificateur d'adapter son action au caractère spécifique de communautés homogènes, et au chercheur d'avoir une base plus solide pour ses investigations (échantillonnages, questionnaires, enquêtes) sur le terrain. De surcroît ces procédés techniques garantissent l'utilisation de mesures standards, indispensables lorsqu'on veut entreprendre des comparaisons entre divers systèmes ou dans le temps.

4 - En effet, la quatrième voie qui s'ouvre au chercheur, et qu'il peut emprunter à tous les stades de son analyse, est celle à laquelle il doit aboutir en fin de compte. C'est celle de l'analyse comparative qui, à travers la recherche de tendances générales, permet l'établissement d'une théorie, nécessaire à la prédiction et à l'action, ou à un approfondissement de la connaissance. Nous avons vu effectivement avec les exemples d'écologie factorielle intra-urbaine, qu'une analyse appuyée sur une théorie peut être extrêmement riche et féconde. En dégageant les mécanismes les plus généraux, elle permet d'une part de constater que l'interaction de ces simples mécanismes suffit déjà à expliquer une grande partie de la réalité ; ce qui laisse d'autre part aux spécialistes le loisir de concentrer leur effort de réflexion sur les distorsions et les exceptions, afin de mieux comprendre cette réalité et peut être de découvrir de nouvelles théories et de nouveaux modèles, qui serviront à leur tour à mettre en évidence de nouvelles exceptions et qu'ainsi progresse la connaissance.

Cependant pour que, en l'absence d'une théorie unificatrice, une comparaison puisse s'opérer entre divers systèmes régionaux, interurbains ou intra-urbains, il faut qu'elle s'appuie sur la même méthode et les mêmes critères. Or nous avons vu qu'une des principales critiques à

l'encontre des applications du modèle factoriel était la grande variété des définitions du champ d'étude, des unités d'observations, de l'ensemble des variables, et le choix de diverses options de la technique dont parfois, de surcroît, certaines contraintes ne sont pas respectées. Des résultats hétéroclites rendent toute comparaison suspecte ou trop générale et ne peuvent pas conduire à l'établissement d'une théorie solide. Les géographes anglo-saxons, "fascinés" par les techniques nouvelles, ont paru surtout intéressés par les expériences méthodologiques servant plus souvent à montrer la validité des méthodes, à la lumière des connaissances géographiques préalables, qu'à expliquer un phénomène géographique inconnu. Depuis plus de 20 ans qu'elles sont appliquées, cependant, la preuve de l'efficacité des méthodes multivariées n'est plus à faire, et il est temps que ces méthodes servent à leur tour à tester des théories ou à formuler de nouvelles hypothèses.

Après ce bilan général des applications d'une méthode, il reste à indiquer les directions qui, à notre sens, devraient s'imposer pour les recherches futures.

1 - la première consiste à introduire au niveau des recherches courantes une plus grande systématisation dans les définitions du champ d'étude, des données et des méthodes, afin qu'elles concordent d'une étude à l'autre et puissent permettre des comparaisons conduisant éventuellement à l'élaboration d'une théorie. En outre, un plus grand respect des contraintes mathématiques du modèle devrait autoriser des comparaisons moins générales concernant des saturations secondaires sur les principaux facteurs, ou des dimensions mineures.

2 - la deuxième direction de recherches souhaitable est l'application, parallèlement à l'analyse factorielle, de toute une batterie de techniques depuis longtemps éprouvées - régressions, corrélations, algorithmes de classification, fonctions discriminantes, analyse canonique, analyse de variance, etc ... - à l'étude d'un même phénomène ; sans que pour autant, cela va de soi, le chercheur abdique son pouvoir de jugement. Celui-ci devrait être au contraire davantage sollicité par une confrontation continuelle de résultats plus nombreux concordant, complémentaires ou contradictoires, à sa propre connaissance du phénomène établie concrètement sur le terrain. Dans cette perspective, le laboratoire bien équipé que constituent pour les sciences humaines l'ensemble des analyses multivariées a été fort peu utilisé.

3 - une troisième direction parallèle, celle qui conduit à la frontière de la recherche scientifique, suppose que les expériences méthodologiques se poursuivent comme, par exemple, l'utilisation de l'analyse factorielle à trois dimensions, proposée par TUCKER (1963), pour rendre compte des variations temporelles, ou bien l'utilisation des méthodes dérivées de la théorie des variables régionalisées de MATHERON (1965), qui tiennent compte de l'autocorrélation spatiale et de la taille modifiable de l'unité d'observation. Cette direction de recherche comprend aussi les expériences conceptuelles touchant la définition du champ d'étude, qui devraient par exemple conduire à l'incorporation de variables nouvelles sur l'environnement physique et 1e comportement des gens, qui apporteraient une information autre que strictement socio-économique et démographique.

Enfin il reste à noter que le domaine de la géographie physique, négligé par les géographes anglo-saxons dans leurs applications de l'analyse multivariée, reste ouvert à l'investigation, quoique des géologues comme KRUMBEIN et IMBRIE (1963) et des botanistes tels que GOODAL (1954) aient déjà apporté d'intéressantes contributions.

TROISIEME PARTIE

Les dimensions socio-économiques de l'état de Pennsylvanie :

Un exemple de régionalisation

CHAPITRE I

Le choix de l'exemple pennsylvanien

LE CONTEXTE DE L'ETUDE

Quelles sont en général les raisons qui guident un chercheur dans le choix d'une étude ? C'est sans doute l'intérêt que porte l'auteur à une région particulière, ce qui veut dire qu'il la connaît déjà, d'une certaine manière. Ou bien ce sont des raisons d'utilité sociale et économique. Dans les deux cas qu'il n'est pas rare de trouver réunis, le but est en général de décrire l'agencement particulier des traits caractéristiques d'une région et, plus ou moins implicitement, de mieux connaître les processus qui conduisent à un tel agencement. Bien sûr, le fait que les données qualitatives ou quantitatives susceptibles d'apporter de l'information sur ces caractéristiques ne sont pas toujours disponibles, influence de façon décisive le choix de la taille de l'unité d'observation et aussi, souvent, contribue à modifier le type d'étude, ou/et le choix de la région. C'est alors seulement qu'apparaît la question de savoir quelle est la méthode la mieux appropriée à l'étude, à la région, et au type de données choisi. Dans le cas présent, le problème est légèrement différent puisque la méthode que nous venons de présenter est à la fois assez puissante et souple pour pouvoir traiter différents types d'études, selon que l'on prend telle ou telle option. Ainsi l'étude que nous allons entreprendre doit répondre aux critères que nous venons de citer (connaissance de l'auteur, utilité sociale ou économique, accessibilité des données) et doit en outre pouvoir servir comme exemple d'exploitation d'une méthode particulière.

Pour que le premier critère soit respecté de façon étroite, il faudrait que l'étude se déroule dans une région française bien connue de l'auteur telle que l'Auvergne, ou bien dans un pays d'Amérique Latine comme le Vénézuela. Ces régions n'ont pas été choisies parce que les données dont on avait besoin pour remplir les conditions du deuxième critère n'y étaient pas immédiatement accessibles. C'est aussi en raison de ce second critère que l'étude d'un centre urbain n'a pas été retenue étant donné qu'il existe un grand nombre d'études de ce genre, du moins dans la littérature de langue anglaise; et aussi parce que l'exemple hypothétique qui illustre la présentation de la méthode est déjà pris dans le domaine urbain.

Une autre étude intéressante du point de vue social et économique est celle des régions déprimées à l'intérieur d'un pays prospère. L'exemple le plus frappant est, à l'intérieur des Etats-Unis, la région Appalachia définie en 1961 comme une région *en détresse* par l'Acte de Redéveloppement Régïonal. Appalachia est la plus grande région d'un seul tenant qui. ait été

désignée pour recevoir une aide du Gouvernement. Elle comprend plusieurs Etats ainsi que des morceaux d'Etats. Elle a été délimitée au niveau du comté, l'équivalent du canton français, d'après des critères de sous-emploi, de revenu familial moyen (inférieur à 40 % de la moyenne nationale) et d'éducation. En 1960, le pourcentage de chômeurs atteignait 7,1 % et auraitsans doute était plus important s'il n'avait été atténué par une forte émigration. Presqu'un tiers des familles avait un revenu annuel inférieur à 3 000 dollars. Un peu moins de 12 % de la population âgée de 25 ans et plus n'atteignait pas le cinquième grade d'éducation, ce qui représente dix années d'école.

Une question se pose cependant : ces variables sont-elles bien représentatives d'une mauvaise santé socio-économique ? I1 est certain qu'en agissant directement sur chacune d'entre elles, on peut améliorer sensiblement leur valeur moyenne dans chaque comté : ainsi, augmenter les allocations de chômage, créer des emplois, relever les salaires, créer des écoles et accorder davantage d'aide sociale aux familles les plus défavorisées pour que leurs enfants puissent poursuivre plus longtemps leurs études. Tout cela ne peut qu'avoir une action positive, et les administrateurs pourront se féliciter en constatant quelques années plus tard une augmentation de la valeur moyenne de ces mêmes caractéristiques dans les régions qui ont reçu une attention particulière**.**

Mais est-ce la manière d'agir la plus efficace ? Les fonds accordés à l'aide pour le re-développement sont en quantité limitée. Il importe donc de savoir exactement où les implanter pour que les résultats soient les meilleurs possibles. Si l'on considère l'interrelation complexe qui existe entre les phénomènes sociaux et économiques, il n'est pas certain qu'une action directe sur plusieurs caractéristiques dont on ignore le degré de dépendance soit la meilleure solution. On peut supposer qu'en utilisant tous les fonds disponibles pour le développement de l'emploi, on réglera plus facilement du même coup les problèmes de. chômage, de pauvreté et d'éducation que si l'on répartit équitablement l'aide entre ces trois rubriques. Ou bien, on peut supposer que c'est l'éducation qui sert de point de départ à tout développement, et que c'est là qu'il faut faire intervenir en priorité l'assistance de l'Etat.

Ce ne sont que des exemples sans doute peu réalistes par leur simplicité mais ils montrent aussi combien tout dépend du but qu'on se propose d'atteindre : amélioration à court terme en vue d'élections proches ou bien amélioration à long terme qui tient compte des processus de *feed back* pour permettre un bien être durable. Tant que nous n'aurons pas cherché à retrouver un peu d'ordre dans l'enchevêtrement inextricable des variables, nous ne saurons jamais exactement quel est l'impact de notre action. D'autre part, le fait que l'unité spatiale soit vaste ne permet pas une aide sélective. La moyenne d'un comté peut aussi bien représenter une combinaison de valeurs fortes et de zones très défavorisées (riches et pauvres par exemple) ou bien tout un ensemble de valeurs uniformément moyennes. Ces considérations nous ont poussé à restreindre notre étude à l'Etat de Pennsylvanie dans le but de découvrir quelle est la structure spatiale de la *santé socio-économique* de cet Etat, lorsqu'on prend des unités d'observation plus fines que le comté, et un grand nombre de caractères. Comme, mise à part l'extrémité Sud-Est, la Pennsylvanie fait partie de la région Appalachia, il peut être particulièrement intéressant d'examiner la zone frontière elle-même, pour vérifier si cette séparation, sans doute justifiée au niveau du comté, l'est aussi à un niveau inférieur.

LES DONNEES

Nous venons de voir que le choix d'unités d'observation à très petite échelle pouvait permettre d'isoler des différences très fines dans les zones critiques de. l'étude telle que la frontière d'Appalachia, et cela devrait permettre aussi de distinguer l'influence de différents niveaux d'urbanisation à l'intérieur des comtés ruraux. C'est pourquoi l'unité administrative définie comme *Minor Civil Division* ou MCD a été choisie comme unité d'observation. Les MCDs sont l'équivalent des communes françaises. Elles sont au nombre de 2 569 exactement en Pennsylvanie. Leur nombre à l'intérieur de chaque comté est très variable, ainsi que leur taille et la densité de la population. C'est un. des obstacles les plus sérieux à l'établissement d'une étude homogène. On devrait s'attendre à ce qu'au moins le total de la population dans chaque unité soit approximativement le même ; or ce n'est pas le cas. La population des MCDs peut varier dans des proportions de 1 à 100 dans les cas extrêmes.

On peut les classer en gros en trois grandes catégories : "*cities", "borough"* et *"townships"* qui correspondent à nos villes, bourgs et villages. Le nom de *township* est donné en général à l'agrégation administrative d'une population dispersée de fermiers et de paysans. Le *borough* exprime un certain degré de concentration spatiale, donc d'urbanisation avec des fonctions commerciales et parfois industrielles. Les *cities* représentent le centre de vastes zones urbanisées telle que Pittsburg et Philadelphie : ce sont les MCDs les plus peuplées et d'étendue spatiale relativement grande. Elles sont entourées par des MCDs plus petites, townships ou boroughs qui se sont développées au fur et à mesure que la ville s'accroissait. Lorsque ces faubourgs sont associés au centre de la ville, ils forment une autre unité définie par le Recensement comme SMSA (Statistical Metropo1itan Area) qui peut être à nouveau re-divisé en *"census tracts"* ou quartiers. Les *"census tracts"* sont les plus petites unités définies par les Bureaux du Recensement et les plus couramment utilisées pour étudier la structure interne des villes. Les critères de définition des *"census tracts"* sont des critères d'homogénéité dans les caractères socio-économiques de la population à l'intérieur d'un tract et de faible variation entre chaque tract quant à la taille de cette population. Bien que ces critères ne soient pas toujours faciles à respecter, les census tracts sont mieux adaptés aux techniques multivariées que les MCDs et c'est en partie pour cette raison de commodité que les études détaillées dans le cadre des SMSAs sont pléthoriques, et qu'il n'existe aucune étude détaillée étendue à une région entière.

En général une autre technique est employée dans les études régionales entreprises au niveau du comté, étant donné que les unités spatiales ne sont pas alors en très grand nombre. C'est par exemple la cartographie des variables prises une à une, et la comparaison des cartes que l'on obtient. Mais même si le nombre de variables cartographiées est restreint, i1 faut beaucoup d'habileté pour comparer et synthétiser les différentes structures spatiales observées. Cela devient de plus en plus difficile au fur et à mesure que le nombre de variables augmente. Or, l'un des principaux avantages des techniques multivariées est justement de permettre une augmentation du nombre de. variables et du nombre d'observations.

Ainsi le choix de l'échelle d'observation de notre étude régionale a été en grande partie déterminé par le fait que jamais jusqu'ici aucune recherche n'a été entreprise à ce niveau de détail. La sélection des variables que l'on a incorporées dans l'étude s'est faite selon deux critères. Le premier, il faut le déplorer, est toujours celui de la disponibilité des données au

niveau d'observation choisi. Le grand nombre des observations nous a fait renoncer à utiliser des variables autres que celles du recensement. Parmi les variables du recensement, notre choix a été guidé par notre jugement personnel des concepts qui décrivent le mieux la prospérité socio-économique, et cela, bien sûr, à la lumière des nombreuses études qui ont été effectuées dans ce domaine. Les 42 variables utilisées ont été tirées des rubriques suivantes : structure de la population, par âge, structure socio-professionnelle, revenu, logement, éducation, mobilité, emploi. Une liste complète de ces variables accompagnées d'une description détaillée est fournie dans l'appendice A.

L'ANALYSE EN COMPOSANTES PRINCIPALES

L'Analyse en Composantes Principales, comme nous l'avons vu dans la Première Partie, sert à synthétiser un grand nombre de variables en un nombre plus petit de composantes sous jacentes. On doit insister ici sur le fait que nous ne l'employons pas pour tester une hypothèse, mais simplement comme un outil d'exploration assez grossier qui permet une description plus économique de concepts aussi complexes que la prospérité socio-économique, ou la pauvreté. Sous cette forme, cette méthode a été très souvent employée dans la recherche géographique, comme nous l'avons vu précédemment, (BERRY,1961, THOMPSON et al. 1962, RAY and BERRY 1964).

L'ensemble des données sur les 2 569 MCDs a d'abord été réduit en une matrice de corrélation entre les 42 variables. Puis, les composantes ont été extraites dans l'ordre de leur capacité à rendre compte de la variance entre ces 42 variables. Plus l'inter-corrélation est grande entre les variables, moins il faut d'axes pour *"expliquer"* une grande proportion de la variance.

Numéro d'ordre des composantes	1	2	3	4	5	% Cumulé
Variance (en %)	20.46	8.87	8.07	4.79	4.05	46.24

TABLE 1 - Pourcentage de la variance expliquée dans les cinq premières composantes principales

Dans la présente analyse, cinq composantes expliquent 46,24 % de la variance qui existe parmi les 42 caractéristiques de départ, mais la première à elle seule rend déjà compte de 20,46 %. Le niveau d'explication apporté par la seconde et la troisième composante tombe respectivement à 8,87 % et 8,07 %. La quatrième composante vient loin derrière et n'est associée qu'à 4,79 % de la variance, ce qui laisse comme pouvoir explicatif à la cinquième composante, 4,05 %. C'est en considérant cette décroissance subite dans l'apport d'information nouvelle que nous avons choisi d'examiner uniquement les trois premières composantes principales.

Le degré de corrélation entre les variables et les composantes principales que l'on désigne aussi sous le terme anglais de *"Loading"* ou saturations, sert de point de départ à l'identification de ces nouvelles bases. Ainsi, une corrélation ou saturation de 0,50 et plus en valeur absolue signifie qu'au moins 25 % de la variance d'une variable donnée est expliquée par la composante. Cette valeur est en général considérée comme un point de rupture commode sur

la voie de l'interprétation. Cela revient à considérer que les variables dont la saturation dépasse le seuil de 0,50 contribuent de la manière la plus significative à la nature de la composante. Cependant, les variables de faible saturation (entre 0,30 et 0,50) ne devront pas être complètement négligées, car elles peuvent servir encore à confirmer la définition des composantes.

