Analyse Statistique de la Contiguïté

(Statistical Analysis of Contiguity)

Publications de l'Institut de Statistique des Universités de Paris

XVIII, p 81-112.

Première partie, pages 81 – 93

(First part, p 81 - 93)

Publ. Inst. Stat. Univ. Paris
1969. XVIII – 81 - 112

5. ANALYSE STATISTIQUE DE LA CONTIGUITE

L. LEBART

Introduction.

Les statisticiens ont souvent affaire à des ensembles de mesures ou d'observations qui ne peuvent être considérées comme des réalisations indépendantes de variables ou de vecteurs aléatoires. En Economie, dans les Sciences Humaines ou Biologiques, les répétitions d'épreuves identiques sont extrêmement rares.

Cependant, le champ des observations suggère souvent la forme des liaisons entre observations. C'est le cas des séries chronologiques et des modèles à erreurs liés dans le temps des économètres. Nous étudierons ici le cas plus général où les observations se réalisent sur un graphe (i.e. le cas où un ensemble de couples d'observations est privilégié : couples d'observations successives pour les séries temporelles, couples d'observations contigües pour les variables régionales ou départementales, couples d'observations appartenant à une même strate, etc...).

Nous verrons comment éprouver la validité de l'hypothèse de dépendance vis-à-vis de la structure de graphe, comment décrire cette dépendance (1ère partie), puis nous verrons quelles en sont les conséquences : enrichissement des analyses factorielles par la mise en évidence de l'échelle et de la localisation des liaisons (2ème partie), rectification des tests habituels sur les coefficients de régression fondés sur l'indépendance des observations (3ème partie, A), perte d'Information et définition d'une pseudo-taille de l'échantillon (la taille effective étant source d'illusion lorsque les observations sont liées) - (3ème partie, B).

Des exemples d'application à des variables socio-économiques (graphe des départements français) illustrent chaque étape.

Première Partie.

Mesure et test du degré de contigüité d'une variable sur un graphe.

A propos d'observations économiques spatiales, R.C.Geary (Ref.4) a proposé un coefficient de contigüité tout à fait analogue au rapport de Von Neumann (ref. 12).

Nous allons rappeler les principaux résultats de Geary, puis, en utilisant le formalisme de la théorie des graphes, introduire la notion de niveau de contigüité, et préciser la loi des coefficients utilisés. Un exemple d'application est donné en fin de chapitre.

I - Le coefficient de contigüité de Geary.

Considérons un pays divisé en n départements ou régions, le département i ayant une frontière commune avec k_i autres départements.

Appelons z_i la mesure d'un caractère du département i . Le coefficient de contigüité c de Geary est donné par la formule :

$$c = \frac{n-1}{2\,n_1} \cdot \frac{\Sigma^1 (z_i - z_j)^2}{\Sigma^i (z_i - z)^2}$$

avec $n_1 = \Sigma_{i=1}^{n} k_i$ = nombre total de « connexions ».

Σ^1 = somme pour i et j tels que les départements i et j soient contigus,

Σ^i = somme pour tout i .

Le coefficient c apparaît donc comme étant le rapport de deux estimations de la variance de la variable Z , dont l'une, le numérateur : $\frac{\Sigma^1 (z_i - z_j)^2}{2\,n_1}$ tient compte des positions respectives des différentes régions.

Deux points de vue, qui donnent des résultats assez voisins, peuvent être adoptés pour expliquer la dispersion du coefficient c .

On peut imaginer que les n valeurs de la variable Z sont des valeurs numériques définies à l'avance, et que l'épreuve qui préside à la réalisation de la variable aléatoire « c » est le choix de l'une des n ! permutations de ces valeurs ; ce qui revient à tirer dans une urne, de façon exhaustive, la valeur zi qui sera affectée au département i .

On peut également considérer les n valeurs de la variable Z comme n réalisations indépendantes d'une variable aléatoire normale.

