Statistique et analyse des données

Suzanne Winsberg

Une utilisation de métrique particulière destinée à l'analyse de fonctions échantillonnées

Statistique et analyse des données, tome 12, n° 3 (1987), p. 97-117.

<http://www.numdam.org/item?id=SAD_1987__12_3_97_0>

Statistique et Analyse des Données
1987 - Vol. 3 n° 3 p. 97-117

UNE UTILISATION DE METRIQUE PARTICULIÈRE DESTINÉE À L'ANALYSE DE FONCTIONS ÉCHANTILLONNÉES

Suzanne WINSBERG

Résumé : *Pour effectuer avec profit une analyse en composantes principales de fonctions échantillonnées il est utile d'employer des métriques destinées à filtrer les données d'une variation connue de manière à découvrir des variations moins évidentes. La discussion est illustrée par des exemples.*

Abstract : *When analyzing data which consists of sampled functions it is often preferable to use metrics designed to filter the data, removing knowun or obvious variation in order to uncover unknown or subtle variation. The discussion is motivated by two exemples.*

Mots clés : *A.C.P. - métrique - filtre - fonctions échantillonnées.*

Manuscrit reçu le 15 novembre 1986
Révisé le 2 novembre 1987

Introduction

Considérons des donnees provenant d'un echantillonnage de n fonctions du temps f_i (t) périodiques ou non. Chaque fonction f_i est échantillonnée en p points, t_j , donnant lieu à un vecteur-ligne X_i formé de valeurs $X_i(t_j) = f_i(t_j)$. Ce vecteur X_i est appelé fonction échantillonnée. Donc les données forment une matrice de valeurs $X_i^j(t_j)$ de dimensions n x p.

Des données présentent souvent une variation évidente qui peut être décrite par un espace linéaire de fonctions qu'on dénote S. Dans un tel cas il peut être important de rechercher une variation additionnelle plus subtile pour mettre en évidence un autre aspect des données. Nous allons montrer que pour analyser ce type de données il est préférable d'enlever la variation évidente ou connue pour mieux découvrir la variation inconnue. Considérons l'ensemble de données périodiques formé des températures mensuelles de trente-deux villes françaises. Les données ont l'aspect de cosinus échantillonnés avec leur maxima en juillet. La variation évidente est donc la moyenne et un multiple du cosinus, c'est-à-dire, $S = \{1, \cos\}$. Il convient alors d'enlever cette variation pour étudier le résidu. Le but est donc de filtrer les données des types de variations connues pour découvrir une variation plus subtile.

Considérons des méthodes pour analyser X comme l'A.C.P., la classification automatique ou l'analyse discriminante. Toutes ces méthodes se basent sur une métrique quadratique entre des vecteurs-lignes de p dimensions. Donc en appliquant toutes ces méthodes nous incorporons un "filtre" dans la métrique pour enlever les variations évidentes afin d'atteindre le but désiré.

Des techniques de filtrage

BESSE et RAMSAY (1986) ont etudié le "filtrage" pour ces techniques et ont developpé une famille de métriques qui atteignent ce but. Ils ont travaille sur l'A.C.P., mais leurs idees s'appliquent également à d'autres méthodes basees sur les métriques. Conceptuellement l'approche de BESSE et RAMSAY est basée sur l'interpolation ou le lissage des données, quoique pratiquement l'interpolation ou le lissage ne soit pas forcément faite. Quand il s'agit de fonctions du temps il peut convenir de recourir à l'interpolation et/ou le lissage sans aucun autre but.

Lorsqu'on dispose d'un ensemble de fonctions echantillonnées, on peut effectuer une A.C.P. ordinaire sur X pour obtenir r fonctions et interpoler les composantes principales. BESSE (1979) a suggéré de renverser les étapes. C'est-à-dire d'interpoler les lignes de X pour former des fonctions continues et de faire l'A.C.P. de ces fonctions pour obtenir r composantes principales. Si l'interpolation est effectuée par des splines, BESSE (1979) a développé une métrique Q telle que une A.C.P. discrète sur X basée sur la métrique Q est équivalente à une A.C.P. sur les fonctions interpolées basée sur la métrique I . En pratique on peut obtenir ces résultats sans renverser les étapes en faisant l'A.C.P. de la matrice X basée sur la métrique Q . On obtient r composantes principales, qui peuvent être par la suite interpolées. Ces arguments peuvent se généraliser pour inclure le lissage avec peu de difficulté.

L'approche de BESSE et RAMSAY commence par le choix d'un opérateur différentiel qui annule les fonctions qu'on désire enlever. Dans l'exemple des températures mentionné ci-dessus l'opérateur approprié est $D + \frac{1}{\omega^2} D^3$ où D est $\frac{d}{dt}$. Travaillant maintenant par normes plutôt que par métriques

l'approche de BESSE et RAMSAY est de calculer et utiliser la norme définie par ces étapes :

(i) interpoler ou lisser la fonction échantillonnee

(ii) appliquer l'opérateur

(iii) utiliser la norme habituelle de l'espace fonctionnel L_2 .

