ECOLE DES MINES DE PARIS
CENTRE DE GÉOSTATISTIQUE

BIBLIOTHÈQUE

Fontainebleau CG N–314

Une méthode de filtrage des données

J.–P. ORFEUIL

Janvier 1973

35, rue Saint–Honoré, 77305 FONTAINEBLEAU CEDEX – Téléphone 01 64 69 47 04 – Télécopieur 01 64 69 47 05

Nous résumons ici une méthode de filtrage de données que nous avons utilisée pour notre stage de D.E.A. (Statistique Mathématique, Pr. J.P. BENZECRI).

1 - Les données

Sur une zone de 40 × 40 degrés carrés (couvrant environ l'Europe Occidentale) on connaît :

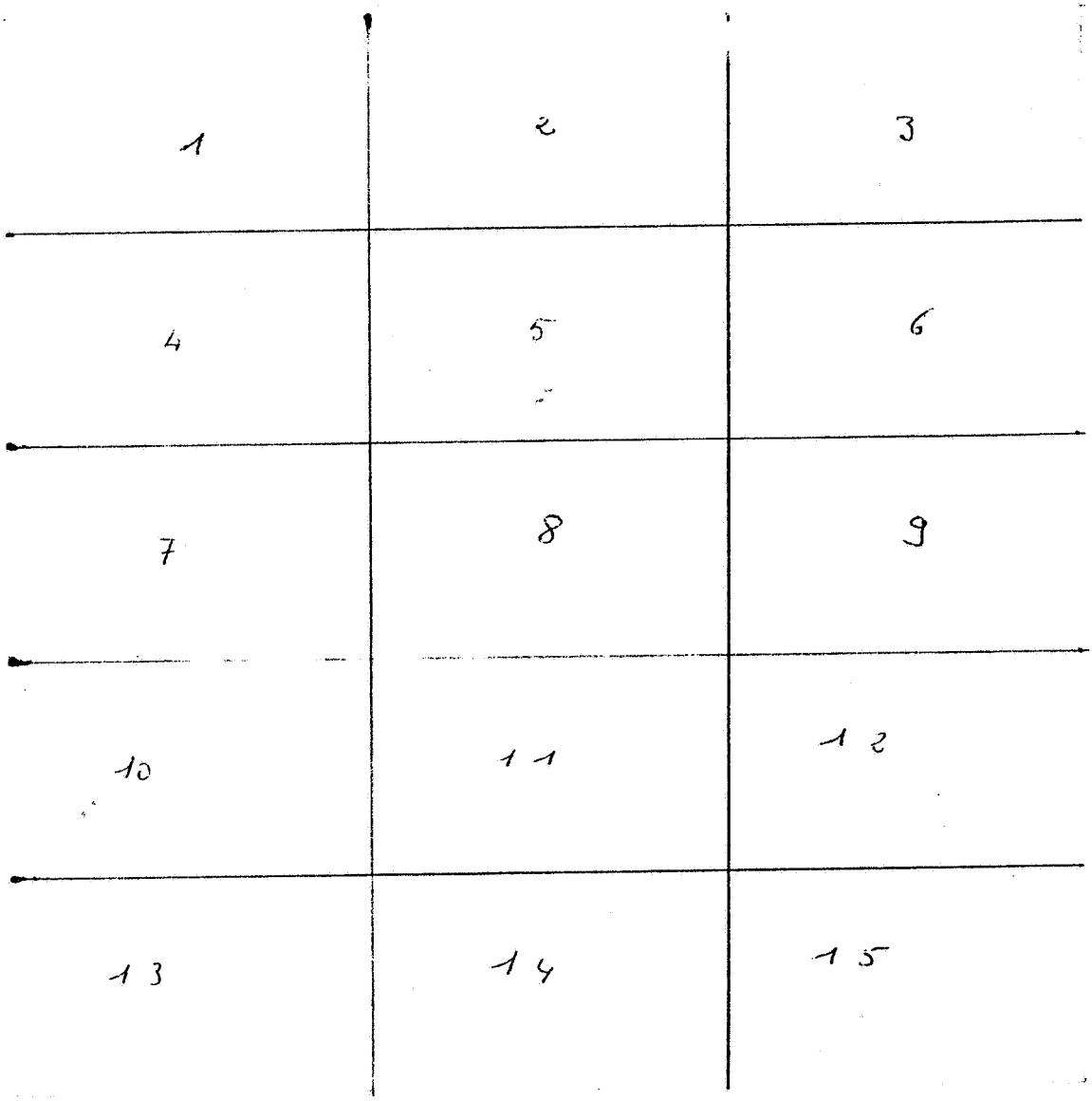
. 15 valeurs de l'Anomalie à l'air libre et 15 valeurs de l'Altitude, prises en 15 points répartis toujours de la même façon sur tous les degrés carrés (cf. figure 1)