Dans les deux cas cas, l'espérance mathématique de la v.a. « c » est :

$$E(c) = 1$$

Dans le cas où les n valeurs de Z sont considérées comme des réalisations de v.a. normales, indépendantes, Geary a calculé les quatre premiers moments de c , et montré pour un exemple que la distribution de c est très voisine d'une distribution normale, donc que l'écart-type $\sigma(c)$ de c constitue un critère adéquat de significativité.

Dans ces conditions, une réalisation de l'échantillon donne une valeur de c qui a 95 chances sur 100 d'être comprise entre les valeurs : $1 - 2\,\sigma(c)$ et $1 + 2\,\sigma(c)$.

II - Généralisation au cas d'un graphe symétrique quelconque.

Après avoir défini les niveaux de contigüité, nous donnerons pour les moments de « c » des formules présentées différemment de celles données par Geary : l'utilisation des matrices associées au graphe constitué par l'ensemble des régions permettra le calcul automatique des moments d'ordre supérieur à 4, mais surtout offrira la possibilité de travailler sur des graphes beaucoup plus importants que ceux étudiés par Geary (26 comtés de l'Irlande) sans nécessiter de dénombrements fastidieux.

Nous commencerons par rappeler quelques définitions relatives aux graphes :

A - Matrice associée à un graphe non orienté

- un graphe non orienté (ou symétrique) est constitué par un ensemble de

sommets (les départements dans les exemples qui vont suivre), certains de ces sommets étant reliés entre eux par une arête (certains de ces départements étant contigüs). Nous supposerons que deux sommets sont reliés par, au plus, une arête (deux départements ne se touchent qu'une fois). Deux sommets reliés par une arête sont dits adjacents. On appelle degré k_i *d'un sommet* i *le nombre de sommets qui lui sont adjacents. Un graphe dont tous les sommets ont même degré* k *, est dit homogène de degré* k .

Les n sommets du graphe G seront numérotés de 1 à n . Nous dirons qu'un échantillon $(x_1, x_2, \ldots, x_i \ldots x_n)$ se réalise sur le graphe G si l'observation x_i est un caractère attaché au sommet i .

– Considérons un graphe G à n sommets et $\frac{n}{2}$ arêtes.

On appelle matrice associée au graphe G *, la matrice carrée a* n *lignes et* n *colonnes «* M *» de terme générique* (m_{ij}) *tel que :* $m_{ij} = 1$ *, si les sommets* i *et* j *sont adjacents,* $m_{ij} = 0$ *si les sommets* i *et* j *ne sont pas adjacents.* La relation d'adjacence étant symétrique, on a : $m_{ij} = m_{ji}$. La matrice M contient donc n_1 éléments différents de 0 . La diagonale de M contient des éléments non nuls que si certains sommets sont adjacents à eux-mêmes. Dans le cas des départements, nous adopterons, sauf spécification contraire, la convention $m_{ii} = 0$ pour tout i.

B - Chemin sur un graphe non orienté

On appelle chemin sur un graphe non orienté toute succession d'arêtes : $(u_1, u_2, \ldots, u_{k-1}, u_k, \ldots, u_p)$ *telle que pour tout* $k \leq p$, u_{k-1} *et* u_k *touchent un même sommet. La longueur du chemin est le nombre des arêtes qui le composent.*

PROPOSITION 1 : *Soient deux graphes non orientés* G *et* G_1 *, ayant les mêmes sommets, de matrices associées* M *et* M_1 *. Soit* a_{ij} *l'élément générique de la matrice produit* $A = MM_1$ *. Le nombre entier* a_{ij} *est le nombre de chemins qui relient le sommet* i *au sommet* j *, formés chacun d'une arête de* G *suivie d'une arête de* G_1 *. En effet :* $a_{ij} = \sum_k m_{ik} m_{1kj}$.

Dire que pour une valeur de k, $m_{ik} m_{1kj} = 1$, cela veut dire que $m_{ik} = 1$ et $m_{1kj} = 1$ donc le sommet k est adjacent à i dans G , et adjacent à j dans G_1 . Il existe donc bien un chemin formé d'une arête de G , puis d'une arête de G_1 » reliant i à j . Le nombre a_{ij} est donc égal au nombre de ce chemin.