Dans leur méthode il faut considérer le noyau reproductif qui est associé à l'opérateur différentiel et le calculer.

BESSE et RAMSAY utilisent des splines qui minimisent la norme L pour l'interpolation. La fonction interpolée f_i pour les points (t_i^j, X_i^j) minimise $\int_0^T [f_i(t)]^2 \, dt$ sous les contraintes $f_i(t_j) = X_i^j$. Notons que l'opérateur différentiel L qu'ils emploient pour l'interpolation est le même opérateur F qu'ils ont choisi comme filtre. Malheureusement les opérateurs nécessaires comme filtres donnent parfois lieu à des fonctions interpolées ayant des propriétés indésirables et malencontreuses. Nous recommandons donc de choisir L séparément de F . ($L = D^2$ donnent lieu à des splines cubiques et cet opérateur serait désirable dans beaucoup de cas). Il est possible de généraliser leur méthode et de trouver Q de façon a englober la situation où L est différent de F ; mais cela devient beaucoup plus compliqué. De plus pour les fonctions périodiques comme dans le cas des températures mensuelles citées ci-dessus il est désirable que la fonction interpolée soit périodique et ait le même ordre de continuité au point de retournement qu'elle possède aux noeuds. Ceci ajoute des complications a leur méthode.

Par contre le filtrage peut s'introduire sans interpolation ou lissage. De plus en géneral l'interpolation seule sans le filtage change tres peu la metrique. WINSBERG et KRUSKAL (1986) stimulés par BESSE et RAMSAY ont développé deux approches pour filtrer beaucoup plus simples sans recours à l'interpolation/ou lissage. Elles s'appliquent egalement bien et facilement à des donnees périodiques ou non. La première approche de WINSBERG et KRUSKAL est l'analogue discrète de l'approche de BESSE et RAMSAY et consiste à filtrer par l'opérateur des différences qui correspond à l'opérateur différentiel de BESSE et RAMSAY. La seconde approche de WINSBERG et KRUSKAL filtre par projection sur l'espace orthogonal aux fonctions qu'on désire enlever.

On peut illustrer la première approche (approche par opérateur de différence) comme suit : pour annuler S = {1, cos} l'opérateur différentiel est $D + \frac{1}{\omega^2} D^3$ qui annule l'espace plus grand {sin, cos, 1}. Il s'agit d'employer le filtre $F = \Delta + \frac{1}{\omega^2} \Delta^3$ où pour les fonctions périodiques ou non-périodiques respectivement quand p = 4

$$\Delta = \begin{bmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \Delta = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

Dans la seconde approche (approche par operateur de projection) on choisit une base pour S . Ces fonctions n'ont pas besoin d'être orthogonales. Ici il s'agit de former la matrice K dont les lignes sont formées des p valeurs des fonctions à enlever que nous appelons les fonctions nulles. Le filtre est alors défini comme $F = I - K^0 \equiv I - K^+ K$. Un avantage de l'approche par opérateur de projection est le suivant : il n'est pas nécessaire de trouver un opérateur qui annule S ; surtout il n'est pas nécessaire de

filtrer des autres fonctions supplémentaires comme il est requis parfois dans l'approche de l'opérateur de différence ou un ensemble de fonctions est annulé. Cet ensemble peut contenir des fonctions autre que celles qu'on désire annuler, c'est-à-dire des fonctions supplémentaires.

Dans ces deux approches la pratique est la suivante : on effectue l'A.C.P. de la matrice des données filtrées XF basée sur la métrique habituelle (I) ou de manière équivalente on effectue l'A.C.P. sur la matrice de données X basée sur la métrique $Q = FF'$. L'A.C.P. par rapport à une métrique Q peut s'ennoncer comme suit : étant donné X (n par p) de rang complet avec $(n \geq p)$ trouvez A et B qui minimisent $\| X - AB' \|$ sous les contraintes

(i) A est n par r , B est p par r , avec $r < p$;

(ii) B est Q-orthogonal (c'est-à-dire $B'QB = I$) ;

(iii) les colonnes de A sont dans un ordre décroissant. Si Q est singulier la solution n'est pas unique. Pour trouver A et B' , mettons $Y = X E \Lambda$ et prenons la décomposition en éléments simples PDU' de Y . Alors $A = PD$ et $B' = U'\Lambda^{-1} E$, où $M = E \Lambda E'$.

Dans l'approche par opérateur de projection la procédure d'annulation permet d'obtenir précisément le résidu la différence entre X et sa projection sur S . Cependant dans l'approche de l'opérateur différentiel et dans l'approche de l'operateur de différence cette procédure ne laisse pas Res mais l'image de Res sous F .