2 - Le problème

On souhaite définir un coefficient de corrélation entre l'Anomalie et l'Altitude, et ce pour un certain nombre de sous-zones - nombre qui n'est pas précisé. En d'autres termes, on souhaite définir quelques zones (entre 10 et 20) telles que, à l'intérieur de chaque zone, la corrélation soit constante.

Le problème majeur est le suivant : les données sont entachées d'erreurs qu'on ne connaît pas mais qu'on sait importantes, mais qui n'ont aucune raison d'être indépendantes d'un degré carré à l'autre, ni aucune raison d'avoir la même variance. En effet dans les zones de plaine on a beaucoup d'informations et les grandeurs sont bien connues. Dans les zones montagneuses et dans les zones maritimes, on a peu de données, et les chiffres qui figurent dans le fichier résultent d'un certain nombre de manipulations dont on connaît mal la nature. Il arrive même que des raisons "politiques" viennent se greffer aux facteurs naturels, car certains pays répugnent à publier leurs données.

Un coefficient de corrélation local, calculé sur les données brutes, n'aurait aucun sens. On sait de plus que, du point de vue théorique, et dans des cas particulièrement favorables (variables gaussiennes), l'évaluation de la corrélation est délicate (cf. [1], page 209). Nous laisserons ce point de côté.

