LES CAHIERS DE L'ANALYSE DES DONNÉES

C. SABATON

Sur l'optimisation d'un système d'observation : application à un réseau de contrôle de la pollution atmosphérique

Les cahiers de l'analyse des données, tome 2, n° 2 (1977), p. 173-192

<http://www.numdam.org/item?id=CAD_1977__2_2_173_0>

Les Cahiers de l'Analyse des Données
Vol. II 1977 n° 2 - p 173 192

SUR L'OPTIMISATION D'UN SYSTÈME D'OBSERVATION : APPLICATION A UN RÉSEAU DE CONTRÔLE DE LA POLLUTION ATMOSPHÉRIQUE

II – Méthodes de prévision et réduction d'un système d'observation

[POLLUTION E.D.F.]

par C. Sabaton (1)
avec compléments théoriques de J.-P. Benzécri (2)

Rappelons le problème particulier qui a suggéré le thème de notre recherche. Les relations entre les diverses stations d'un réseau de surveillance de la pollution en région parisienne (*) ont été mises en évidence précédemment (cf. § 3.1. (**)). Le réseau comporte un noyau de postes centraux fortement corrélés entre eux. Les postes périphériques sont un peu particuliers. Dans l'ensemble cependant, les liaisons existant entre les différentes stations de mesure semblent assez importantes. Peut-on, à partir des taux d'acidité mesurés en un certain nombre de postes, obtenir une bonne estimation du taux d'acidité en un poste j du réseau ? Si oui, le poste j pourra être supprimé.

Nous appliquerons successivement plusieurs méthodes conçues pour choisir rationnellement le sous-réseau des postes à conserver. Ces méthodes ont toutes en commun un même principe, qui est d'associer à chaque poste (encore appelé station) une série chronologique de pollution (ou fonction sur l'ensemble I des jours d'observation) ; et de fonder l'extrapolation de la série du poste j (i.e. l'estimation de la pollution en j pour un jour supplémentaire $i_S \notin I$) sur la seule observation de quelques postes dont les séries sur I sont voisines de celles de j (la pollution au jour i_S étant connue en ces postes et non en j). Mais les méthodes diffèrent quant aux critères d'approximation et plus encore quant au codage et à l'analyse des informations recueillies.

Dans la première méthode (§ 3.2.) les taux de pollution sont codés sous forme logique : la pollution observée devient comme la réponse à une question fermée admettant 10 réponses qui sont les 10 niveaux définis en découpant l'histogramme des pollutions en tranches d'égal effectif (cf. § 2.1.). La pollution en une station j éliminée pourra être estimée d'après celle observée aux stations conservées par une régression faite après analyse de correspondance ; régression dont un coefficient de corrélation mesure l'efficacité.

Dans les deuxième et troisième méthodes (cf. §§ 3.3. et 3.4.) les taux de pollution sont traités comme des variables continues : mais là, (§ 3.3.) selon l'esprit de l'analyse en composantes principales, la série chronologique de chaque poste est transformée pour être centrée et normée (moyenne 0 ; variance 1), le critère d'approximation étant la stabilité des facteurs ; tandis qu'ici (§ 3.4.), en vue d'une analyse

---

(1) Ingénieur à la Direction des Etudes et Recherches de l'Electricité de France.

(2) Professeur de statistique à l'Université Pierre et Marie Curie (Paris VI).

(*) Réseau du Laboratoire d'Hygiène de la Ville de Paris - en abréviation : L.H.V.P.

(**) Cf. : Les cahiers de l'Analyse des Données ; Vol. II n° 1.

de correspondance, on distingue dans chaque série son profil et son poids : l'extrapolation est fondée sur la similitude des profils, le critère d'approximation tient compte des poids.

Somme toute, les méthodes 1 et 3 (§ 3.2. et 3.4) conduisent l'une et l'autre à proposer un réseau simplifié acceptable . Tandis que la méthode 2 (§ 3.3.) apparaît sujette à retenir d'autant plus des informations qu'elles sont plus redondantes ; fait paradoxal mais qu'un contre-exemple schématique (§ 3.3.3.) explique bien.

Au § 4, nous concluons en suggérant des voies pour comparer ces méthodes et préciser leur validité.

Le présent article, surtout méthodologique, pourra intéresser un lecteur n'ayant de l'article précédent qu'une connaissance superficielle ; la lecture de quelques alinéas auxquels il est fait référence étant seule essentielle.

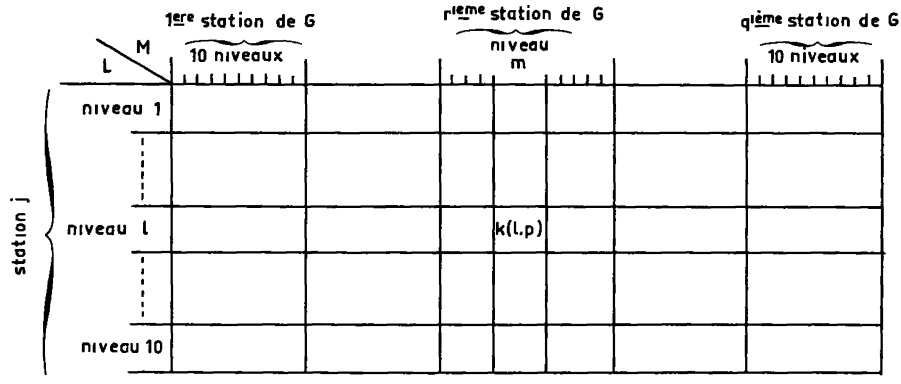
3.2. *Première méthode d'optimisation : estimation basée sur une régression faite après analyse du tableau de contingence recensant des données codées sous forme binaire*

3.2.1. *Exposé de la première méthode*

Soit j une station du réseau.

On cherche à estimer la pollution en cette station à l'aide des observations faites en un groupe G d'autres stations.

Pour cela, on construit le tableau suivant :



. G est le groupe des q stations "explicatives".

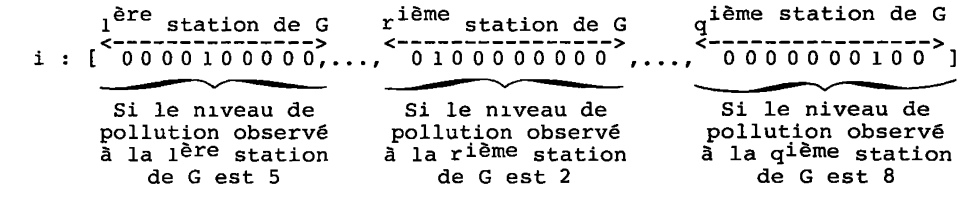
. les taux d'acidité sont, comme précédemment, répartis en 10 classes d'égal effectif.

. $k(\ell,p) = k[\ell, 10(r-1) + m]$: nombre de jours où l'on a observé à la fois le niveau de pollution ℓ à la station j et le niveau de pollution m à la $r^{ième}$ station du groupe G.

L'analyse des correspondances de ce tableau permet de dégager les facteurs principaux de la correspondance entre la station j et le groupe de stations G.

Soit i et i' deux jours d'observation tels que les taux mesurés aux postes de G soient globalement voisins. Si la station j est bien expliquée par le groupe G, à ces deux jours doivent correspondre des taux de pollution voisine à la station j.

Un jour d'observation i peut-être caractérisé par le vecteur-ligne suivant :



La démarche est la même qu'au § 2.2. :

Les 773 "vecteurs-jours" (traités comme lignes supplémentaires adjointes au tableau K_{LM}) sont projetés dans l'espace des premiers facteurs de l'analyse du tableau [Dans notre étude, l'inertie expliquée par les deux premiers étant, pour chaque tableau, suffisante, les jours sont projetés dans le plan 1 x 2 issu de l'analyse du tableau K_{LM}].

Le taux de pollution à la station j le jour i est alors estimé d'après la position de i dans le plan 1 x 2 ; comme au § 2.2., on peut soit faire une régression linéaire (voire polynomiale) de la pollution par rapport aux facteurs $F_1(i)$ et $F_2(i)$; soit prendre la moyenne des taux mesurés à la station j les 10 jours les plus proches de i dans cet espace (régression par boule). Le coefficient de corrélation ρ entre le taux effectivement mesuré à la station j et son estimation permet de définir la part ρ^2 de la variance de la pollution à la station j expliquée par les taux observés sur le groupe G.

La station j sera supprimée si cette variance expliquée atteint un niveau suffisant défini soit a priori en fonction de la précision désirée, soit a posteriori en fonction du nombre maximum de stations qu'on accepte de conserver.

3.2.2. *Application au niveau du LHVP*

La démarche devait être la suivante :

- chaque station est expliquée par les 29 autres stations du réseau.

