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° 1 (1977), p. 79-96

<http://www.numdam.org/item?id=CAD_1977__2_1_79_0>

Les Cahiers de l'Analyse des Données
Vol. II - 1977 - n° 1 - p. 79-96

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

I. - Pollution et météorologie; corrélations entre stations

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

Introduction

Soit X(I,J) un tableau de données où J est un ensemble de variables observées sur un ensemble I d'individus. Peut-on, à partir de l'étude de ce tableau et en fonction de ces observations, déterminer s'il est possible dans l'avenir de ne conserver qu'un sous-ensemble J_r de variables d'après lesquelles l'ensemble J pourrait être reconstitué approximativement ?

Nous avons rencontré ce problème dans l'étude d'optimisation d'un réseau de mesure de la pollution atmosphérique :

- J est alors un ensemble de postes ou stations de contrôle déjà en fonctionnement.

- I est l'ensemble des jours où l'on a réalisé une observation (une observation complète comprenant une mesure de pollution en chacun des postes).

Le réseau initial pourrait-il être remplacé par un réseau moins dense permettant, à moindre coût, de répondre aux mêmes objectifs.

Le réseau de surveillance de l'agglomération parisienne (§ 1) a servi de base à cette étude. Il a paru intéressant de déterminer tout d'abord la part de pollution déjà expliquée en un poste par certains paramètres sociaux ou météorologiques : en pourcentage de variance, cette part atteint 50 %, d'où une formule approchée déjà intéressante pour la pratique (§ 2.2.1. ; fig. 3'). L'analyse multidimensionnelle des données disponibles permet ensuite de mettre en évidence les liaisons existant entre les différents postes de mesure. Le sous-ensemble J_r de postes à conserver sera déterminé par l'étude de ces liaisons. Le choix est essentiellement guidé par une conception des informations que l'on désire garder. Trois méthodes, conçues à partir de critères différents, seront éprouvées sur nos données.

Dans le présent article, on décrit les liens entre pollution observée et variables météorologiques ; ainsi que les corrélations entre les diverses stations du réseau considéré. Un prochain article sera consacré à l'optimisation du réseau ; problème qui offre matière à des confrontations de méthodes dont les conclusions pourront servir à des études très diverses.

---

(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).*

1. Description du réseau utilisé

Le réseau utilisé pour cette étude est celui des 30 postes* du laboratoire d'Hygiène de la ville de Paris (LHVP), réseau de surveillance de l'agglomération parisienne. Ces postes sont équipés d'appareils permettant d'obtenir les teneurs moyennes journalières des deux principaux indices de pollution atmosphérique actuellement mesurés :

- l'acidité forte de l'air, exprimée en µg d'anhydride sulfureux par m^3 d'air.

- la teneur en fumées noires, exprimée en µg de particules fines de fumées par m^3 d'air.

Les mesures disponibles s'étalent sur une période de 10 ans (Octobre 1962 - Septembre 1972).

A titre indicatif les tableaux suivants donnent, pour chacun des 30 postes, les moyennes et variances respectives des taux d'acidité et de fumées observés pendant cette période, ainsi qu'un exemple de jour de forte pollution et un exemple de jour de faible pollution.

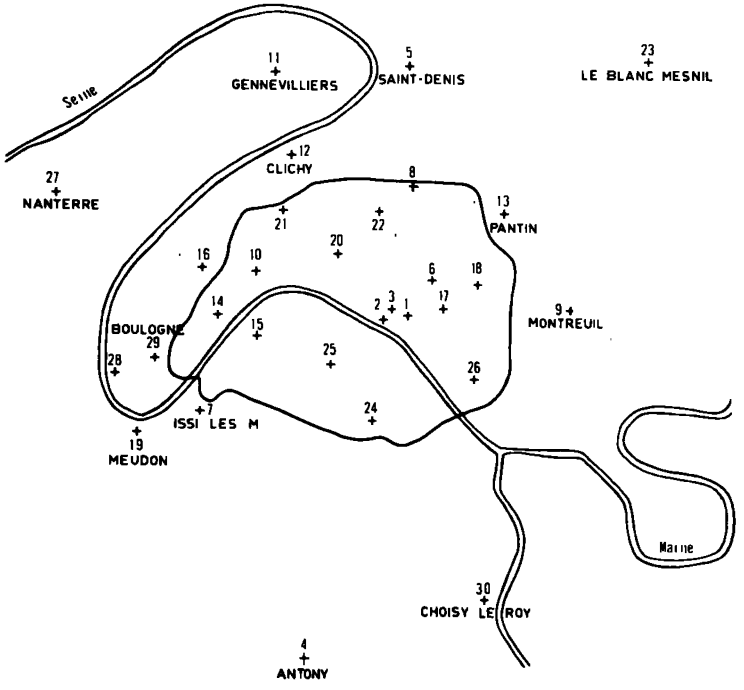


Fig 1

(*) *Dans la suite de cette étude nous emploierons indifféremment pour désigner ces points de mesure les termes de "poste" ou de "station".*

Postc	ACIDITE Moyenne	TPIIT^ S"£ type	Exemple forte pollution	Exemple faible pollution	FUMEES Moyenne	Ecart-type	Exemple forte pollution	Exemple faible pollution
1	131	117	435	^	^		^	^
Paris, 4ème	Mg/«3		»g/ra3	ng/m3	ug/m3	60*5	iiq/m^	pq/m*
2.ème	115	102	180	20	82	53	182	32
rarxs TI 4								
Paris 4ème	108	103	410	5	67	52	166	25
Antony	75	88	359	l	52	4*		^
SaintUenis	141	113	483	l9	78	58	2l1	26
paris 6 llème 7	148	130	438	19	146	70	277	56
Issy-les-Moulineaux	133	118	466	22	66	56	136	13
Paris 8 l8ème Montreuil	115	91	397	27	114	68	297	42
10.,ème	107	80	271	26	74	46	283	23
ralls 11 ID	175	145	570	24	70	59	147	13
Genne 12 Uiiers	115	93	315	l5	58	49	l61	t3
Clichy 13	151	li4	438	2?	70	60	176	l9
Pantin	138	108	448	29	83	67	20o	20 "^
Paris 14 èroe 15	183	146	506	18	67	57	165	13
Paris i5ème	144	132	573	4	66	54	157	14
tt*6\ ois ae Boulogne	70 ^/m3	82	225 ug/m^	23 „7/«*	5€ pg/m3	52	126 „7/«*	23 „g/»i
Paris l7 uême	154	134	481	34	86	66	206	23
Paris 18 20ème 19 Meudon	128 103	115 101	407 366	21 15	72 66	56 57	102 i63	17 l0
2 Paris°9ème	136	115	579	17	81	48	l85	39
SParis 2l l7ëme	151	115.	434	22	97	60	207	36
Paris 22 l8ême	151	130	563	12	74	54	191	21
2 LeBlanc3Mesnil	84	7*	28Q	24	58	44	243	24
Paris 24 l3ërae	127	96	433	22	74	55	188	20
Paris 25 l4ème	139	117	543	5	65	43	180	18
Paris 26 12ème	123	92	368	22	63	44	258	24
27 Hanterre	"	89	28o /	7	5l	**	^	"
28 Boulogne	l27	117	337	l0	60	53	U5	14
29 Boulogne	141	13Q	393	4	65	57	147	15
30 Choisy-le-Ko \	92	76	2"	^	! !	L	! i	3o

La figure 1 montre la disposition géographique des postes dans l'agglomération. leur répartition est peu homogène : les postes "intra-muros" sont trop regroupés par rapport aux postes périphériques qui sont beaucoup plus isolés.