En examinant la table des saturations (table 2), nous constatons que la plupart des caractéristiques de prospérité sont fortement et positivement corrélées avec la première composante : haut degré d'éducation, employés "à col blanc", professions libérales, revenu familial élevé, rentiers et actionnaires, valeur élevée du logement. Les caractéristiques opposées sont corrélées négativement avec cette composante : faible et moyenne éducation, travailleurs "à col bleu", faible revenu familial, bénéficiaires de la Sécurité Sociale, proportion de jeunes et de vieux, faible valeur de logement. Ainsi., c'est en examinant les caractéristiques qui lui sont le plus étroitement associée et en considérant ce qu'elles peuvent avoir de commun entre elles, que l'on a pu donner à la première composante le nom de *dimension du bien-être socio-économique.*

Les corrélations très faibles (entre + 0,30 et - 0,30) signifient que les variables en question n'interviennent pas ou très peu dans la définition du phénomène socio-économique décrit par la dimension. Ainsi, il peut être tentant de supprimer ces variables qui introduisent une sorte de bruit de fond au cours d'une seconde analyse dans l'espoir d'obtenir un phénomène plus clairement défini. Cependant nous avons hésité à entreprendre cette nouvelle sélection avant d'avoir accompli la rotation des axes; car il se pourrait que certaines de ces variables peu saturées, comme la proportion des services de santé ou le nombre de personnes par pièce d'habitation, aident à définir un groupe particulier de variables.

Nous rappelons que le calcul des poids locaux ou scores permet d'attribuer à chaque observation une. valeur qui définit sa position sur la nouvelle composante. Dans l'espace engendré par tous les nouveaux axes, les scores représentent donc les coordonnées des observations. La carte construite d'après les scores sur la première composante (figure 1) décrit l'Etat de Pennsylvanie en termes de bien-être social et économique.

Avant d'analyser et de donner une interprétation à la structure spatiale observée, il nous a paru intéressant d'examiner brièvement la deuxième composante. Celle-ci est géométriquement orthogonale c'est à. dire théoriquement non corrélée à la première. En effet, nous constatons que la dichotomie introduite par la première composante entre les zones prospères et les zones déprimées trouve son complément dans la seconde. Des saturations positives élevées sur des caractéristiques moyennes de logement., d'éducation, de revenu et d'activité, s'opposent à des saturations négatives élevées sur des caractéristiques extrêmes de richesse et de pauvreté. Ainsi la deuxième composante révèle une nouvelle dichotomie, celle qui existe entre des zones homogènes de prospérité moyenne et des zones hétérogènes où les extrêmes coexistent.

La cartographie et l'interprétation des scores sur la deuxième composante peut être particulièrement utile en désignant par exemple aux planificateurs les zones où leur intervention doit se faire de manière discriminatoire. Ce travail qui n'a pas été entrepris ici, pourrait être l'objet d'une étude future.

Variables	Numéro	Composante 1	Composante 2	Composante 3
OLD	1	- 0.16	- 0.54	0.52
YOUG	2	- 0.18	0.38	- 0.57
FOR	3	0.24	- 0.20	0.38
IMIG	4	0.32	- 0.04	- 0.04
MOB 1	5	0.12	0.24	0.08
MOB 2	6	0.19	- 0.10	- 0.19
MOB 3	7	0.31	- 0.17	- 0.26
ED 1	8	- 0.44	0.07	- 0.20
ED 2	9	- 0.50	0.29	0.37
ED 3	10	0.83	- 0.31	- 0.17
EDF	11	0.74	- 0.29	- 0.17
LAB	12	0.48	0.47	0.35
FLAB	13	0.45	0.31	0.51
HEALTH	14	0.30	- 0.03	0.33
SALE	15	0.49	- 0.16	0.07
WCOL	16	0.79	- 0.27	0.01
BCOL	17	- 0.45	0.24	- 0.00
FARM	18	- 0.34	0.02	- 0.56
SERV	19	- 0.00	- 0.06	0.27
MINE	20	- 0.30	- 0.19	0.05
TRADE	21	0.32	- 0.07	0.16
REC	22	0.21	- 0.08	0.04
PROF	23	0.48	- 0.31	0.15
UNEMP	24	- 0.29	- 0.18	0.07
FAM 1	25	- 0.56	- 0.55	- 0.02
FAM 2	26	- 0.38	0.57	0.33
FAM 3	27	0.78	- 0.15	- 0.28
FINC	28	0.43	0.23	0.44
WAGE	29	0.11	0.67	- 0.05
NFS	30	0.10	- 0.05	- 0.18
FSELF	31	- 0.34	0.10	- 0.57
SOC	32	- 0.47	- 0.57	0.39
WELF	33	- 0.38	- 0.31	0.09
OTHER	34	0.59	- 0.21	0.11
FAMO	35	- 0.50	- 0.48	- 0.15
POLD	36	- 0.47	- 0.15	- 0.03
POUNG	37	- 0.47	- 0.32	- 0.11
PERU	38	- 0.26	0.05	- 0.24
VAL 1	39	- 0.65	- 0.02	0.35
VAL 2	40	0.53	0.30	- 0.13
VAL 3	41	0.73	- 0.18	- 0.35
VACU	42	- 0.17	- 0.09	- 0.01

TABLE 2 - Corrélations entre les variables originales et les trois premières composantes principales

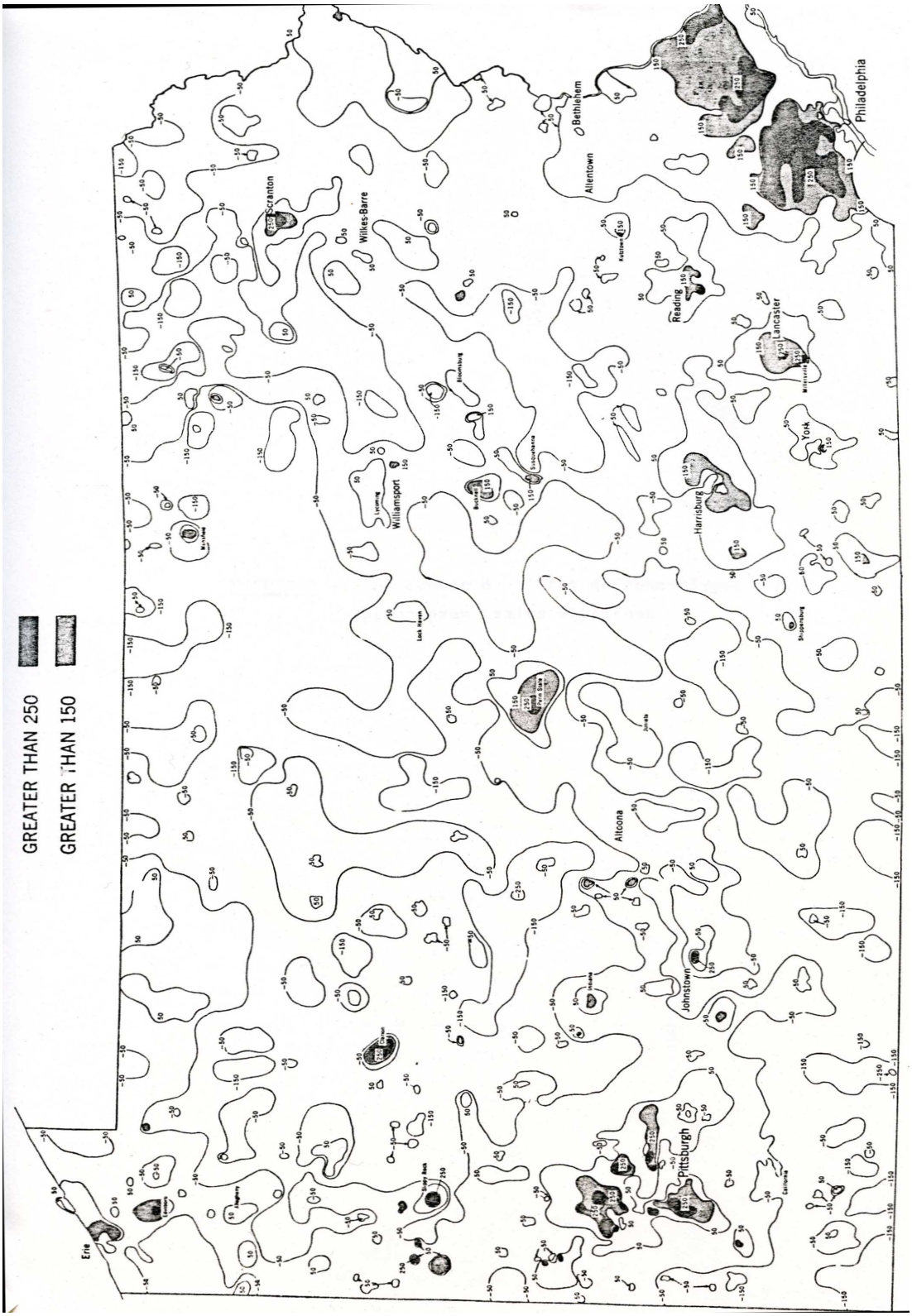


Figure 1 – Répartition spatiale du bien-être socio-économique en Pennsylvanie

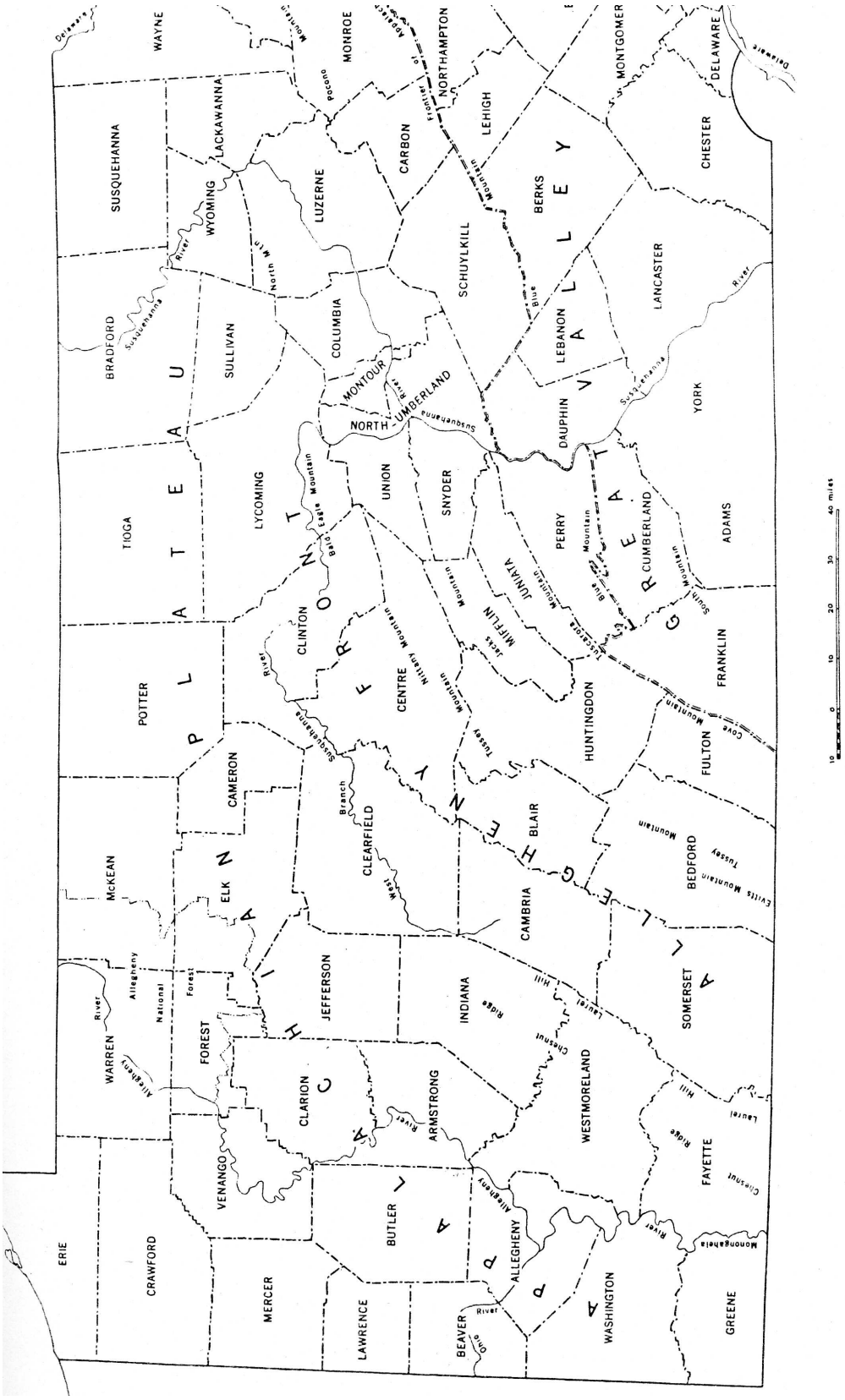


Figure 2 – Les comtés de l'Etat de Pennsylvanie et les principaux traits physiques

CHAPITRE II

L'ETUDE REGIONALE

RÉPARTITION SPATIALE DE LA SANTÉ SOCIO-ÉCONOMIQUE EN PENNSYLVANIE

L'éventail des scores va de 0,403 pour un faubourg particulièrement riche et fermé de Pittsburg à - 0,248 pour Markleysburg, un petit bourg isolé dans la zone montagneuse des Chesnut Hïlls. Ce sont là des valeurs extrêmes qui n'ont pas beaucoup de signification dans un contexte général. D'après la fréquence d'apparition des différentes valeurs, on a choisi de représenter les scores par six niveaux d'isolignes allant des valeurs inférieures à - 0,150 aux valeurs supérieures à 0,250 avec des intervalles de 0,100[20].

Les grandes tendances régionales

Lorsqu'on suit la division régionale de l'État de Pennsylvanie d'après les critères physiques habituels (Figure 2), on observe une certaine correspondance avec la tendance générale des scores (Figure 1). Les Basses Terres en bordure du lac Erie forment une base de scores moyens déformés ici et là par des buttes de scores plus élevés et par des fossés de faibles scores. La disposition des scores les plus hauts, qui s'accorde parfaitement semble-t-il à celle des villes, sera étudiée séparément plus loin. La base générale, qui seule nous intéresse ici., reflète une tendance rurale. Les scores de valeur moyenne dont elle est formée confirment la prospérité relative de cette région agricole dont nous. savons par ailleurs qu'elle produit des légumes et des fruits sur ses sols sablonneux.

Le Plateau Appalachien , avec ses vallées profondes et étroites, séparées par de larges interfluves, est plus généralement composé de faibles scores (de - 150 à - 50), les scores dépassant - 50 étant une exception. Il faut cependant mettre à part une large zone autour de Pittsburg qui comprend des comtés densément peuplés et urbanisés : Allegheny, Westmoreland, Washington, Beaver and Butler. Pour tous les autres comtés, la plupart des scores sont inférieurs à - 50 ce qui indique encore qu'un faible bien-être socio-économique est associé à une population rurale. Cependant cette population ne se consacre pas uniquement aux activités agricoles. Seuls, les comtés suivants vivent principalement d'une agriculture pauvre : Crawford, Jefferson, Indiana, Greene, Tioga, Bradford. La plupart des autres comtés sont aussi spécialisés dans les industries extractives comme Clarion, Amstrong, Mercer, Clearfield avec les mines de charbon, ou Forest, Warren, McKean et Potter avec l'exploitation des forêts et de petits

[20] Pour des raisons pratiques, seuls les chiffres après la virgule sont indiqués sur la carte et servent de référence dans le commentaire qui l'accompagne.

gisements de pétrole et de gaz. Les scores inférieurs à - 150 reflètent en général des zones isolées, pas très accessibles comme par exemple les ensembles de collines et les crêtes de la région centrale et de la partie sud-est du plateau Appalachien. I1 est intéressant de remarquer que certaines zones forestières sont privilégiées. Ce sont par exemple, dans le sud des comtés de Potter et de Tioga, des régions organisées pour la récréation et le tourisme, Forêts ou Parcs d'Etat, dont les scores dépassent le niveau - 150. La Forêt Nationale d'Alleghany qui recouvre une grande partie des comtés de Warren, McKean et Forest est encore plus avantagée avec des zones supérieures à - 50.

Dans la *Région des Crêtes Appalachiennes*, l'ensemble des scores suit l'alternance des crêtes et des vallées qui. s'incurvent du Sud-Ouest au Nord-Est au pied du Front d'Alleghany, dernier rebord du Plateau Appalachien. Comme pour l'instant nous ne considérons que les traits généraux, nous noterons deux principaux ensembles : une bande de scores moyens (de - 50 à 50) qui s'étend au contact du Front, suivie parallèlement à l'Est par une large bande de faibles scores (de - 150 à - 50) qui correspond aux chaînes des Evitts, Jacks and Tuscarora Mountains. La partie Est de la région des Crêtes Appalachiennes, où se trouvent les gisements d'anthracite de Pennsylvanie, est beaucoup moins nettement divisée. En effet, l'ordonnance en bandes parallèles est troublée par l'apparition d'un second ruban de scores élevés le long des branches Est et Ouest de la Susquehanna River.

Les montagnes qui séparent les Crêtes Appalachiennes de la Région de la Grande Vallée servent aussi de frontière à la région d'Appalachia. Il faut rappeler que la Région d'Appalachia a été définie en 1965 comme une zone déprimée d'après des critères principalement économiques de revenu et d'emploi. Elle comprend treize Etats contigus dont une grande partie de la Pennsylvanie. La frontière d'Appalachia a été établie au niveau du comté. On remarquera en effet que, de part et d'autre de cette frontière, les comtés ont une base de scores différente, cette base étant systématiquement plus élevée dans les comtés qui ne font pas partie d'Appalachia. Cependant, l'étude présente, entreprise au niveau inférieur des MCDs, devrait faire apparaître des nuances à l'intérieur de ces comtés. La moitié Est du comté de Franklin qui fait partie de la Grande Vallée bénéficie de scores moyens, entre - 50 et + 50, associés à. de riches terres agricoles. C'est sans doute ce qui. a suffi à justifier l'exclusion de ce comté de la Région d'Appalachia.