La station j_1 la mieux expliquée (correspondant au plus grand coefficient ρ) est éliminée.

- chacune des 29 stations restantes est expliquée par les 28 autres. La station j_2 la mieux expliquée est éliminée, etc...

Le processus s'arrête quand le plus grand coefficient de corrélation ρ obtenu est inférieur au seuil qui lui est fixé.

Il faut prendre soin, à chaque étape, de vérifier que chacune des stations éliminées reste bien expliquée par l'ensemble des stations restantes.

En fait, pour des problèmes de temps de calcul sur ordinateur, le réseau a été divisé en groupes par une méthode de classification ascendante hiérarchique (cf. figures 10 et 10 bis).

A titre d'exemple, on a construit un réseau réduit en supprimant les stations expliquées à au moins 85 % par les stations restantes de

CLASSIFICATION DES 30 STATIONS DU L.H.V.P

(à partir des données quotidiennes d'acidité)

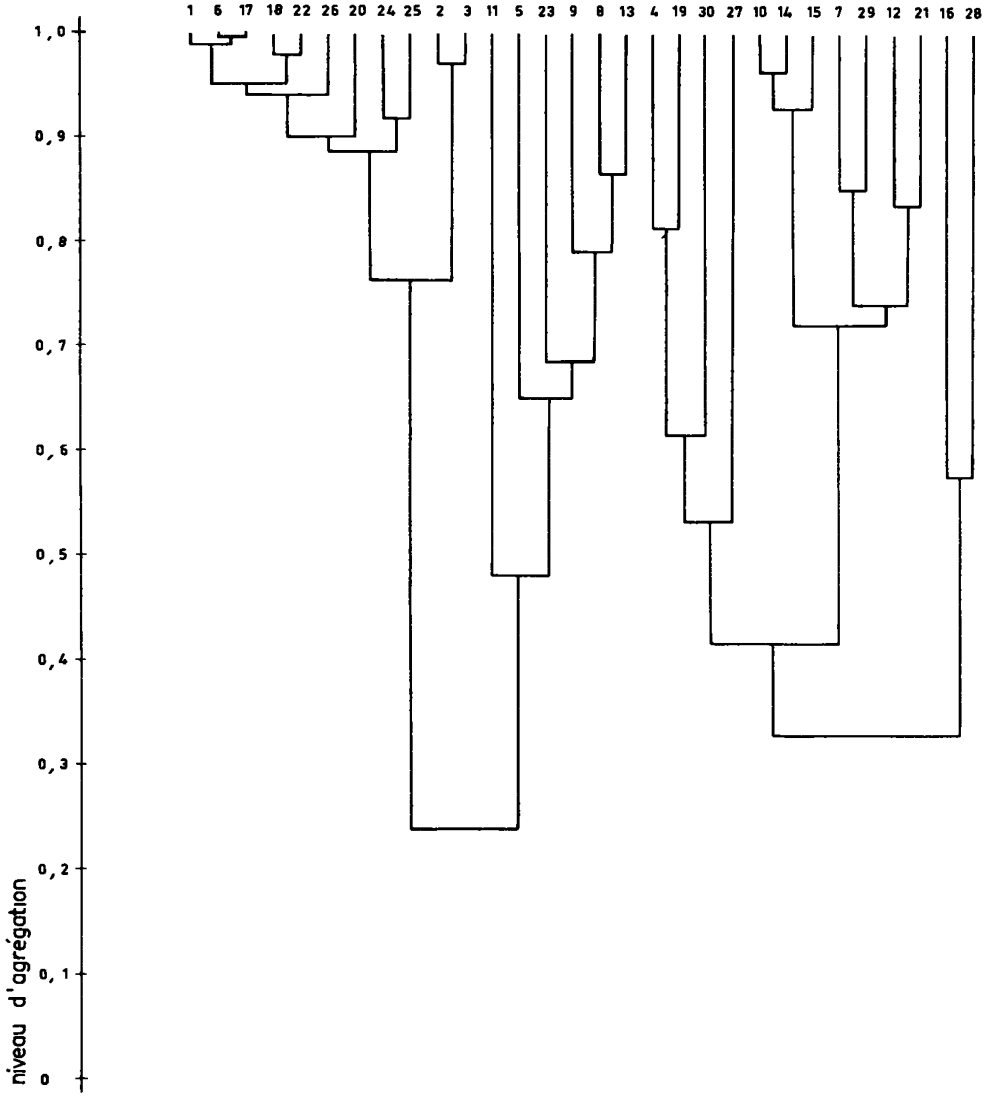


FIGURE · 10

RESEAU DE MESURE (30 postes) DE LA POLLUTION ATMOSPHERIQUE

DE L'AGGLOMERATION PARISIENNE

(réseau du laboratoire d'hygiène de la ville de Paris)

Groupes formés en vue de l'optimisation du réseau à partir

de la classification des 30 stations

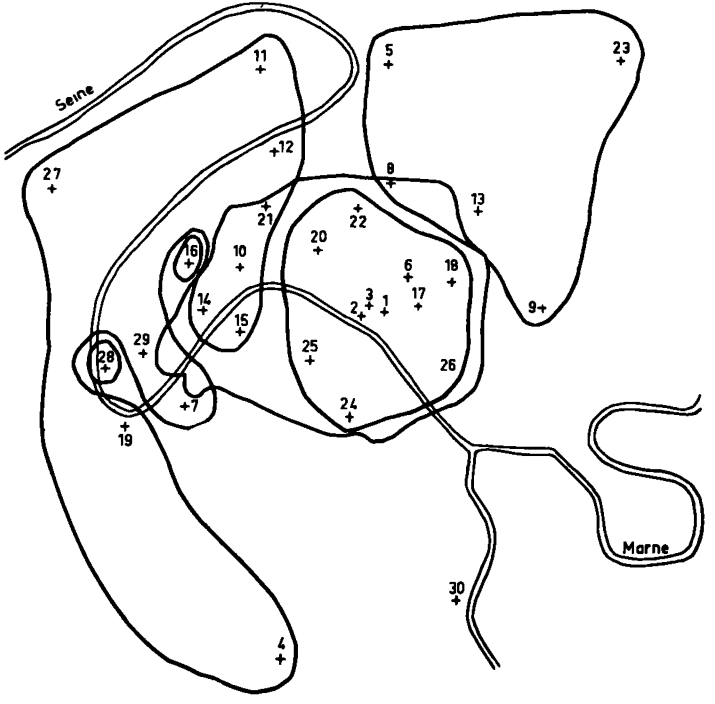


FIGURE : 10 bis

RESEAU OBTENU A L'AIDE DE LA PREMIERE METHODE D'OPTIMISATION

Estimation par régression après analyse des correspondances

O station conservée

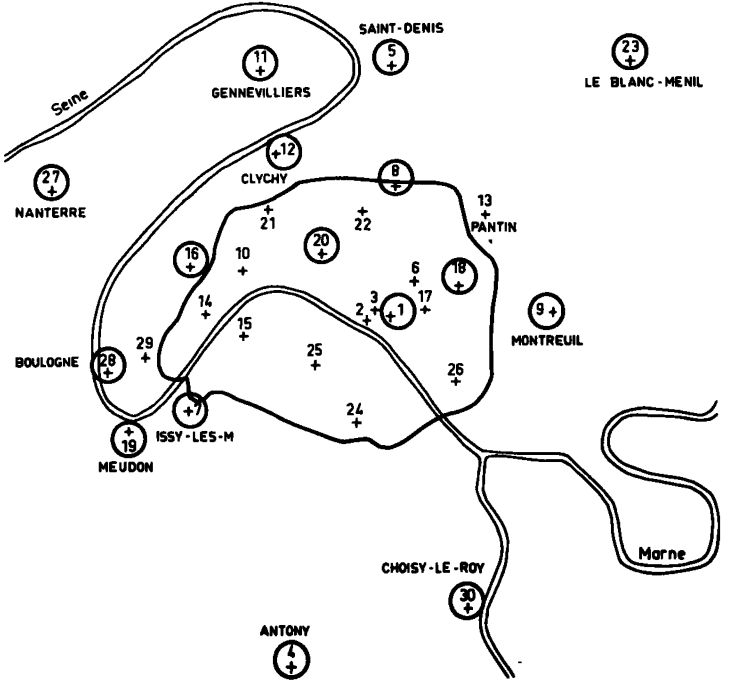


FIGURE : 11

leur groupe [$\rho \geq 0,92$; ρ étant calculé par la régression par boule : voir note (*) en bas de page].

La figure 11 montre le sous-réseau obtenu. Ce réseau comporte 15 stations. Il contient peu de postes intra-muros. Ces derniers semblent présenter peu de particularités. La plupart des stations périphériques sont, par contre, plus singulières ; souvent mal expliquées par le reste du réseau, elles ne peuvent, dans l'optique d'optimisation choisie être supprimées.

