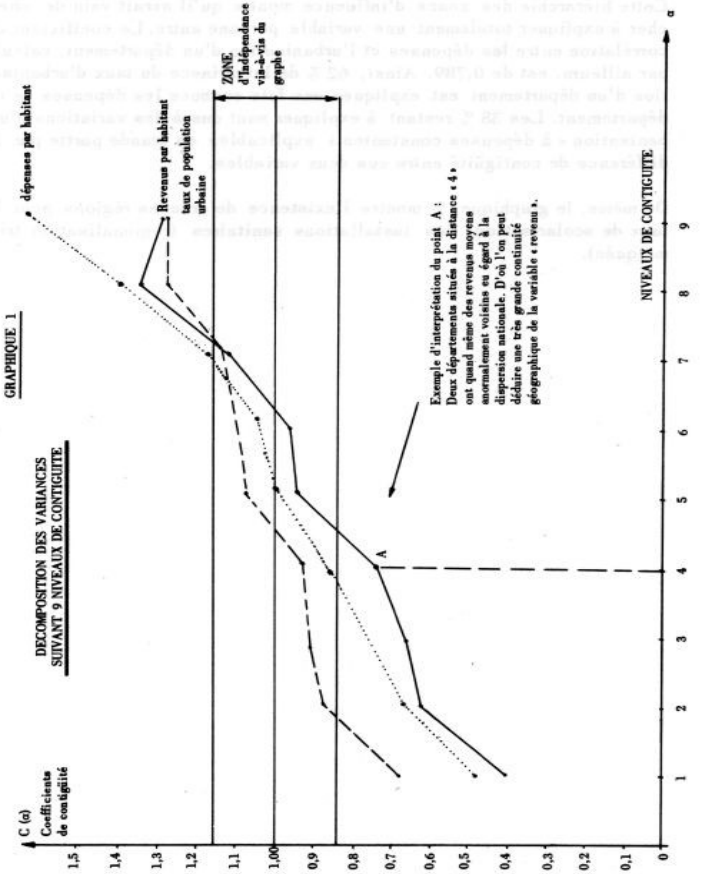
-

Analyse Statistique de la Contiguïté *(Statistical Analysis of Contiguity)* [Publications de l'ISUP, XVIII, P 81 - 112]

Seconde partie, pages 94 – 103 *(Second part, p 94 - 103)*

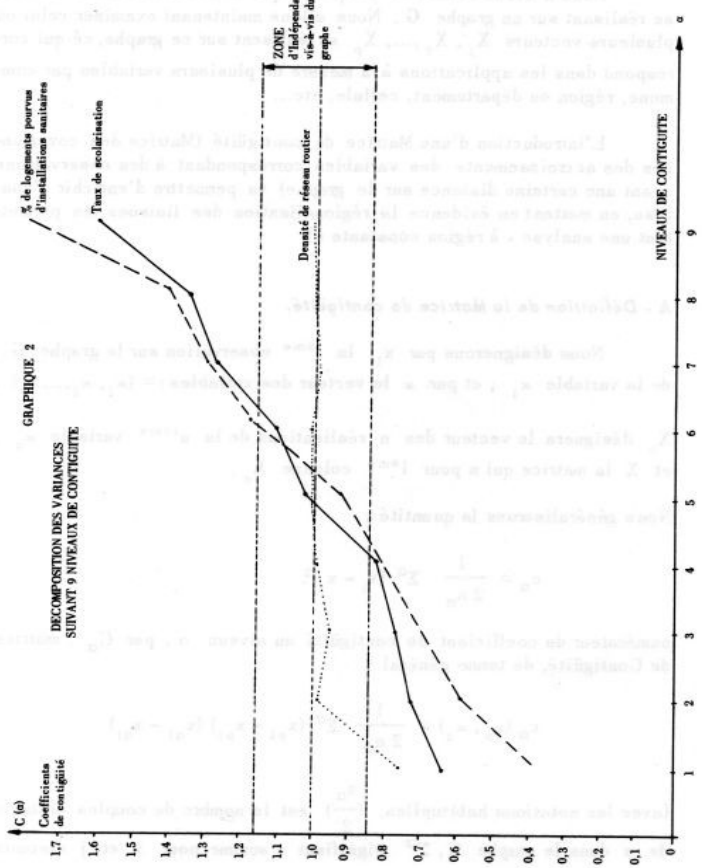
GRAPHIQUE 1

DECOMPOSITION DES VARIANCES SUIVANT 9 NIVEAUX DE CONTIGUITE



GRAPHIQUE 2

DECOMPOSITION DES VARIANCES SUIVANT 9 NIVEAUX DE CONTIGUITE



Deuxième Partie.