PROPOSITION 2 : *Soit un graphe* n-O.G *, de matrice associée* M *. Soit* a_{ij} *, le terme générique de la matrice* $A = M^\alpha$ *: il y a sur le graphe* G, a_{ij} *chemins de longueur* α *joignant le sommet* i *au sommet* j .

Il suffit d'appliquer $(\alpha - 1)$ *fois la proposition 1 en prenant successivement pour matrice* M_1 *les matrices* $M, M^2 \ldots M^{\alpha-1}$.

Si au lieu de faire un produit matriciel ordinaire, on fait un produit suivant l'arithmétique booléenne, ce qui revient à remplacer toute valeur > 0 par la valeur 1 , dans les multiplications de matrices, on a le résultat suivant : la matrice $M^{(\alpha)}$ (M multiplié selon les règles booléennes α fois par elle-même) est la matrice associée au graphe G où deux sommets sont reliés par une arête, si, dans le graphe G , il existe au moins un chemin de longueur α reliant ces deux sommets.

C - Distance de deux sommets

On appelle distance de deux sommets i *et* j *d'un graphe* G n.o *, la longueur* d (i, j) *du plus court chemin joignant* i *à* j .

Cette distance possède évidemment les trois propriétés suivantes :

$$d(i, j) = 0 \iff i = j$$

$$d(i, j) = d(j, i)$$

$$d(i, j) \leq d(i, k) + d(k, j)$$

Soit M la matrice associée à G, I la matrice unité, posons $M_1 = M + I$

(M_1 est la matrice associée au graphe G si l'on suppose que chaque sommet est adjacent à lui-même).

PROPOSITION 3 : *La matrice* $M_1^{(\alpha)} - M_1^{(\alpha-1)}$ *où* $M_1^{(u)}$ *désigne la* $u^{ième}$ *puissance booléenne de* M_1, *est la matrice associée au graphe* G *tel que :* i *et* j *sont adjacents dans* G *si, et seulement si* $d(i,j) = \alpha$ *dans* G.

En effet, on voit facilement que dans le graphe associé à $M^{(u)}$ sont adjacents tous les couples i et j tels que $d(i,j) \leqslant u$ dans G.

Faire la différence $M_1^{(\alpha)} - M_1^{(\alpha-1)}$ revient à ne laisser que les arêtes correspondant à des couples de sommets tels que $d(i,j) = \alpha$.

D - Notations

Nous noterons dans la suite M_α la matrice $M_1^{(\alpha)} - M_1^{(\alpha-1)}$ et n_α le nombre d'éléments de la matrice M (deux fois le nombre de couples distants de α dans G). Nous noterons U la matrice de terme générique (u_{ij}) avec $u_{ij} = 1$ pour tout i et j.

Nous noterons également $k_{\alpha i}$, (ou k_i si la valeur de α est fixée et connue) la somme des éléments de la $i^{ème}$ ligne de M, donc le nombre de sommets situés à la distance α du sommet i (degré du sommet i).

Nous désignerons par diag (A) la matrice diagonale qui a les mêmes éléments diagonaux que la matrice A.

Nous désignerons également par N_α la matrice diagonale qui a pour $i^{ème}$ élément diagonal $k_{\alpha i}$ (on peut vérifier que diag $(M_\alpha^2 = \text{diag}\,(M_\alpha U) = N_\alpha)$.

E - Niveaux de contigüité

Le coefficient de contigüité a été défini comme le rapport :

$$c = \frac{\Sigma^1 (z_i - z_j)^2}{2 n_1} \qquad \frac{(n-1)}{\Sigma (z_i - z)^2} = \frac{\Sigma^1 (z_i - z_j)^2}{2 n_1 s^2}$$

où s^2 désigne la variance expérimentale de la variable Z. Le numérateur s'écrit :

$$\Sigma^1 (z_i - z_j)^2 = 2 \Sigma_i k_i z_i^2 - 2 \Sigma^1 z_i z_j$$

Appelons Z le vecteur des z_i et utilisons la matrice diagonale N_1, définie au paragraphe précédent.

$$\Sigma^1 (z_i - z_j)^2 = 2 (Z' N_1 Z - Z' M_1 Z)$$
$$= 2 . Z' (N_1 - M_1) Z$$

où M_1 est la matrice associée au graphe G.