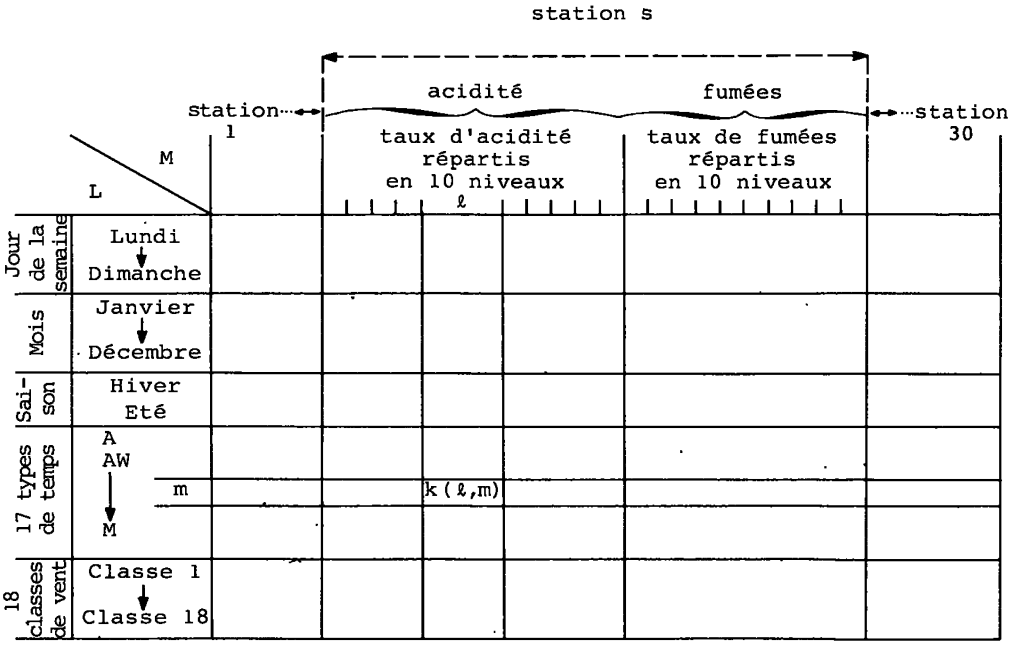
2. *Part de la pollution expliquée par certains paramètres sociaux ou météorologiques*

Certains paramètres météorologiques, voire -dirons nous- sociaux, ont une influence notable sur la pollution atmosphérique : tels sont les types de temps (caractéristique globale d'une situation météorologique : eg. anticycloniques d'Est) ; la vitesse et la direction du vent ; et aussi le jour de la semaine (la pollution industrielle est réduite le dimanche) ; la saison (le chauffage en dépend) ; et même le mois (l'industrie s'arrête au mois d'août).

Dans quelle mesure la connaissance de ces paramètres un jour donné permet-elle d'estimer en chaque poste du réseau les taux d'acidité et de fumées noires. Si l'approximation était satisfaisante, on pourrait purement et simplement supprimer le réseau de pollution et ne consulter que le calendrier et les bulletins de la Météorologie Nationale ! Au terme de ce § 2, nous verrons toutefois que les paramètres globaux n'expliquent que la moitié environ de la variance des taux de pollution.

2.1. *Etude de l'influence des paramètres retenus sur la pollution*

L'influence globale sur la pollution de ces paramètres a été étudiée par l'analyse des correspondances [référence 1] du tableau suivant :



. L'ensemble M contient 600 éléments : chacun d'eux représente pour une station et un polluant donnés, un certain niveau de pollution. Pour chaque polluant et pour chaque station, les taux de pollution sont répartis en 10 classes ou niveaux d'effectif égal (les premiers niveaux correspondant aux taux les plus faibles). Ce qui fait donc 2 x 10 x 30 éléments numérotés de 1 à 600.

L'ensemble L contient 56 éléments :

- les 7 jours de la semaine
- les 12 mois de l'année
- 2 saisons : . "l'hiver" comprenant les mois de Novembre à Mars.
  . "l'été" comprenant les mois de Mai à Septembre.

Les mois d'Octobre et d'Avril ne sont pas pris en compte car, suivant les années il y a ou non émissions dues au chauffage.

- 17 types de temps (i.e : anticyclonique d'Est...)
- 18 classes de vent (vents classés suivant leur vitesse et leur direction).

Si ℓ représente, par exemple, le $5^{ème}$ niveau d'acidité de la station s et si m représente le type de temps 6, alors :

$k(\ell, m)$ = nombre de jours où l'on a observé à la fois un type de temps 6 et un taux d'acidité à la station s se classant dans le $5^{ème}$ niveau.

Résultats de l'analyse

Le premier facteur représente, à lui seul, 71 % de l'inertie totale. A l'aide des deux premiers facteurs, c'est-à-dire avec une représentation plane, on obtient 79 % de l'inertie totale.

Le graphique 2 donne la représentation des états 1 de la pollution : les classes se répartissent, sur le premier axe, des faibles pollutions (classes 1) aux fortes pollutions (classes 10). Les stations se regroupent très nettement par niveau de pollution. L'effet polluant global de l'ensemble des paramètres pris en compte semble donc être le même pour tous les postes.

Le graphique 3 donne la représentation des paramètres étudiés. Le premier axe indique l'effet polluant de ces paramètres : ils s'y échelonnent des moins polluants aux plus polluants.

A l'extrémité négative, l'été, les mois de Juin, Juillet, Août : il n'y a plus d'émissions dues au chauffage et pendant les mois de vacances (Août particulièrement) la vie industrielle est ralentie. On retrouve des temps d'été : temps orageux, anticyclonique d'Ouest, puis les vents de Nord-Ouest et les forts vents de Sud.

Du côté positif, les vents faibles (vents de 0 et 1 m/s plus polluants que les vents de 2 m/s), les vents d'Est, Nord-Est, Nord, et les temps perturbés de même nom, les anticycloniques purs. A l'extrémité positive, les anticycloniques Sud, les anticycloniques Est, voisinant avec les mois les plus polluants (Décembre, Janvier, Février ; ce sont les mois les plus froids, le chauffage y est maximum).

Dans la partie médiane, on trouve les jours de la semaine : le dimanche est le moins pollué (la vie industrielle est ralentie) ; le taux de pollution augmente du lundi au vendredi pour redécroître en début de week-end.