Cependant, les scores de toute la partie Ouest sont inférieurs à - 50. Ces faibles scores, qui sont la règle dans les comtés voisins de Fulton et Huntington de l'autre côté de la frontière, reflètent un manque en termes de bien-être socio-économique. I1 semble, ici, que ce manque soit associé à l'agriculture médiocre qui caractérise les régions montagneuses. De même que le comté de Franklin ne devrait pas être complètement exclu de la Région d'Appalachia, le comté de Perry ne devrait pas entièrement en faire partie. Seule la moitié Ouest de Perry, avec des scores de- 50 à 50 est rurale, principalement agricole et pauvre. Cette impression de dénuement est renforcée par des scores inférieurs à - 150 dans l'extrême Ouest montagneux. Mais toute la moitié Est de Perry qui se trouve dans la zone de croissance suburbaine de Harrisburg connaît un sort tout différent comme le prouvent des scores bien supérieurs.

On peut encore identifier trois autres comtés litigieux le long de la frontière d'Appalachia. Ce sont Carbon, Monroe et Pike, qui tous les trois font partie d'Appalachia mais dont les scores moyens dans l'ensemble (de - 50 à 50), donc peu discriminants, justifient qu'on

les examine de plus près. Ainsi, on voit que le comté de Carbon ne devrait pas être séparé de ses voisins producteurs de charbon, les comtés de Schuylkill et de Luzerne, puisqu'il est couvert comme eux de scores moyens engendrés sans doute par les problèmes communs de la crise minière. Le cas des comtés agricoles de Monroe et de Pike est différent. On constate que ces comtés montrent par endroits des scores particulièrement élevés (dans l'intervalle de 50 à 150) correspondant à des régions relativement montagneuses et isolées, exploitées comme Parcs d'Etat. L'extension de la zone de récréation et de résidence secondaire de la population lplus riche de Philadelphie est sans doute l'interprétation la plus vraisemblable de cette localisation inattendue de scores élevés, et peut conduire à remettre en question l'appartenance de ces comté à la région d'Appalachia.

La région appelée Région de la Grande Vallée est composée des vallées de Cumberland, de Lebanon et de Lehigh, connues pour leurs terres fertiles. Comme il fallait s'y attendre, on constate qu'elle est favorisée par un niveau général de scores moyens (- 50 à 50) avec davantage d'exceptions parmi les scores plus élevés que parmi les plus bas. A l'Est de la Grande Vallée s'étendent encore des plaines ondulées, des basses collines et des vallées fertiles, qui correspondent en gros aux comtés de Lancaster et de York. C'est une région agricole de peuplement flamand dont les sols sont parmi les plus riches des Etats-Unis. En conséquence, on voit des scores élevés (dans l'intervalle de 50 à 150) en occuper la plus grande partie, et on ne note aucun score inférieur à - 50. Le coin Sud de l'Etat fait partie de la région naturelle de la Plaine Côtière Atlantique, basse, plate et fertile. C'est là qu'on trouve la plus grande étendue de scores élevés de toute la Pennsylvanie (150 et au-delà). Cependant, on pense que l'influence de la croissance urbaine de Philadelphie entre pour une plus grande part que la richesse de la terre dans la mesure exceptionnelle de bien-être que ces scores représentent. En effet, à part le comté de Chester, toute cette région est fortement urbanisée.

.

Nous venons de montrer les relations qui existent entre la répartition des scores de bien-être et le découpage de l'Etat de Pennsylvanie en grandes régions naturelles. Nous avons vu que souvent la présence des villes pouvait troubler la netteté de ces relations. Ainsi, maintenant, nous allons essayer de découvrir et d'examiner à un niveau plus fin d'analyse les rapports qu'il peut y avoir entre certains scores et certains types de villes.

La structure urbaine

Dans les Basses Terres d'Erie, sur les bandes qui bordent le Front d'Alleghany ou qui longent la rivière Susquehanna, dans le Sud-Est de l'Etat, partout où le niveau général des scores va de - 50 à 50, on a essentiellement affaire à une population rurale dispersée observée dans le cadre des *"townships"*. Les scores qui s'élèvent au-dessus de ce niveau se trouvent à l'emplacement des bourgs où une population plus concentrée commence à fournir des fonctions urbaines. En voici d'Ouest en Est quelques exemples : Wateford dans le comté d'Erie, Carrolltown dans le comté de Columbia, Mifflinsburg dans le comté d'Union, Quarryville dans le comté de Lancaster, etc...

Lorsque les *"townships"* juxtaposés forment une base de scores de - 150 à - 50,comme c'est le cas de la partie centrale et du Nord-Est du Plateau Appalachien, alors ce sont les scores de la classe supérieure (entre - 50 et 50) qui dénotent la présence des bourgs. Ce sont par exemple Clitonville dans le comté de Venango, Reynoldsville dans le comté de Jefferson, Snow

Shoe dans le comté de Center, etc... Finalement, lorsque des pointes de scores de - 150 à - 50 percent une étendue de scores inférieurs à - 150, nous sommes en présence de bourgs tels que Coalport dans le comté de Clearfield, New Baltimore dans le comté de Sommerset, Austin dans le comté de Potter ou Ekland dans le comté de Tioga.

En résumé, les bourgs se distinguent de la région rurale qui les entoure par des scores légèrement plus forts. Bien plus, si l'on monte dans la hiérarchie urbaine, on voit que les villes et les grands centres régionaux peuvent présenter des scores de deux à trois classes plus élevés que ceux de la base générale où elles se trouvent. En effet, la plus vaste étendue de. scores élevés (supérieurs à 150) entoure les deux plus grandes villes de l'Etat : Philadelphie et Pïttsburg. Ailleurs aussi., des taches et des points formés de scores élevés ont pu être identifiés. Ce sont des centres urbains importants tels que Lancaster, Reading et Scranton, au-delà de 250 ; ou bien, entre 150 et 250, ce sont avec Harrisburg, Newcastle, Erie, des sièges administratifs de comtés qui, en outre, jouent un rôle régional. Cependant, on doit arrêter là toute généralisation parce qu'il n'y a pas toujours correspondance entre la taille des villes et celle des scores. En tenant compte davantage de leur fonction que de leur taille on a pu regrouper les villes en trois types dont la description suit.

Banlieues ou villes satellites des grands centres. régionaux

Un examen plus détaillé des villes que l'on vient de citer confirme un caractère urbain bien connu à l'échelle nationale. I1 s'agit de la pauvreté du centre des grandes villes qui s'oppose à la richesse des banlieues résidentielles. Les scores sur la composante de bien-être reflètent cette tendance. On ne trouve pas les scores les plus forts dans le centre des villes mais dans la zone qui les entoure. Par exemple, en tant que MCDs, les villes suivantes ne présentent que des scores moyens : Pittsburg (80), Philadelphie et Scranton (53), Lancaster (37), Reading (20), Johnston (-14) ; alors que les communes de banlieue présentent des scores au-delà de 250. La banlieue proche de Erie (57), celle d'Harrïsburg (50) et celle de York (12) sont dans la classe de 150 à 250.

Ce phénomène n'est pas surprenant aux Etats-Unis. Cependant, il interfère dans l'étude présente avec le problème de la taille de l'unité d'observation définie par le recensement. Pour les communes les plus importantes, au-delà d'un certain niveau d'agrégation, l'opposition entre les districts riches et les districts pauvres, entre les districts à différents niveaux d'éducation, et la discrimination spatiale des diverses fonctions urbaines disparaissent. On pense que les scores moyens que l'on observe ne reflètent pas un bien-être socio-économique moyen, mais plutôt la combinaison de valeurs fortes et de valeurs faibles. qui coexistent à l'intérieur d'une grande ville. De façon inverse, les banlieues sont découpées en petites unités de recensement, bourgs pour 1a plupart et petites villes à l'intérieur desquelles, étant donné leur taille, il est peu vraisemblable de trouver une grande variété de fonctions et de types de résidences; ce qui. permet de mieux saisir les différences qui existent entre elles.

En effet, bien que les riches communes résidentielles soient les plus nombreuses et forment des taches remarquables de scores élevés autour des grandes villes, l'éventail des scores apparaît beaucoup plus large après un examen minutieux et dénote la présence de banlieues plus défavorisées telles que certaines banlieues industrielles. Ainsi, par exemple, on trouve à l'Ouest de Philadelphie des communes de scores très faibles (de - 150 à - 50) le long de la Delaware

River. Par opposition, des communes dont les noms seuls indiquent déjà l'agrément d'un site élevé, sont désignées par les scores les plus forts (au-delà. de 250). Ce sont parmi tant d'autres : Clifton Heights, Drewel Hill, Rosemont, etc ... La ville de Pittsburg et ses alentours sont tout aussi caractéristiques. La commune de Pittsburg elle-même présente des scores moyens. Les scores les plus faibles se trouvent le long des rivières, Monongaheli et Ohio, comme à Dusquesne, Homestead et Mc Kees Rocks. Les communes favorisées par leur localisation élevée comme Westview, Bellevue, Carnegie, Dormont or Bethel Park ont les plus hauts scores.

Il faut signaler ici, que l'examen des scores sur la deuxième composante qui n'a pas été entrepris dans l'étude actuelle pourrait justifier une étude future. D'après la définition que l'on a donnée plus haut de cette seconde composante, la cartographie des scores devrait montrer plus clairement l'opposition entre l'hétérogénéité du centre des grandes villes et l'homogénéité des petites villes satellites de banlieue.

Les villes minières et les villes manufacturières

Le phénomène que nous venons d'analyser n'apparaît pas systématiquement dans toutes les grandes ville. Certains centres locaux importants, parfois plus grands que quelques unes des villes qui ont été identifiées précédemment, ne sont pas entourés de banlieues ou de villes satellites présentant des scores élevés. Ce sont en particulier : Allentown, Bethlehem, Easton, Wilkes Barre, Williamsport, Sharon. Les scores du centre de la ville n'atteignent pas 100 avec 98 pour Bethlehem et 95 pour Allentown qui sont les plus élevés, et ne descendent pas au-dessous de 50. Les scores des communes alentours sont soit dans la même classe soit dans une classe inférieure. On remarque que Wilkes Barre, qui est par ailleurs le siège du comté de Luzerne et une ville manufacturière importante, est aussi le centre d'une région productrice d'anthracite. Parallèlement, deux autres villes manufacturières, Aliquippa dans le comté de Beaver et Sharon dans celui de Percer, sont aussi des centres de régions charbonnières.

Connaissant le déclin rapide de l'extraction du charbon depuis vingt ans, on peut penser que la prospérité et le dynamisme de ces villes en ont subi l'impact. De la même manière, on suppose que Williamsport, capital du Lycoming, un comté qui vit essentiellement de la manufacture du bois, a beaucoup souffert de la crise de l'industrie forestière. Cependant, Easton, siège du comté de Northampton spécialisé dans l'industrie manufacturière, et dans le même comté, Bethlehem dont l'activité principale est la production d'acier, ne sont pas directement concernés par la crise des industries extractives I1 est encore plus surprenant de constater qu'Allentown, grand centre de fabrication de machines et d'appareillage électrique n'apparaît pas plus dynamique. Ainsi, l'on est tenté de conclure que l'activité manufacturière elle-même n'est pas vraiment un générateur de bien-être dans le sens où ce terme a été défini par la première composante. En effet, nous devons souligner que si l'industrie manufacturière est la principale activité de la ville, cela entraîne nécessairement une plus forte proportions de *"cols bleus"* parmi la population active; et nous rappelons que cette variable a un poids négatif sur la composante de bien-être et se trouve associée avec des caractéristiques telles qu'un faible revenu familial, une faible valeur du logement, une éducation moyenne.

Apparemment, les villes dont les activités sont plus diversifiées, même lorsque l'industrie manufacturière est l'une de ces activités, présentent des scores plus élevés, sinon dans la ville elle-même, du moins dans ses faubourgs résidentiels. Ce sont les grands centres

régionaux que nous avons déjà cités : Philadelphie, Pittsburg, Erie, Lancaster, Reading, dont on attend, de par leur taille, un éventail étendu de fonctions. Ce sont aussi des agglomérations plus petites comme York, Johnston, Hazleton, Beaver, Ligonier, situées respectivement dans les comtés de York, Cambria, Luzerne, Beaver et Westmoreland, dont la principale activité est encore l'industrie manufacturière. On pense qu'ici, les scores élevés sont dus à l'importance relative du secteur tertiaire, c'est-à-dire du commerce et des services, par rapport au secteur secondaire.

Les villes de Collège et d'Université

Une nouvelle infirmation de la relation entre la taille des villes et la valeur des scores est donnée par la présence de scores étonnement élevés dans de petites agglomérations comme Clarion, State College, Slippery Rock, Edinboro, pour les scores supérieurs à 250 et comme Lewisburg, Mansfield, Kutzton, Shippenburg, Indiana, pour les scores de 150 à 250 (Fig. 1). Bien qu'elles soient de taille très différentes, toutes ces villes comprennent des Institutions d'Education Supérieure, soit des Universités comme Penn State à State College ou Bucknell à Lewisburg, soit simplement des collèges d'Etat comme Mansfield ou Slippery Rock.

Finalement, il n'est pas surprenant de trouver ces villes de Collège et d'Université dans les tout premiers rangs des scores sur la composante de bien-être. Si l'on se reporte à la table des poids (Table 2), on voit que la variable qui contribue le plus à la définition de la première composante est précisément ED3, c'est à dire la proportion de personnes ayant au moins un an d'éducation universitaire. La corrélation de ED3 avec la composante de bien-être est égal à 0,84, ce qui veut dire que 70 % de la variance dans cette variable d'éducation supérieure est *"expliqué"* par la composante. Toutes les autres variables qui sont étroitement associées à la première composante témoignent de la prospérité et du dynamisme des villes de Collège et d'Université. En effet, nous rappelons que la composante de bien-être rend compte d'au moins 50 % de la variance dans les caractéristiques de prospérité : revenu élevé, logements coûteux ; et de10 à 50 % de la variance dans les variables d'emploi : population active, population féminine active, activités commerciales ; et d'immigration : population venant d'autres Etats ou de pays étrangers.

En résumé, on a dû apporter certaines nuances au caractère urbain de la première composante qui est d'abord apparu comme déterminant dans la répartition spatiale des scores. Une étude détaillée montre que la fonction des villes joue un rôle important qui recouvre en partie seulement celui de la taille. D'après la théorie des Places Centrales confirmée par de nombreuses observations empiriques, on sait bien qu'à une taille donnée correspond un ensemble de fonctions; mais certaines fonctions sont moins sensibles que d'autres à l'effet de la taille, et c'est précisément le cas pour les Institutions d'Education Supérieure aux Etats-Unis. Mise à part l'éducation, d'autres caractéristiques fonctionnelles d'activité tertiaire, telles que professions libérales et services, apparaissent décisives pour l'évaluation du niveau de bien-être des villes. En conséquence, ce sont les villes universitaires, les banlieues résidentielles des centres régionaux aux fonctions très diversifiées, qui sont les plus favorisées, par opposition aux villes plus exclusivement minières ou manufacturières.

Le phénomène d'altération avec la distance

Il reste à examiner un dernier phénomène qui relie le "bien-être" au degré d'urbanisation. En règle générale, les auréoles de scores décroissants qui entourent les villes dont les scores sont. élevés varient en fonction inverse de la distance à mesure qu'on s'éloigne du centre. C'est une fonction exponentielle, dont la pente est plus ou moins forte selon que le niveau moyen des scores de la région aux alentours immédiats est plus ou moins bas. Deux exemples serviront à illustrer ce phénomène : State College, dont le comté agricole du Centre est favorisé par des scores supérieurs à 250, comme nous l'avons signalé précédemment. Autour de la ville elle-même s'étend une zone en forme d'ellipse dont le plus grand axe mesure environ 13 miles avec des scores de 150 à 250. Vient ensuite une zone de scores dans la classe de 50 à 150 qui s'étend en gros dans un rayon de 10 miles à partir du centre de la ville. Puis, commence la base du comté avec des scores moyens de - 50 à 50.

Clarion, un comté rural pauvre avec quelques industries extractives de charbon et de bois, n'apparaît pas dans l'ensemble aussi prospère que le comté du Centre. Dans la base générale des scores qui vont de -150 à - 50, Clarion, le centre administratif du comté du même nom, est une exception frappante. La ville elle-même dépasse le niveau 250 sur une étendue de 4 miles de rayon, plus vaste que State College. Mais ensuite, les scores décroissent brusquement jusqu'à un niveau inférieur à - 50, ne laissant presque aucune place pour des scores intermédiaires. Ces deux exemples montrent comment l'étendue de l'influence qu'exerce une ville sur son hinterland dépend davantage du niveau général de prospérité de la région que de la prospérité de la ville elle-même.