Le réseau obtenu permet en fait de reconstituer au mieux la pollution en chacun des postes éliminés.

Une autre optique qui devrait conduire à des résultats analogues est de chercher un sous-ensemble de postes permettant non plus de surveiller chaque zone particulière, mais de conserver les grands traits généraux du phénomène de pollution dans l'agglomération. Tel est l'objet du § suivant.

3.3. *Deuxième méthode : critère de l'invariance des composantes principales*

La méthode est exposée par ses auteurs (cf. Y. Escouffier ; puis J.M. Braun) dans le langage des opérateurs. Pour les lecteurs peu familiers avec ce langage, on donne d'abord au § 3.3.1. à la fois les principes et les formules de la méthode.

L'application au réseau de pollution conduit à des résultats peu satisfaisants (§ 3.3.2.) : cet échec s'explique par l'analogie de cas modèles très simples, où le critère mathématique adopté conduit à un choix des stations conservées qui n'est optimal en aucun sens raisonnable du terme (§ 3.3.3.).

3.3.1. *Exposé sommaire de la méthode*

Reprenons en les précisant les notations du § 3.1.3.

I : ensemble de 773 jours, i ; on note : Card I = n = 773 ;

S : ensemble de 30 stations, s ; on note : Card S = p = 30 ;

x(i,s), noté aussi x_s^i : pollution acide mesurée au jour i en la station s ;

$x_s^I = \{x_s^i \mid i \in I\}$: la pollution en s, considérée comme une fonction sur I ; ou encore comme un vecteur dont les 773 composantes sont indicées par $i \in I$;

$a_s^I = \{a_s^i \mid i \in I\}$: la pollution en s, centrée et réduite par une transformation linéaire : $a_s^i = \alpha_s x_s^i + \beta_s$; (α_s et β_s sont choisis pour que sur I, a_s^I ait moyenne nulle et variance 1 ; de plus, $\alpha_s > 0$) ;

$C_\alpha^I = \{C_\alpha^i \mid i \in I\}$: composante principale de rang α : C_α^I est une fonction sur I, ou vecteur de $R^I = R^{773}$, ayant comme les a_s^I, moyenne nulle et variance 1. Les composantes principales successives sont choisies pour fournir la meilleure approximation de l'ensemble des a_s^I (pour $s \in S$)

(*) A titre indicatif disons que pour la station 13, on trouve un taux d'explication par les autres stations de son groupe égal à $\rho^2 = 0,96$ si l'on utilise une régression par boule ; mais le niveau est sensiblement inférieur : $\rho'^2 = 0,93$ si l'on se borne à une formule de régression linéaire.

au sens des moindres carrés. De façon précise les C_α^I sont vecteurs propres de la matrice suivante t^{II} :

$$t^{II} = \{t^{ii'} | i,i' \in I\} ;$$

$$t^{ii'} = (1/(30 \times 773)) \Sigma \{a_s^i \, a_s^{i'} | s \in S\} ;$$

$$t^{II} \, C_\alpha^I = \mu_\alpha \, C_\alpha^I \text{ ,au sens suivant :}$$

$$\Sigma \{t^{ii'} \, C_\alpha^{i'} | i' \in I\} = \mu_\alpha \, C_\alpha^i ;$$

dans cette formule, on n'a pas distingué entre indices supérieurs et inférieurs, parce que sur R^{773} on prend constamment la structure euclidienne usuelle (carré de norme égal à la somme des carrés des coordonnées ; ou, plus exactement, à cette somme divisée par 773).

$Fac_\alpha(s)$: valeur du facteur α pour la station s ; i.e. coordonnée du point a_s^I de R^{773}, sur l'axe défini par le vecteur C_α^I ; ou encore produit scalaire :

$$Fac_\alpha(s) = (1/773) \Sigma \{a_s^i \, C_\alpha^i | i \in I\} ;$$

ou encore : corrélation entre a_s^I et C_α^I, considérés comme variables aléatoires dont on possède 773 réalisations ; etc...

$$Fac_1(s) \, C_1^I + Fac_2(s) \, C_2^I + Fac_3(s) \, C_3^I =$$

meilleure approximation de a_s^I en combinaison des trois premiers axes ; ceux-ci étant eux-mêmes justement choisis pour que la somme des carrés des écarts de l'approximation sur tous les s de S soit minima (on aurait pu de même prendre 4 axes, 5 axes, etc...).

En somme, la connaissance des (α_s, β_s) (coefficients de passage des acidités mesurées x_s^i, aux acidités normalisées a_s^i) des $Fac_\alpha(s)$ et des composantes principales C_α^i permet de reconstituer les données primaires (reconstitution approchée e.g. au rang 3, si on se borne à trois facteurs) : car par les $Fac_\alpha(s)$, on passe des C_α^i aux a_s^i ; puis par les (α_s, β_s) on passe des a_s^i aux x_s^i.

Le principe de la réduction est donc le suivant : tenter à l'aide d'un petit nombre de stations de déterminer approximativement les C_α^I d'où découlera finalement aussi la connaissance approchée des x_s^I relatifs aux stations écartées. L'approximation vaudra non seulement pour le passé (les 773 pour lesquels on a déjà des mesures complètes) mais pour l'avenir : car les $C_\alpha'^I$ issus de l'analyse factorielle (ou composantes principales) des données relatives à un ensemble S' de p' station (p' < p = 30 ; S' < S) sont nécessairement dans le sous-espace vectoriel de R^{773} engendré par les $\{a_{s'}^I | s' \in S'\}$, d'où une formule :

$$C_\alpha'^I = \Sigma \{\text{coeff } (\alpha,s') \, a_{s'}^I | s' \in S'\}$$

qui permet aussi de calculer $C_\alpha'^t$ pour un jour supplémentaire t en fonction des $a_{s'}^t$ (ou des $x_{s'}^t$, acidités primaires, non réduites, mesurées au jour t) et finalement de remonter, approximativement toute la chaîne des informations.

Le principe étant vu, reste à l'exprimer en critère numérique, et à fonder un optimum sur ce critère. ESCOUFIER et BRAUN remarquent en substance que l'ensemble réduit S' conduit aux mêmes composantes principales que S tout entier si il y a identité entre t^{II} (défini ci-dessus) et t_2^{II} :

$$t_2^{ii'} = (1/(p' \times 773)) \Sigma \{a_{s'}^i \, a_{s'}^{i'} | s' \in S'\}$$

RESEAU OBTENU A L'AIDE DE LA DEUXIEME METHODE D'OPTIMISATION

Critère de l'invariance des composantes principales ;
Première variante : sans tenir compte du nombre des stations conservées