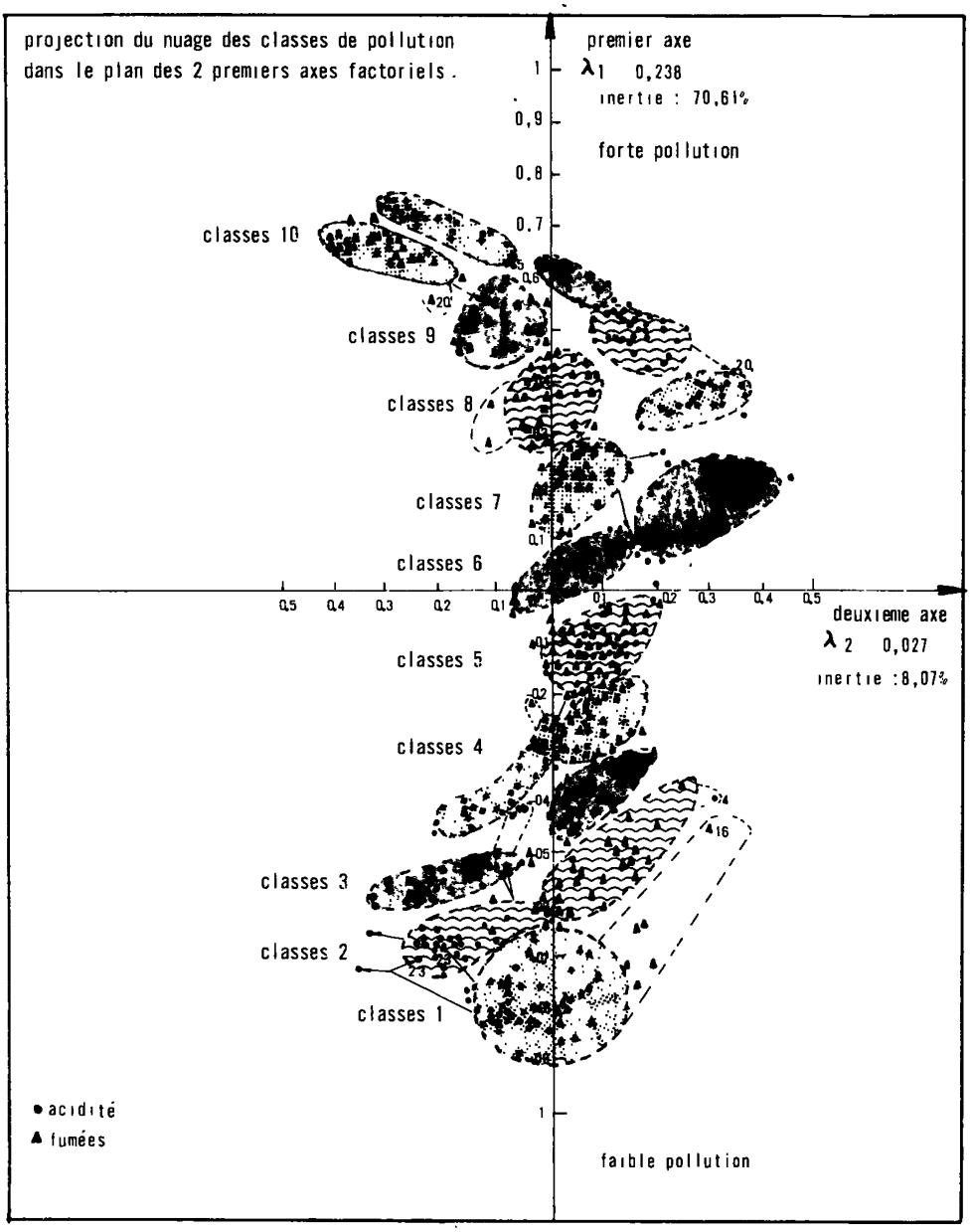


TABLEAU DE CONTINGENCE

{TAUX DE POLLUTION AUX 30 STATIONS} * {JOURS, MOIS, SAISONS, FACTEURS METEOROLOGIQUES}

Fig : 2

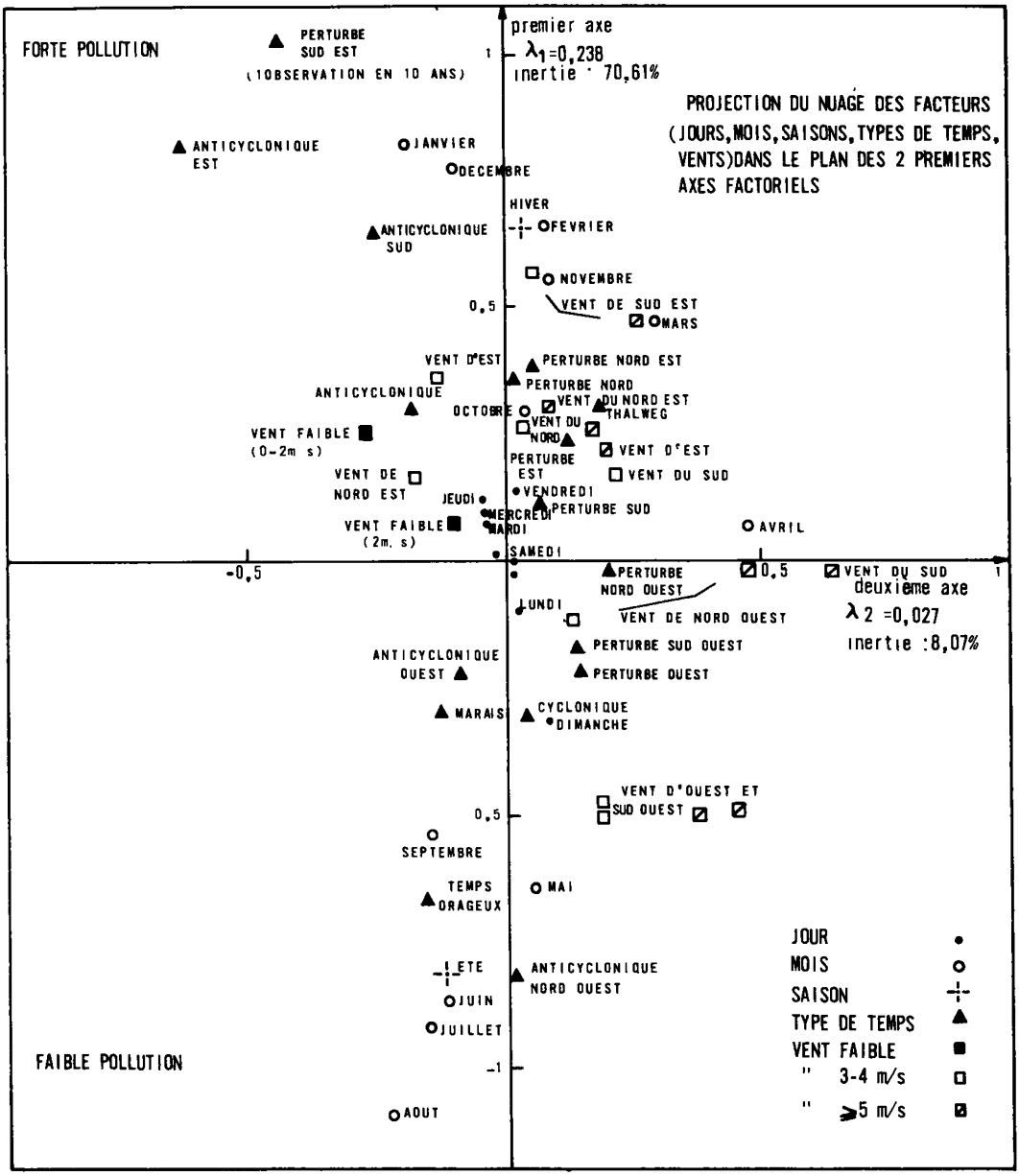


TABLEAU DE CONTINGENCE

{TAUX DE POLLUTION AUX 30 STATIONS} * {JOURS, MOIS, SAISONS, FACTEURS METEOROLOGIQUES}

Fig:3

Remarque :

La distinction entre les degrés de pollution en hiver -période de chauffage- et en été est très nette. La comparaison des effets polluants de deux situations météorologiques pour d'égales émissions est difficile si ces deux situations n'ont pas les mêmes répartitions saisonnières. Pour nous affranchir de ce problème, les tableaux obtenus avec les données recueillies d'une part en hiver, d'autre part en été, ont été analysés. Les résultats sont très comparables à ceux obtenus sur l'année entière, la hiérarchie des divers paramètres sur l'axe principal d'inertie est identique.