ANALYSE DES PRINCIPAUX FACTEURS APRES ROTATION

Nous venons de voir comment l'analyse en composantes principale engendrait des dimensions descriptives. La première et la plus grande de ces dimensions, qui explique environ 20 % de la variance totale des variables introduites dans l'analyse, est une mesure agrégée de "bien-être" avec laquelle une grande partie des variables sont fortement corrélées, en particulier les variables qui sont considérées dans d'autres études socio-économiques comme des indices de prospérité. I1 est question maintenant d'étudier les résultats de la rotation des composantes. On a fait subir une rotation orthogonale à successivement deux, puis trois, puis cinq des 42 composantes originales extraites par l'analyse précédente, et on a obtenu ainsi de nouvelles dimensions que l'on désignera désormais sous le nom de facteurs. La rotation a été effectuée par la procédure automatique Varimax d'après le critère de simple structure dont on ne reviendra pas sur le principe, déjà exposé dans la première partie. Les trois rotations, effectuées sur un nombre différent de dimensions, ont redistribué la variance totale sur ces dimensions de telle manière que les trois nouveaux ensembles de facteurs obtenus en expliquent maintenant les pourcentages suivants :

Facteurs	Nombre de facteurs ayant subi une rotation		
En %	**2**	**3**	**5**
Facteur 1	17.05	14.95	14.06
Facteur 2	12.28	12.35	11.62
Facteur 3		10.10	7.15
Facteur 4			7.22
Facteur 5			6.19
Total	**29.33**	**37.40**	**46.24**

TABLE 3 : % de variance expliquée par trois groupes de facteurs de taille différente après rotation

On notera que chacun des trois ensembles de facteurs explique le même pourcentage de variance que les ensembles de composantes à partir desquels ils sont dérivés, c'est à dire respectivement : 29,33 °/, 37,40 % et 46,24 %. Cependant, l'un des effets recherchés par la rotation est une contribution plus équitable de chacun des facteurs à cette explication. Ainsi les derniers facteurs gagnent un certain pouvoir explicatif aux dépends des premiers. Dans la rotation sur deux facteurs, le facteur 2 explique 12,28 % de la variance contre 8,87 % expliqué précédemment par la Composante II, tandis que le facteur 1 explique maintenant 17,05 % contre 20,46 % pour la composante II. Dans la rotation sur cinq facteurs, on passe respectivement de 4,79 % et 4,05 % de la variance expliquée par les composantes IV et V à 7,22 % et 6,19 % pour les facteurs 4 et 5. A ce propos on doit remarquer que la solution de rotation sur trois facteurs donne les meilleurs résultats, avec une répartition mieux équilibrée de l'explication entre les trois facteurs. On reviendra plus tard en détail sur ce groupe particulier. Pour l'instant, nous allons d'abord examiner rapidement les résultats de la rotation sur deux facteurs. En partant de la matrice des corrélations de chacune des 42 variables avec chacun des 5 facteurs, on a isolé les variables dont les corrélations sont les plus fortes ou variables-clefs (Table 4).

En examinant la première colonne de saturations, après rotation sur deux facteurs, on remarque que les facteurs 1 et 2 sont composés de nombreuses variables que l'on a déjà vu participer à la composante de bien-être. La compréhension de cette notion complexe est cependant améliorée ici, par l'isolement sur deux facteurs séparés de deux groupes distincts de variables. Avec des caractéristiques telles que haut niveau d'éducation, revenu élevé, grande valeur du logement, opposées à revenu moyen, éducation moyenne, faible valeur du logement, le facteur 1 est identifié comme un facteur de prospérité et de statut social.

Le facteur 2 qui est représenté par des caractéristiques telles que la population active et les salariés d'un côté, la population jeune et vieille et les bénéficiaires de pensions de l'autre, peut se concevoir comme un facteur d'activité. Ainsi, cette partition claire en deux facteurs distincts, statut social et activité, non corrélés, sinon indépendants, peut paraître satisfaisante par

sa simplicité, mais elle est cependant bien insuffisante en ce qu'elle rend compte d'à peine un quart de la variance, à travers toute la Pennsylvanie, des caractéristiques socio-économiques considérées. Or, on a vu (Table 1) qu'en ajoutant une troisième dimension, on obtient avec 37,40 % une amélioration sensible du niveau d'explication de la variance. Aussi, nous avons remarqué que le pouvoir explicatif est mieux réparti lorsque la rotation est faite sur trois facteurs. En déplaçant la structure orthogonale rigide composée des trois premières dimensions, on a retrouvé trois grands groupes de variables étroitement associées.

Quel est dans chacun des groupes le phénomène commun qui relie ces variables ? L'examen minutieux de la matrice des saturations sur chacun des trois facteurs (Appendice B) a permis de les identifier comme suit :

Facteur 1 = *statut social*
Facteur 2 = *activité*
Facteur 3 = *urbanité*

On remarque que les deux premiers facteurs sont les mêmes que ceux qui étaient apparus lorsque la rotation était faite sur deux axes seulement. En effet, ils désignent exactement le même ensemble de variables dont seules les saturations respectives sont légèrement modifiées par l'effet du troisième facteur.

Le terme de *Statut Social* convient au facteur 1, fortement et directement corrélé avec les catégories supérieures des caractéristiques d'éducation, de revenu, de logement et de profession, inversement corrélé avec les catégories moyennes ou inférieures des mêmes caractéristiques, comme le montre la Table 3. Si l'on compare les résultats des rotations sur deux et sur trois facteurs, on remarque qu'une importance plus grande est donnée à l'éducation et au revenu dans le dernier cas; tandis que la profession, représentée par les employés à col blanc opposés aux travailleurs à col bleu, a moins de poids. Cette tendance est confirmée par l'examen des corrélations suivantes dans l'ordre de magnitude.

Les saturations des professions libérales et des employés du commerce passent de 0,50 dans le premier cas à 0,40 dans le second, ceux de la population jeune et des employés agricoles vont de - 0,30 aux environs de zéro. Il est évident qu'il y a une baisse de contribution des variables de type professionnel et cela, sans doute, au bénéfice de l'un des deux autres facteurs que nous allons maintenant examiner.

Le second facteur est fondé clairement sur l'activité et l'emploi. Les variables qui lui sont le plus étroitement associées comprennent la population active totale, les salariés, la population féminine active, les logements de valeur moyenne pour les corrélations positives ; et pour les corrélations négatives opposées : les bénéficiaire de la sécurité sociale, la population vieille et la population jeune, c'est à dire essentiellement la population *"inactive"*.

Variable clefs			Corrélations avec les facteurs 2	3	5
Facteur 1	Riche	ED3	0.87	0.85	0.87
		WCOL	0.81	0.70	0.70
		FAM3	0.74	0.78	0.72
		VAL3	0.71	0.80	0.75
	Pauvre	FAM2	- 0.63	- 0.72	- 0.70
		ED2	- 0.58	- 0.69	- 0.72
		VAL1	- 0.54	- 0.65	- 0.58
		BCOL	- 0.51	- 0.45	- 0.47
Facteur 2	Actif	LAB	0.66	0.66	0.51
		WAGE	0.63	0.62	0.18
		VAL2	0.54	0.55	0.60
		FLAB	0.51	0.51	0.36
	Non-actif	FAM1	- 0.77	- 0.78	- 0.78
		SOC	- 0.74	- 0.74	- 0.55
		OLD	- 0.55	- 0.55	- 0.32
		POUNG	- 0.53	- 0.53	- 0.57
Facteur 3	Urbain	FLAB		0.55	0.82
		OLD		0.54	0.29
		FOR		0.49	0.03
		HEALTH		0.42	0.43
	Rural	YOUNG		- 0.68	- 0.50
		FSELF		- 0.67	- 0.69
		FARM		- 0.64	- 0.20
		ED1		- 0.38	- 0.14

Table 4 – Corrélations entre les variables clefs et les trois premiers facteurs

Viennent ensuite des corrélations négatives moins fortes qui peuvent encore servir à confirmer l'identification précédente. Ce sont les bénéficiaires de pensions (0,47), les employés des mines (- 0,33), les sans-emploi (- 0,31). La première et la dernière de ces variables concernent encore la population inactive; et leur association avec les employés des mines n'a rien de surprenant. Elle souligne simplement les conséquences désastreuses bien connues de la crise de l'industrie minière. Les corrélations positives secondaires sont toutes liées aux niveaux de revenu : revenu féminin (0,43), revenu élevé (0,30), revenu moyen (0,27).

Elles soulèvent le problème de la corrélation entre les facteurs. Bien que les facteurs soient géométriquement orthogonaux, c'est à dire théoriquement non corrélés, il est difficile de trouver pour les désigner des concepts qui soient complètement indépendants. Ainsi, par l'intermédiaire des variables de revenu, on voit comment les concepts de Statut Social et d'Activité sont liés puisqu'il est logique de penser que les régions les moins actives sont aussi les moins prospères et de statut social moins élevé. On peut envisager deux approches, d'ailleurs compatibles, à ce problème délicat. L'une est de considérer qu'il n'est pas possible de définir

avec précision les facteurs par un simple concept ; et que lorsqu'on parle de Statut Social ou d'Activité cela représente davantage une tendance qu'une rigoureuse définition.

L'autre approche revient à considérer que la part d'influence du revenu qui entre dans le facteur Statut Social n'est pas la même que celle qui entre dans le facteur Activité, la variable revenu étant abstraitement séparée en deux parties distinctes. C'est là précisément l'effet de l'opération mathématique de projection d'une variable sur deux axes orthogonaux. D'ailleurs, il suffit d'examiner le rôle que joue le revenu dans chacun des deux facteurs pour concevoir que cette variable peut appartenir à deux notions très différentes : dans le facteur 1, les revenus élevés sont clairement opposés aux revenus moyens. Dans le facteur 2, les revenus élevés et moyens sont opposés de façon moins nette aux revenus les plus fables et aux revenus de pensionnés et chômeurs.

On vient de voir combien il était délicat de découvrir un motif clair derrière les deux premiers facteurs. Le phénomène qui lie ensemble les variables associées avec le troisième facteur est encore plus difficile à cerner, à première vue. La plus forte corrélation (Table 4), qui correspond à la proportion de femmes dans la population active, est suivie par la proportion de personnes au-delà de 65 ans. Une variable que l'on a vu corrélée de façon positive avec le facteur d'activité se trouve maintenant proche d'une variable corrélée négativement avec ce même facteur. S'il est certain que cette nouvelle association cache un concept complètement différent, le sens de ce dernier n'est pas évident.

Les corrélations positives suivantes, comme la proportion d'étrangers et la proportion d'employés des services de santé, ne nous aident guère dans la recherche d'une cause sous-jacente commune. L'examen des corrélations négatives nous éclaire davantage. Vient d'abord la population au-dessous de 18 ans, puis la proportion d'agriculteurs, propriétaires et employés, suivie par la variable de faible éducation.

Cet ensemble de variables, dont deux en particulier ont trait à une occupation de type rural, et dont aucune ne fait de référence directe au revenu ou à l'emploi, est suffisamment consistant pour nous conduire à l'hypothèse que le troisième facteur est un facteur Urbain-Rural. Une analyse rapide des variables clefs à corrélation positive (table 4) confirme cette hypothèse. C'est un aspect économique bien connu qu'aux EtatsUnis, la proportion des femmes dans la population active est plus grande dans les villes qu'à la campagne. Aussi, et en particulier en Pennsylvanie, la proportion des gens âgés est relativement plus forte dans les villes que dans les zones rurales où les familles ont encore tendance à avoir de nombreux enfants. La plus forte proportion d'étrangers dans les villes peut s'expliquer par la politique américaine d'immigration qui encourage davantage les emplois urbains. Le manque de services de santé dans les campagnes est souligné par la contribution de cette variable à l'aspect urbain du facteur, tandis qu'il n'est pas surprenant de voir figurer dans le côté rural opposé le faible niveau de l'éducation. Incorporées dans ce groupe de variables clefs, on trouve, avec des corrélations un peu plus faibles, des professions typiquement urbaines telles que les professions libérales (0,41), les cols blancs (0,40), les actionnaires (0,39) et des variables encore plus significatives comme la proportion de la population employée dans le commerce (0,31).

Un indice de "confort" (- 0,31), mesurant le nombre de personnes par pièce n'a pas pu être utilisé pour identifier le troisième facteur parce qu'il contredit apparemment la notion

intuitive de manque d'espace à la ville par rapport à la campagne. Mais c'était sans tenir compte de la taille des pièces et du fait que, comme nous venons de le voir, le nombre d'enfants est plus élevé dans les familles rurales. C'est aussi parce qu'un grand nombre de variables qui lui sont liées contribuent en même temps au premier facteur, que la signification sous-jacente du troisième facteur est particulièrement difficile à dégager.

On retrouve là le problème de la corrélation entre les facteurs, car il se trouve que les occupations de haut statut social qui engendrent une certaine prospérité sont des occupations de type urbain. En partie mobilisées par le facteur urbain, les variables d'occupation professionnelle sont moins fortement corrélées avec le facteur de statut social après rotation sur trois facteurs, qu'elles ne le sont lorsque la rotation est faite sur deux facteurs seulement. La différence est particulièrement sensible pour les variables clefs du facteur urbain. La proportion d'étrangers, par exemple, dont la corrélation avec le premier facteur est encore notable (0,32) dans le cas de la rotation sur deux facteurs, n'a plus qu'un rôle négligeable (0,09)dans l'autre cas. Les corrélations avec les propriétaires et employés agricoles passent respectivement de - 0,35 à - 0,03, et de - 0,30 à - 0,01. Nous pouvons dire que d'un type d'analyse à l'autre, le troisième facteur a puisé une partie de son pouvoir explicatif dans le premier facteur.

I1 nous paraît intéressant d'examiner maintenant l'impact de l'addition d'une quatrième et d'une cinquième dimension à la structure qui subit la rotation, pour savoir si cela peut nous aider à mieux comprendre les différentes facettes du phénomène de bien-être socio-économique. Après rotation sur cinq facteurs, on remarque tout d'abord que la variance expliquée est loin d'être également répartie parmi les facteurs (Table 3). Des trois facteurs que l'analyse précédente avait dégagés, le troisième apparaît le plus affecté par l'addition de deux nouveaux facteurs. I1 perd presque 3 % de son pouvoir explicatif, alors que le pourcentage de variance expliquée par 1e premier et par le deuxième facteur a seulement très peu diminué.

La comparaison des matrices de corrélation des variables originales avec respectivement trois et cinq facteurs (Table 4) devrait nous renseigner sur la nature de ces changements. On voit un déplacement de l'ensemble de la structure factorielle assez important pour créer un recouvrement des facteurs de Statut Social, d'Activité et d'Urbanité qui ne sont plus perçus en tant que tels aussi clairement. Le premier facteur qui peut encore, par certains aspects, représenter le Statut Social dépend toutefois plus exclusivement des caractéristiques d'éducation. Le second facteur ne mérite plus guère de définir l'Activité puisque les corrélations des variables d'activité ont diminué au bénéfice de certaines variables de statut social comme la valeur du logement. C'est le troisième facteur qui remplace en partie le second dans son rôle lié à l'activité avec le renforcement de variables concernant la population active. Ce faisant, il perd son caractère d'urbanité avec la chute des corrélations sur les variables témoins du degré d'Urbanisme. La plupart de ces variables à connotation urbaine ou rurale se trouvent réparties entre le quatrième et le cinquième facteur (Table 5). Le quatrième facteur souligne le type d'occupation par sa relation directe étroite avec les cols blancs, opposée à une relation inverse avec les agriculteurs. Le cinquième qui oppose la population âgée à la population jeune représente la structure d'âge.

Dans cette étude particulière, il apparaît que l'addition d'un quatrième facteur et d'un cinquième n'est pas recommandée. Elle ne conduit pas à la découverte de nouveaux ensembles de variables intéressants, mais elle fait simplement éclater les groupes, déjà identifiés, en

groupes plus petits dont la signification, dépendant d'un très petit nombre de variables, n'a plus autant de cohérence. Cependant, cette dernière expérience n'est pas inutile en ce qu'elle nous assure que les résultats de la rotation sur trois facteurs sont les plus satisfaisants.

En résumé, le processus de rotation nous a conduit à identifier trois groupes distincts de variables, chacun mesurant un aspect particulier d'un même phénomène général. *Statut Social, Activité, Degré d'Urbanisme* décrivant de manière plus simple et plus précise le concept vague de santé socio-économique. Parce qu'il ne sont simplement que les différentes facettes d'un même phénomène, il est normal que ces trois facteurs ne soient pas indépendants. C'est la principale faiblesse du procédé de rotation orthogonale. En imposant une structure orthogonale aux facteurs, il n'est pas possible de les faire coïncider parfaitement avec un groupe de variables. I1 en résulte que les variables les plus étroitement corrélées avec le facteur ne sont pas nécessairement celles qui se trouvent au centre du groupe, et cela ne fait que compliquer la tâche déjà délicate d'identification des facteurs.

Variables clefs			Corrélations avec les facteurs
Facteur 4	Pseudo rural	FARM ED1	0.84 0.35
	Pseudo urbain	FOR WCOL	- 0.57 - 0.36
Facteur 5	Pseudo non-actif	OLD SOC	0.71 0.59
	Pseudo actif	WAGE YOUNG	- 0.87 - 0.48

Table 5 – Corrélations entre les variables clefs et les quatrième et cinquième facteurs après rotation.

C H A P I T R E III

COMPARAISON DE L'ANALYSE INITIALE ET DES DIVERSES OPTIONS

EFFETS DE L'UTILISATION D'ECHELLES DE MESURES NON-METRIQUES

L'expérience suivante a pour but de vérifier si les résultats de l'analyse en composantes principales et de l'analyse factorielle sont affectés par le type d'échelle utilisé pour mesurer les données. I1 faut auparavant rappeler qu'il existe trois principaux types d'échelle de mesure : l'échelle de rapport, l'échelle ordinale et l'échelle binaire. L'échelle de rapport est celle que l'on a utilisée pour mesurer les caractéristiques socio-économiques de l'étude précédente. Elle est composée d'un ensemble de nombres réels, théoriquement infini, mais limité en général par l'arrondissement à deux ou trois chiffres après la virgule, allant de zéro à une valeur maximum déterminée ou non. Ces nombres représentent, dans le cas qui nous intéresse, des comptes en valeur absolue ou en pourcentage de personnes, de familles ou de logements possédant des caractéristiques particulières dans une unité d'observation donnée.