Nous définirons *le coefficient de contigüité au niveau* α comme le rapport :

$$c_\alpha = \frac{\Sigma^\alpha (z_i - z_j)^2}{2 n_\alpha s^2}$$

où Σ^α désigne une somme pour i et j distants de α.

(et comme précédemment : $\frac{n_\alpha}{2}$ nombre de couples distants de α).

Ce coefficient est tout à fait analogue au coefficient c, mais concerne le graphe des couples d'éléments distants de α (associé à la matrice M_α). Avec les notations précédentes, on peut écrire :

$$c = \frac{Z' (N_\alpha - M_\alpha) Z}{n_\alpha s^2}$$

Alors que le premier coefficient c permet de tester l'hypothèse d'une influence significative entre départements voisins, les coefficients $c_2, c_3, \ldots, c_p$ vont permettre de tester jusqu'où cette influence est significative.

III - Loi des coefficients « c_α ».

Le coefficient « c_α » est le rapport de la forme quadratique :

$$\frac{Z'(N_\alpha - M_\alpha)Z}{2n_\alpha} = \frac{Q_1}{2n_\alpha}$$

à la variance :

$$\frac{1}{(n-1)} Z'(I - \frac{U}{n})Z = \frac{1}{n-1} Q_2$$

Geary a montré, dès 1933, que les moments du rapport d'une forme quadratique à la variance sont les quotients des moments du numérateur par ceux du dénominateur. Ce résultat généralisé depuis (Cf. 10,11) est le seul que nous utiliserons.

Le principe de la démonstration est le suivant :

La loi de l'échantillon s'écrit, les variables étant supposées centrées et réduites :

$$P(Z) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} Z'Z\right\} dZ$$

On fait une transformation orthogonale qui diagonalise Q_2 :

On sait que dans ces conditions : $Q_2 = \sum_{i=1}^{n} (z_i - z)^2 = \sum_{1}^{n-1} u_i^2$

La loi des u_i étant $\frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} U'U\right\} dU$

Par un changement de variable tel que $Q_2 = \sum_{i=1}^{n-1} u_i^2 = r^2$

c'est-à-dire : $U_1 = r \sin \varphi_1$

$U_2 = r \cos \varphi_1 \sin \varphi_2$

$U_{n-1} = r \cos \varphi_1 \dots \cos \varphi_{n-2}$

La loi de l'échantillon s'écrit :

$$\frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} r^2\right\} r^{n-2} A(\varphi_1 \dots \varphi_{n-2}) \, dr \, d\varphi_1 \dots d\varphi_{n-2}$$

avec $Q_1 = \Sigma^{\alpha} (Z_i - Z_j)^2 = r^2 B(\varphi_1 \dots \varphi_{n-2})$

La quantité $\frac{Q_1}{Q_2}$ ne dépend que de $\varphi_1 \dots \varphi_{n-2}$

$$E(Q_1^{\alpha'}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} r^2\right\} r^{(n+2\alpha'-2)} A(\Phi) . B^{\alpha'}(\Phi) \, dr \, d\Phi$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} r^2\right\} r^{(n+2\alpha'-2)} dr \, A(\Phi) . B^{\alpha'}(\Phi) . d(\Phi)$$

$$= E(Q_2^{\alpha'}) . E\left[\left(\frac{Q_1}{Q_2}\right)^{\alpha'}\right]$$

Or, Q_2 suit une loi du χ^2 à (n-1) degrés de liberté ; ses moments sont donc connus.

La forme Q_1 suit la loi de $\sum_{i=1}^{p} q_i u_i^2$ où q_i désigne la $i^{ème}$ valeur propre de Q_1 et u_i une variable aléatoire normale indépendante de u_j ou $i \neq j$.