2.2. Part de la pollution expliquée par les paramètres étudiés

Peut-on, connaissant le jour, le mois, le type de temps, la classe de vent, avoir une bonne approximation des taux de pollution dans l'agglomération ?

Soit i et i' deux jours tels que les modalités prises par ces paramètres soient globalement voisines. Les taux de pollution observés les jours i et i' sont-ils sensiblement égaux ?

L'étude a été faite sur les taux d'acidité et de fumées observés à la station 1.

2.2.1 Méthode utilisée

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

$$i = \begin{vmatrix} k(1,i) \\ \vdots \\ k(\ell,i) \\ \vdots \\ k(56,i) \end{vmatrix} = \begin{vmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix}$$

Jour de la semaine	le jour i est un mardi
Mois	le jour i est en janvier
Saison	le jour i est en hiver
Type de temps	on y observe le type de temps 1
Classe de vent	le vent moyen est de direction Sud-Ouest et de vitesse 3 ou 4 m/s (classe 3)

Ce vecteur peut-être ajouté, comme colonne supplémentaire, au tableau de correspondance précédemment étudié.

Les paramètres -parmi ceux pris en compte- qui influent le plus sur la pollution sont bien résumés par l'ensemble des deux premiers facteurs principaux de l'analyse du tableau. Les jours d'observation sont donc projetés dans ce plan.

Soit $G_1(i)$ et $G_2(i)$ les coordonnées du jour i dans ce plan. Afin d'estimer la pollution ce jour en un poste s, on utilisera seulement

les informations fournies par $G_1(i)$ et $G_2(i)$ (c'est ce qu'on appelle faire une régression après analyse factorielle).

Une première voie simple est de déterminer le taux de pollution en s au jour i (taux d'acidité ou de fumées noires) par une régression linéaire :

$$Pol(i,s) = P_{s_1}\, G_1(i) + P_{s_2}\, G_2(i) + q_s$$

On se souviendra que les facteurs peuvent être calculés facilement, même sans le secours de l'ordinateur, une fois que le tableau K_{LM} a été analysé. On a la formule :

$$G_\alpha(i) = \lambda_\alpha^{-1/2} \Sigma \{(k(\ell,i)/k(i))\ F_\alpha(\ell) \mid \ell \in L\}$$

où $F_\alpha(\ell)$ est le facteur, issu de l'analyse de K_{LM}, correspondant à la valeur propre λ_α.

Ici :

$$G_1(i) = (0{,}238)^{-1/2} \Sigma \{(k(\ell,i)/k(i))\ F_1(\ell) \mid \ell \in L\} \qquad (1)$$

$$G_2(i) = (0{,}027)^{-1/2} \Sigma \{(k(\ell,i)/k(i))\ F_2(\ell) \mid \ell \in L\} \qquad (2)$$

Le calcul est d'autant plus facile que k(i), total de la colonne i, vaut 5 nombre de questions, et que les k(ℓ,i) valent 0 ou 1.

On obtient, pour les taux d'acidité la formule de régression suivante :

$$x(i,s) = 139{,}2\ G_1(i) + (-32{,}6)\ G_2(i) + 133{,}3 \qquad (3)$$

Le coefficient de corrélation ρ entre cette valeur calculée et le taux effectivement mesuré est égal à 0,75.

Pour les taux de fumées :

$$Fum(i,s) = 65{,}6\ G_1(i) + (-28{,}2)\ G_2(i) + 85{,}7 \qquad (4)$$

Le coefficient de corrélation ρ entre cette valeur calculée et le taux mesuré est égal à 0,70.

Pour un jour donné i, les formules (1), (2), (3) et (4) permettent d'obtenir, si l'on possède les informations "socio-météorologiques" (jour, mois, type de temps, vent), une estimation de l'état de la pollution affectant le poste s ; le taux d'explication moyen espéré est ρ^2, soit 56 % pour l'acidité et 49 % pour les fumées.

Remarque : Sur le $\alpha^{ème}$ axe factoriel, le point de coordonnée $(\lambda_\alpha)^{+1/2}\ G_\alpha(i)$ est le barycentre des 5 points de coordonnées $\{F_\alpha(q(i)) \mid q = 1,5\}$ où q(i) désigne la modalité prise par i pour le paramètre explicatif q ($1 \leq q \leq 5$).

L'estimation, pour un jour donné i, de la pollution peut se faire à l'aide de la figure 3' de la manière suivante :

Prenons par exemple le 2 Janvier 1967. C'est un lundi, le type de temps observé est un Perturbé de Nord Ouest, le vent moyen a une vites se de 2 m/s. Soit M(i) le barycentre des 5 points : Lundi, Janvier, Hiver, Perturbé de Nord Ouest, Vent faible (2 m/s) (voir figure 3'). Soit $M_1(i)$ et $M_2(i)$ ses coordonnées sur les 2 premiers axes.