Prendre des mesures sur une échelle ordinale revient à attribuer à chacune des variables le rang où se place une unité d'observation donnée sur la variables considérée. Ainsi, la valeur 1 est la valeur minimum que peut prendre une variable, suivie par les valeurs 2 puis 3, etc..., jusqu'à ce que les n unités d'observations aient été classées sur cette variable. Il est fréquent que de telles valeurs soient la seule source d'information disponible. C'est le cas par exemple, en Géographie du comportement, lorsqu'on demande à un groupe de gens de classer par ordre de préférence comme lieu de résidences certaines régions ou certaines villes. I1 est indispensable de savoir si de telles mesures peuvent être soumises à une analyse factorielle ou plutôt, puisque l'analyse peut toujours être mathématiquement effectuée, dans quelle mesure les résultats de cette analyse sont pertinents.

On peut chercher la réponse à cette question dans la comparaison de deux analyses qui soient exactement semblables, quant aux variables et aux unités d'observation choisies et qui ne diffèrent que par l'échelle de mesure des données. On utilise l'échelle binaire pour indiquer par exemple l'absence ou la présence d'un phénomène. Conventionnellement, 1 marque la présence, zéro marque l'absence. L'analyse factorielle a-t-elle encore un sens dans ce cas ? Là encore, une comparaison entre deux analyses identiques, dont les données seraient mesurées sur des échelles différentes, devrait nous fournir des indications précieuses sur l'utilisation possible des échelles de mesure les. plus élémentaires.

Analyses en composantes principales sur les données non métriques

Pour réaliser les expériences que l'on vient de recommander, les données de rapport qui ont été utilisées dans l'étude précédente ont dû être mesurées à nouveau sur des échelles

ordinales et binaires. Dans la première expérience, les 2 569 MCDs de l'Etat de Pennsylvanie ont été mesurées d'après leur rang sur les 42 variables socio-économiques. Dans la seconde expérience, on a utilisé la médiane, c'est à dire la valeur des variables au rang 1285 comme point de rupture. Pour chaque variable, les 1284 observations de valeur plus petite ont reçu la valeur zéro. Les 1284 observations de valeur plus grande ont pris la valeur 1.

Variables	métriques	ordinales	binaires
ED3	0.83	0.79	0.68
WCOL	0.79	0.77	0.66
FAM3	0.78	0.76	0.66
EDF	0.74	0.73	0.57
VAL3	0.73	0.70	0.50
OTHER	0.60	0.58	0.50
VAL2	0.53	0.62	0.57
SALE	0.50	0.54	0.48
PROF	0.49	0.49	0.42
LAB	0.48	0.54	0.47
FLAB	0.45	0.49	0.45
FINC	0.43	0.47	0.41
TRADE	0.33	0.40	0.39
IMIG	0.32	0.52	0.54
MOB3	0.31	0.36	0.30
HEALTH	0.30	0.41	0.39
REC	0.22	0.43	0.50
VAL1	-0.66	-0.70	-0.64
FAM1	-0.56	-0.63	-0.58
ED1	-0.50	-0.49	-0.43
SOC	-0.47	-0.46	-0.36
POLD	-0.47	-0.51	-0.48
POUNG	-0.47	-0.47	-0.46
BCOL	-0.46	-0.48	-0.43
ED2	-0.45	-0.46	-0.37
FAM2	-0.38	-0.36	-0.31
WELF	-0.38	-0.38	-0.38
FARM	-0.34	-0.26	-0.32
MINE	-0.31	-0.22	-0.11
UNEMP	-0.30	-0.28	-0.31

Table 6 - Corrélations les plus fortes entre les variables et la première composante principale. Comparaison entre les variables métriques et les variables non-métriques.

A partir des deux nouvelles matrices des données, on a obtenu deux nouvelles matrices de corrélation des 42 variables. Si l'on compare successivement chacune de ces matrices de corrélation avec la matrice de corrélation dérivée d'après les données de rapport originelles, il n'y a pas à première vue de grande différence. Dans l'expérience qui emploie des données

ordinales, certains coefficients de corrélation sont légèrement plus forts que dans l'expérience d'origine, d'autres sont un peu plus faibles, et dans l'ensemble il y en a autant de plus grands que de plus petits. I1 est intéressant de remarquer cependant que les plus grandes différences positives correspondent à des variables qui ne sont pas fortement corrélées avec aucune autre dans l'analyse d'origine, comme si les mesures d'après le rang donnaient relativement plus d'importance aux variables dont la contribution est plus faible d'après les mesures de rapport.

Dans le cas binaire, à part quelques rares exceptions, les corrélations sont toujours plus faibles que celles que l'on observe à l'origine. On doit s'attendre à une telle régularisation des résultats puisqu'en passant des mesures de rapport très précises aux mesures ordinales et binaires plus élémentaires, on a perdu de l'information. A ce stade de l'analyse, il ne semble pas que le degré d'information perdu soit très important. Peut-être apparait-il plus évident si l'on entreprend une analyse en composantes principales sur les données ordinales, puis sur les données binaires et si l'on compare les résultats de ces deux analyses à ceux que l'on a présenté dans le chapitre I de cette étude (Table 6).

Avec les données ordinales, le pourcentage de la variance expliquée par la première composante est un peu plus élevé qu'avec les données de rapport : 21,84 % au lieu de 20,46 %. Par conséquent, il faut s'attendre à trouver des saturations plus fortes dans le premier cas. En fait (Table 6), les six caractéristiques les plus corrélées avec la première composante, avec des saturations supérieures à 0,60, telles que forte éducation, revenu élevé, forte valeur du logement, cols blancs, voient diminuer leur corrélation dans l'expérience ordinale. Ces différences sont équilibrées par une légère augmentation des corrélations pour les six caractéristiques qui suivent dans l'intervalle de 0,40 à 0,60 et sont largement compensées par un accroissement net des corrélations moyennes qui concernent des variables telles que services de santé, services de loisirs, commerces, immigrants. D'autre part, les corrélations négatives, petites ou moyennes qui contribuent à l'aspect rural de la première composante, c'est à dire les agriculteurs et les mineurs, sont encore plus faibles en valeur absolue. Cependant, à part ces quelques exceptions, les saturations négatives sont de manière générale plus élevées.

Avec les données binaires la première composante rend compte d'un pourcentage plus petit de la variance : 17,72 % au lieu de 20,46 %. Ainsi, il n'est pas surprenant de trouver des corrélations plus faibles (Table 6). Bien que l'ordre d'importance des variables soit assez bien respecté, ce sont les six corrélations les plus fortes qui subissent la plus grande diminution (autour de - 0,13) ; les six corrélations suivantes décroissent beaucoup moins sensiblement, d'environ - 0,02 unités, tandis que les petites et moyennes corrélations augmentent d'en moyenne 0,10 unités. Sur le côté négatif de l'axe, l'effet n'est pas aussi marqué. La plupart des corrélations restent très semblables avec seulement une décroissance nette de trois d'entre elles : employés des mines, éducation moyenne, bénéficiaire de la Sécurité Sociale.

Il faut rappeler avant de conclure que, dans la première analyse avec des données de rapport, la distribution des saturations sut l'axe principal est très irrégulière. Pour les seize variables qui contribuent le plus à la direction positive, l'éventail des corrélations est très ouvert (de 0,30 à 0,83),avec beaucoup de valeurs extrêmes et très peu de valeurs intermédiaires. Sur la direction négative, l'éventail est plus fermé (de - 0,30 à - 0,66) et rassemble autour des valeurs moyennes. Ainsi, l'effet de l'utilisation des données ordinales semble être une régularisation de la distribution des saturations sur la première composante, avec un rétrécissement de l'étendue

des valeurs positives et un élargissement de l'étendue des valeurs négatives. Cet effet de régularisation observé avec les données ordinales est préservé et même renforcé lorsqu'on utilise des données binaires. L'impression que l'on avait tirée de la comparaison rapide des matrices de corrélation est donc confirmée après l'utilisation de méthodes plus complexes. Cependant dans l'effet régularisateur observé, n'entrent en jeu que des différences assez petites. Avec chacun des trois différents types de mesures, la première composante est définie par exactement les mêmes variables dont la contribution, si elle varie de façon notable, reste cependant dans 1e même ordre d'importance.

Facteurs	**Echelle métrique**	**Echelle ordinale**	**Echelle binaire**
Facteur 1	14.95	17.28	13.99
Facteur 2	12.35	11.95	10.04
Facteur 3	10.10	9.45	8.59
Pourcentage cumulé de la variance expliquée	37.40 %	38.68 %	32.62 %

Table 7 – Pourcentage de la variance expliquée par les trois premiers facteurs après rotation : comparaison entre données métriques, ordinales et binaires.

Variables	**métriques**	**ordinales**	**binaires**
ED3	0.85	0.83	0.71
VAL3	0.80	0.68	0.49
FAM3	0.78	0.69	0.52
EDF	0.77	0.77	0.64
WCOL	0.70	0.77	0.68
OTHER	0.49	0.56	0.46
PROF	044	0.58	0.53
MOB3	0.44	0.48	0.37
SALE	0.41	0.58	0.54
VAL2	0.31	0.52	0.43
IMIG	0.28	0.53	0.58
TRADE	0.20	0.42	0.44
FOR	0.09	0.42	0.49
FAM2	-0.72	-0.57	-0.42
ED2	-0.69	-0.53	-0.37
VAL1	-0.65	-0.60	-0.48
BCOL	-0.45	-0.44	-0.37
ED1	-0.27	-0.44	-0.43
POLD	-0.26	-0.32	-0.29

Table 8 – Corrélations entre les variables et le facteur 1 : comparaison entre données métriques, ordinales et binaires.

Rotation sur des données non métriques

I1 nous a paru intéressant de voir dans quelle mesure le type d'échelle sur laquelle sont mesurées les variables pouvait affecter la découverte, par rotation, de phénomènes sous-jacents. On a vu que dans l'analyse sur les données de rapport, la rotation sur trois axes est celle qui donne les résultats les plus satisfaisants. C'est pourquoi seule cette version a été retenue dans l'analyse comparative qui suit, entre données métriques et données non-métriques (Table 7).

La comparaison de la proportion de variance expliquée par chaque facteur dévoile déjà des différences notables entre les trois types de données. Il suffit de se reporter à la Table 7 pour constater que, d'après le critère de simple structure, les résultats de l'analyse factorielle sur les données de rapport sont les meilleurs avec respectivement 14,95 %, 12,35 % et 10,10 % de la variance expliquée par les trois facteurs. Dans l'analyse des données ordinales, l'explication de la variance n'est pas aussi bien répartie entre les trois facteurs. Avec 17,28 %, le premier facteur domine largement les deux autres qui n'expliquent que 11,95 % et 9,45 % de la variance. Avec les données binaires, le pourcentage de la variance expliquée par le premier facteur est plus petit (13,99 %) qu'avec les données de rapport mais, comme avec les données ordinales, il est dis proportionnellement élevé comparativement aux pourcentages expliqués par les deux facteurs suivants : 10,04 % et 8,59 %.

Variables	métriques	ordinales	binaires
LAB	0.66	0.71	0.70
WAGE	0.62	0.48	0.31
VAL2	0.54	0.41	0.33
FLAB	0.51	0.57	0.63
FINC	0.43	0.50	0.57
FAM3	0.30	0.37	0.38
FAM1	-0.78	-0.72	-0.63
SOC	-0.74	-0.59	-041
OLD	-0.55	-0.34	-0.12
POUNG	-0.53	-0.60	-0.56
WELF	-0.47	-0.55	-0.49
POLD	-0.38	-0.44	-0.43
VAL1	-0.37	-0.44	-0.36
MINE	-0.33	-0.36	-0.25
UNEMP	-0.31	-0.42	-0.42

Table 9 – Corrélations entre les variables et le facteur 2 : comparaison entre données métriques et non-métriques.

Si l'on compare maintenant les saturations sur le premier facteur (Table 8), on constate qu'avec les données métriques, comme avec les données non-métriques, les variables de revenu, d'éducation, de logement sont prééminentes. La seule différence vient de ce que dans le dernier cas, ces variables perdent un peu de leur importance aux dépens des variables d'emploi. L'effet de rétrécissement que l'on a remarqué sur la direction positive de la première composante

affecte maintenant les deux directions positives et négatives du premier facteur. Cette tendance est renforcée dans le cas binaire. Cependant, d'une analyse à l'autre, le premier facteur garde sa signification comme facteur de statut social. La comparaison des saturations sur le facteur 2 fait ressortir des différences dont la distribution ne semble pas être due au hasard (Table 9). Elles comprennent une diminution relative des valeurs les plus grandes qui touchent le type de revenu et la structure par âge (salariés, bénéficiaires de la Sécurité Sociale, population âgée) ; tandis que le rôle originellement modéré des indices d'activité (aide aux nécessiteux, chômeurs) et les indices de prospérité (revenu, logement) se trouve renforcé. La seule exception à cette règle concerne la population active dont la contribution déjà élevée dans l'analyse d'origine augmente dans les deux dernières analyses. A part cela, on peut ramener à nouveau l'ensemble des différences à un rétrécissement de l'éventail des saturations sur les deux directions opposées de l'axe factoriel. Ce déplacement relatif affecte si peu l'ordre de contribution des variables que le facteur 2 en tant que facteur d'activité est encore tout à fait reconnaissable; même lorsque les données sont mesurées sur une échelle très élémentaire comme l'échelle binaire.

Variables	métriques	ordinales	binaires
FLAB	0.55	0.44	0.30
OLD	0.54	0.68	0.68
FINC	0.50	0.40	0.27
FOR	0.49	0.40	0.25
HEALTH	0.42	0.30	0.18
PROF	0.41	0.26	0.13
WCOL	0.40	0.23	0.10
OTHER	0.39	0.24	0.09
LAB	0.38	0.26	0.14
SALE	0.31	0.17	0.10
SOC	0.31	0.53	0.59
VAL1	0.06	0.28	0.46
YOUNG	-0.67	-0.74	-0.66
FSELF	-0.67	-0.70	-0.63
FARM	-0.64	-0.69	-0.61
ED1	-0.38	-0.28	-0.17
PERU	-0.33	-0.24	-0.17
VAL2	-0.02	-0.23	-0.41
VAL3	-0.02	-0.22	-0.40

Table 10 - Corrélations entre les variables et le facteur 3 : comparaison entre données métriques et non-métriques.

En ce qui concerne le troisième facteur, il suffit d'un coup d'oeil sur les trois colonnes de la table 10 pour percevoir des différences remarquables entre les trois types d'analyse. Dans l'analyse ordinale, la plupart des saturations les plus élevées sont réduites d'environ 0,12 unités, affaiblissant le caractère urbain du facteur. Dans l'analyse binaire, la diminution est en gros deux fois plus forte et le rôle des caractéristiques urbaines s'en trouve d'autant plus réduit. Seule une variable forte dans le cas métrique, la population âgée, est renforcée dans les cas non-métriques et continue à contribuer à la définition du troisième facteur. Les saturations négatives n'ont presque pas changé. Elles augmentent un peu dans le cas ordinal et diminuent

légèrement dans le cas binaire. Ainsi des variables encore fortes comme la population jeune et les agriculteurs permettent de conserver l'aspect rural du facteur 3. Cependant, sur chacune des directions opposées de cet axe factoriel apparaissent quelques contributions nouvelles, typiques du facteur d'activité (bénéficiaires de la Sécurité Sociale, aide aux nécessiteux, logements moyens, salariés). Elles ont comme effet de transformer le facteur urbain-rural d'origine en un facteur "urbain peu actif-rural actif". La présence des indices notables de logement riche du côté rural et de logement pauvre du côté urbain renforce cette nouvelle identification.

En résumé, une analyse factorielle peut conduire à des résultats différents selon l'échelle sur laquelle sont mesurées les données. La distorsion des résultats suit deux tendances générales : 1)- elle est d'autant plus importante que le facteur considéré est plus petit ; 2)- elle augmente au fur et à mesure que l'on passe à des échelles de mesure moins précises. Ceci, il faut le reconnaître, ne constitue pas une découverte particulièrement originale. Reste cependant le fait que l'étude comparative qui précède apporte sa contribution dans une connaissance plus précise des mécanismes qui conduisent à de tels résultats. Ainsi, avec les échelles les plus élémentaires, les facteurs ont un pouvoir discriminant moins grand. Leur corrélation avec les variables montrent moins de valeurs extrêmes et davantage de valeurs moyennes.

Puisque les variables ont une saturation modérée sur plusieurs facteurs, on peut en induire que les groupes de variables sont moins distincts. Dans l'étude actuelle, cela peut avoir deux causes :

1)- l'échelle ordinale est, par exemple, caractérisée par un ensemble plus petit de modalités. Le choix des valeurs va seulement de 0à 2 569 alors que dans l'analyse d'origine, les valeurs en pourcentage vont de 0,000 à 99,999, avec cinq chiffres significatifs. Ainsi, la différence entre les variables est moins précise dans un cas que dans l'autre, et cela est encore plus vrai lorsque le nombre d'unités d'observations est plus petit.

2) - la plus petite différence entre deux valeurs ordinales successives est toujours la même, c'est-à-dire1, tandis que les différences entre une certaine valeur et la valeur suivante peuvent varier considérablement sur des échelles de rapport.

I1 est évident que lorsqu'on passe des données ordinales aux données binaires, on perd encore davantage de précision dans l'information de départ, perte qui se répercute dans les résultats, d'autant plus que l'analyse est plus complexe et les facteurs moins importants. Cependant, le degré d'information perdue dépend principalement de la distribution réelle des données de rapport et de combien elle diffère de ses versions ordinales et binaires. Bien sûr, les mesures ordinales et binaires ne sont utilisées que lorsque les données disponibles ne peuvent pas être mesurées sur une échelle de rapport.. Mais lorsqu'on les emploie, on ne devrait pas oublier leurs faiblesses et considérer le moindre écart dans les résultats comme significatif. Si, par expérience dans son propre domaine, un chercheur sait que d'une observation à l'autre, les variables qu'il étudie peuvent varier beaucoup, il doit consentir à perdre cette précision dans les variations en utilisant des échelles élémentaires qui, incidemment, peuvent mieux convenir à une étude de généralisation.