Analyse factorielle d'un ensemble d'observations géographiques.

Nous n'avons considéré jusqu'ici, que le cas d'un seul vecteur X se réalisant sur un graphe G. Nous allons maintenant examiner celui où plusieurs vecteurs $X_1, X_2, \ldots, X_p$ se réalisent sur ce graphe, ce qui correspond dans les applications à la mesure de plusieurs variables par commune, région ou département, cellule, etc...

L'introduction d'une Matrice de contigüité (Matrice des covariances des accroissements des variables correspondant à des observations ayant une certaine distance sur le graphe) va permettre d'enrichir l'analyse, en mettant en évidence la régionalisation des liaisons (en permettant une analyse « à région constante »).

A - Définition de la Matrice de contigüité.

Nous désignerons par x_{1i} la $i^{ème}$ observation sur le graphe G, de la variable α_1 ; et par α le vecteur des variables : $= (\alpha_1, \alpha_2, \ldots \alpha_p)$.

X_u désignera le vecteur des n réalisations de la $u^{ième}$ variable α_u, et X la matrice qui a pour $i^{ème}$ colonne X_u.

Nous généraliserons la quantité :

$$c_\alpha = \frac{1}{2\, n_\alpha} \Sigma^\alpha (x_i - x_j)^2$$

numérateur du coefficient de contigüité au niveau α, par C_α, matrice de Contigüité, de terme général :

$$c_\alpha (\alpha_p, \alpha_q) = \frac{1}{2\, n_\alpha} \Sigma^\alpha (x_{pi} - x_{pj})(x_{qi} - x_{qj})$$

(avec les notations habituelles, $(\frac{n_\alpha}{2})$ est le nombre de couples distants de α dans le graphe G, Σ^α signifiant « somme pour i et j distants

de α »). Si l'on introduit la matrice M_α associée au graphe des distances α, et $N_\alpha = \text{diag}\,(M_\alpha^2)$, on a :

$$c\ (\alpha_p, \alpha_q) = \frac{1}{n_\alpha} X'_p (N_\alpha - M_\alpha) X_q$$

La matrice de contigüité C_α s'écrit :

$$C_\alpha = \frac{1}{n_\alpha} X' (N_\alpha - M_\alpha) X$$

Notons que la matrice des covariances expérimentales du vecteur α s'écrit :

$$C_o = \frac{1}{n\,(n-1)} X' (nI - U) X$$

B - Principe de la Représentation.

Dans une analyse factorielle en facteurs communs et spécifiques, on représente les diverses variables par des points ayant pour coordonnées sur le $i^{ème}$ axe leur coefficient de corrélation avec le $i^{ème}$ facteur.

D'autre part, lorsque les différentes variables sont réparties au hasard sur le graphe, les matrices C_α définies en (A) sont toutes des estimations centrées de la matrice C, matrice des covariances du vecteur α.

La configuration des points représentant les variables ne devrait donc pas changer si l'on substitue à la matrice C_o la matrice C_1, qui est également un estimateur de C dans l'hypothèse d'indépendance vis-à-vis du graphe. Par contre, si cette condition d'indépendance n'est plus vérifiée, on doit assister à une déformation du nuage de points.

Cette déformation sera surtout une contraction vers l'origine, plus ou moins forte selon l'importance des zones d'influence des différentes variables (Contraction dûe au fait qu'un élément diagonal de C_1 sera inférieur à l'élément correspondant de C, puisque dans le cas où les observations

sont contigües, la quantité $\frac{1}{2n_1} \Sigma^1 (x_i - x_j)^2$ sous estime la variance).

On peut ainsi, à l'aide d'une figure dans le plan des deux premiers facteurs représenter simultanément les liaisons entre les différentes variables, leur degré de contigüité, et les liaisons entre les accroissements de ces variables à différents niveaux.