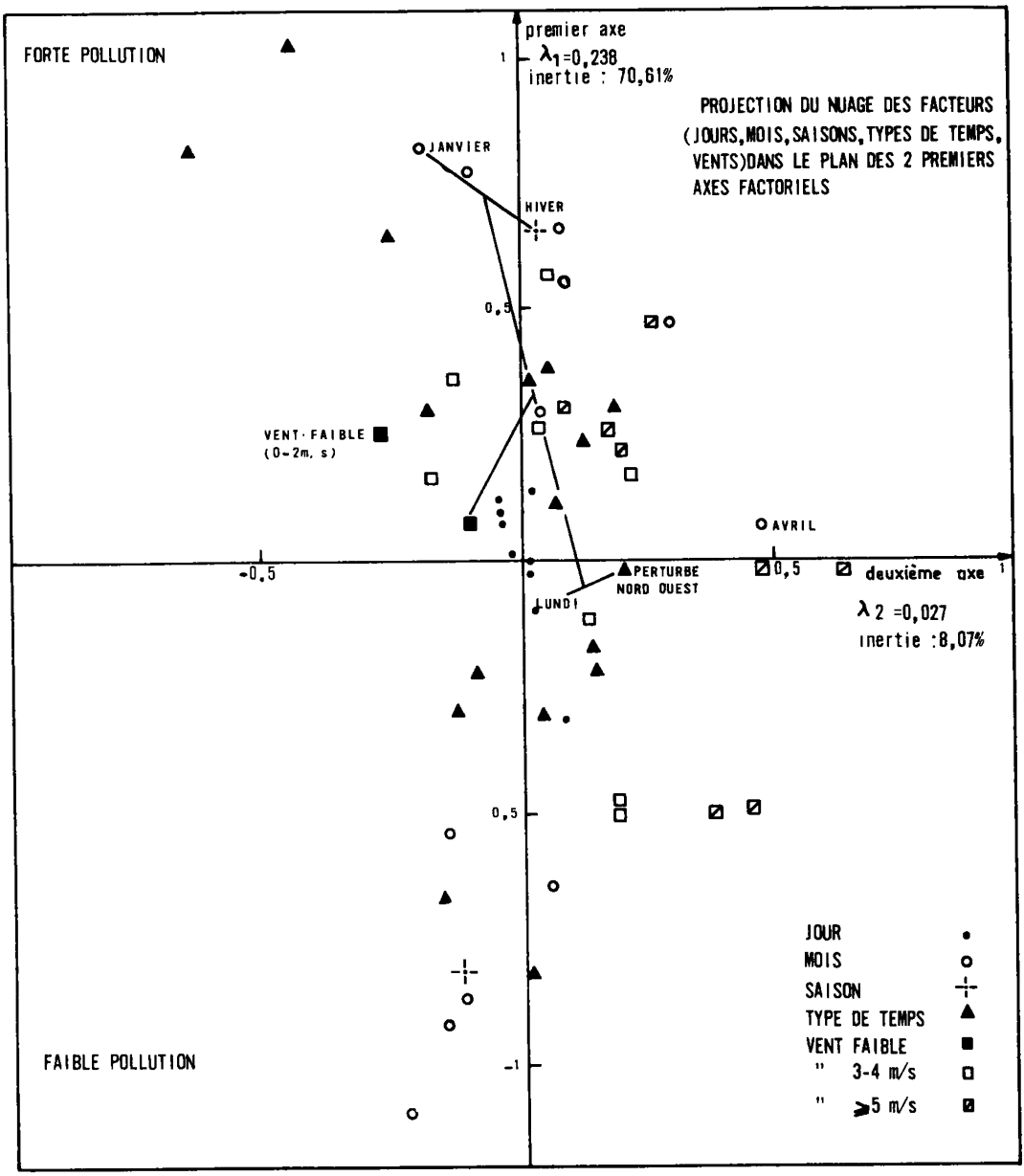


TABLEAU DE CONTINGENCE

{TAUX DE POLLUTION AUX 30 STATIONS} * {JOURS, MOIS, SAISONS, FACTEURS METEOROLOGIQUES}

Estimation rapide de la pollution un jour i, d'après les seules variables socio-météorologiques : schéma de la construction d'un barycentre pour calculer $G_1(i)$ et $G_2(i)$; on a supprimé les légendes relatives aux modalités non réalisées au jour i.

Fig: 3'

Alors :

$$G_1(i) = (0,238)^{-1/2}\, M_1(i) \; ; \; G_2(i) = (0,027)^{-1/2}\, M_2(i)$$

Les formules (3) et (4) permettent l'estimation suivante de la pollution pour ce jour :

$x(i,s) = 153\ \mu g/m^3$ correspondant à $181\ \mu g/m^3$ observé

$Fum(i,s) = 81\ \mu g/m^3$ correspondant à $81\ \mu g/m^3$ observé

On peut également songer à une régression polynominale : exprimer $Pol_S(i)$ comme un polynôme du $2^{ème}$ degré en $G_1(i)$ et $G_2(i)$. Nous appliquerons maintenant une méthode plus directe : chercher dans le plan factoriel 1 x 2 les jours i' les plus proches de i ; la pollution en i sera estimée par la moyenne de ce qu'elle a été en ses voisins (c'est ce qu'on appelle la régression par boule). En principe, si l'on s'intéresse à un jour i particulier, rien ne s'oppose à ce que la méthode soit pratiquée manuellement, pourvu que la carte du plan 1 x 2 soit tracée et que l'on ait un fichier en ordre des pollutions mesurées pour chacun des quelque 800 jours pour lesquels on a des observations complètes.

Ici, afin d'éprouver la validité de cette méthode d'estimation par les voisins, on l'a appliquée par ordinateur à un échantillon de 200 jours pour une station : le temps de calcul est alors plus long qu'avec une formule linéaire obtenue une fois pour toutes, car pour chaque point, il faut entreprendre une recherche des plus proches voisins, mais (et c'est le cas au § 3.2.1. du prochain article, sinon ici) l'approximation est souvent meilleure.

Pour chaque jour i de l'échantillon, la moyenne des taux de pollution observés à la station 1 chacun des 10 jours les plus proches de i dans le plan est comparée au taux effectivement mesuré à la station 1 le jour i [le choix du nombre 10 est arbitraire : on aurait pu prendre 20 ou 15 ; l'essentiel est que l'effectif total de l'échantillon permette de trouver un assez grand nombre de voisins proches de i]. Soit ρ le coefficient de corrélation entre la moyenne des taux des jours voisins et le taux mesuré. La variance du taux de pollution expliquée par la moyenne des voisins est égale à ρ^2 fois la variance totale, ce qui chiffre la part de pollution expliquée par les différents paramètres pris en compte dans la régression par boule.

Le calcul du coefficient ρ est fait d'après un échantillon de jours dont les données déjà recensées sont à la base de notre analyse ; mais comme on l'a vérifié empiriquement, les résultats se stabilisent dès que l'échantillon suffit à représenter la variabilité du phénomène étudié. Dès lors, si pour un jour nouveau i' on possède les informations "socio-météorologiques" (jour, mois, type de temps, vent), on place i' en élément supplémentaire dans le plan 1 x 2 et on calcule pour lui, d'après ses voisins, une estimation de l'état de la pollution qui l'affecte ; et le taux d'explication moyen espéré est ρ^2.

2.2.2. Résultats de la régression par boule

Les jours d'observations ont tous été projetés dans le plan factoriel. ρ a été déterminé à partir de 200 jours tirés au hasard parmi eux.

Les calculs ont été menés à bien grâce au programme "Reboul" mis au point par Mademoiselle LEBEAUX (Laboratoire de Statistique Mathématique Université PARIS VI - Référence 2).

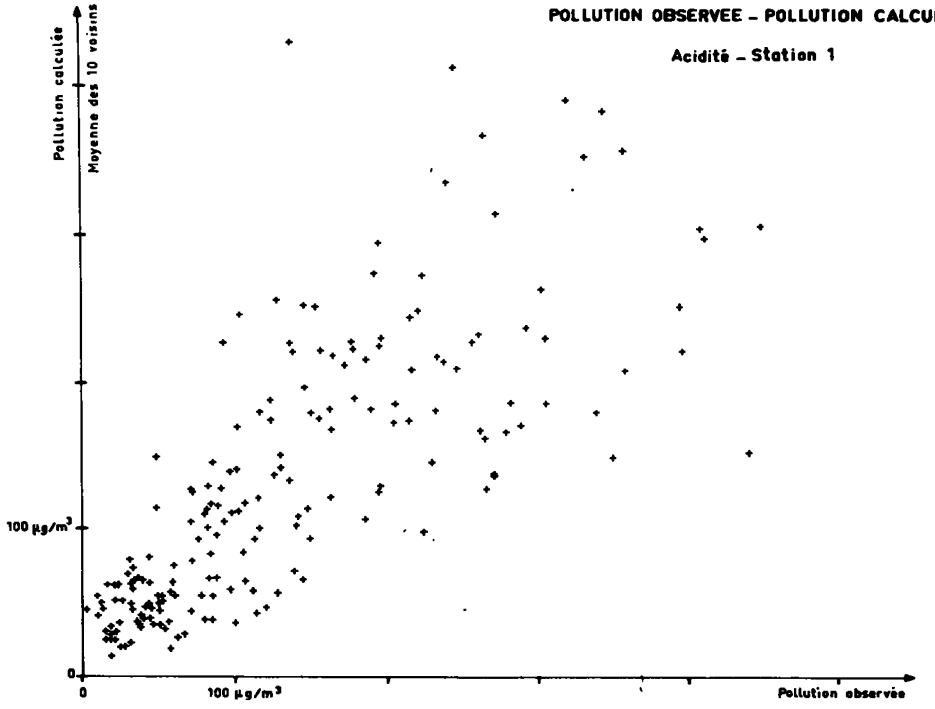

POLLUTION OBSERVEE – POLLUTION CALCULEE
Acidité – Station 1
Pollution calculée
Moyenne des 10 voisins
100 µg/m³
0
0
100 µg/m³
Pollution observée