En conclusion, on voit que dès le départ, selon. le type de données, le but de l'étude et le degré de généralisation, le jugement du chercheur est sollicité pour choisir une échelle de

mesure particulière ou pour dégager les implications d'un choix imposé. Dans tous les cas, ces implications ou les raisons de ce choix devraient être sans cesse rappelées et soulignées au cours de l'étude.

CONCLUSION

L'étude choisie comme exemple d'exploitation de l'analyse en composantes principales et de l'analyse factorielle souligne les difficultés qui apparaissent quand on a affaire à un cas concret. Un plan de recherche classique commence par 1a définition du but de l'étude et par la délimitation de son extension spatiale. Le choix de la taille des observations et des données qui devraient servir au mieux le but proposé vient ensuite. Finalement, on est peut-être amené si nécessaire à modifier le caractère des premières phases selon les possibilités offertes par les dernières.

Un tel plan nous a conduit a choisir l'étude de la répartition spatiale et de la structure interne du bien-être socio-économique en Pennsylvanie. On a pensé que l'étude d'un Etat situé de part et d'autre de la frontière de la Région Appalachia pouvait aider à améliorer les critères utilisés pour définir cette Région; et par conséquent permettre l'amendement des programmes qui lui sont appliqués.

I1 faut rappeler que "l'Acte sur les Zones à Redévelopper" de 1961 prévoit de fournir des fonds et d'autres incitations à l'investissement pour les zones les plus déprimées. Plus tard, la politique de "l'Acte de Développement de la région Appalachia" de 1965 prévoit plutôt de concentrer l'aide là où les investissements privés ont le plus de chance de devenir rapidement profitables, c'est à dire probablement dans les endroits qui sont déjà les plus favorisés. Maintenant, il est frappant de constater qu'en définissant leurs critères d'action, ces deux programmes, dont les buts sont assez différents, tiennent compte seulement des manifestations les plus élémentaires de la pauvreté, principalement sur la base du revenu familial moyen. Ni l'un ni l'autre ne tient compte des phénomènes complexes qui engendrent ces manifestations. Ajouté à cela, on a de bonnes raisons de croire que la base du comté, sur laquelle la région Appalachia a été délimitée, ne constitue pas une grille assez fine pour séparer les zones qui ont besoin d'assistance de celles qui peuvent se suffire elles-mêmes. C'est pourquoi, pour mieux comprendre à une échelle plus fine le phénomène de santé socio-économique, on a entrepris une analyse en composantes principales au niveau des M.C.D. sur 42 caractéristiques socio-économiques, choisies avec soin d'après l'importance et la signification de leur rôle dans les études socio-économiques précédentes.

La première composante identifie l'ensemble des indices - éducation, revenu, logement, occupation professionnelle - qui expriment le mieux la variation spatiale de la santé socio-économique à travers les communes de l'Etat de Pennsylvanie. Lorsqu'on cartographie les scores sur cette première composante, on voit apparaître deux tendances spatiales principales, dont l'une suit les grandes divisions physiques régionales, et l'autre souligne le degré d'urbanisation. La première, sur la base du "Township" ou village, oppose les régions rurales

fertiles consacrées à une agriculture prospère aux pauvres régions montagneuses et forestières dont la vie économique dépend largement des industries extractives (charbon et bois), et un peu d'une agriculture de subsistance. La seconde montre comment le degré de bien-être varie selon la taille des agglomérations (bourgs, villes, centres régionaux)et surtout le type de fonctions qu'elles remplissent. En allant des scores les plus hauts aux scores les plus bas, on trouve dans l'ordre quatre différents types d'agglomérations urbaines :

1) les villes de banlieue ou villes satellites des deux grands centres régionaux, Pittsburg et Philadelphie ;
2) les villes de Collège et d'Université comme Clarion et State College ;
3) les centres régionaux eux-mêmes, et des villes manufacturières plus petites comme York ou Reading, favorisées par une grande variété de fonctions servant un arrière pays rural prospère ;
4) finalement des villes plus exclusivement consacrées à l'industrie, peu stimulées par leur environnement rural, comme Bethlehem, ou des villes minières frappées par la crise de l'industrie extractive comme Wilkes Barre.

I1 est clair maintenant que de telles considérations sur les manifestations variées d'un bien-être régional et local ne seraient pas possibles si l'on avait comme seul critère, le revenu moyen de la population agrégée des comtés; et par conséquent, il ne serait pas possible d'appliquer les mesures correctives nécessaires aux zones les plus affectées. En outre, la finesse des unités d'observation permet de montrer les réajustements qui pourraient être réalisés de part et d'autre de la frontière d'Appalachia.

L'étape suivante consiste à essayer de dégager les différentes facettes de cette composante de bien-être, dont en vient d'examiner la répartition spatiale. Les mêmes scores élevés se trouvent dans certaines banlieues de grandes villes et certaines petites villes universitaires. On rencontre des scores très bas à la fois dans des zones rurales et dans des villes minières. Ces résultats conduisent à penser qu'il existe différents types de bien-être ou d'absence de bien-être associés à différents groupements de variables. La rotation effectuée sur les trois premières composantes a permis de dégager trois facteurs distincts de bien-être. Le problème du nombre le mieux approprié de facteurs sur lesquels doit se faire la rotation n'a pas été éludé. On a vu que l'addition d'un quatrième et d'un cinquième facteur n'ajoutait aucune signification intéressante dans le cas particulier de notre étude.

Le premier facteur associé, avec des caractéristiques comme le revenu, l'éducation, le logement, a reçu la dénomination de *facteur de statut social*. Le second, lié à l'emploi, a été défini comme un *facteur d'activité*. Le troisième, qui accuse l'importance des femmes dans la population active et montre celle des services en opposition avec les propriétaires et les ouvriers agricoles, est vu comme un *facteur urbain-rural*. Ces résultats montrent dans quelle mesure l'éducation et le revenu d'un côté, le niveau de participation à la population active d'un autre, et le degré d'urbanisation d'un troisième point de vue, peuvent aider à cerner et rendre plus "saisissable" ou discernable la notion assez vague de bien-être. Ils nous aident à mieux comprendre les éléments communs ou opposés des divers groupes de M.C.D. tels que les montrent les scores sur la composante de bien-être.

Dans la dernière partie, notre intention étant de voir comment l'échelle de mesure des

données affecte les résultats, la même analyse a été reproduite avec deux types de données non-métriques. On a reconnu deux effets principaux : l'un, est un effet de régularisation, visible en ce que l'on trouve moins de valeurs extrêmes et davantage de valeurs moyennes parmi les corrélations des variables avec la composante principale et avec les trois premiers facteurs après rotation. On pense que cet effet est dû en grande partie au fait que l'éventail des modalités qui servent à évaluer les variables est plus restreint sur les échelles non-métriques que sur l'échelle métrique. Le second effet concerne l'intensité de la différence qui existe d'un cas à l'autre entre les contributions de variables aux facteurs. A mesure que le facteur décroît en importance, la différence croît, les facteurs les plus petits étant relativement plus sensibles à une perte d'information initiale. Ces deux effets caractérisent chacune des analyses faites respectivement avec des données ordinales et des données binaires. Aussi, comme on s'y attendait, on constate qu'ils sont plus forts dans le dernier cas avec une tendance très nette à une moindre contribution des variables. Cependant, on doit être très prudent si l'on veut étendre ces résultats particuliers à d'autres études dans un souci de généralisation. En effet, on pense que, d'un résultat à l'autre, l'intensité des distorsions dépend surtout de combien la distribution des données réelles utilisées initialement diffère de la distribution de leur version ordinale et binaire, différence qui, bien sur, peut varier largement d'une étude à l'autre.

En conclusion, tous les résultats résumés ci-dessus paraissent tout à fait vraisemblables et conformes à notre attente. Bien sûr, de tels résultats ne peuvent pas aller à l'encontre de notre intuition, parce qu'ils sont déterminés de manière assez intuitive par les choix initiaux du chercheur. Tout d'abord et principalement, le choix des variables est une des décisions subjectives les plus importantes que l'on soit obligé de faire. On pense que n'importe quelle augmentation, si large soit-elle, du nombre de variables est une solution aveugle incompatible avec 1a connaissance à priori, au moins implicite, que nous avons du phénomène étudié. On peut toujours trouver dans les recensements un très grand nombre de données prêtes à alimenter une étude. Mais il ne faut pas oublier qu'il existe une forte redondance, c’est à dire que la plupart des caractéristiques ont beaucoup de substituts. Dans ce cas, le rôle de l'analyse en composantes principales ou de l'analyse factorielle mobilisées dans la suppression de cette forte redondance, ne peut pas parvenir à faire ressortir la participation réelle des données à la structure du phénomène étudié. Pour éviter un tel piège, nous avons essayé de structurer notre choix autour des caractéristiques qui, d’après les expériences précédentes, sont sensées représenter au mieux le bien-être socio-économique. En outre, de notre point de vue, le nombre et la taille des observations est aussi une décision importante. Grâce à leur grand nombre, on a pu, dans notre étude, retirer davantage d'information sur le comportement relatif des variables entre elles et leur variations spatiale très fine a pu servir de base à la précision des résultats visuels sur la carte des scores.

Finalement, dans le contexte de l'étude actuelle, étant donné le grand nombre d’observations et 1a distribution particulière des variables choisies, on a pu montrer que l'échelle de mesure n'est pas une solution méthodologique décisive à la complexité du problème. Ainsi, la connaissance des implications qui découlent de la sélection des variables, de la manière de les mesurer et du choix des unités d'observation permet de conserver, au cours de l'étude, le maximum de l'information introduite dès le départ ; car si ces méthodes multivariées sont puissantes, l'utilité de leurs résultats dépend surtout de la manière dont elles sont alimentées.

CONCLUSION

Au cours des quinze dernières années, l'analyse en composantes principales, l'analyse factorielle et les analyses multivariées voisines ont été employées dans de nombreuses études et ont envahi la littérature géographique de langue anglaise. A cause de leur vulgarisation même, ces techniques ont été trop souvent utilisées de façon peu sélective, et ont acquis peu à peu une mauvaise réputation. Depuis deux ou trois ans, elles sont l'objet d'une grande méfiance de la part des chercheurs, parfois teintée d'un certain mépris. Or elles ne méritent ni "cet excès d'honneur, ni cette indignité".

Rappelons que le rôle d'un modèle factoriel consiste à rechercher dans quelle mesure les variations qui existent entre des unités spatiales, et telles qu'elles sont exprimées par des variables originales, peuvent être expliquées par un nombre plus petit de nouvelles variables indépendantes appelées "composantes principales" ou "facteurs". Leurs avantages, par rapport aux méthodes répandues de régression et de corrélation multiple et partielle, comprennent :

- 1) la possibilité d'examiner sur un très grand nombre d'observations les relations d'un grand nombre de variables inextricablement liées entre elles, sans qu'il soit nécessaire de supposer l'indépendance d'une ou de plusieurs de ces variables ;

- 2) ce qui, par conséquent, permet l'élimination des dangers d'une explication causale qui peut varier selon le nombre de variables et l'ordre de leur introduction dans l'analyse.

Cependant, la présentation en détail des procédés de calcul montre les faiblesses que les modèles factoriels partagent avec les méthodes de corrélation et de régression. Ce sont en particulier des contraintes de linéarité, de normalité, d'indépendance à travers l'espace et le temps, imposées sur la distribution des données. D'autre problèmes plus particuliers aux modèles factoriels ont été soulevés : les communautés, le nombre de facteurs, le type de rotation, le choix des données, les échelles de mesures, tous soulignant le nombre et la variété des différentes options ou des combinaisons d'options qui sont offertes. Les applications présentées donnent une indication de l'étendu des domaines déjà traités, et montrent que beaucoup de possibilités techniques et de sujets d'études restent encore à explorer.

Jusqu’à présent, en une quinzaine d'années d'utilisation intensive, les méthodes d'analyse factorielle ont permis surtout de retrouver de "vieilles évidences" et des généralités déjà bien établies. Il n'y a pas lieu d'être déçu par cette concordance, car on ne peut pas raisonnablement s'attendre à ce que des méthodes, mêmes puissantes, contredisent les résultats de plus d'un siècle d'observations et de patientes monographies.

Cependant les méthodes factorielles ont aussi prouvé que leurs potentialités étaient grandes, et sans doute davantage à la mesure de la rapidité des évolutions de l’époque contemporaine. Cc n'est pas par hasard que ces méthodes se sont d'abord développées aux Etats-Unis, et ont démontré leur plus grande utilité dans l'étude des villes américaines, là où les

arguments historiques sont faibles par rapport à la vitesse des transformations. Ces transformations, pour être contrôlées, exigent de promptes décisions qui peuvent être fondées sur les résultats d'analyses quantitatives. C'est là aussi qu'apparaissent les plus grands dangers de l'utilisation de méthodes comme l'analyse factorielle, dangers de construire l'avenir sur des rapports faux parce qu'incomplets et trop schématiques, que l'on présente souvent comme argument décisif contre ces méthodes.

Or, il est nécessaire de souligner que ces dangers ne sont pas inhérents à l'utilisation des modèles mathématiques, car les mathématiques n'ont jamais eu la prétention de tout pouvoir enfermer dans des rapports simples ou de référer à un ordre qui serait *naturel* et unique. Si l'utilisation de méthodes mathématiques est dangereuse, c'est parce qu'en général on confère à ces dernières un pouvoir magique qu'elles n'ont pas. Plus elles seront pratiquées et moins elles seront dangereuses, dans la mesure où il est plus facile à ceux qui connaissent leur fonctionnement de dénoncer la "mauvaise pratique" qu'on peut en faire. C'est ainsi qu'HARVEY, aux Etats-Unis, a pu dénoncer la philosophie par trop déterministe de BERRY et de ses disciples. HARVEY reconnaît l'utilité des théories et des mathématiques qui permettent de les élaborer et de les tester ; mais il reproche à BERRY de leur prêter un caractère normatif qui conduit à prédire par exemple que "les pauvres doivent vivre nécessairement 1à où ils ont le moins le moyen de vivre". Pour HARVEY, il faut faire en sorte que de telles théories deviennent fausses :

> "La surenchère compétitive pour l'utilisation de 1a terre est un mécanisme qui a servi à construire une théorie et qu' il faut éliminer" (HARVEY, 1972)

HARVEY et ses disciples, de plus en plus nombreux parmi les nouvelles générations de géographes anglo-saxons, pensent qu'une révolution philosophique doit accompagner la révolution quantitative. Vivement intéressés par l'approche dialectique, ils ne démentiraient certainement pas les propos de Pierre GEORGE et ses collaborateurs (1964, p. 27) que par ailleurs ne contredit pas l'approche factorielle :

> "Le problème spécifique de la géographie est d'étudier, à l'intérieur d'un espace défini, tous les rapports de causalité entre les phénomènes de consommation au sens le plus large du terme - y compris 1'occupation des logements et le recours aux services - et les phénomènes de production ; de déterminer les groupes homogènes d'évolution synchrone et corrélative ; de les isoler des simples faisceaux de coïncidences circonstancielles, et de faire apparaître les contradictions et les survivances inhibitrices. "

En France, où cette approche philosophique a fait depuis longtemps école, une ouverture s'est produite récemment vers les méthodes mathématiques. Depuis deux ou trois ans, on voit apparaître des études faisant appel à l'analyse factorielle, comme *L'analyse factorielle appliquée à la région milanaise* (DALMASSO et alii, 1973) ou *L'analyse écologique des structures de l'utilisation du sol* (ALLAIRE et alii, 1973) ; et à l'analyse des correspondances comme *Composantes et types socio-professionnels des campagnes du Languedoc-Roussillon* (AURIAC F et BERNARD M, 1974) ou *La composition par âge de 141 villes touristiques du littoral français* (CRIBIER et alii, 1974).

APPENDICE A

DESCRIPTION DES VARIABLES

Variables	Description	
Structure par âge, mobilité, éducation		
OLD	Personnes âgées de 65 ans et plus	(a)
YOUNG	Personnes âgées de moins de 18 ans	(a)
FOR	Personnes dont l'un des parents au moins est né à l'étranger	(a)
IMIG	Personnes immigrées depuis 1945	(a)
MOB1	Personnes dont la résidence à l'intérieur du même comté était différente en 1965	(b)
MOB2	Personnes qui résidaient dans un autre comté en 1965	(b)
MOB3	Personnes qui résidaient dans un autre Etat en 1965	(b)
ED1	Personnes ayant reçu une éducation limitée au premier degré	(c)
ED2	Personnes ayant reçu 4 ans d'éducation du second degré	(c)
ED3	Personnes ayant reçu un an au moins d'éducation dans un Collège[21]	(c)
EDF	Femmes ayant reçu un an au moins d'éducation dans un Collège	(c)

Variables	Description	
Emploi par type d'occupation et par type d'industrie		
LAB	Personnes dans la population active	(d)
FLAB	Femmes dans la population active	(d)
HEALTH	Personnes employées dans les professions médicales et la santé	(e)
SALE	Personnes employées dans le commerce : vendeurs, représentants, …	(e)
WCOL	Cols Blancs : Professions libérales, techniciens, administrateurs,	(e)
BCOL	Cols Bleus : artisans, contremaîtres, ouvriers	(e)
FARM	Personnes employées dans l'agriculture : propriétaires, ouvriers	(e)

[21] La première année de collège correspond à la classe terminale d'un lycée français.