Pratiquement, on réalisera les analyses Factorielles de C et de C_1, et l'on cherchera à interpréter les éventuelles différentes de résultats, dues essentiellement à la répartition non aléatoire des variables sur le graphe.

C - Exemple d'application.

1) Sur le graphe à 88 sommets des départements français, (avec les conventions du chapitre I), nous nous proposons d'étudier la répartition simultanée des caractéristiques départementales suivantes : (pour la période 1951-1954)

1) Revenus par habitant (R)

2) Taux d'urbanisation (U)

3) Dépenses par habitant (D)

4) Taux de scolarisation (Sc.)

5) Pourcentage des logements munis d'installations sanitaires (I.S.)

6) Taux d'industrialisation (pourcentage de la population active occupée dans l'industrie) (Ind.)

7) Nombre de médecins par habitant (Méd.)

Les variables ont été préalablement réduites ; nous ferons l'analyse de la matrice des corrélations et de la matrice de contigüité au niveau 1 (ses éléments diagonaux sont donc les coefficients de contigüité des différentes variables).

Matrice de corrélation (C)

	R	U	D	Sc	I.S.	Ind.	Méd.
R	1,000						
U	0,625	1,000					
D	0,615	0,789	1,000				
Sc.	0,201	0,583	0,648	1,000			
IS.	0,455	0,652	0,794	0,474	1,000		
Ind.	0,816	0,663	0,544	0,157	0,472	1,000	
Méd.	0,246	0,574	0,746	0,666	0,661	0,212	1,000

Matrice de contigüité au niveau 1 (C_1)

	R	U	D	Sc.	I.S.	Ind.	Méd.
R	0,411						
U	0,399	0,698					
D	0,325	0,486	0,481				
Sc.	0,298	0,504	0,423	0,644			
IS.	0,217	0,359	0,326	0,291	0,370		
Ind.	0,267	0,353	0,240	0,220	0,170	0,391	
Méd.	0,208	0,394	0,393	0,397	0,331	0,167	0,725

Sur la figure 1, les points représentant les positions des variables dans le plan des deux premiers facteurs sont situés sur deux contours polygonaux : le contour polygonal en tirets contient les points résultat de l'analyse de C (Matrice de corrélation). Le contour polygonal en pointillé contient les points issus de l'analyse de C_1 (Matrice de contigüité).

Les flèches indiquent comment se déplacent chacune des variables lorsqu'on remplace C par C_1 dans l'analyse Factorielle.

Les facteurs ont été extraits par la méthode de Joreskög (réf. 7), qui donne une solution très proche de celle du maximum de vraisemblance de Lawley.

Figure (1)

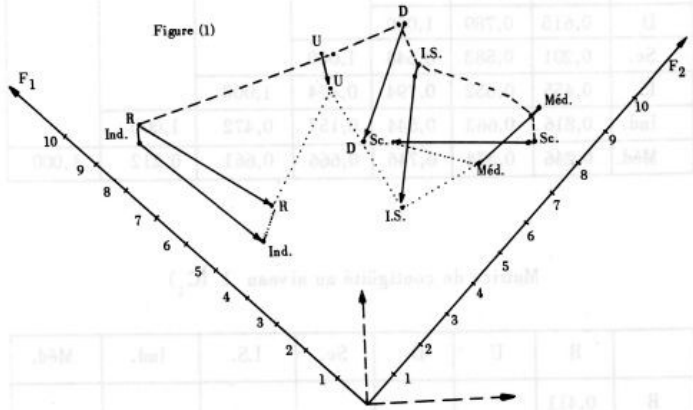


Fig. 1

Analyse de C, matrice de corrélation :

Le facteur F_1 semble varier comme l'industrialisation et la prospérité ; le facteur F_2 comme la scolarisation et « l'équipement en médecins » ; les dépenses, l'urbanisation, et l'équipement sanitaire des logements recevant la contribution des deux facteurs F_1 et F_2.

Nous ne tenterons pas ici d'explication (pour intéressante qu'elle soit) des particularités de la répartition géographique des différents facteurs.

Analyse de C_1, matrice de contigüité :

Le polygone « en pointillé » est bien une image contractée vers

l'origine du polygone « en tiret », à l'exception du point « Scolarisation », qui change de direction.

La longueur des flèches caractérise l'intensité de la contigüité de chacune des variables (en effet, la distance à l'origine d'une variable est voisine de sa variance ; le rapprochement de l'origine montre à quel point une mesure locale de la variance sous-estime celle-ci).