Fig. 4

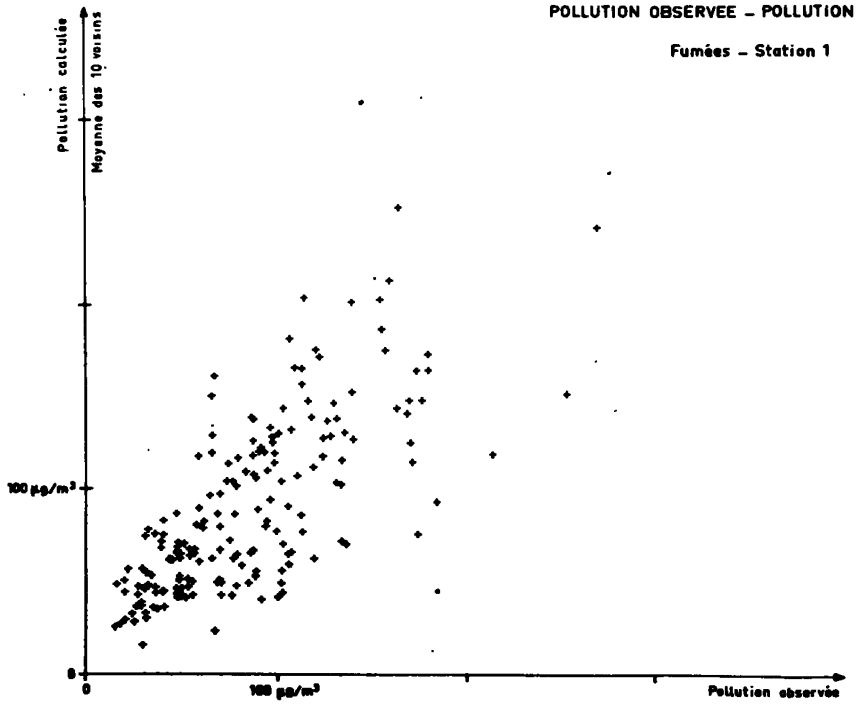

POLLUTION OBSERVEE – POLLUTION CALCULEE
Fumées – Station 1
Pollution calculée
Moyenne des 10 voisins
100 µg/m³
0
0
100 µg/m³
Pollution observée

Fig : 5

Les figures 4 et 5 présentent, pour chacun des deux polluants en une station, le nuage des points représentant les couples "valeur calculée, valeur mesurée" afférents aux 200 jours considérés. Le coefficient de corrélation obtenu est égal à 0,76 pour l'acidité et à 0,69 pour les fumées. 58 % de la dispersion du taux d'acidité apparaît donc expliquée, au poste 1, par l'ensemble des paramètres pris en compte ; 48 % de la dispersion pour la teneur en fumées. Ce qui donne des résultats analogues à ceux obtenus par régression linéaire par rapport aux deux premiers facteurs. Toutefois, au § 3.2. du prochain article, on verra (cf. note) un exemple où la régression par boule l'emporte sur la régression linéaire.

3. *Méthodes pour établir un sous-réseau optimal*

Si l'on ne se contente pas d'un calendrier et d'un bulletin de la Météorologie Nationale pour estimer la pollution selon la méthode du § 2.2.1., on se demandera quel est le sous-ensemble J_r de postes conservant au mieux l'information fournie par le réseau initial.

L'étude a été faite à partir des taux d'acidité observés. Parmi les 10 ans de données disponibles, seuls les 773 jours d'observation complète seront utilisés.

L'analyse générale de ces taux a tout d'abord permis de mettre en évidence les corrélations entre les divers postes du réseau (§ 3.1.). Puis on a appliqué diverses méthodes à la recherche d'un sous-réseau optimal (§ 3.2. ; 3.3. ; 3.4.) : l'exposé de ces méthodes et leur confrontation font l'objet d'un article ultérieur.

3.1. *Comparaison des séries chronologiques de la pollution acide observée aux différentes stations*

Les représentations globales du nuage des postes fournies par diverses méthodes d'analyses concordent dans l'ensemble [figures 6, 8, 9]. Nous les présenterons successivement en les comparant.

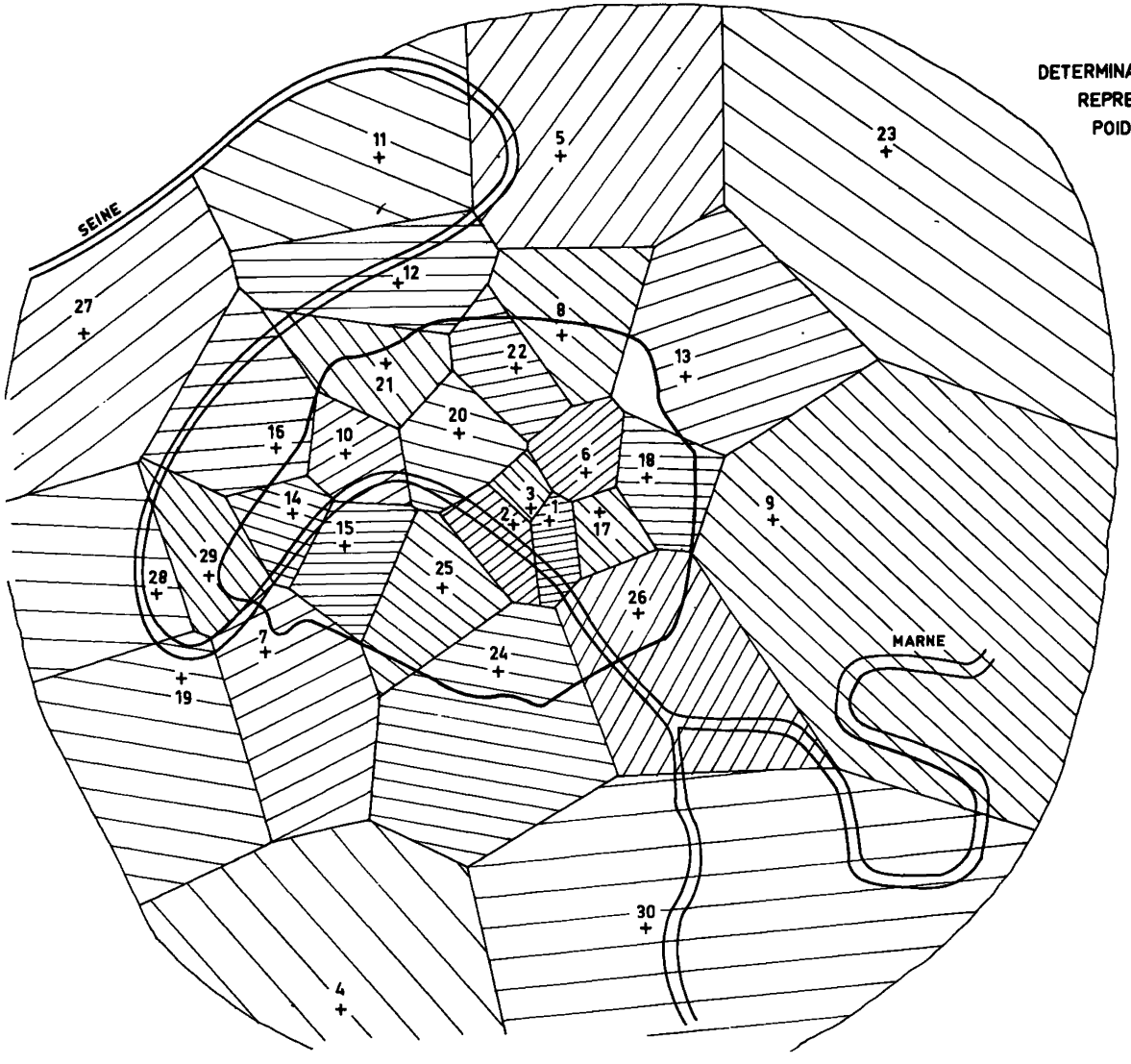
3.1.1. *Analyse en composantes principales normées des données*

On a fait l'analyse en composantes principales [référence 3] du tableau des taux centrés et réduits d'acidité forte (on reviendra au § 3.1.3. sur cette méthode).