Exemples

Pour presenter ces procédures de filtrage en pratique nous recourons a trois ensembles de données dont deux sont des données artificielles construites pour étudier les techniques et le troisième est l'ensemble de données formé des températures mensuelles des trente-deux villes de France mentionné plus haut. Le premier ensemble de données artificielles est formé de 14 fonctions $a_i \cos\omega t$, échantillonnées à 12 points avec le maximum au 7ème point auxquelles on ajoute des trapèzes de hauteurs variables dont certaines sont égales à zéro, débutant au point 1 , terminant au point 6 , avec un maximum entre le 3ème et le 4ème point. Le deuxième ensemble se déduit du premier en ajoutant un bruit gaussien dont l'écart-type est de 15 % de l'écart-type des données. Pour le premier ensemble les deux approches retrouvent également bien la forme du trapèze (voir figure 1a) et la position de chaque objet (voir figure 1b). Ici il faut se rappeler qu'ayant éliminé le cosinus et la constante seule la première composante principale qui devrait avoir la forme du trapèze est intéressante. Pour le deuxième ensemble de données artificielles qui contient du bruit on peut remarquer (voir figures 2 et 3) que la méthode par projection est nettement supérieure. Pour ces mêmes données on peut aussi remarquer que l'A.C.P. ordinaire n'est pas aussi efficace que la méthode de filtrage par projection pour découvrir la variation inconnue (voir figures 2 et 4). Donc en considérant les données réelles sur les températures mensuelles nous ne présentons que les résultats pour l'approche de l'opérateur par projection. L'ensemble des donnees est présentée dans le Tableau 1. Comme on peut voir en regardant la figure 5a, une fois qu'on a enlevé les composantes évidentes notamment la constante et le cosinus on remarque trois phénomènes. La première composante principale correspond à un

printemps chaud et ce phenomene se trouve caracteristique des villes loin des zones littorales plutôt que celles qui se trouvent dans les zones côtières comme on le voit dans la figure 5b. La seconde composante principale correspond à un printemps précoce et ce phenomene est caractéristique de villes qui se trouvent dans une bande de lattitude étroite notamment Embrun, Grenoble, Toulouse, Bordeaux et Biarritz, opposé par une autre bande de latitude plus au nord contenant Brest, Nancy, Lille, Rouen et Paris (voir figure 5c). La troisième composante principale correspond à une chute aigue de température de novembre à décembre. Cette chute sépare Embrun, la seule ville d'une altitude élevée du reste (voir figure 5b ou 5c). Pour mieux situer ces villes une carte de la France est montrée en figure 6. En conclusion le filtrage s'avère utile quand le but est de découvrir de la variation inconnue dans un ensemble de données formé de fonctions lisses echantillonnées et que la variation connue peut être modélisée facilement.

Tableau 1 Température mensuelle moyenne (.1° C) de villes françaises.

Ville	Symbole	Jan	Fev	Mar	Avr	Mai	Jun	Jui	Aoû	Sep	Oct	Nov	Dec
Ajaccio	aj	77	87	105	126	159	198	220	222	203	163	118	87
Angoulême	ang	42	49	79	104	136	170	187	184	161	117	76	49
Angers	an	46	54	89	113	145	172	195	194	169	125	81	53
Besançon	bes	11	22	64	97	136	169	187	183	155	104	57	20
Biarritz	bia	76	80	108	120	147	178	197	199	185	148	109	82
Bordeaux	bor	56	66	103	128	158	193	209	210	186	138	91	62
Brest	bre	61	58	78	92	116	144	156	160	147	120	90	70
Cler-Ferr	clf	26	37	75	103	138	173	194	191	162	112	66	36
Dijon	dij	13	26	69	104	143	177	196	190	159	105	57	21
Embrun	em	5	16	57	90	130	164	189	183	153	101	46	5
Grenoble	gre	15	32	77	106	145	178	201	195	167	114	65	23
Lille	lil	24	29	60	89	124	153	171	171	147	104	61	35
Limoges	lim	31	39	74	99	133	168	184	178	153	107	67	38
Lyon	lyo	21	33	77	109	149	185	207	201	169	114	67	31
Marseille	mar	55	66	100	130	168	208	233	228	199	150	102	69
Montpellier	mon	56	67	99	128	162	201	227	223	193	146	100	65
Nancy	ncy	8	16	55	92	133	165	183	177	147	94	52	18
Nantes	nan	50	53	84	108	139	172	188	186	164	122	82	55
Nice	nic	75	85	108	133	167	201	227	225	203	160	115	82
Nîmes	nim	57	68	101	130	166	208	236	229	197	146	98	65
Orléans	orl	27	36	69	98	134	166	184	182	156	109	66	36
Paris	par	34	41	76	107	143	175	191	187	160	114	71	43
Perpignan	per	75	84	113	139	171	211	238	233	205	159	115	86
Reims	rei	19	28	62	94	133	164	183	179	151	103	61	30
Rennes	ren	48	53	79	101	131	162	179	178	157	116	78	54
Rouen	rou	34	39	68	95	129	157	176	172	150	110	68	43
St-Quen.	stq	20	29	63	92	127	156	174	174	150	105	61	31
Strasbourg	str	4	15	56	98	140	172	190	183	151	95	49	13
Toulon	tou	86	91	112	134	166	202	226	224	205	165	126	97
Toulouse	tls	47	56	92	116	149	187	209	209	183	133	86	55
Tour	tou	35	44	77	106	139	174	191	187	162	117	72	43
Vichy	vic	24	34	71	99	136	171	193	188	160	110	66	34
Moyenne		39	48	81	109	143	177	198	195	169	123	79	48