SERV	Personnes employées dans les services	(e)
MINE	Personnes employées dans l'industrie minière	(e)
TRADE	Personnes employées dans le commerce de gros et de détail	(e)
REC	Personnes employées dans les services de loisirs	(e)
PROF	Personnes exerçant des professions libérales	(e)
UNEMP	Personnes qualifiées en chômage	(e)

Revenu et type de revenu

FAM1	Familles dont le revenu est inférieur à 3000 $[22]	(f)
FAM2	Familles dont le revenu est compris entre 3000 et 15000 $	(f)
FAM3	Familles dont le revenu est égal ou supérieur à 15000 $	(f)
FINC	Femme dont le revenu est supérieur à 3000 $	(e)
WAGE	Salariés	(e)
MSF	Patrons	(e)
FSELF	Exploitants agricoles	(e)
SOC	Bénéficiaires de la Sécurité Sociale[23]	(e)
WELF	Bénéficiaires de l'Assistance Publique[24]	(e)
OTHER	Personnes ayant d'autres types de revenu : actionnaires, rentiers, …	(e)
FAMO	Familles dont le revenu est inférieur à 0.75 % du seuil de pauvreté[25]	(f)
POLD	Personnes âgées de 65 ans et plus avec un revenu inférieur au seuil de pauvreté	(g)
POUNG	Personnes âgées de moins de 18 ans dont le revenu familial est inférieur au seuil de pauvreté.	(h)

[22] A l'époque du recensement, le dollar valait environ 5 francs, soit 0.75 €

[23] Paiements effectués par l'administration de la SS dans le cadre du Programme National de Secours aux Personnes frappées d'incapacité de Travail : personnes âgées, veuves, malades, …

[24] Paiements de l'Assistance Publique aux familles pauvres avec enfants à charge, aux personnes âgées, aux aveugles, etc.

[25] A l'époque du recensement, était fixé à 4 000 $ pour une famille avec 4 enfants à charge.

Variables	Description	
Valeur et type de logement		
PERU	Logements comprenant plus de 1.5 personnes par pièce	(i)
VAL1	Logements dont la valeur est inférieure à 10 000 $	(i)
VAL2	Logements dont la valeur est comprise entre 10 000 et 25 000 $	(i)
VAL3	Logements dont la valeur est égale ou supérieure à 25 000 $	(i)
VACU	Logements inoccupés depuis six mois et davantage	(i)

NB : Toutes les variables énumérées ci-dessus sont exprimées en pourcentages d'un ensemble ; particulier d'individus, de familles ou de logements. Les indices entre parenthèses permettent de se reporter pour chaque variable, à l'un des ensembles dont la définition suit :

(a) Population totale
(b) Population âgée de 5 ans et plus
(c) Population âgée de 20 à 50 ans
(d) Population âgée de 14 ans et plus
(e) Population active
(f) Nombre total de familles
(g) Population âgée de 65 ans et plus
(h) Population âgée de moins de18 ans
(i) Nombre total de logements

APPENDICE B

MATRICES DES CORRELATIONS CALCULEES

D'APRES LES VARIABLES D'ORIGINE

Corrélations entre les variables originales et les deux premiers facteurs après rotation.

Variables	n°	Facteur 1	Facteur 2
OLD	1	0,15	-0,55
YOUNG	2	- 0,36	0,22
FOR	3	0,31	-0,04
IMIG	4	0,29	0,13
MOB1	5	-0,03	0,27
MOB2	6	0,21	0,02
MOB3	7	0,35	0,02
ED1	8	-0,41	-0,18
ED2	9	-0,58	-0,02
ED3	10	0,87	0,19
EDF	11	0,78	0,15
LAB	12	0,14	0,66
FLAB	13	0,21	0,51
HEALTH	14	0,27	0,13
SALE	15	0,51	0,13
WCOL	16	0,81	0,19
BCOL	17	-0,51	- 0,04
FARM	18	-0,29	- 0,16
SERV	19	0,02	- 0,06
MINE	20	-0,15	- 0,33
TRADE	21	0,31	0,11
REC	22	0,22	0,05
PROF	23	0,58	0,00
UNEMP	24	-0,14	- 0,31
FAM1	25	-0,16	- 0,77
FAM2	26	-0,63	0,27
FAM3	27	0,74	0,29
FINC	28	0,23	0,43
WAGE	29	-0,27	0,62
NFS	30	0,11	0,01
FSELF	31	-0,34	-0,09
SOC	32	-0,08	- 0,74
WELF	33	-0,14	- 0,47
OTHER	34	0, 62	0,14
FAMO	35	-0,16	- 0,68
POLD	36	-0,31	- 0,38
POUNG	37	-0,22	- 0,53
PERU	38	-0,24	- 0,09
VALl	39	-0,54	- 0,37
VAL2	40	0,29	0,54
VALS	41	0,71	0,24
VACU	42	-0,09	- 0,16

Corrélations entre les variables originales et les trois premiers facteurs près rotation

Variables	n°	Facteur 1	Facteur 2	Facteur 3
OLD	1	- 0,10	- 0,54	0,53
YOUNG	2	- 0,04	0,21	- 0,67
FOR	3	0,09	- 0,03	0,48
IMIG	4	0,28	0,13	0,10
MOB1	5	- 0,06	0,27	0,05
1oB2	6	0,28	0,02	- 0,06
MOB3	7	0,44	0,02	- 0,06
ED1	8	- 0,26	- 0,18	- 0,38
ED2	9	- 0,68	- 0,02	0,04
ED3	10	0,84	0,19	0,26
FDF	11	0,77	0,16	0,22
LM	12	- 0,04	0,66	0,38
FLAB	13	- 0,06	0,51	0,55
HEALTH	14	0,08	0,13	0,42
SALE	15	0,41	0,13	0,30
WCOL	16	0,70	0, 20	0,40
BCOL	17	- 0,44	- 0,04	- 0,25
FARM	18	0,00	- 0,17	- 0,63
SERV	19	- 0,10	- 0,06	0,25
MINE	20	- 0,16	- 0,33	- 0,02
TRADE	21	0,19	0,11	0,29
REC	22	0,17	0,05	0,14
PROF	23	0,43	0,00	0,41
UNEMP	24	- 0,16	- 0,31	- 0,00
FAM1	25	- 0,13	- 0,77	- 0,10
FAM2	26	- 0,71	0,27	- 0,00
FAM3	27	0,78	0,30	0,10
FINC	28	- 0,01	0,43	0,50
NAGE	29	- 0,21	0,62	- 0,17
NFS	30	0,18	0,01	- 0,10
FSELF	31	- 0,03	- 0,10	- 0,67
SOC	32	- 0,26	- 0,74	0,31
WELF	33	- 0,17	- 0,47	0,01
OTHER	34	0,49	0,14	0,39
FAM0	35	- 0,06	- 0,68	- 0,21
POL	36	- 0,25	- 0,38	- 0,18
POUNG	37	- 0,13	- 0,53	- 0,20
PERU	38	- 0,10	- 0,09	- 0,33
VAL1	39	- 0,64	- 0,37	0,05
VAL2	40	0,31	0,54	0,02
VAL3	41	0,79	0,24	0,02
VACU	42	- 0,07	- 0,16	- 0,05

Corrélations entre les variables originales et les cinq premiers facteurs après rotation.

Variables	n°	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5
OLD	1	- 0,14	- 0,32	0,29	- 0,12	0,71
YOUNG	2	- 0,05	0,12	- 0,50	0,25	- 0,47
FOR	3	0,05	0,06	0,02	- 0,56	0,21
IMIG	4	0,26	0,18	0,01	- 0,10	0,02
MOB1	5	- 0,07	0,22	0,10	- 0,03	- 0,10
MOB2	6	0,34	- 0,07	0,04	0,04	- 0,15
MOB3	7	0,46	0,00	- 0,06	- 0,00	- 0,08
ED1	8	- 0,24	- 0,23	- 0,14	0,35	- 0,10
ED2	9	- 0,72	- 0,02	- 0,00	- 0,07	0,03
ED3	10	0,86	0,20	0,16	- 0,21	0,00
EDF	11	0,81	0,12	0,20	- 0,14	- 0,04
LAB	12	- 0,00	0,51	0,68	0,01	- 0,14
FLAB	13	- 0,00	0,35	0,81	- 0,06	- 0,03
HEALTH	14	0,13	0,04	0,43	- 0,21	0,00
SALE	15	0,33	0,29	- 0,01	- 0,31	0,22
WCOL	16	0,69	0,25	0,17	- 0,35	0,09
BCOL	17	- 0,47	- 0,07	- 0,37	- 0,10	- 0,19
FARM	18	- 0,03	- 0,05	- 0,19	0,83	0,12
SERV	19	- 0,01	- 0,20	0,33	- 0,10	- 0,04
MINE	20	- 0,13	- 0,35	- 0,23	- 0, 22	- 0,02
TRADE	21	0,13	0,24	- 0,01	- 0, 33	0,18
REC	22	0,18	0,05	0,04	- 0,16	0,01
PROF	23	0,52	- 0,09	0,37	- 0,26	0,00
UNEMP	24	- 0,10	- 0,40	- 0,12	- 0, 19	- 0,09
FAMl	25	- 0,05	- 0,77	- 0,03	0,21	0,23
FAM2	26	- 0,69	0,10	0,09	- 0,08	- 0,29
FAM3	27	0,71	0,44	- 0,02	- 0,07	0,09
FINC	28	0,02	0,34	0,69	- 0,05	0,03
WAGE	29	- 0,07	0,17	0,16	- 0,08	- 0,86
NFS	30	0,12	0,14	- 0,04	0,25	0,19
FSELF	31	- 0,06	- 0,02	- 0,19	0,84	0,03
SOC	32	- 0,28	- 0,54	0,00	- 0,15	0,59
WELF	33	- 0,14	- 0,45	- 0,14	- 0,10	0,14
OTHER	34	0,37	0,40	0,01	- 0,30	0,39
FAMO	35	0,02	- 0,73	- 0,10	0,22	0,08
POL	36	- 0,15	- 0,54	0,01	0,17	- 0,12
POUNG	37	- 0,07	- 0,56	- 0,13	0,17	0,04
PERU	38	- 0,07	- 0,18	0,27	0,08	- 0,23
VAL1	39	- 0,57	- 0,49	0,07	- 0,06	- 0,00
VAL2	40	0,24	0,59	0,02	- 0,02	- 0,09
VAL3	41	0,74	0,36	- 0,09	- 0,06	0,03
VACU	42	- 0,08	- 0,12	- 0,04	0,10	0,10

APPENDICE C

MATRICES DES CORRELATIONS CALCULEES

D'APRES LES VARIABLES TRANSFORMEES

EN DONNEES NON-METRIQUES

Corrélation entre les variables ordinales et les trois premières composantes principales

Variables	n°	Composante 1	Composante 2	Composante 3
OLD	1	- 0,12	0, 74	0,04
YOUNG	2	- 0,20	- 0,67	- 0,25
FOR	3	0,33	0,45	- 0,14
IMIG	4	0,52	0,04	- 0,17
MOBl	5	0,14	- 0,05	0,08
MOB2	6	0,22	- 0,15	- 0,18
MOB3	7	0,36	- 0,08	- 0,34
ED1	8	- 0,48	- 0,24	0,00
ED2	9	- 0,46	0,16	0,32
ED3	10	0,79	0,09	- 0,28
EDF	11	0,72	0,05	- 0,29
LAB	12	0,54	- 0,02	0,55
FLAB	13	0,49	0,18	0,51
HEALTH	14	0,40	0,28	- 0,01
SALE	15	0,54	0,19	- 0,19
WCOL	16	0,77	0,23	- 0,18
BCOL	17	- 0,47	- 0,15	0,05
FARM	18	- 0,25	- 0,59	- 0,30
SERV	19	0,07	0,29	0,00
MINE	20	- 0,22	0,20	- 0,21
TRADE	21	0,40	0,19	- 0,11
REC	22	0,43	0,13	- 0,15
PROF	23	0,48	0,32	- 0,24
UNEMP	24	- 0,28	0,25	- 0,21
FAM1	25	- 0,62	0,26	- 0,37
FAM2	26	- 0,36	- 0,04	0,47
FAM3	27	0,76	- 0,15	- 0,13
FINC	28	0,46	0,18	0,43
WAG,E	29	0,09	- 0,44	0,34
NFS	30	0,14	- 0,20	- 0,14
FSELF	31	- 0,26	- 0,61	- 0,29
SOC	32	- 0,46	0,68	- 0,08
WELF	33	- 0,38	0,34	- 0,28
OTHFR	34	0,58	0,22	- 0,09
FAMO	35	- 0,54	0,12	- 0,44
POLD	36	- 0,51	0,08	- 0,17
POUNG	37	- 0,47	0,13	- 0,39
PERU	38	- 0,10	- 0,12	- 0,28
VAL1	39	- 0,69	0,36	0,11
VAL2	40	0,62	- 0,31	- 0,07
VAL3	41	0,69	- 0,23	- 0,24
VACU	42	- 0,09	0,14	- 0,28

Corrélations entre les variables ordinales et les trois premiers facteurs après rotation

Variables	n°	Facteur 1	Facteur 2	Facteur 3
OLD	1	- 0,03	0,68	- 0,33
YOUNG	2	- 0,11	- 0,74	- 0,02
FOR	3	0,41	0,40	- 0,10
IMIG	4	0,53	0,04	0,13
MOBl	5	0,06	-0,00	0,16
MOB2	6	0,26	- 0,18	0,04
MOB3	7	0,48	- 0,17	- 0,02
ED1	8	- 0,43	- 0,28	- 0,16
ED2	9	- 0,53	0,22	- 0,08
ED3	10	0,83	0,07	0,18
EDF	11	0,77	0,02	0,15
LAB	12	0,14	0,26	0,71
FLAB	13	0,14	0,44	0,57
HEALTH	14	0,38	0,30	0,09
SALE	15	0,58	0,17	0,07
WCOL	16	0,70	0,20	0,19
BCOL	17	- 0,44	- 0,18	- 0,15
FARM	18	- 0,12	- 0,69	- 0,12
SERV	19	0,04	0,26	- 0,10
MINE	20	- 0,04	0,07	- 0,35
TRADE	21	0,42	0,18	0,05
REC	22	0,46	0,11	0,06
PROF	23	0,57	0,26	- 0,04
UNEMP	24	- 0,08	0,11	- 0,41
FAM1	25	- 0,28	0,01	- 0,72
FAM2	26	- 0,56	0,10	0,17
FAM3	27	0,68	- 0,10	0,37
FINC	29	0,17	0,39	0,49
WAGE	29	- 0,16	- 0,24	0,48
NFS	30	0,17	- 0,22	0,05
FSELF	31	- 0,13	- 0,70	- 0,11
SOC	32	- 0,25	0,53	- 0,58
WELF	33	- 0,12	0,14	- 0,55
OTHER	34	0,56	0,24	0,16
FAM0	35	- 0,19	- 0,13	- 0,67
POLD	36	- 0,32	- 0,05	- 0,44
POUNG	37	- 0,15	- 0,09	- 0, 60
PERU	38	0,05	- 0,24	- 0,22
VAL1	39	- 0,59	0,28	- 0,44
VAL2	40	0,51	- 0,23	0,41
VAL3	41	0,68	- 0,22	0,29
VACU	42	0,09	0,00	- 0,31

BIBLIOGRAPHIE

ABLER R., ADAMS J.S., GOULD P.R. (1971) *Spatial Organization, the Geographer's View of the World*, Englewood Cliffs, Prentice Hall.

ABU-LUGH0D J. (1969), "Testing the theory of Social Area Analysis : the Ecology of Cairo, Egypt" *American Sociological Review*, 34, pp. 198-212

AHMAD Q. (1965), "Indian Cities : Characteristics and Correlates", *Research Paper 102*, Department of Geography, University of Chicago.

ALLAIRE G, PHIPPS M, STOUPY M. (1973), *Espace Géographique,* n° 3

ANDERSON T. W (1958). *Introduction to Multivariate Statistical Analysis*, New-York : John Wiley et Sons, Inc.

ANDERSON T.R. and BEAN L.L (1961), "The Shevky-Bell Social Areas : Confirmation of Results and Reinterpretation", *Social Forces*, Vol. 40, pp. 119-124.

ANDERSON T.R. and EGELAND J. S (1961.), "Spatial Aspects of Social Area Analysis", *American Sociological Review*, vol. 26, pp. 392-98.

AURIAC F. et BERNARD M.C. (1974), "Composantes et types socio-professionnels des campagnes du Languedoc-Roussillon", *Bulletin de la Société Languedocienne de Géographie*, Tome 8, Fascicule 1, Montpellier.

BARBUT Marc, (1968) *Mathémathiques des Sciences Humaines*, P.U.F., Paris

BELL W (1953), "The Social Areas of the San Francisco Bay Region" *American Sociological Review*, XVIII, pp. 29-47

BELL W. and E. SHEVKY (1955), *Social Area Analysis*, Stanford University Press.

BELL W. (1955), "Economic, Family and Ethnic Status : An Empirical Test", *American Sociological Review*, XX, pp. 45-52.

BENZECRI J.P. et al. (1973) *L'analyse des Données*, Dunod, Paris, 2 vol.

BERRY B.J.L (1960) "An Inductive Approach to the Regionalization of Economic Development", *Research Paper 62*, Department of Geography, University of Chicago.

BERRY B.J.L (1961 a) "Basic Patterns of Economic Development" in *Atlas of Economic Development*, University of Chicago Press, pp. 110-119.

BERRY B.J.L (1961b), "A method for deriving Multi-Factor Uniform Regions", *Polish Geographical Review*, 13 : 263-282.