Le premier facteur représente une part d'inertie très importante (82 %). Les stations sont toutes fortement corrélées avec la première composante principale. Celle-ci reflète le niveau de la pollution de fond dans l'agglomération. Le deuxième facteur (6,5 % de l'inertie) est d'origine géographique : sur le deuxième axe, les stations se répartissent suivant une opposition Nord-Est ≠ Sud-Ouest. De même, le troisième facteur représente une opposition Sud-Est ≠ Nord-Ouest. En sorte que le plan 2 x 3 restitue approximativement la carte des stations. Sur la figure 6 on a donc représenté la projection des stations dans le plan des composantes principales 2 x 3. La coordonnée d'une station sur chacun des axes est égale au coefficient de corrélation entre la station et la composante principale correspondante. En vue de comparaison avec l'analyse des correspondances, on a joint dans leur ordre les stations périphériques par une ligne brisée [cf. fig. 6, 8, 9].

3.1.2. *Remarque sur l'interprétation des résultats - Analyse des correspondances des tableaux de pollution - Pondération des postes*

Les trente stations du réseau ne sont pas réparties régulièrement dans l'agglomération. Un assez grand nombre d'entre elles sont serrées



DETERMINATION DES AIRES GEOGRAPHIQUES
REPRESENTEES PAR CHAQUE POSTE
POIDS PROPORTIONNEL ASSOCIE

Poste	Poids
1	1
2	1
3	1
4	20
5	20
6	2
7	7
8	4
9	38
10	2
11	9
12	6
13	9
14	1
15	2
16	5
17	1
18	2
19	11
20	2
21	3
22	2
23	27
24	11
25	3
26	10
27	15
28	8
29	3
30	18

Fig : 7

au centre de la ville, ce qui peut fausser l'interprétation des résultats de l'analyse précédente et expliquer en partie l'importance du premier facteur. On a donc cherché si l'on obtenait des résultats comparables en donnant aux postes géographiquement isolés un poids plus important que celui affecté aux postes plus groupés. L'analyse des correspondances du tableau de données a permis de faire cette étude.

Cette analyse a été faite, à titre comparatif, d'une part sans effectuer cette pondération, d'autre part en affectant chaque poste d'un poids proportionnel à l'aire géographique qu'il est censé représenter (cette aire est délimitée par les médiatrices des segments joignant le poste considéré aux postes limitrophes : (figure 7)).

Les résultats de la première analyse (sans pondération des postes) sont donnés sur la figure 8 ; ceux de la deuxième (avec pondération) sont sur la figure 9. Il apparaît d'abord clairement qu'à ce niveau de l'analyse (il n'en sera pas de même ensuite, cf. § 3.4.) les coefficients de pondération ne jouent qu'un rôle secondaire : les nuages des figures 8 et 9 diffèrent peu. De plus ces nuages reproduisent approximativement la disposition des stations sur la carte de la région parisienne (figures 1, 6 et 7) : on s'en assurera en suivant sur les figures 8 et 9 le contour {19, 4, 30, 9, 23, 5, 11, 27, 28} des stations périphériques : sur la figure 9 (analyse pondérée) l'accord est particulièrement satisfaisant entre analyse factorielle et disposition géographique.

Les groupements de stations observés dans le plan 1 x 2 sont de plus confirmés par la classification automatique présentée au § 3.2.2. [figures 10 et 10 bis ; qui illustrent un prochain article].

3.1.3. Comparaison entre méthodes d'analyse

Quant à la comparaison avec l'analyse en composantes principales du § 3.1.1., on voit d'abord que la représentation géographique apparue sur le plan 2 x 3 au § 3.1.1. [figure 6], se retrouve exactement sur le plan 1 x 2 issu des deux analyses de correspondances [figures 8 et 9]. Pour comprendre ce qu'est devenu en analyse des correspondances le facteur général (1er facteur) de l'analyse en composantes principales, il faut avoir présents à l'esprit les modèles différents auxquels se réfèrent ces deux analyses.

En composantes principales, on associe à chaque station s une variable $a_s(i)$ (fonction du jour i), obtenue en transformant linéairement les taux x(i,s) d'acidité mesurés en s afin que $a_s(i)$ ait moyenne nulle et variance 1 (sur l'ensemble des jours). Les composantes principales $\{C_1(i), C_2(i), \ldots, C_{30}(i)\}$ sont des fonctions du jour i deux à deux non corrélées ayant moyenne nulle et variance 1 ; elles sont obtenues en combinant linéairement les variables $a_s(i)$, de telle sorte que dans leur ensemble les $a_s(i)$ soient approchés au mieux par des combinaisons linéaires des premières composantes principales ; les coefficients de cette approximation étant directement lisibles sur les axes factoriels : i.e si la station s a pour coordonnée $Fac_1(s) = 0,7$ sur le 1er axe et $Fac_2(s) = 0,4$ sur le 2ème axe, cela signifie que la meilleure représentation de $a_s(i)$ en combinaison linéaire des deux premières composantes principales est $0,7\ C_1(i) + 0,4\ C_2(i)$. Dans cette formule, le coefficient $Fac_1(s)$ n'est autre que la corrélation entre $a_s(i)$ et $C_1(i)$; de même pour $Fac_2(s)$, $Fac_3(s)$, etc... ; l'on a :

$$a_s(i) \approx Fac_1(s)\ C_1(i) + Fac_2(s)\ C_2(i) + Fac_3(s)\ C_3(i)$$

Dans un tel modèle, on conçoit que le 1er facteur soit ordinairement un facteur général ; c'est-à-dire que, par exemple dans le cas

N.B. Dans les figures ci dessous, on a joint dans leur ordre les stations périphériques

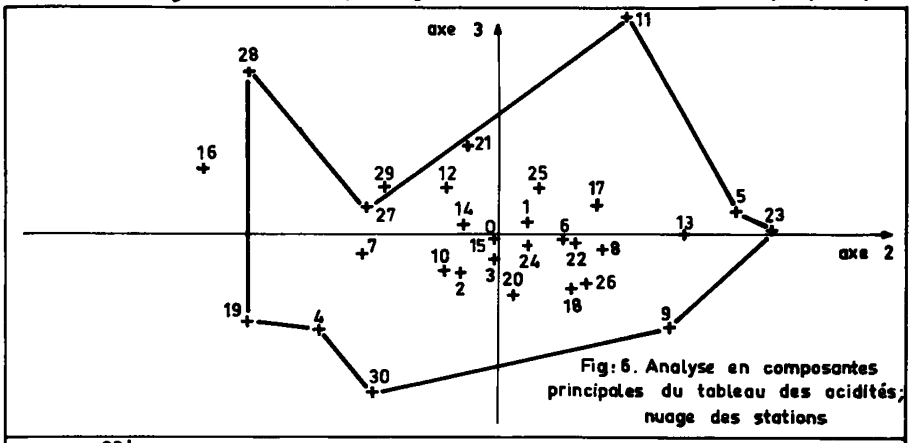


Fig: 6. Analyse en composantes principales du tableau des acidités; nuage des stations