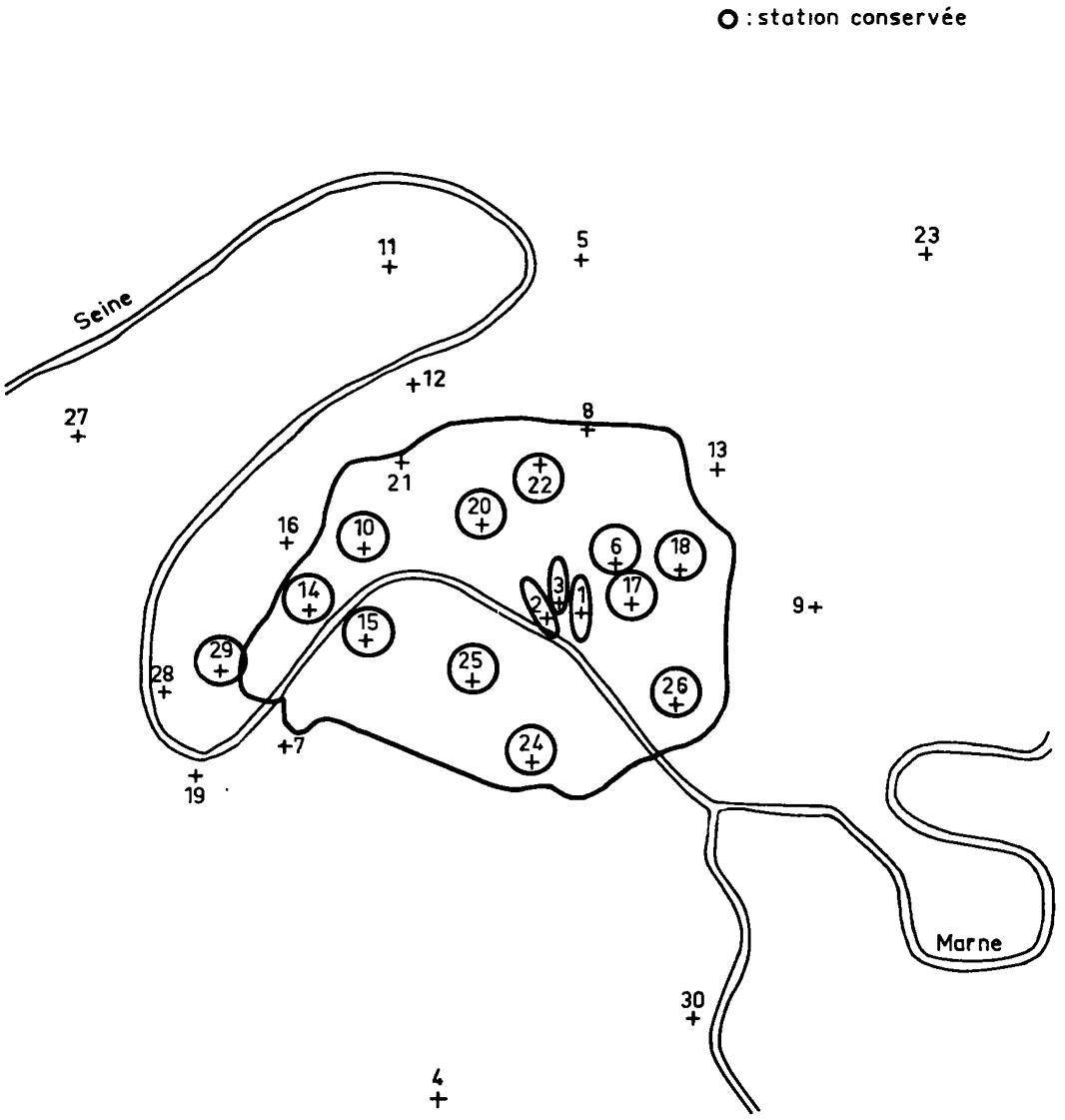


FIGURE : 12

RESEAU OBTENU A L'AIDE DE LA DEUXIEME METHODE D'OPTIMISATION

Critère de l'invariance des composantes principales ;
deuxième variante : en divisant par le nombre des stations conservées

O : station conservée

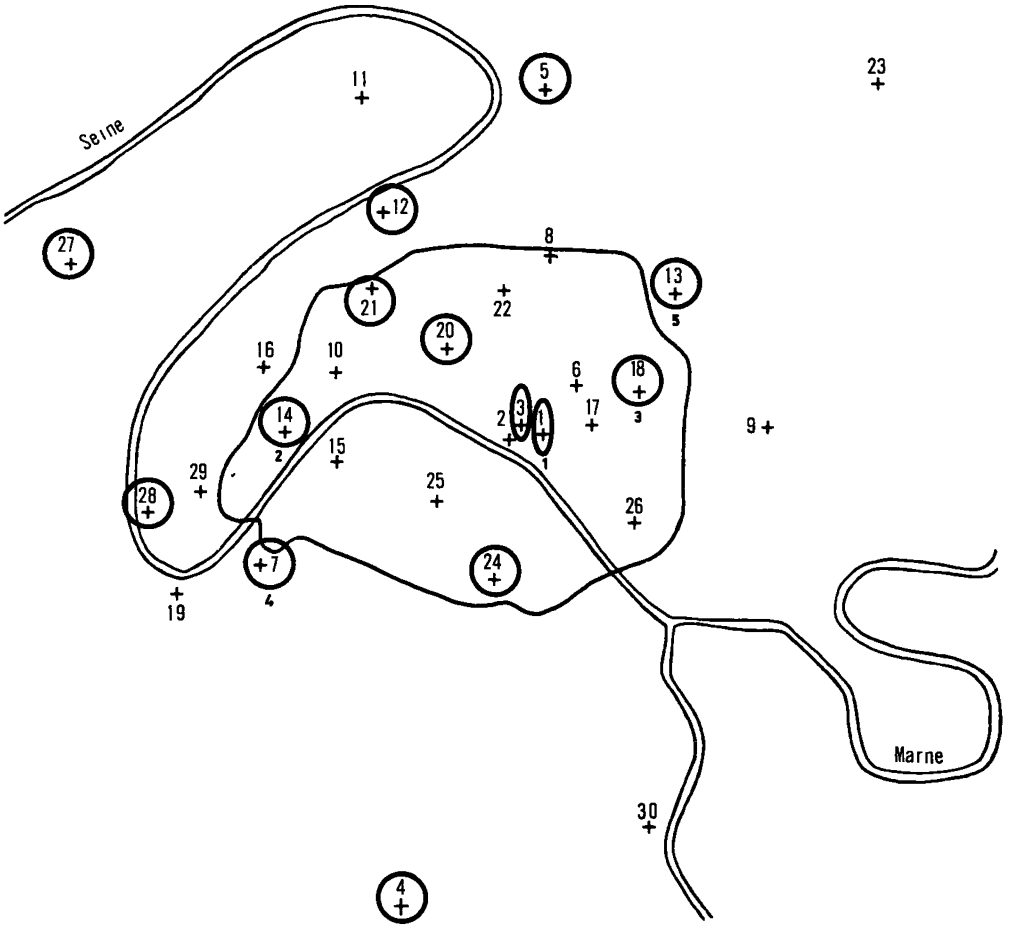


FIGURE : 12bis.

Ici, on a noté p' = card S', le nombre de stations conservées. Dans certaine version de la méthode, on ne divise pas par p' mais toujours par le même nombre p (ici : 30), ce qui on le verra est tout à fait fâcheux.

Le critère d'écart entre t^{II} et t_2^{II} est tout simplement :

$$\Sigma \{(t_2^{ii'} - t^{ii'})^2 / i \in I,\ i' \in I\} = d^2 (S, S')$$

Cette quantité n'est autre que la norme au carré de l'opérateur associé à la différence $(t^{II} - t_2^{II})$; ou encore la norme au carré de $(t^{II} - t_2^{II})$ considéré comme élément symétrique du produit tensoriel de l'espace euclidien R^{773} par lui-même. Nous n'insisterons pas ici sur cette interprétation, utile toutefois pour déduire immédiatement de principes d'invariance les valeurs du critère dans les contre-exemples du § 3.3.3.

En toute rigueur, la recherche du sous-ensemble S' de cardinal p' donné (pour ne garder que p' stations) rendant minimum d^2 (S, S'), requiert l'examen de très nombreuses combinaisons. On se borne communément dans la pratique à procéder pas à pas soit par adjonction soit par élimination de stations (ou de variables dans un autre problème de même format) :

* adjonction : 1) choisir la station s_1 telle que d^2 (S, $\{s_1\}$) soit minimum (cas : p' = 1) ;

2) s_1 étant adoptée une fois pour toutes, choisir s_2 dans $\{S - s_1\}$, telle que d^2 (S, $\{s_1, s_2\}$) soit minimum ; etc...

* élimination : 1) éliminer s_{30} telle que d^2 [S, $\{S - s_{30}\}$] soit minimum (cas : p' = 29) ;

2) s_{30} étant rejetée une fois pour toutes, choisir s_{29} telle que d^2 (S, $\{S - s_{30} - s_{29}\}$) soit minimum ; etc...

Dans la présente étude (cf. § 3.3.3.) on a procédé par adjonction, pour construire un sous-réseau de 15 stations. Une formule de récurrence simple permet d'effectuer rapidement les calculs.

3.3.2. *Le réseau réduit obtenu*

La méthode a été d'abord appliquée en divisant par 30 et non par p' le nombre des stations conservées, cf. supra ; sur la figure 12, on voit que seules subsistent les stations centrales, pourtant fort redondantes entre elles (en particulier les stations 1, 2 et 3 qui sont quasi-identiques) ; tandis que la diversité des stations périphériques est éliminée. Il apparaîtra sur un exemple type (§ 3.3.3.) que la méthode peut en effet systématiquement conserver les stations redondantes et éliminer au contraire radicalement les stations originales. La méthode a été appliquée une deuxième fois en divisant par p' : le réseau des 15 postes obtenu (figure 12 bis) ne présente qu'une seule redondance frappante : les postes 1 et 3 sont conservés. Il est curieux de noter toutefois que les cinq premières stations retenues (numérotées de 1 à 5 sur la fig. 12 bis) s'alignent assez exactement sur le premier axe de l'analyse factorielle : ce qui est encore une sorte de redondance.