Figure 1 Solution obtenue pour une A.C.P. des données artificielles sans bruit basée sur la metrique Q = FF' où F = $I-K^0K$ et S = {1, cos}.

Figure 2 Solution obtenue pour une A.C.P. des données artificielles avec bruit basée sur la métrique Q = FF' où F = $I - K^0 K$ et S = {1, cos}.

Figure 3 Solution obtenue pour une A.C.P. des données artificielles avec bruit basée sur la métrique Q = FF' où F = $\Delta + \frac{1}{\omega^2} \Delta^3$.

Figure 4 Solution obtenue pour une A.C.P. des données artificielles avec bruit basée sur la métrique Q = I .

Figure 5 Solution obtenue pour une A.C.P. des données formee des températures mensuelles basée sur la métrique Q = FF' où F = $I - K^0 K$ et S = {1, cos}.

Figure 6 Une carte de la France.

Références

BESSE, P. (1979). Etude descriptive des processus : Approximation et interpolation. Thèse de 3è cycle. Université Paul Sabatier. Toulouse.

BESSE, P. et RAMSAY J.O. (1986). Principal components analysis of sampled functions. Psychometrika, 51, 285-311.

WINSBERG, S. et KRUSKAL J.O. (1986). Easy to generale metrics for use with sampled functions. COMPSTAT 86. Proceedings in computational statistics. Physica-Verlag.