Ainsi, Revenu et Industrialisation sont très contigus, par contre, l'urbanisation l'est assez peu (un département fortement peuplé a plutôt tendance à dépeupler ses voisins...).

La déformation des positions respectives des différents points (ici déplacement de la variable « scolarisation » dans une direction non radiale) indique une différence entre les covariations locales, puisque la Matrice C_1 n'est autre qu'une matrice des covariances des accroissements locaux des variables.

Ainsi, le degré de scolarisation d'une région est, en moyenne sur l'ensemble du territoire, indépendant de la richesse de cette région.

Localement, cependant, cette variable est beaucoup plus liée au niveau de vie (variables : dépense, urbanisation, installations sanitaires). Cette particularité est confirmée par des études monographiques du sujet (Réf. 3).

Troisième Partie.

Dépendance des observations et information.

A - Rectification des tests relatifs aux coefficients de Régression.

Le modèle de régression linéaire multiple qui permet d'étudier la dépendance d'une variable endogène X, vis-à-vis, de variables exogènes, aléatoires ou non, $Z_1, Z_2, Z_3 \ldots$, suppose une spécification de la matrice des covariances théoriques des résidus en une matrice du type : $\sigma^2 I$, (ce qui suppose en particulier l'indépendance de ces résidus).

Lorsque ces variables se réalisent sur un graphe, et que les positions respectives des observations sont connues, il est difficile de négliger les liaisons qui peuvent exister entre les résidus correspondant à des observations voisines...

Il importe donc de donner une spécification plus plausible de la matrice des covariances des résidus, ce qui modifiera notamment notre information sur les coefficients de régression, et leur éventuelle signification.

a) Rappel du cas où les résidus ont une matrice des covariances du type : $\sigma^2 I$,

Soit X le vecteur des différentes observations de la variable endogène x.

Z la matrice ayant pour $i^{ème}$ colonne les observations de la $i^{ème}$ variable exogène Z_i, $i = 1$ à p.

Le modèle de régression multiple s'écrit :

$$X = Z\beta + \beta_o + \varepsilon$$

Si l'on suppose que les différentes variables sont centrées :

$$X = Z\beta + \varepsilon$$

Les équations normales s'écrivent :

$$Z'X = Z'Z\hat{\beta}$$

Par suite $\hat{\beta} = (Z'Z)^{-1} Z'X$

$\hat{\beta}$ étant une fonction linéaire de X suit une loi normale de matrice des covariances :

$$V(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta') = (Z'Z)^{-1} Z'(V(X))\, Z(Z'Z)^{-1}$$

Comme $v(X) = V(\varepsilon) = \sigma^2 I$, on retrouve le résultat habituellement utilisé.

$$V(\hat{\beta}) = \sigma^2 (Z'Z)^{-1}$$

La quantité $\frac{1}{\sigma^2} (\hat{\beta} - \beta)' (Z'Z) (\hat{\beta} - \beta)$ est donc un χ^2 à p d° de liberté.

D'autre part, la quantité $\frac{\varepsilon'\varepsilon}{\sigma^2}$ suit une loi du χ^2 à N - p - 1 d° de liberté.

Le quotient de ces χ^2, qui sont indépendants, et qui ne font pas intervenir de quantités inconnues, est donc un F de Fischer avec p et n-p-1 d° L.

L'inégalité : $\frac{(\hat{\beta} - \beta)' (Z'Z) (\hat{\beta} - \beta)}{\varepsilon'\varepsilon} < F_\alpha$

donne, dans l'espace à p dimensions un ellipsoïde de confiance au seuil α pour les coefficients β.

b) Cas où les résidus ont une matrice des covariances V quelconque.

L'estimation à dispersion minimale de β s'obtient maintenant en minimisant la quantité :

$$(X - Z\beta)' V^{-1} (X - Z\beta)$$

En dérivant, on est conduit à l'équation matricielle :

$$Z' V^{-1} = (Z' V^{-1} Z) \hat{\beta}$$

D'où $\hat{\beta} = (Z' V^{-1} Z)^{-1} Z' V^{-1} X$ et $V(\hat{\beta}) = (Z' V^{-1} Z)^{-1}$

Fin de la seconde partie