La deuxième fonction caractéristique de Q_1 est :

$$\Psi(t) = -\frac{1}{2} \sum_{i=1}^{\infty} \text{Log}(1 - 2q_i\theta) \quad \text{avec} \quad \theta = it$$

Par suite, les différents cumulants de la forme Q_1 s'écrivent :

$$K_s = 2^{s-1}\ (s-1)!\ (\Sigma\ q_i^s)$$
$$= 2^{s-1}\ (s-1)!\ \ tr\ (N_\alpha - M_\alpha)^s$$

Posons $N_\alpha - M_\alpha = A$, $I - \frac{U}{N} = V$

Ainsi $K_1 = tr\ A$

$K_2 = 2\ tr\ A^2$

$K_3 = 8\ tr\ A^3$

$K_4 = 48\ tr\ A^4$

Grâce aux relations :

$m_2 = K_2 + K_1^2$

$m_3 = K_3 + 3\ K_2\ K_1 + K_1^3$

$m_4 = K_4 + 4\ K_3\ K_1 + 3\ K_2^2 + 6\ K_2\ K_1^2 + K_1^4$

On obtient les différents moments de c

$E\ (c_\alpha) = 1$

$$E\ (c_\alpha^2) = \frac{(n-1)^2}{(tr\ A)^2}\ \ \frac{2\ tr\ A^2 + (tr\ A)^2}{2\ tr\ V^2 + (tr\ V)^2}$$

Remarquons que $tr\ V^u = tr\ (I - \frac{U}{n})^u = tr\ V = (n-1)$

nous écrirons néanmoins les formules avec V pour conserver leur symétrie :

$$E\ (c_\alpha^3) = \frac{(n-1)^3}{(tr\ A)^3}\ \ \frac{8\ tr\ A^3 + 6\ tr\ A^2\ tr\ A + (tr\ A)^3}{8\ tr\ V^3 + 6\ tr\ V^2\ tr\ V + (tr\ V)^3}$$

$$E\ (c^4) = \frac{(n-1)^4}{(tr\ A)}\ \ \frac{48\ tr\ A^4 + 32\ tr\ A^3\ tr\ A + 12\ (tr\ A^2)^2 + 12(tr\ A^2)\ (tr A)^2 + (tr A)^4}{48\ tr\ V^4 + 32\ tr\ V^3\ tr\ V + 12(tr\ V^2)^2 + 12(tr\ V^2)\ (tr\ V)^2 + (tr V)^4}$$

Le calcul des valeurs numériques des différents moments lors des applications montre que la loi de C est en général très voisine d'une loi normale.

Ainsi pour le coefficient de contigüité au niveau 1 du graphe de l'exemple ci-dessous, les caractéristiques sont les suivantes : A est une matrice (88×88)

$tr\ A = 422$

$tr\ A^2 = 2664$

$tr\ A^3 = 18638$

$tr\ A^4 = 138108$

Par suite $E\ (c_1) = 1$

$\mu\ 2 = 0{,}006511226$ $(\sigma = 0{,}0807)$

$\mu\ 3 = 0{,}2549\ .\ 10^{-4}$

$\mu\ 4 = 0{,}12637\ .\ 10^{-3}$

Le coefficient d'asymétrie de K. Pearson est :

$$\beta_1 = \mu_3^2 / \mu_2^3 = 2{,}3\ .\ 10^{-3}$$

Le coefficient d'aplatissement :

$$\beta_2 = \mu_4 / \mu_2^2 = 2{,}98$$

La loi de c_1 est donc extrêmement proche d'une loi normale, puisque $\beta_1 \simeq 0$ et $\beta_2 \simeq 3$.

IV - Exemple d'application.

Sur le graphe à 88 sommets formé par les départements français, (les départements Seine, Seine et Oise, d'une part, et Territoire de Belfort,

Haute Saône, d'autre part, ayant été groupés pour des raisons d'homogénéité géographique) ont été calculés les coefficients de contigüité de 6 variables sur 9 niveaux de contigüité.