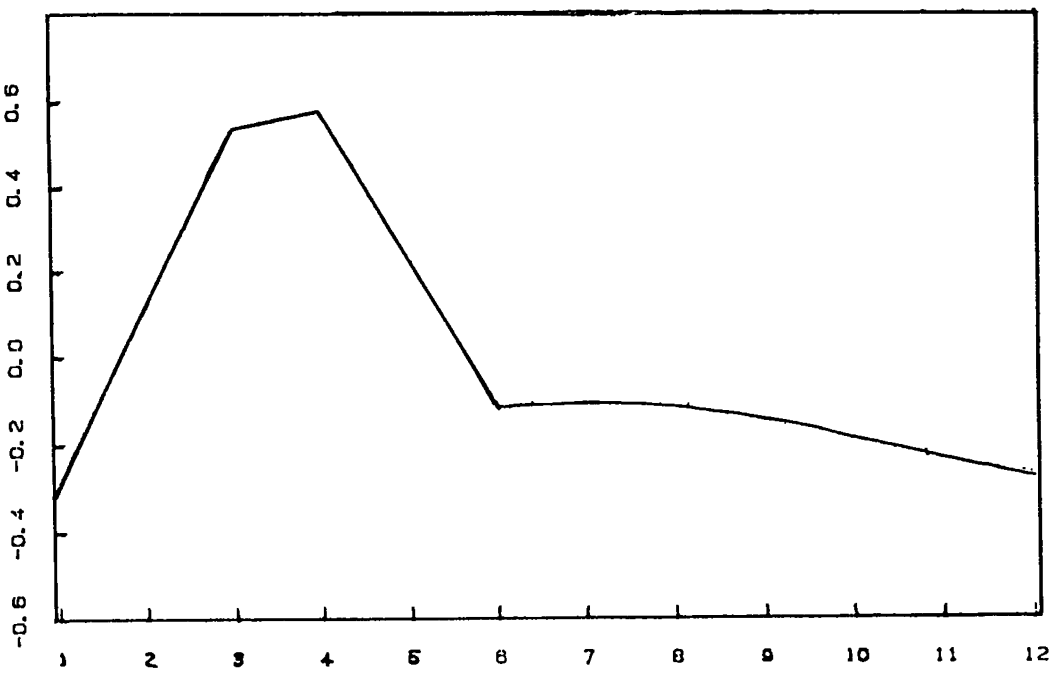

Figure 1a

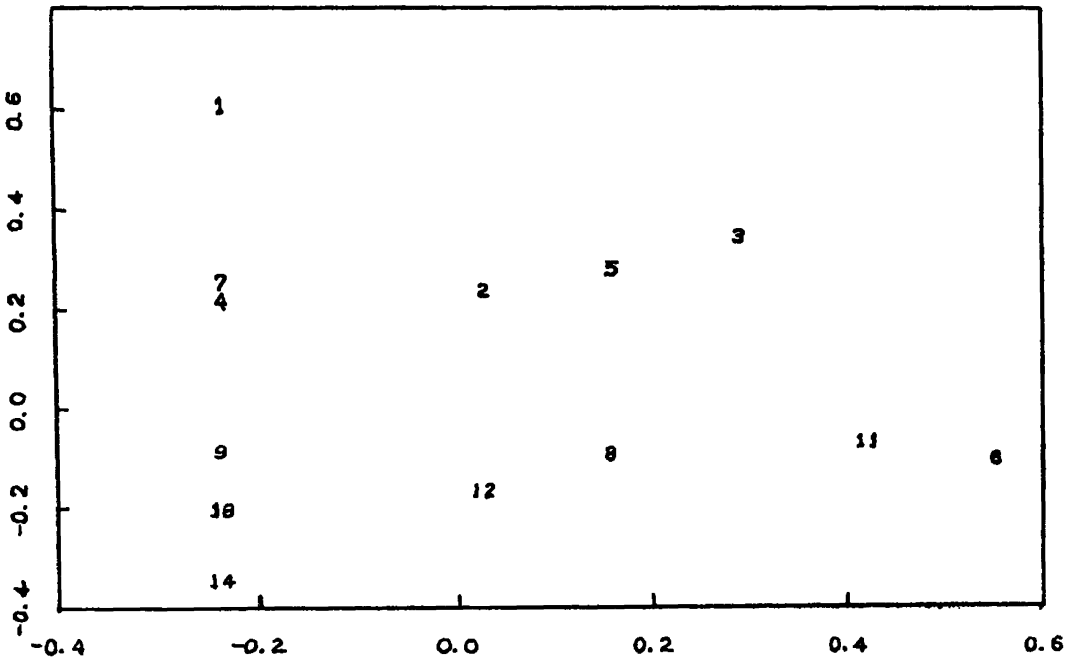

Figure 1b

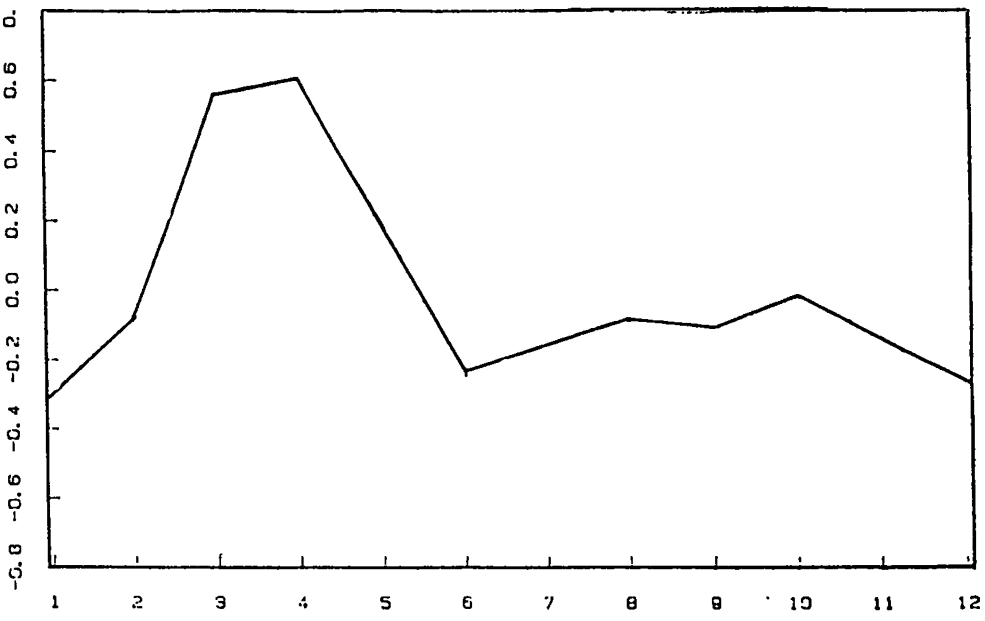

Figure 2a

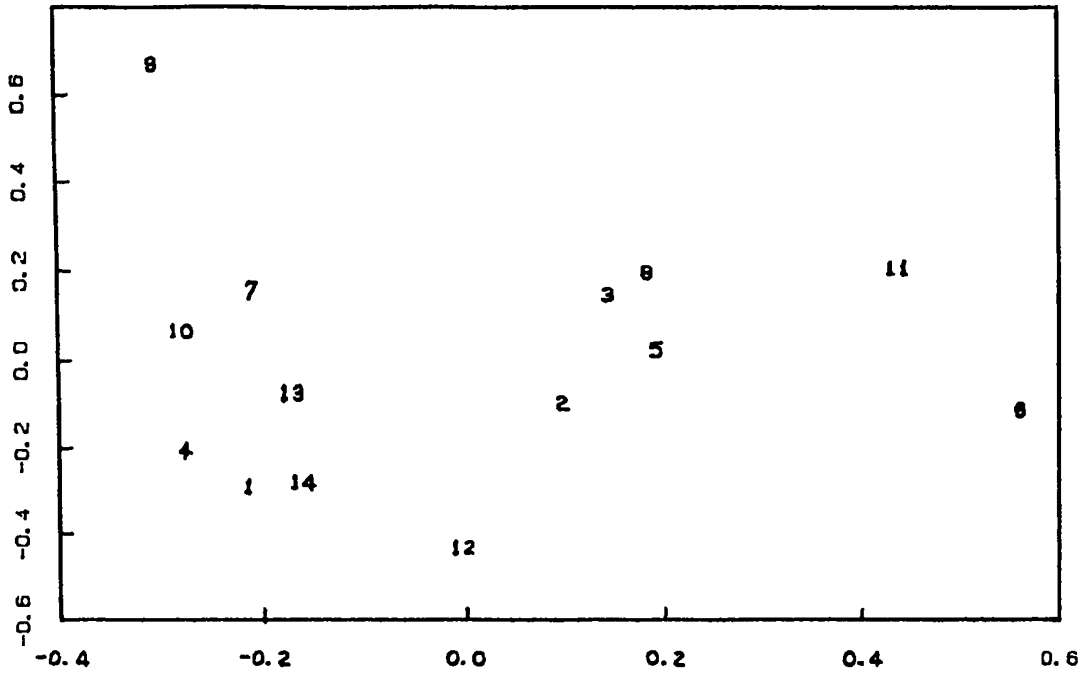