3.3.3. *Contre-exemple*

En principe, la méthode du § 3.3.1. vise à obtenir la stabilité des composantes principales (composantes principales centrées normées C_α^I) ; en fait, le critère adopté -conservation de l'opérateur t^{II}- met en jeu non seulement la conservation des C_α^I, mais celle des valeurs

propres μ_α. Or, ce n'est pas même cela qui importe, mais la stabilité du sous-espace engendré par les premiers axes factoriels qui seule est significative. Au fond, il s'agit de conserver le maximum d'information : dans certains cas extrêmes, cette expression cesse d'être ambigüe ; on peut sans appareil mathématique décider quelles stations doivent être conservées. Evidemment, la méthode mathématique générale doit dans ces cas extrêmes s'accorder avec ce qu'impose le bon sens. Tel n'est cependant pas le cas. Voici un contre-exemple.

Supposons que card S = 10 ; et que les stations se répartissent en deux groupes : d'une part la station v, isolée ; d'autre part 9 stations identiques à u (ou identiques aux erreurs de mesure près) ; pour simplifier les calculs, admettons que a_u^I et a_v^I sont non corrélés. On a alors pour résultats de l'analyse factorielle :

$$C_1^I = a_u^I \quad ; \quad \mu_1 = 0{,}9$$

$$C_2^I = a_v^I \quad ; \quad \mu_2 = 0{,}1$$

Soit maintenant S' $\subset$ S ; si v n'appartient pas à S', on a d^2 (S, S') = 0,02 : car en bref, si on fait dans R^I un changement de base orthonormée avec pour deux premiers vecteurs axiaux a_u^I et a_v^I, t^{II} et t_2^{II} (ce dernier calculé en divisant par le nombre de postes retenus) sont diagonaux et s'écrivent :

$$t^{II} = \begin{vmatrix} 0{,}9 & & & \\ & 0{,}1 & & 0 \\ & & 0 & \\ & & & \ddots \\ & 0 & & & 0 \end{vmatrix} \quad ; \quad t_2^{II} = \begin{vmatrix} 1 & & & \\ & 0 & & 0 \\ & & \ddots & \\ & & & \ddots \\ & 0 & & & 0 \end{vmatrix}$$

Des principes généraux d'invariance tensorielle, il résulte que l'écart (carré de norme de $t^{II} - t_2^{II}$) n'est autre que la somme des carrés des différences des composantes, soit :

$$d^2 = (0{,}9 - 1)^2 + (0{,}1)^2 = 0{,}02$$

Si $v \in S'$ et card S' = p', on a :

$$d^2 = [0{,}9 - ((p' - 1)/p')]^2 + [0{,}1 - (1/p')]^2 = 2\,[0{,}1 - (1/p')]^2$$

Donc selon le critère d^2 (S, S'), il ne sera possible de conserver la station v qu'à partir de card S' = p' = 5 ; pour p' = 4, on devra prendre 4 fois la station u (exactement garder 4 stations identiques à u), ce qui est absurde alors qu'avec deux stations, u et v, on a un réseau réduit qui est parfait !

On peut plus généralement bâtir des contre-exemples analogues avec plusieurs classes de stations (classe u, classe v, classe w) dont chacune s'identifie à une station type : avec le critère d^2, on tendra à prendre d'abord des informations redondantes dans la classe contenant le plus de stations, alors que le bon sens requiert qu'on prenne une station dans chaque classe.

Pour donner plus de force au contre-exemple, on s'est placé ici dans la variante du calcul (diviser par card S') qui est apparue la plus satisfaisante (cf. figure 12') : et l'on n'en a pas moins trouvé un réseau réduit peu acceptable. Si dans la définition de t_2^{II} on ne divise pas

par p' mais par p, les résultats sont encore plus extrêmes : il se peut par exemple que le tableau des données comporte des stations allant par paires identiques, et que la réduction qui consiste à garder un représentant de chaque paire ne soit pas reconnue optimale ; le critère d^2 n'étant pas nul dans ce cas !

En un certain sens, ce paradoxe est lié à ce que, comme on l'a déjà dit au § 3.1.3., on ne tient pas compte des pondérations des stations : ce qui, pour le réseau étudié, équivaut à favoriser le centre où les stations sont reserrées. Mais on peut concevoir que les stations redondantes (classe u dans l'exemple) totalisent véritablement l'essentiel du poids (e.g. les 9/10) ; et alors il faudra tout de même garder une station isolée légère [e.g. v, de poids 1/10] plutôt que de répéter des mesures identiques entre elles.

3.4. *Troisième méthode : ajustement d'un nuage au support d'un sous-nuage*

On donne d'abord le principe géométrique de la méthode (§ 3.4.1.) ; puis les calculs algébriques du critère d'ajustement (§§ 3.4.2. et 3.); enfin l'application au réseau de pollution.

3.4.1. *Principe géométrique*

Pour fixer les notations, nous considérons le tableau des données comme un tableau de correspondance k_{IS} : ce point de vue est justifié par l'importance des pondérations dans toute étude d'approximation et l'avantage du principe d'équivalence distributionnelle ; mais les constructions géométriques et le critère d'ajustement s'étendent à tout nuage de points plongé dans un espace euclidien. Comme dans l'application exposée au § 3.4.4., nous appelons I ensemble des jours, et S ensemble des stations ; le problème est de choisir une partie Sréd de S, (encore appelé ensemble réduit, ou ensemble des stations conservées), tel que la connaissance des {k(is, s)/s∈Sréd}, pour un jour supplémentaire is ($is \notin I$) aux stations conservées, permette une reconstitution approchée des k(is, s) pour les stations éliminées ($s \in S - Sréd$).

Dans l'espace R_I, (espace des mesures sur I), il correspond à chaque station s un profil de pollution f_I^s, qui sera affecté (selon l'usage) de la masse f_S : on a ainsi le nuage $N(S) \subset R_I$. Dans R_I, N(S) a pour support affin (plus petite sous-variété linéaire contenant N(S) : e.g., si S ne comporte que deux points, le support affin est la droite qui les joint) une sous-variété linéaire $L_I(S)$ incluse dans l'hyperplan H_I (ensemble des mesures de masse totale 1) :

$$L_I(S) \subset H_I \subset R_I \ ;$$

sauf dépendance linéaire entre les stations, $L_I(S)$ a pour dimension 29 (cardS - 1). Nous dirons qu'un sous-ensemble Sréd fournit une bonne approximation de S tout entier si le nuage N(S) s'écarte peu du support affin $L_I(Sréd)$: le profil $f_I^{s'}$ d'une station éliminée ($s' \notin Sréd$) sera estimé par sa projection orthogonale $fest_I^{s'}$ sur le support $L_I(Sréd)$: $fest_I^{s'}$, comme tout point de ce support est une combinaison linéaire des $\{f_I^s/s \in Sréd\}$. Les coefficients de cette combinaison linéaire ont pour somme, 1 : c'est ce qu'on appelle des coordonnées barycentriques :

$$fest_I^{s'} = \Sigma \{coef_s^{s'} \ f_I^s | \ s \in Sréd\} \ ;$$

$fest_I^{s'}$ est barycentre du système des profils des stations conservées affectés des masses $coef_s^{s'}$. La construction de fest est schématisée sur la figure 13.

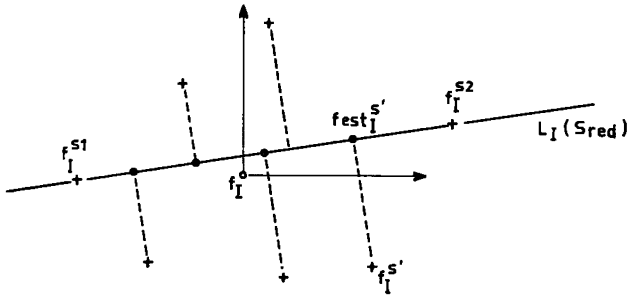


Figure 13 : Schéma de la reconstitution des données d'après un réseau réduit ; ici Sréd = {s1,s2} ; L_I(Sréd) est une droite ; on peut supposer que la figure est faite dans le plan des axes 1x2 issu de l'analyse de correspondance.

Les formules classiques du calcul des coordonnées barycentriques d'un point projeté seront rappelées aux §§ 3.4.2. et 3. ; montrons d'abord que les coefficients $coef_s^{s'}$ résolvent notre problème d'approximation ; et suggérons comment choisir Sréd pour que cette approximation soit optimale.

Soit is un jour supplémentaire où l'on a mesuré les {k(is, s)/s∈Sréd}. On posera pour tout s∈Sréd :

$$s \in Sréd : f_{is}^{s} = k(is, s)/k(s) ;$$

où le dénominateur k(s) est celui fourni par la marge du tableau principal k_{IS}. Et pour une station s'éliminée :

$$s' \in S\text{-}Sréd : fest_{is}^{s'} = \Sigma \{coef_s^{s'} \; f_{is}^{s} / s \in Sréd\} ,$$

formule où les coefficients $coef_s^{s'}$ sont ceux calculés pour la projection orthogonale de $f_I^{s'}$ sur L_I(Sréd), (toujours d'après le tableau principal k_{IS}). D'où la reconstitution cherchée :

$$\forall \; s' \in S\text{-}Sréd : kest(is, s') = fest_{is}^{s'} \; k(s') ,$$

(où k(s') provient du tableau k_{IS}).