BERRY B. J. L. and BARNUM H. G, (1962), "Agregate Relations and Elemental Components of Central Place Systems", *Journal of Regional Science*, Vol. 4, pp. 35-68

BEPRY B.J. L. (1965), "Identification of Declining Regions : An Empirical Study of the Dimensions of Rural Poverty" in R.S. THOMAN and W.D. WOOD (eds.), *Areas of Economic Stress in Canada*, Kingston : Queen's University Press, Ontario, pp. 22-66

BERRY B.J.L. and TENNANT R.J. (1965), "Socio-economic Classification of Municipalities in Northeastern Illinois Metropolitan Area", *Commercial Structure*, (Chicago : Northeastern Illinois Metropolitan Area Planning Commission)

BERRY B.J. L. (1966), "Essays on Commodity Flows and the Spatial Structure of the Indian Economy", *Research Paper 111*, Department of Geography, University of Chicago

BERRY B.J. L. (1968), "Tnterdependency of Spatial Structure and Spatial Behavior : A General Field Theory Formulation", *Papers of the Regional Science Association*, Vol. XXI, pp.205-227.

BERRY B.J. L. (1969), "Relationships between Regional Economic Development and the Urban System : The Case of Chile", *Tijdschrift voor Economische en Sociale Geografie*, Vol. 60, pp 283-307

BERRY B.J. L. and REES P.H (1969), "The Factorial. Ecology of Calcutta", *American Journal of Sociology*, LXXIV : 5, pp. 445-491

BERRY B.J. L. and HORTON F.E (1970), *Geographic Perspectives in Urban Svstems*, N.Y. : Prentice Hall Inc.

BERRY B.J. L. ed (1971), "Comparative Factorial Ecology", *Economic Geography*, 47, (Supplement)

BERRY B.J. L. and SMITH K.B. eds., (1972), *City Classification Handbook : Methods and Applications*, New-York : John Wiley and Sons.

BROWN S.E. and TROTT Ch. E (1968), "Grouping Tendencies in an Economic Regionalization of Poland", *Annals, Association of American Ceographers*, Vol. 58, n° 2, pp. 327-342

BROWN L. A and LONGBRAKE D.B (1970), "Migration Flows in Intraurban Space : Place Utility Considerations", *Annals, Association of American Ceographers,* Vol. 60, n° 2,

op. 368-384

BUNGE W. (1962), *Theoretical Geography*, Lund Studien in Geography, Series C, Vol. 1, 210 p

BURGESS W (1925), "The growth of the City" in R.E. PRK, E.W. BURGESS and R.D. Mc KENZIE, *The City*, Chicago : University of Chicago Press.

CAREY G. W (1966), "The Regional Interpretation of Manhattan Population and Housing Patterns Through Factor Analysis", *The Geographical Review*, Vol.. 56, pp. 551-569.

CARROLL J.B. (1953), "An Analytical Solution for Approximating Simple Structure in Factor Analysis", *Psychometrika*, 18, pp. 23-38

CATTELL R.B. (1952), *Factor Analysis : An Introduction and Manual for the Psychologist and Social Scientist*, New-York, Harper and Row.

CATTELL R.B. (1965), "Factor Analysis, An Introduction to Essentials" *Biometrics*, Vol. 21, pp. 190-215

CAUVIN C, DALMASSO E., FALLER M, PRUVOT M., RIMBERT S., SCHAUB G (1973), « Analyse Factorielle appliquée à la région milanaise », Equipe de recherche associée au CNRS, n° 214, *UER de Géographie*, Université Louis Pasteur, Strasbourg

CHADULE (Groupe) (1974) *Initiation aux méthodes statistiques en géographie*, Masson.

CHORLEY R.J and HAGGETT P, eds (1967) *Models in Geography*, London, Methuen.

CLARK W.A.V. and CADWALLADER M (1973), "Locational Stress and Residential Mobility", *Environment and Behavior*, n° 1, Vol. 5, pp. 29-41

COX K.R (1968), "Suburbia and Voting Behavior in the London Metropolitan Area", *Annals, Association of American Geographers*, Vol. 58, n° 1, pp. 111-127

CRIBIER F, DENIAU C, KYCH A & LEPAPE L (1974), "Composition par Age de 141 villes touristiques du littoral français", *Population*, Juin 1974, pp. 465-490

C.U.S. (1968), *Papers from the Center for Urban Studies*, University of Chicago.

DAVIES W.K.D (1971), "Varimax and the Destruction of Generality", *Area*, 3, 112-118 ; 254-259

DAVIES W.K.D (1972), « Varimax and Generality : a Second Reply », *Area*, 4, pp 207-8.

DENIAU C. et LEBART L (1969), *Introduction à l'Analyse Factorielle*, CREDOC, Paris

DOGAN M and ROKKAN S, eds (1969), *Quantitative Ecological Analysis in the Social Sciences,* M.I.T. Press, Cambridge, Mass.

FISCHER J. C (1966), *Yugoslavia, A Multinational. State*, San-Francisco, Chandler Publ Co.

GARRISON W and MARBLE D.F, eds. (1967), *Quantitative Geography*, Northwestern University Studies in Ceography, n° 13

GEORGE P. (1965), *Sociologie et Géographie*, P.U.F, Paris.

GEORGE P, GUGLIELMO, KAYSER, LACOSTE (1964), *La géographie active*, P.U.F., Paris

GITTUS E. (1964), "The Structure of Urban Areas", *The Town Planning_Review*, 35, pp. 5-20

GOLLEDGE R.G., BRIGGS R. and DEMKO D. (1969), "Configurations of Distances in Intra-Urban Space", *Proceedings of the Association of American Geographers,* Vol. 1, pp. 60-65

GOODALL D.W. (1954), "Objective Methods for the Classification of Vegetation, III, An Essay in the Use of Factor Analysis", *Australian Journal of Botany*, Vol. 2, pp. 304-324

GOULD P. (1965), *On Mental Maps*, Ann Arbor, Michigan Inter-University Community of Mathematical Geographers, Michigan.

GOULD P. R (1967), "Structuring Information on Spatial-Temporal Preferences", *Journal of Regional Science*, Vol.. 7, n° 2 (Supplement)

GOULD P. R. (1969a), "The Structure of Space Preferences in Tanzania", *Area*, N° 4, pp. 29-35.

GOULD P. R. (1969b), "Problems of Space Preferences, Measures and Relationships", *Geographical Analvsis*, Vol.1., n° 1, pp. 31-44.

GOULD P. R. (1969c), "Methodological Developments since the Fifties", *Progress in Ceography*, Vol. 1

GOULD P. R (1972), « The Black Boxes of Jönköping » in DOWNS R. M. and STEA D. (eds), *Cognitive Mapping Images of Spatial Environment*, Aldine, Chicago

GOULD P. R and OLA D. (1970), "The Perception of Residential Desirability in the Western Region of Nigeria", *Environment and Planning*, Vol. 2, pp. 73-87

GOULD P. R. and WHITE R. (1968), "The Mental Maps of British School Leavers", *Regional Studies*, Vol. 2, pp 161-182

GOULD P.R. and WHITE R. (1974), "The Mental Maps" in *Pelican Geography and Environment Studies*, Ed. Peter HALL, Penguin Books Ltd, Harmondsworth, England

HADDEN J. K and BORGATTA E.F (1965), *American Cities : Their Social Characteristics*, Chicago : Rand McNELLY et Co

HAGGETT P. (1965), *Locational Analysis in Human Geography,* New-York, St Martin's Press, Inc

HAGOOD M. J. (1943), "Statistical Methods for Delination of Regions Applied to Data on Agriculture and Population", *Social Forces*, Vol. 21, pp. 287-297

HARMAN H (1960) *Modern Factor Analysis*, Chicago : University of Chicago Press.

HARRIES K. D. (1973), "Spatial Aspects of Violence and Metropolitan Population", *The Professional Geographer*, Vol. 25, n° l., pp 1-6

HARRIS CHAUNCY D and ULLMAN E. L (1945), "The Nature of Cities", *The Annals of the American Academy of Political and Social Science*, CCXLII, p. 7-17

HARVEY D (1972), Perspectives in Geography, *Geography of the Ghetto, Perceptions, Problems and Alternatives*, Harold M. R , Ed. Northern Illinois University Press

HENSHALL J.D. and KING J.L (1966), "Some Structural Characteristics of Peasant Agriculture in Barbados", *Economic Geography*, Vol. 42, pp. 75-84

HERBERT D.T. (1968), "Principal Components Analysis and British Studies of Urban Social Structure", *Professional Geographer*, Vol. 20, pp. 280-283

HERBERT D.T. (1970), "Principal Component Analysis and Urban Social Structure : A Study of Cardiff and Swansea", *Urban Essays : Studies in the Geography ofWales*, ed. Carter and Waine

HODGE G (1968), "Urban Structure and Regional Development", *Regional Science Ass. Papers*, Vol. 21, pp. 101-123

HOLZINGER.Y and HARMAN H. H (1941), *Factor Analysis*, Chicago : University of Chicago Press.

HOYT, H (1939), *The Structure and Growth of Residential Neighhorhoods in American Cities*, Washington, Federal Housing Administration

JEFFREY D, CASETTI E, KING L. (1969), "Economic Fluctuations in a Multiregional Setting, a bi-factor Analytic Approach", *Journal of Regional Science*, Vol. 9, n° 3, pp. 397-404

KAISER H. F. (1958), "The Varimax criterion for Analytic Rotation in Factor Analysis", *Psychometrika*, 23, 187-200

KENDALL M. and STUART A. (1961), *The Advanced Theory of Statistics*, New-York : Hafner, Vol. II, *Inference and Relationship*

KING L.J. (1961), "A Multivariate Analysis of the Spacing of Urban Settlements in the United

States", *Annals, Association of American Geographers*, Vol. 51, pp 222-233

KING L.J. (1966), "Cross-Sectional Analysis of Canadian Urban Dimensions : 1951 and 1961", *Canadian Geographer,* Vol. 10, no. 205-224

KING L.J. (1969), *Statistical Analysis in Geography*, Englewood Cliffs, New-York : Prentice Hall

KLOVAN J.E (1968), "Selection of Target Areas by Factor Analysis", *Western Miner*, p. 44-54

KRUMBEIN W.C and IMBRE J. (1963), "Stratigrafic Factor Maps", *American Association of Petroleum Geologist, bulletin*

KRUSKAL J.B. (1964) "Multi-dimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis", *Psychometrika*, Vol. 29, pp. 1-27 ; 115-129

MABOGUNJE A. L. (1965), "Urbanization in Nigeria : A Constraint on Economic Development", *Economic Development and Cultural Change*, Vol. 13, pp. 413-438

MAHALANOBIS P.C. (1936), "On the Generalized Distance in Statistics", *Proceedings Nat. Inst. Science India*, Vol. 12, pp. 49-55

MARCHAND B. (1974), *Etude des voies et des activités piétonnières autour de la gare de Saint-Maur-des-Fossé*, Institut de Recherche sur les Transports, Arcueil.

MATHER P.M. (1971), "Varimax and Generality », *Area*, 3, 252-254

MATHER P.M. (1971), "Varimax and Generality », *Area*, 4, 27-30

MATHERON G. (1965), *Les Variables Régionalisées et leur estimation*, Paris, Masson.

McCONNELL H, CHAPMAN K. P & KNOX J. C (1967), "A Multivariate Analysis of Topographic Slope in Selected Loess-Mant1ed Second Order Basin in Illinois", Paper presented at the 62nd annual meeting of *Annals, Association of American Geographers*, Toronto, August 1966, p.183

McNULTY L. L (1972), "Urban Structure and Development : The Urban System of Ghana", *Journal of Developing Areas*

McQUITTY L. L. (1957), "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies", *Educational and Psychological Measurement*, 17, 207-229

MEGEE M. (1965), "Economic Factors and Economic Regionalization in the United States", *Geografiska Annaler*, Vol. 47 B, pp. 125-137

MORRILL L.R and WOHLENBERG E. H (1.971), *The Geography of Poverty*, E. J Taaffe,

Series Editor, McGraw-Hill Book Company, New-York.

MOSER C.A and SCOTT W (1961), *British Towns : A Statistical Study of their Social and Economic Differences*, London : Oliver et Boyd Ltd

MURDIE R. A (1968), "The Factorial Ecology of Metropolitan Toronto, 1951-1961 : Essay in the Social Geography of the City", Chicago, University of Chicago, *Department of Geography Research Paper, n° 116*

NEUHANS J. O and WRIGLEY C. (1954), "The Quartimax Method : An Analytical Approach to Orthogonal Simple Structure", *British Journal of Statistical Psvchology,* 7, pp. 81-91

OLSEN B. M and GARB B (1965), An Application of Factor Analysis to Regional Economic Growth", *Journal of Regional Science*, 6, pp. 51-56

PARK R.E. (1916), "The City : Suggestions for the Investigations of Human Behavior in the Urban Environment", *The American Journal of Sociology*, XX, pp. 577-612

PARK R.E. (1936), " Human Ecology", *American Journal of Sociology*, XLII, n° 2, 1-15

PEDERSEN P.O (*1967),* "An Empirical Model of Urban Population Structure in Copenhagen", *Proceedings of the First Scandinavian-Polish Science Seminar* (Warsaw : Polish Scientific Publishers)

PEDERSEN P.O (1968), "Central Places and Functionnal Regions in Denmark, Factor Analysis of Telephonc Traffic", *Saertryk af Geografisk Tijdschrift*, 67 bind., pp. 1-18

PINCHEMEL P (1968), "Redécouvrir la Géographie", *Annales de l'Université de Paris*, n° 3, pp. 350-60

PINZKA C and SAUNDERS D.R (1954), *Analytic Rotation to Simple Structure II, Extension to an Oblique Solution*, Princeton, New-York , Educational Testing Service Research Bulletin

PRICE D.O. (1942), "Factor Analysis in the Study of Metropolitan Centers", *Social Forces*, 20, pp. 449-455

RAY D. M and BERRY B. J. L (1965), "Multivariate Socio-Economic Regionalization, A Pilot Study in Central Canada" in T. Rymes and S. Ostry, eds., *Regional Statistical Studies*, Toronto, University of Toronto Press, pp. 1-48

REES P. H (1970), "The Factorial Ecology of Metropolitan Chicago, 1960" in Berry B. J. L and Frank Horton, eds, *Geographic Perspectives on Urban Systems*, Englewood Cliffs, New-York, Prentice Hal, Inc.

ROBINSON A. H (1956), "The Necessity of Weighting Values in Correlation of Areal Data", *Annals, Association of American Geographers*, Vol. 46, pp. 233-236

ROBSON R.T. (1969), *Urban Analysis : A Study of City Structure*, London, Cambridge University Press

ROZEBOOM W. W (1966), *Foundations of the Theory of Prediction*, Dorsey Press, Hemewood, Illinois.

RUMMEL R.J. (1970), *Applied Factor Analysis*, New-York University Press.

SCHMID C. F (1960), "Urban Crime Areas", *American Sociological Review*, 25, pp 527-542

SCHUESSLER K.F and DRIVER H (1956), "A Factor Analysis of Primitive Societies", *American Sociological Review*, Vol. 21, pp. 493-499

SCHWIND P. J (1971), "Spatial References of Migrants for Regions : The Example of Maine", *Proceedings of the Association of American Geographers*, Vol. 3, pp. 150-156

SHEPARD R.N (1962) "The Analysis of Proximities : Multidimensional Scaling with an Unknown Distance Function, I and II", *Psychometrika*, 27, 125-39 ; 219-46

SHEVKY E. and BELL W. (1955), *Social Area Analysis*, Stanford : Stanford University Press

SHEVKY E and WILLIAMS M. (1949), *The Social Areas of Los Angeles : Analysis and Typology*, Berkeley and Los Angeles : University of California Press

SOJA E. W (1963), *The Geography of Modernisation in Kenya*, Syracuse Geographical Series, n° 2, Syracuse University Press

SPENCE N. A (1968) "A Multi-Factor Uniform Regionalization of British Counties on the Basis of Employment Data for 1961", *Regional Studies*, II, pp 87-104.

STRAHLER A. N (1954) "Statistical Analysis in Geomorphic Research", *Journal of Geology*, Vol.. 62, pp. 1-25

SWEETSER F. L (1965), "Factorial Eco1ogy : Helsinki 1960", Boston University, *Department of Sociology, Research Paper n° 2*

SWEETSER F.L (1965), "Factor Structure and Ecological Structure in Helsinki and Boston", *Acta Sociologica*, 8, pp. 205-225

THOMAS E. N. and ANDERSON D.L (1965), "Additional Comments on Weighting Values in Correlation Analysis of Areal Data", *Annals, Association of American Geographers*, Vol. 55, pp. 492-505

THOMPSON J. H, SUFRIN S.C., GOULD P. and BUCK M.A (1962), "Toward a Geography of Economic Health : The Case of New-York State", *Annals, Association of American Geographers*, Vol. 52, pp. 1-20

THURSTON L.L. (1954), "An Analytical Method for Simple Structure", *Psychometrika*, 19, pp. 173-182

TORGERSON W.S. (1952) "Multidimensional Scaling I : Theory and Method", *Psychometrika*, 17, pp 401-419

TORGERSON W.S. (1958), *Theory and Method of Scaling*, John Wiley, New-York

TUCKER L.R. (1963), "Implications of Factor Analysis of Three-Way Matrices for Measurement of Change", in *Problems in Measuring Change*, ed. C.W Harris, Madison : University of Wisconsin Press, pp. 122-137

VAN ARSDOL M., CAMILLERI F. and SCHMID C.F. (1958) "The Generality of Urban Social Area Indices", *American Sociological Review*, XXIII, n° 2, pp. 277-84

WEAVER J.C. (1954) "Crop-Combination Regions in the Middle-West", *Geographical Review*, Vol. 44, pp. 175-200

WRIGHT B. and EVITTS M.S. (1961), "Direct Factor Analysis in Sociometry", *Sociometry*, 24, pp. 82-98