Figure 2b

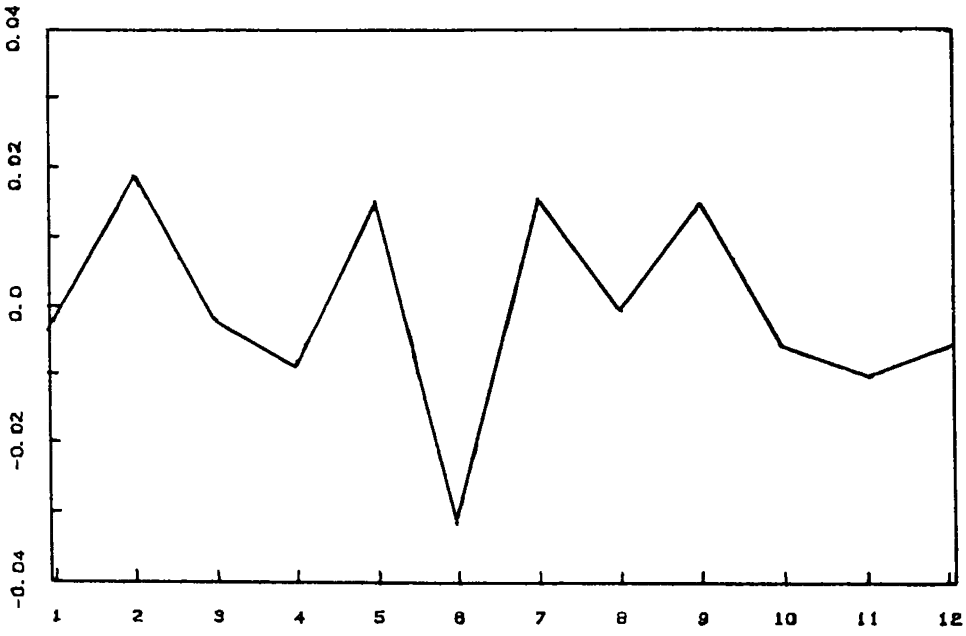

Figure 3a

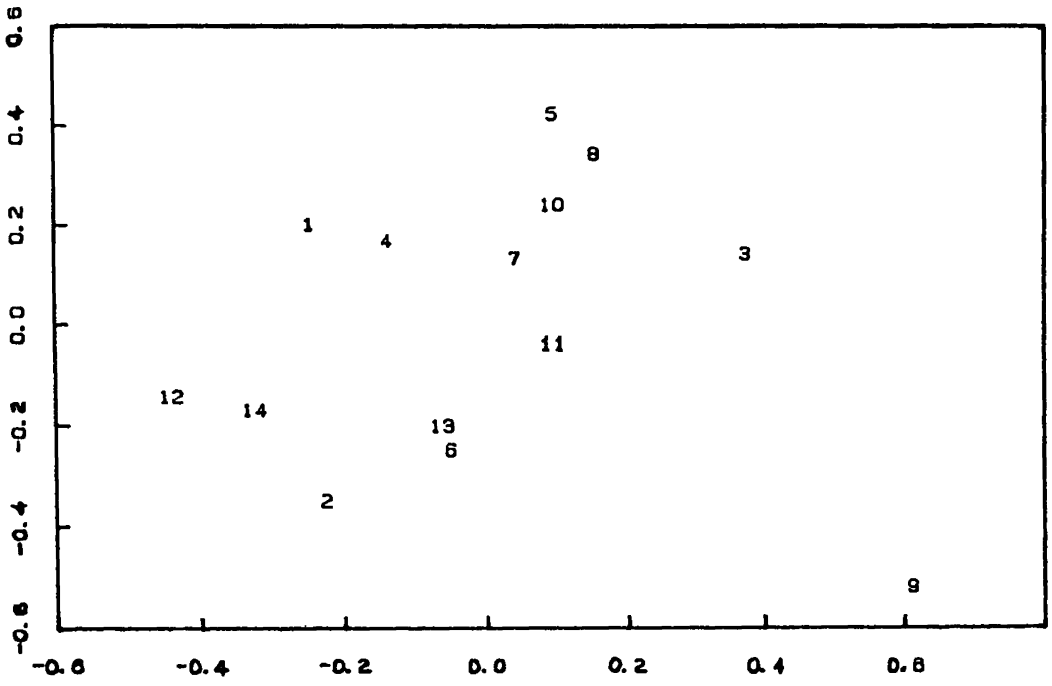

Figure 3b

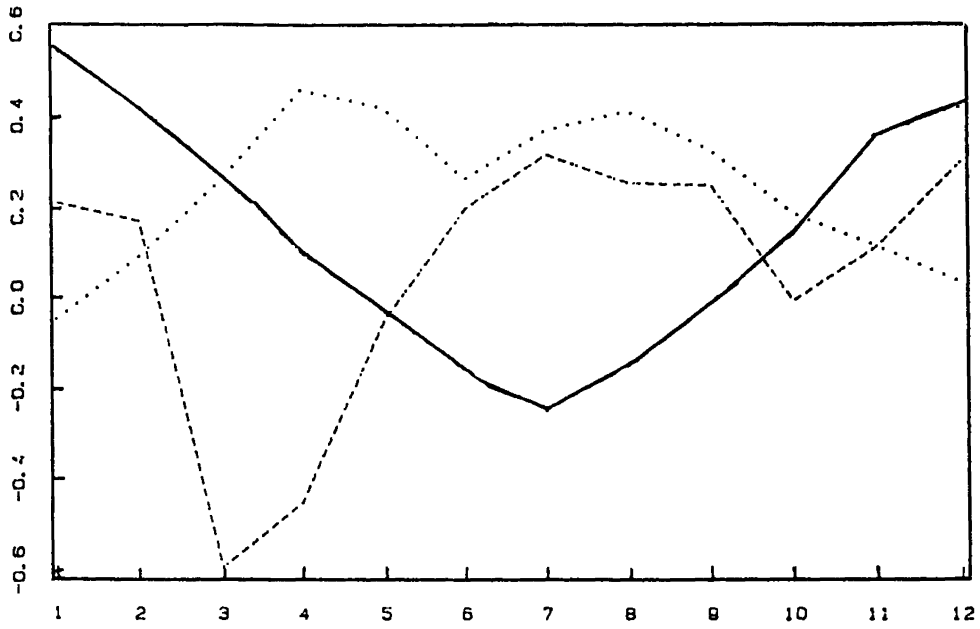

Figure 4

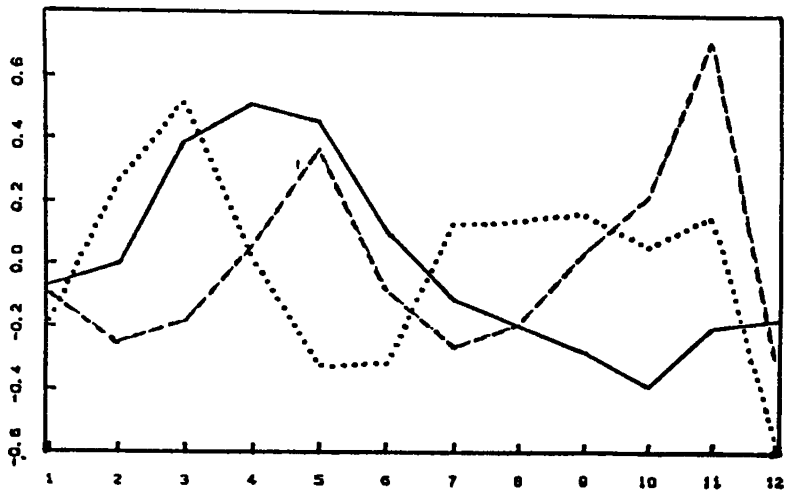