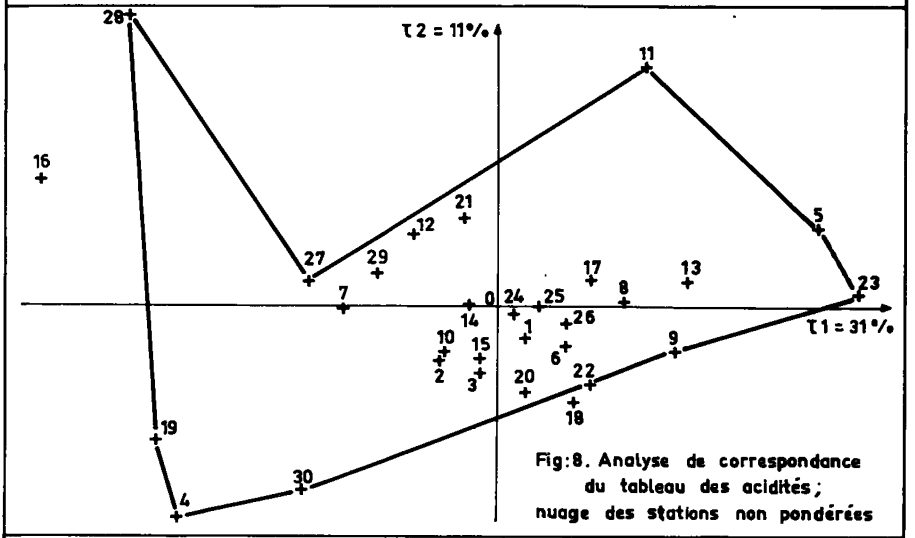


Fig: 8. Analyse de correspondance du tableau des acidités; nuage des stations non pondérées

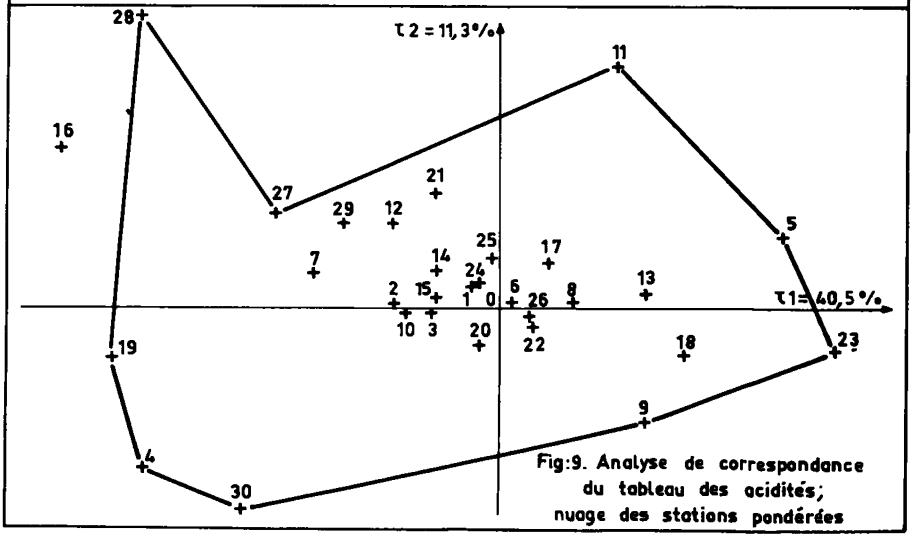


Fig: 9. Analyse de correspondance du tableau des acidités; nuage des stations pondérées

particulier traité ici, $C_1(i)$ sera le taux de pollution moyen au jour i (à une transformation linéaire près destinée à donner à $C_1(i)$ moyenne et variance 1 sur l'ensemble des jours). Les facteurs de rang supérieur ou égal à 2 rendent ensuite compte de la diversité entre stations.

En analyse de correspondance, la formule d'approximation s'écrit :

$$k(i,s) = [k(i)\ k(s)/k][1 + \lambda_1^{-1/2}\ F_1(i)\ G_1(s) + \lambda_2^{-1/2}\ F_2(i)\ G_2(s) + \ldots]$$

Ici, le taux de pollution moyen au jour i (ou poids du jour i) est représenté par la loi marginale [k(i)/k] ; il n'apparaît donc pas dans les facteurs qui expriment exclusivement la diversité des profils. De même, le poids de la station s -k(s)- intervient en tête de la formule d'approximation : c'est la symétrie bien connue entre lignes et colonnes, avec séparation nette des effets de poids et des effets de forme (ou de profils). En analyse en composantes principales (du moins si la normalisation est faite comme ici), le poids de la station disparaît de l'analyse (c'est l'essence même de la normalisation : moyenne 0 et variance 1 pour toute station s) ; mais il peut être rétabli dans une formule finale de reconstitution des données car, puisque

$$a_s(i) = \alpha_s\ x(i,s) + \beta_s \quad \text{(transformation linéaire)}$$

on a aussi :

$$x(i,s) = (a_s(i) - \beta_s)/\alpha_s$$

Dans la présente étude, l'importance que revêt le facteur général dans l'analyse en composantes principales (pourcentage 82 %) exprime la similitude de profil entre les stations, particulièrement celles du bloc central. En analyse de correspondance, cette similitude de profil se manifeste par la faiblesse relative des valeurs propres (0,04 ; 0,01 ; 0,009, etc...) ; de plus les stations centrales sont ici comme là très proches sur les axes. Et après le niveau général, vient plus ou moins heureusement exprimée sur les graphiques, la diversité de profils entre stations : axes 2, 3, etc..., de l'analyse en composantes principales et axes 1, 2, etc... de l'analyse de correspondance.

Ce n'est pas le lieu de reprendre dans toute son ampleur le débat entre écoles d'analyse factorielle (pour un exposé chronologique de ce débat, cf. Histoire et Préhistoire de l'Analyse des Données ; Cahier N° 4 ; V, Vol II, Cahier N° 1). Mais d'une part il fallait expliquer pourquoi, en première approximation, on doit attendre les facteurs de formes Fac_α (de rang $\alpha \leq 2$) issus de l'analyse en composantes principales au rang $(\alpha - 1)$ de l'analyse de correspondance. D'autre part le problème des pondérations, qu'on retrouvera au § 3.4., est d'une telle importance ici (puisqu'il faut représenter une région par un échantillon de stations) qu'on doit voir clairement où et comment celles-ci sont introduites dans l'analyse et influent sur les résultats.

BIBLIOGRAPHIE

[1] BENZECRI et Collaborateurs (1973) :

L'analyse des données. DUNOD.

[2] LEBEAUX (1974) :

Programmes de régression et de classification utilisant la notion de voisinage.(*)
Thèse de 3ème cycle. Université de PARIS VI.

[3] CAILLIEZ - MAILLES - NAKACHE - PAGES (1971) :

Analyse des données multidimensionnelles. Centre d'Etudes Economiques d'Entreprises.

(*) La notice d'une version de ce programme sera publiée dans le prochain cahier.