La qualité de cette formule de reconstitution peut être estimée en l'appliquant aux données que l'on possède, c'est-à-dire en calculant kest(i, s') pour i∈I. Si l'on adopte le principe des moindres carrés, (c'est-à-dire l'estimation de l'ordre de grandeur d'une erreur par sa variance), la quantité critère pour la reconstitution de la station s' sera le carré de la distance $\| f_I^{s'} - fest_I^{s'} \|^2$; et pour l'ensemble des stations on cherchera à rendre minimum la somme de ces carrés de distance pondérés par les masses $f_{s'}$, c'est-à-dire le moment d'inertie Mt_2 du sous-nuage N(S-Sréd), (ou de N(S) tout entier, ce qui revient au même, car si s∈Sréd, la contribution de s à l'erreur est nulle) par rapport au sous-espace L_I(Sréd) :

$$Mt_2 = \Sigma \{f_{s'} \, \| f_I^{s'} - fest_I^{s'} \|^2 | \, s' \in S\text{-}Sréd\}$$

Il est également possible de calculer d'autres quantités critères, telles que le sup des écarts $\| f_I^{s'} - fest_I^{s'} \|^2$, pour s'∈S-Sréd ; ou

encore calculer comme au § 3.2. des corrélations entre pollution réelle et pollution estimée etc... Ici, on se bornera au critère Mt_2 (cf. § 3.4.3.) ; l'application systématique du principe des moindres carrés a l'avantage de mettre en harmonie toutes les études : analyse factorielle, reconstitution approchée, classification automatique ; comme on le verra dans le choix de Sréd.

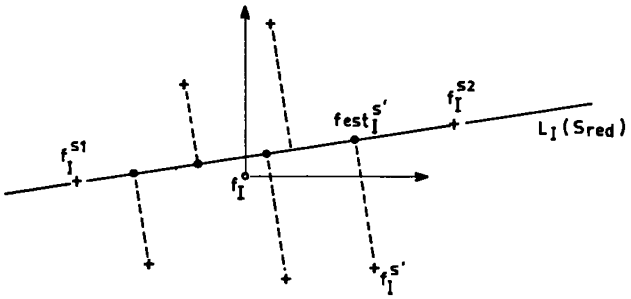


Figure 13 : Schéma de la reconstitution des données d'après un réseau réduit ; ici Sréd = {s1,s2} ; L_I(Sréd) est une droite ; on peut supposer que la figure est faite dans le plan des axes 1x2 issu de l'analyse de correspondance.

Pour le choix de Sréd, le cas le plus simple est celui où l'on ne veut conserver qu'une seule station : ce sera celle dont le profil est le plus proche du profil moyen f_I, qui est aussi le centre de gravité du nuage N(S). Avec card(Sréd) = 2, on cherchera deux stations s_1, s_2 définissant une droite aussi proche que possible du premier axe issu de l'analyse de correspondance (cf. figure 13) ; c'est-à-dire une droite faisant un angle faible avec l'axe et passant aussi près que possible du centre f_I. De plus les stations conservées s_1, s_2 devront avoir des profils nettement distincts, afin que l'orientation constatée de (f_I^{s1}, f_I^{s2}) suivant le premier axe ne soit pas le résultat des fluctuations d'échantillonnage et des erreurs de mesure. On rencontre ici le problème de la stabilité dans l'évaluation des erreurs d'approximation : on a comparé kest à k d'après l'ensemble I des jours d'observation actuellement disponibles ; ici avec 773 jours, l'évaluation semble sûre, mais il conviendrait en toute rigueur de vérifier la stabilité, principalement par des calculs de simulation : en se restreignant à un sous-ensemble I' de I, pour éprouver la formule de prédiction sur les jours de I - I' qui n'ont pas servi à construire la formule ; et aussi en perturbant le tableau des données par des erreurs de mesures simulées aléatoirement.

En général, pour Sréd de cardinal élevé (plus de 3 ou 4 points), on ne peut se contenter d'examiner les cartes et les listages de l'analyse de correspondance : il faut calculer le critère Mt_2. Comme avec toute autre méthode (cf. e.g. § 3.3.1. *in fine*) il est impossible d'essayer toutes les combinaisons de stations. On pourra procéder par adjonction : choisir les premières stations, d'après l'analyse factorielle (complétée par la classification automatique : prendre une station dans chaque classe principale) ; puis adjoindre successivement les stations choisies à chaque pas, pour donner la plus forte diminution de Mt_2.

3.4.2. Projection orthogonale sur un sous-espace vectoriel défini par une base : ici on note R_I l'espace euclidien ambiant ; $\|x_I\|^2$ et $\langle x_I, y_I \rangle$ sont respectivement le carré de norme et le produit scalaire ; nous avons en vue la métrique du χ^2 de centre f_I, mais les formules obtenues valent pour toute métrique. Soit $\{u_I^\alpha / \alpha \in A\}$ un système fini,

linéairement indépendant, de vecteurs de R_I ; notons L_I le sous-espace vectoriel de R_I engendré par ce système, et $pr(x_I)$ la projection orthogonale sur L_I d'un vecteur quelconque x_I de R_I.

Le vecteur $pr(x_I)$ est une combinaison linéaire des u_I^α caractérisée par la propriété que $(pr(x_I) - x_I)$ est orthogonal à tous les $u_I^{\alpha'}$:

$$pr(x_I) = \Sigma \{\ell_\alpha(x_I)\ u_I^\alpha \mid \alpha \in A\} ;$$

$$\forall \alpha' \in A : \langle pr(x_I), u_I^{\alpha'} \rangle = \langle x_I, u_I^{\alpha'} \rangle ;$$

Cette dernière condition permet de calculer les $\ell_\alpha(x_I)$; on a le système :

$$\forall \alpha' \in A : \Sigma \{\ell_\alpha(x_I) < u_I^\alpha, u_I^{\alpha'} > \mid \alpha \in A\} = \langle x_I, u_I^{\alpha'} \rangle ;$$

d'où il résulte que les $\ell_\alpha(x_I)$ sont des formes linéaires, combinaisons des $\langle x_I, u_I^{\alpha'} \rangle$ avec pour coefficients les $t_{\alpha\alpha'}$ obtenus en inversant la matrice des produits scalaires $\langle u_I^\alpha, u_I^{\alpha'} \rangle$:

$$pr(x_I) = \Sigma \{t_{\alpha\alpha'} \langle x_I, u_I^\alpha \rangle\ u_I^{\alpha'} \mid \alpha \in A, \alpha' \in A\} ;$$

$$\forall \alpha, \alpha'' : \Sigma \{t_{\alpha\alpha'} \langle u_I^{\alpha'}, u_I^{\alpha''} \rangle \mid \alpha' \in A\} = \delta_\alpha^{\alpha''} \text{ ; i.e. :}$$

$$t_{AA} = (scal^{AA})^{-1} \text{ ; où } scal^{AA} = \{\langle u_I^{\alpha'}, u_I^{\alpha''} \rangle \mid \alpha', \alpha'' \in A\}.$$

Connaissant $pr(x_I)$, on calcule le carré de sa norme, comme le produit scalaire d'une combinaison linéaire des u_I^α par elle-même : en développant le calcul, il apparaît le composé des matrices t_{AA} et $scal^{AA}$, ce qui simplifie le résultat final ; on a :

$$\| pr(x_I) \|^2 = \langle \Sigma \{t_{\alpha\alpha'} \langle x_I, u_I^\alpha \rangle u_I^{\alpha'}\}, \Sigma \{t_{\alpha''\alpha'''} \langle x_I, u_I^{\alpha''} \rangle u_I^{\alpha'''}\} \rangle$$

$$= \Sigma \{t_{\alpha\alpha'}\ t_{\alpha''\alpha'''} \langle u_I^{\alpha'}, u_I^{\alpha'''} \rangle \langle x_I, u_I^\alpha \rangle \langle x_I, u_I^{\alpha''} \rangle \mid \alpha, \alpha', \alpha'', \alpha''' \in A\}$$

$$= \Sigma \{t_{\alpha\alpha''} \langle x_I, u_I^\alpha \rangle \langle x_I, u_I^{\alpha''} \rangle \mid \alpha, \alpha'' \in A\}.$$

Quant à la composante de x_I orthogonale à L_I, elle a pour norme $\|x_I\|^2 - \|pr(x_I)\|^2$: c'est le théorème de Pythagore.