Figure 5a

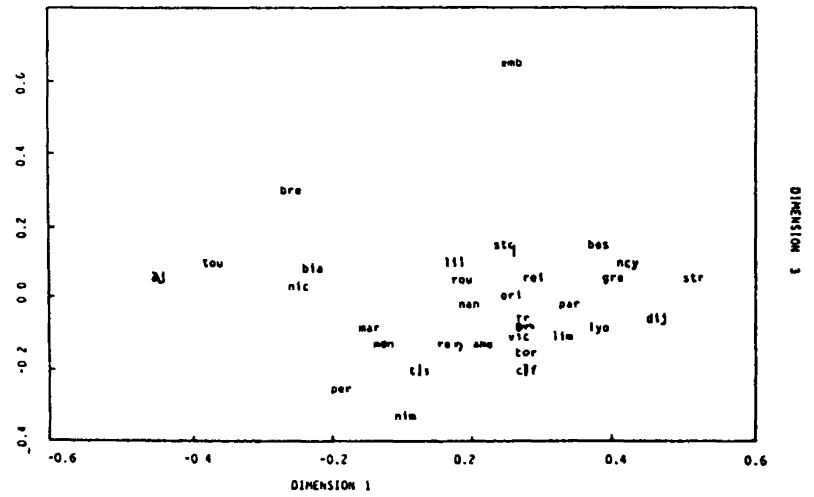

Figure 5b

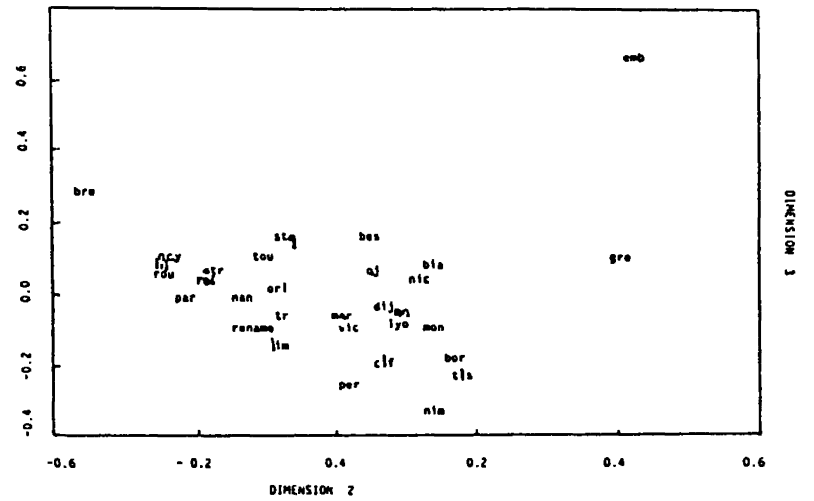

Figure 5c

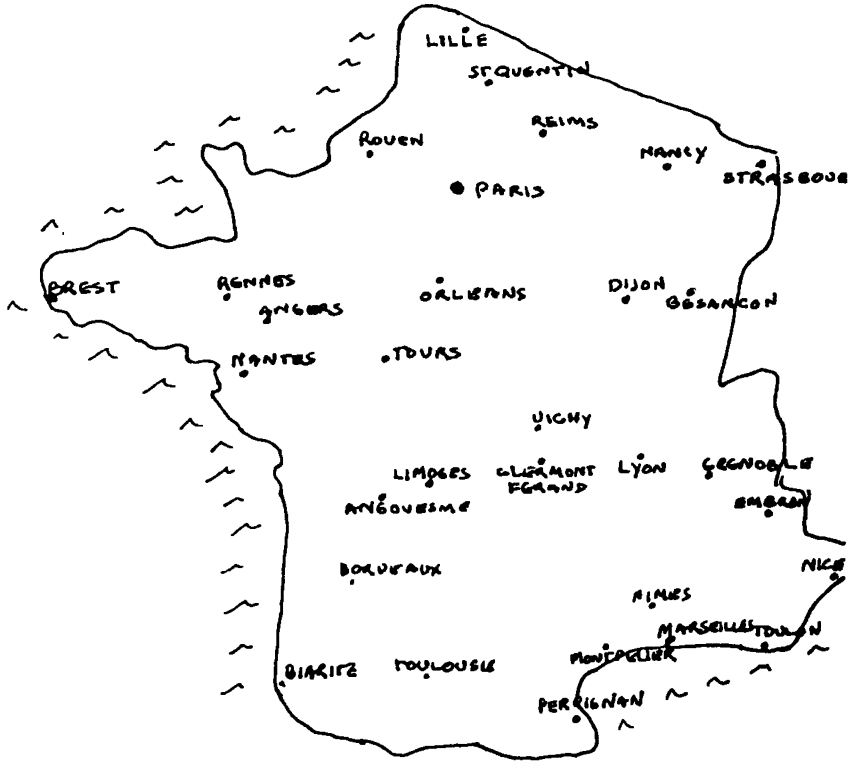
LILLE
ST QUENTIN
REIMS
ROUEN
NANCY
STRASBOURG
PARIS
BREST
RENNES
ORLEANS
DIJON
BESANCON
NANTES
TOURS
VICHY
LIMOGES
CLERMONT FERRAND
LYON
GRENOBLE
ANGOULEME
NICE
BORDEAUX
NIMES
MARSEILLE
TOULON
BIARRITZ
TOULOUSE
MONTPELLIER
PERPIGNAN

Figure 6