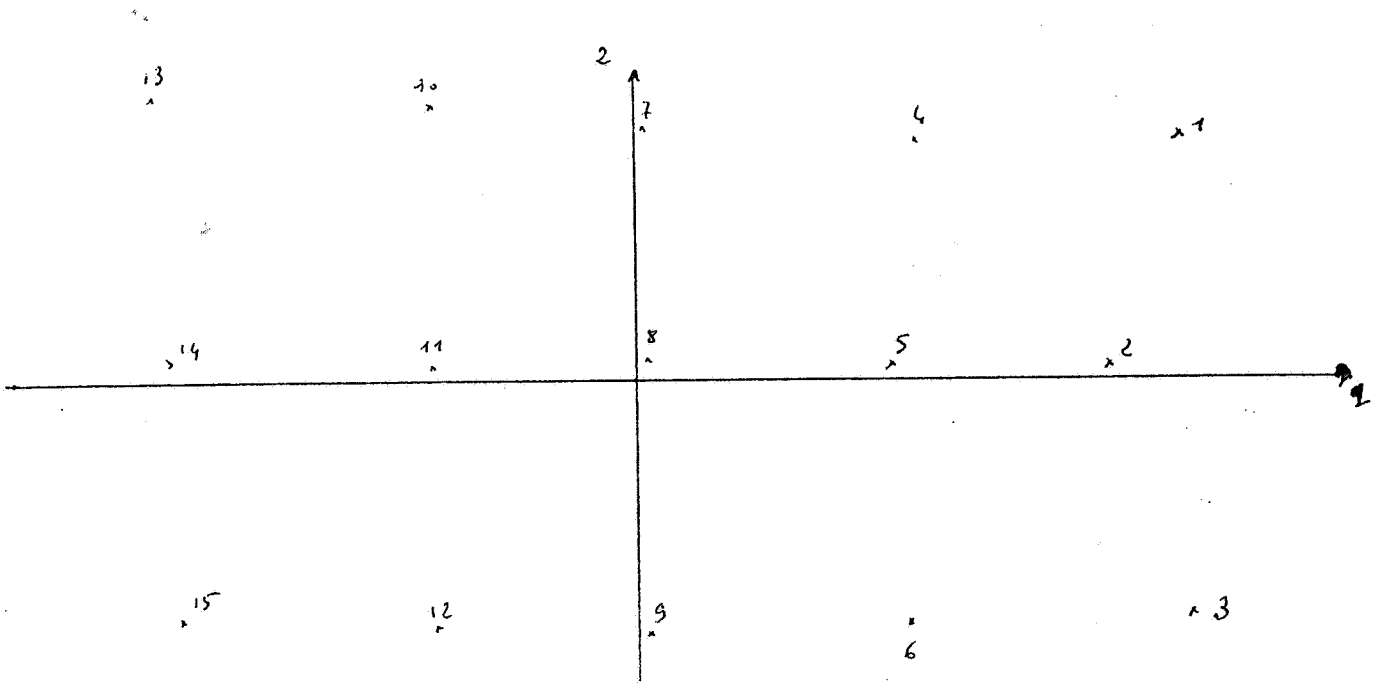

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

3 - Plan pour une solution du problème

On considère les 1600 degrés carrés comme autant d'individus caractérisés par un profil de 15 mesures (soit d'Anomalie, soit d'Altitude) : on a donc 1.600 points dans R^{15}, ou de même 15 points dans R^{1600}. L'Analyse des Correspondances (cf.[2]) permet de bâtir une suite de filtres de moins en moins fins : on assimile tour à tour le nuage de points à sa représentation sur une droite, sur un plan, dans R^3, R^4, etc...

Il faut choisir la taille du filtre. Pour cela un critère classique est la part d'inertie expliquée par une représentation : si le plan des deux premiers axes restitue 90 % de l'inertie, on se limite généralement à lui.

Nous disposons en fait d'un critère beaucoup mieux adapté au problème : on peut en effet représenter les 15 points de "prélèvement" sur le premier facteur, sur les deux premiers axes. etc ... Nous considérons que le filtre d'ordre i est le filtre à choisir (on assimile le nuage à sa représentation dans l'espace des i premiers axes) lorsque cet espace restitue les proximités géométriques des implantations. Il se trouve que le plan des deux premiers axes (qui restitue environ 70 % de l'inertie) redonne ces proximités. La figure y est en effet la suivante :



Le troisième axe oppose les points moyens du degré carré et ses points extrêmes : c'est un facteur de liaison entre deux degrés carrés à frontière commune : les points Nord d'un degré carré sont proches des points Sud du degré carré situé immédiatement au-dessus de lui.

Les axes suivants ne paraissent pas interprétables.

Nous pensons qu'on peut assimiler le nuage à ses représentations bi ou tridimensionnelles. Les calculs ayant été faits pour les deux cas donnent d'ailleurs des résultats très semblables.

Les calculs effectués sont les suivants :

. Analyse du fichier des Altitudes, les Anomalies étant projetées en éléments supplémentaires.

. Analyse du fichier des Anomalies, les Altitudes étant projetées en éléments supplémentaires.

. Calcul des distances entre point Anomalie et point Altitude, pour chaque degré carré (sur le plan des deux premiers axes) pour les deux analyses. On constate que les distances sont très voisines d'une Analyse à l'autre. Une distance faible indique une bonne corrélation, une distance forte une mauvaise. On peut classer les zones.

. Reconstitution des données à l'aide des deux premiers facteurs : Si φ_1^i, φ_2^i, ϕ_1^j, ϕ_2^j désignent les facteurs centrés normés à 1, la valeur filtrée de la jème mesure sur le ième degré carré est :

$$f_{ij} = f_i\, f_j\, (1 + \sqrt{\lambda_1}\, \varphi_1^i\, \phi_1^j + \sqrt{\lambda_2}\, \varphi_2^i\, \phi_2^j)$$

. Corrélation Anomalie-Altitude des grandeurs filtrées pour chaque degré carré.

. Classement des degrés carrés par corrélation - Anomalie Altitude croissante. On constate qu'il concorde avec le classement par distance croissante.

. Répartition en 20 classes de corrélation et détermination des zones correspondantes.

. Confrontation avec les idées des géophysiciens. Bon accord dans l'ensemble.

****

Nous remercions J.P. BORDET (C.I.G.) qui nous a fourni les programmes qui nous ont été utiles, et qui nous a aidé dans notre travail.

****

BIBLIOGRAPHIE

[1] G. MATHERON - Théorie des Variables Régionalisées et Application. Fascicule 5 des Cahiers du Centre de Morphologie Mathématique. Ecole des Mines. FONTAINEBLEAU.

[2] J.P. BENZECRI - Publication de l'ISUP Université Paris V° Quai Saint-Bernard.
En particulier - La Pratique de l'Analyse des Correspondances.