Soit maintenant $\{(x_I^\beta, m_\beta) \mid \beta \in B\}$ un nuage indicé par $\beta \in B$, de points x_I^β de R_I affectés de masses m_β. On calculera par les formules trouvées le moment d'ordre 2 Mt_2 (ou moment d'inertie) de ce nuage par rapport au sous-espace vectoriel L_I ; i.e. la somme pondérée par les m_β des carrés des distances à L_I :

$$Mt_2 = \Sigma \{m_\beta (\|x_I^\beta\|^2 - \|pr(x_I^\beta)\|^2) \mid \beta \in B\} ;$$

ce calcul peut se faire même si le centre de gravité du nuage est distinct de l'origine de R_I, et même s'il n'est pas dans L_I.

3.4.3. Critère d'ajustement d'un nuage à un sous-nuage

On a dit au § 3.4.1. que le nuage N(S) des profils f_I^s afférents aux 30 stations, a pour support affin un sous-espace $L(S_I)$ de R_I dont la dimension est 29 (sauf dépendance linéaire entre les stations) : $L(S_I)$

n'est pas un sous-espace vectoriel de R_I (sous-espace linéaire passant par l'origine), car $L(S_I)$ est inclus dans l'hyperplan H_I des mesures de masse totale 1 ; à plus forte raison , le support L_I(Sréd) du nuage réduit n'est pas un sous-espace vectoriel. Mais les formules de projection obtenues ci-dessus pour un sous-espace vectoriel L_I, s'adaptent vite au cas d'un sous-espace linéaire quelconque : il suffit de se rapporter à une origine prise dans le sous-espace sur lequel on projette.

Soit donc B l'ensemble des stations éliminées : $B \subset S$; notons s_o l'une quelconque des stations conservées choisie arbitrairement pour origine, et A l'ensemble des autres stations conservées :

Sréd = A U {so} ; so $\notin$ A ; B = S - Sréd ; et posons :

$$\forall\ \beta \in B : x_I^\beta = f_I^\beta - f_I^{so} \ ; \ m_\beta = f_\beta \ ;$$

$$\forall\ \alpha \in A : u_I^\alpha = f_I^\alpha - f_I^{so} \ ;$$

On peut maintenant appliquer telles quelles les formules du § 3.4.2. La projection"pr"fournit un estimateur de f_I^β (profil d'une station éliminée) par une combinaison linéaire des profils des stations conservées (à coefficients dont la somme est 1) :

$$f_I^\beta \approx f_I^{so} + pr(x_I^\beta) = fest_I^\beta$$

$$\approx f_I^{so} + \Sigma\{t_{\alpha\alpha'} \ <x_I^\beta, u_I^\alpha> \ (f_I^\alpha - f_I^{so}) \mid \alpha \in A\}$$

La quantité critère à minimiser n'est autre que Mt_2 :

$$\text{critère} = Mt_2 = \Sigma\{f_\beta \ \| f_I^\beta - fest_I^\beta \|^2 \mid \beta \in B\}$$

On notera que les coefficients de projection ainsi que le critère Mt_2 peuvent être calculés si l'on connaît les masses marginales f_β et la matrice 30 x 30 des produits scalaires $<f_I^s, f_I^{s'}>$: il est inutile de travailler dans l'espace R_I (ici R^{773}) ; une fois faite l'analyse de correspondance, on dispose dans le support $L_I(S)$ (de dimension 29) du nuage des stations d'un système de coordonnées orthonormées qui sont les facteurs $Cr_\alpha(s)$ sur l'ensemble S : par exemple on a :

$$<f_I^s , f_I^{s'}> = \Sigma\{Cr_\alpha(s) \ Cr_\alpha(s') \mid \alpha = 1, \ldots, 29\}$$

C'est cette simplification qui rend praticable la recherche de Sréd.

3.4.4. *Application au réseau du laboratoire d'hygiène de la ville de Paris*

Le tableau de base k_{IS} n'est autre que le tableau avec pondération déjà analysé au § 3.1.2. :

$$k(i,s) = \sigma_s \ x(i,s) \ ,$$

où x(i,s) est le taux de pollution acide en s au jour i, et σ_s le coefficient de surface choisi pour la station s (cf. figure 7 ; § 3.1.2. ; article I) ; en sorte que k(i,s) représente la masse de polluant dans le volume aérien de la zone centrée en s. Le nuage N(S) n'est autre que celui représenté sur la figure 9 (article I).

Les r postes conservés sont choisis pour rendre minimum le critère Mt_2. En fait comme on l'a dit plus haut (§ 3.4.1.) il n'est pas nécessaire de calculer les inerties Mt_2 correspondant aux C_{30}^r combinaisons

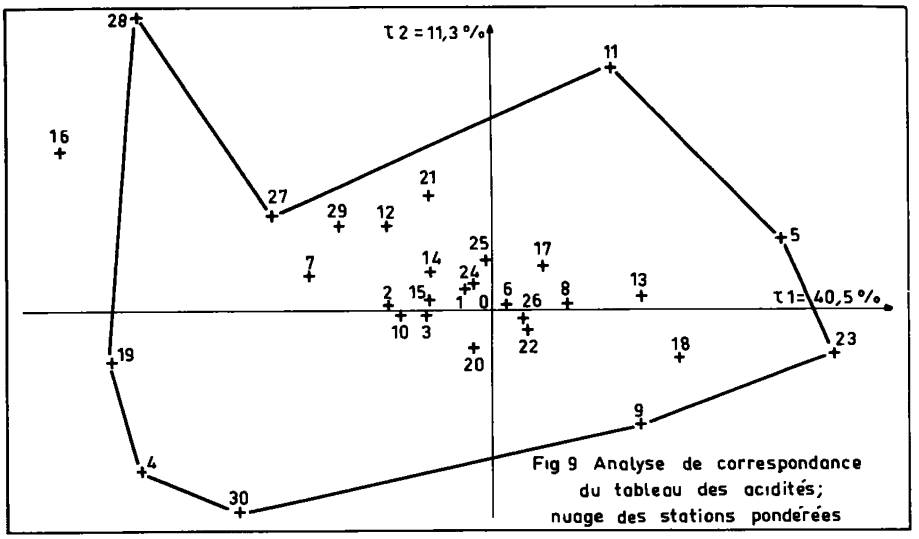
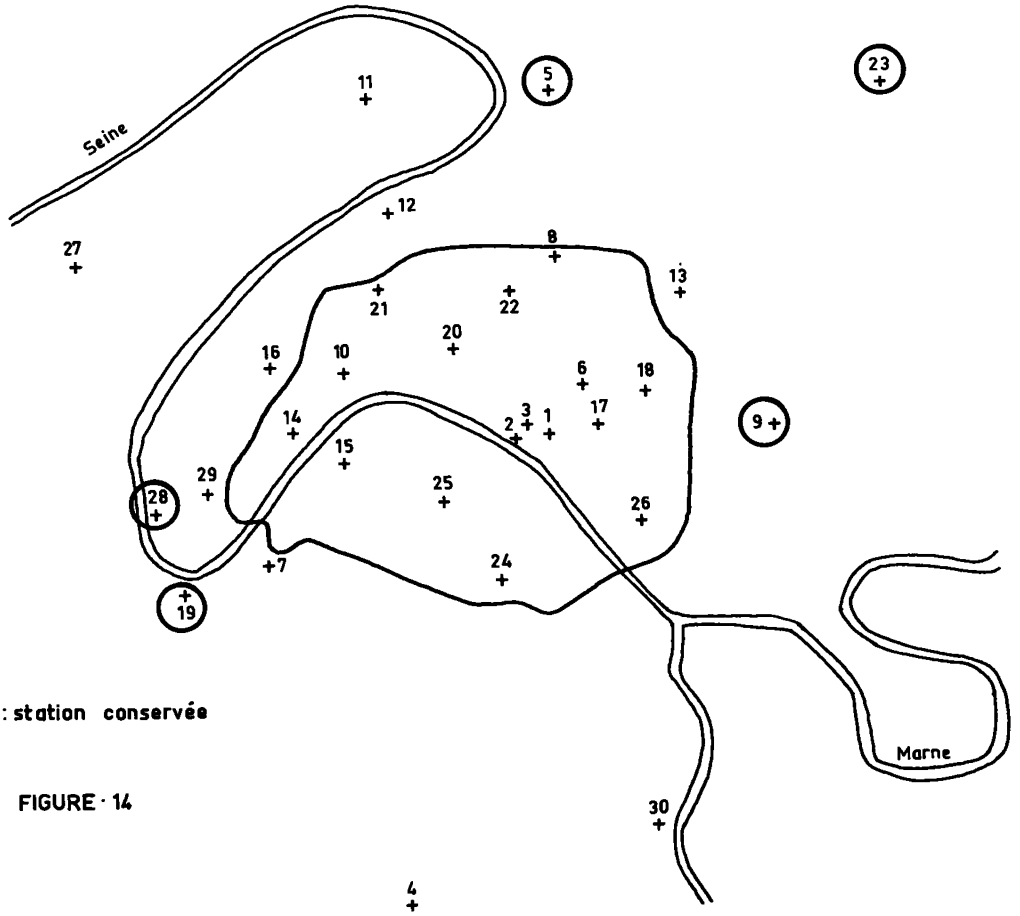


Fig 9 Analyse de correspondance du tableau des acidités; nuage des stations pondérées

RESEAU OBTENU A L'AIDE DE LA TROISIEME METHODE D'OPTIMISATION

Critère d'ajustement d'un nuage à un sous nuage



O : station conservée

FIGURE · 14

possibles de 30 postes pris r à r : les r postes conservés seront choisis parmi ceux les mieux représentés sur les premiers axes de l'analyse du nuage N(S).