Nous obtenons donc 6 graphiques de 9 points chacun, qui caractérisent les niveaux auxquels se disperse la variable étudiée, et qui précisent l'échelle géographique des phénomènes économiques.

Le calcul de ces 54 coefficients, à partir des diverses puissances booléennes de matrices (88 × 88) a été fait sur C.D.C. 3600.

Ces 6 variables sont : (année 1954)

1 – Revenu départemental par habitant ;

2 – Taux de population urbaine ;

3 – Dépenses par habitant ;

4 – Taux de scolarisation de l'enseignement secondaire ;

5 – Densité de réseau routier ;

6 – Pourcentage des logements munis d'installation sanitaire.

– Lecture des graphiques :

La zone rectangulaire comprise entre les ordonnées 0,84 et 1,16 représente approximativement la zone qui devrait contenir les graphiques en cas d'indépendance à tous les niveaux (voir graphiques).

L'écart-type des différents c_α est en effet voisin de 0,08.

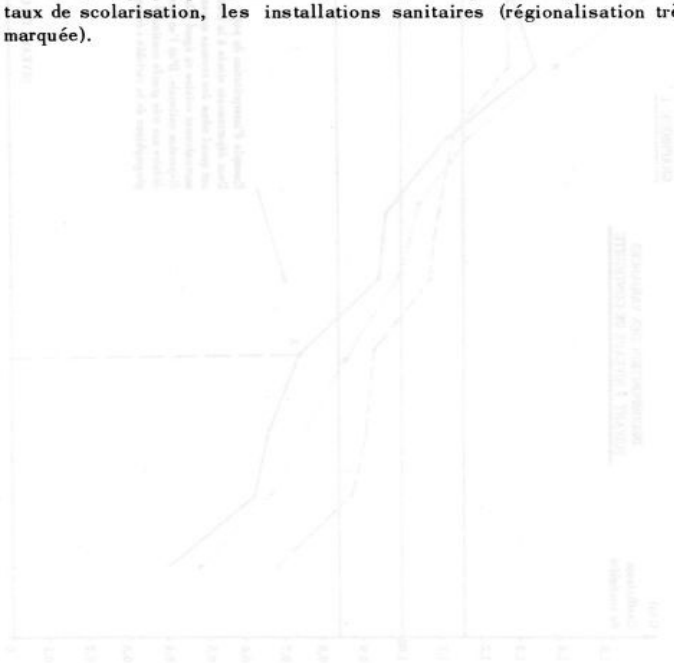
Un point d'abscisse α situé dans la zone inférieure indique une influence significative entre les valeurs des variables distantes de α ; il indique une répulsion s'il est situé dans la zone supérieure.

Toutes les variables considérées ont un caractère de contigüité très marqué, mis à part la densité de réseau routier dont l'influence se limite au département immédiatement voisin.

Les revenus ont une zone d'influence s'étendant sur un rayon de trois départements, autour d'un département donné. La contigüité des dépenses est constamment moins marquée ; les dépenses sont donc en moyenne plus dispersées géographiquement que les revenus. Le taux de population

urbaine a une zone d'influence plus faible encore. La valeur en un point ne dépend que faiblement de celles des départements limitrophes. Ainsi, un groupe de départements assez important et homogène du point de vue des revenus peut contenir des régions dont les dépenses sont déjà plus diversifiées, elles-mêmes contenant des zones inégalement urbanisées. Cette hiérarchie des zones d'influence montre qu'il serait vain de chercher à expliquer totalement une variable par une autre. Le coefficient de corrélation entre les dépenses et l'urbanisation d'un département, calculé par ailleurs, est de 0,789. Ainsi, 62 % de la variance du taux d'urbanisation d'un département est expliquée une fois connues les dépenses de ce département. Les 38 % restant à expliquer sont dus à des variations d'urbanisation « à dépenses constantes » explicables en grande partie par la différence de contigüité entre ces deux variables.

De même, le graphique 2 montre l'existence de grandes régions pour le taux de scolarisation, les installations sanitaires (régionalisation très marquée).



Fin de la première partie