Si l'on ne conserve qu'un seul poste, celui-ci devra être le plus proche du centre de N(S) : on prendra donc 1 ou 17. Prenons par exemple 1 : le critère Mt_2 est alors le moment d'inertie du nuage par rapport à f_I^1 ; i.e. d'après le théorème de Huyghens la somme du moment d'inertie par rapport au centre f_I (i.e. de la trace : $\lambda_1 + \lambda_2 + \ldots + \lambda_{29}$) et du carré de la distance $\| f_I^1 - f_I \|^2$ (multiplié par la masse totale Σf_s : mais celle-ci est 1).

$$Mt_2 = \text{trace} + \| f_I^1 - f_I \|^2 = 0{,}107 + 0{,}034 = 0{,}141$$

Pour un sous-réseau de deux postes, l'inertie minimum est obtenue avec le couple {13, 19}. Le critère Mt_2, ou moment d'inertie du nuage par rapport à l'axe $(f_I^{13}\ f_I^{19})$ vaut 0,099, ce qui est déjà inférieur à la trace 0,107.

Le nombre des postes a été augmenté jusqu'à ce que le critère Mt_2 soit inférieur à $\lambda_3 + \ldots + \lambda_{29}$ = trace - λ_1 - λ_2 ; ce qui revient à dire que le moment d'inertie du nuage N(S) par rapport au sous-espace L_I(Sréd) (lequel ne passe pas par le centre f_I) est inférieur au moment d'inertie par rapport au plan des axes 1 x 2. Avec les trois postes {19, 23, 9 }, Mt_2 = trace x 0,74 ; avec {19, 23, 9,5}, Mt_2 = trace x 0,575 avec {19, 23, 9,5, 28}, Mt_2 = trace x 0,45 : c'est déjà mieux que trace - λ_1 - λ_2 = trace x 0,5.

On a arrêté la sélection à cet ensemble de cinq stations, qui sont toutes périphériques.

4. Conclusion ; validité comparée des méthodes de prévision et de réduction des systèmes d'observation : le problème qui fait l'objet du présent article rentre dans le cadre général de la reconstitution des données manquantes. En effet, réduire à Jréd l'ensemble J des observations effectuées à l'avenir sur chaque individu is, c'est accepter de créer désormais des lignes où manqueront systématiquement les informations des colonnes de J - Jréd ; cette réduction ne se faisant sans perte d'informations que si les informations contenues dans le tableau de base k_{IJ} permettent de reconstituer assez bien ces données manquantes. Dans le cas où les lacunes sont disposées aléatoirement la formule usuelle de reconstitution des données en fonction des facteurs s'est montrée efficace : l'ont montré notamment les thèses de F. MUTOMBO, Ch. NORA et B. TALLUR ; et il est apparu dans la thèse de A. BENSABER que cette formule sert encore pour un tableau dont les lacunes forment des blocs entiers.

La particularité du problème, objet du présent article, est que les données manquantes forment des blocs que l'expérimentation dispose librement, d'après l'étude préalable de données complètes. Ici encore, les études antérieures ne sont pas totalement absentes : le point de vue adopté jusqu'ici en analyse des correspondances (cf. la thèse de S. EDDE ; depuis madame ACHKAR, référence 3) est voisin de celui du § 3.3. : on vise à la stabilité des facteurs sur l'ensemble I des individus : un sous-ensemble Jréd de mesures (ou de modalités de questions) est jugé satisfaisant s'il révèle aussi bien que J tout entier, les mêmes facteurs principaux de diversification sur l'ensemble des individus ; le choix de Jréd se faisant principalement parmi les individus qui apportent de fortes contributions aux facteurs. Même si cette méthode évite la très fâcheuse tendance à la redondance systématique démontrée au § 3.3.3., elle n'en est pas moins en butte à une difficulté au fond analogue. Reprenons avec les notations du présent §, une phrase de l'étude [Liban 60-70], T II (n° 5), § 1 :

"L'analyse du tableau k_{IJ} (1960) est déjà faite ; on peut dans un but de comparaison, analyser aussi le sous-tableau $k_{IJréd}$ relatif à 1960. En fait cette analyse a précédé l'enquête 1970, et c'est d'après elle qu'on a conclu non sans réticence, qu'il serait tolérable de ne garder des 155 questions initiales que 59, choisies pour leur contribution importante aux facteurs issus de k_{IJ}".

"Non sans réticence" ! Réticence justifiée : il est apparu à l'analyse que des questions qui en 1960 ne correspondaient pas à des facteurs importants de diversification (parce que, en bref, il s'agissait d'équipements socio-culturels alors très peu répandus) auraient grandement éclairé l'état du développement rural du Liban en 1970. Le choix du questionnaire réduit avait favorisé la redondance dans l'expression de certains facteurs (ceux qui diversifiaient l'ensemble des villages en 1960) et supprimé des aspects qui périphériques en 1960 étaient devenus centraux en 1970. A la vérité, qui pratique l'analyse des données n'acceptera jamais volontiers qu'on réduise les observations ! Ainsi dans la présente étude, le fort gradient de pollution entre les stations 16 (air pur : bois de Boulogne) et 10 (16ème arrondissement ; proche de l'Etoile) mériterait d'être particulièrement surveillé ; ce qui inciterait à maintenir ces deux stations, quel que soit le critère général adopté.

Ces principes généraux étant rappelés on voit que la présente étude se distingue par l'homogénéité des variables mesurées : les niveaux de pollution aux 30 stations ; et que les méthodes proposées aux §§ 3.2. et 3.4. sont nouvelles, tandis que celle du § 3.3. apparaît affectée d'un biais qu'on n'avait pas jusqu'ici remarqué, croyons-nous. Il reste à étudier (principalement par simulation avons-nous dit ; cf. § 3.4.1.) la stabilité des méthodes quand l'échantillon I est faible ; et à les comparer entre elles en calculant pour toutes les mêmes quantités critères (notamment les coefficients de corrélation entre valeurs réelles et valeurs prédites).

BIBLIOGRAPHIE

[1] ESCOUFIER (1970) :

Echantillonnage dans une population de variables aléatoires réelles.

Thèse d'Etat. Université de MONTPELLIER.

[2] BRAUN (1973) :

Etude des séries chronologiques multiples par l'analyse des données.

Thèse de 3° cycle. Université de PARIS VI.

[3] EDDE (1973) :

Réduction d'un ensemble de caractères en analyse des correspondances.

Thèse de 3° cycle. Université de PARIS VI.

[4] Bibliographie complémentaire.

Une étude antérieure appliquant la régression par l'analyse des correspondances a été faite dans le domaine de la pollution par M M. Berline et Bordet ; voici le résumé de cette étude publié au XIV° Journées de l'Hydraulique à Paris.

Modèle statistique des teneurs en acidité forte dans la région de Fos-sur Mer

A.P. Berline et J.P. Bordet
Société ARLAB, Valbonne.

Le but était la réalisation d'un modèle statistique des teneurs en SO_2, utilisant l'historique de la pollution et des conditions météorologiques et les prévisions météorologiques à court terme.

Le modèle réalisé à partir de données très incomplètes permettait la prévision avec une bonne précision des teneurs en SO_2 vingt quatre heures à l'avance.

La méthode utilisée est la régression factorielle aléatoire, qui consiste à explorer l'historique des situations (pollution, météorologie, émissions) recensées, pour trouver les situations qui se rapprochent le plus de celle qui est prévue. On en réduit, par un calcul simple, la teneur en acidité la plus probable au lieu considéré.

On se souviendra que la 1° application de l'analyse des correspondances à la régression par boule, conçue par J.P. Bordet a été faite dans sa thèse en 1973 sur des données de densité de haute atmosphère. Nous espérons publier bientôt une nouvelle étude sur le même thème, fondée sur des données nouvellement recueillies par satellite.