

REVUE DE STATISTIQUE APPLIQUÉE

A. MORINEAU

A.-E. SAMMARTINO

M. GETTLER-SUMMA

C. PARDOUX

Analyses des données et modélisation des séries temporelles. Application à la prévision des ventes de périodiques

Revue de statistique appliquée, tome 42, n° 4 (1994), p. 61-81

http://www.numdam.org/item?id=RSA_1994__42_4_61_0

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSES DES DONNÉES ET MODÉLISATION DES SÉRIES TEMPORELLES Application à la prévision des ventes de périodiques

A. Morineau(1), A-E. Sammartino(1), M. Gettler-Summa(2), C. Pardoux(2)

(1) CISIA Centre International de Statistique et d'Informatique Appliquées

(2) LISE-CEREMADE, Université Paris IX-Dauphine

RÉSUMÉ

Une même variable est observée aux mêmes dates sur un grand nombre de points d'observations, constituant ainsi autant de séries chronologiques. On pourrait chercher un modèle de prévision pour chaque série. Considérant le principe de parcimonie et le fait que l'évolution d'une série peut être expliquée par ses valeurs passées et celles des séries corrélées, on prend le parti de synthétiser l'ensemble des séries par un petit nombre de séries représentatives. On commence par une classification des points d'observation en utilisant les facteurs résultant d'une analyse des correspondances qui élimine l'effet taille. Une analyse en composantes principales sur la matrice de covariances de chaque classe permet de trouver les chroniques représentatives. Des prévisions peuvent être obtenues pour chaque point d'observation à partir d'un modèle SARIMA. L'application porte sur la prévision des ventes d'un hebdomadaire, les points d'observation des ventes étant les grossistes.

Mots-clés : *Analyse en composantes principales, Analyse factorielle des correspondances, Classification, Modèle SARIMA, Prévision, Série chronologique.*

SUMMARY

In order to develop a forecasting model on multiple time series (all of which involve the same variable and have been observed at the same dates), a forecasting model can be derived for each of the individual series. According to the parsimony principle as well as the idea that temporal changes can be explained by previous values and values of other correlated time series, it should be interesting to have only a few well chosen representative models. First, a cluster analysis is performed, using the results of a correspondence analysis, to separate the N individuals in K groups. Then a principal components analysis on the covariance matrix of each group allows us to find a few time series representative for each group. A forecasting method using a SARIMA model is applied to these time series. Our application is carried out on newspaper sales.

Keywords : *Correspondence Analysis, Clustering, Forecasting, Principal Components Analysis, SARIMA Model, Time Series.*

Introduction

L'objet de ce travail est l'élaboration d'un outil d'aide à la prévision pour un grand nombre d'unités statistiques sur lesquelles on dispose d'une connaissance venant du domaine d'application : l'existence de classes homogènes quant à leur évolution temporelle. Les caractéristiques de ces données impliquent la mise en œuvre répétée des méthodes de l'analyse des données d'une part, de la modélisation des séries temporelles d'autre part.

L'étude comporte donc plusieurs phases afin de répondre aux préoccupations de synthèse, de simplification, puis de modélisation concernant les données :

- recherche des codages les mieux adaptés pour révéler la structure de classes,
- synthèse des données par analyse factorielle,
- définition de sous-problèmes grâce à une classification validée par les variables illustratives,
- recherche de chroniques résumées pour chaque classe, par une analyse en composantes principales,
- recherche de la méthode de prévision la mieux adaptée pour les chroniques synthétiques de chaque classe afin d'obtenir des prévisions pour chaque unité statistique.

L'application permettra de montrer l'importance de l'agencement des différentes phases dans l'élaboration de l'outil d'aide à la prévision.

Une des difficultés du problème réside dans la prise en compte de l'ordre induit par le temps sur les variables actives de l'étude, car les matrices d'inertie brutes ne permettent ni de différencier les unités temporelles, ni de différencier les unités statistiques, en raison de la force de la liaison temporelle.

1. Présentation des données

L'étude concerne un hebdomadaire pour lequel on dispose des ventes réalisées par 1577 grossistes sur une période de 157 semaines. Il s'agit de 3 années entières comprenant une semaine notée 0 précédant la 1^{ère} semaine de la 1^{ère} année. Les données relatives à ces ventes se présentent sous la forme d'un tableau rectangulaire comportant 1577 lignes correspondant chacune à un grossiste et 157 colonnes correspondant aux numéros successifs de l'hebdomadaire. Chaque ligne décrit donc le suivi des ventes hebdomadaires d'un grossiste sur la période étudiée (tableau 1).

Quatre colonnes de description s'ajoutent à ce tableau des ventes : un numéro identifiant le grossiste, l'indication de son secteur géographique, la «nature» de ce grossiste (type de dépôt) et enfin une caractéristique de saisonnalité de ses ventes (voir le tableau 4). On dispose d'autres informations concernant les numéros de l'hebdomadaire ou les date correspondantes (nombre de jours de ventes du numéro en question, période éventuelle de vacances scolaires, existence d'une campagne promotionnelle, etc). Ces informations peuvent être ajoutées dans des lignes supplémentaires du tableau.

Les données, trop nombreuses pour être publiées ici, sont schématisées dans le tableau 1.

TABLEAU 1

$x_t(\omega_i)$ représente le nombre d'hebdomadaires n°t vendus par le grossiste ω_i

date	n°0	...	n°t	...	n°156
grossiste					
grossiste ω_1					
⋮					
grossiste ω_i		...	$x_t(\omega_i)$		
⋮					
grossiste ω_n					

2. Typologie des grossistes

L'objectif de cette première phase de l'étude est d'opérer une réduction du volume des données en vue de limiter le nombre de modèle de prévisions de ventes à construire. En constituant des classes de grossistes ayant des chroniques de ventes comparables, on espère pouvoir définir un modèle de prévision valable pour tous les éléments de chaque classe.

Pour que la typologie des grossistes tienne compte au mieux des évolutions les plus récentes, on trouve préférable (après d'autres expériences) de la construire à partir des données de la dernière année d'observation seulement, c'est-à-dire sur les 52 dernières semaines. Le tableau utile comporte donc en ligne les 1577 grossistes et en colonne les 52 dernières semaines d'observation.

2.1. Analyse des correspondances préalable

Pour la constitution de classes de grossistes susceptibles d'avoir le même modèle de prévision, on doit tenir compte, non pas des volumes vendus, mais du «profil» des ventes dans le temps : ces profils sont-ils semblables ? Dans le tableau des données utiles défini ci-dessus, on amène tous les grossistes à avoir le même volume de ventes annuelles en divisant chaque case par la somme des données en ligne. On définit ainsi le profil-ligne des ventes du grossiste (voir le tableau 2).

De plus on souhaite que, dans la comparaison des profils, deux dates d'observation jouent à peu près le même rôle même si les volumes moyens de ventes à ces deux dates sont très différents. On risquerait sinon de ramener la comparaison des profils à la seule comparaison des pics des profils.

TABLEAU 2

Principe du codage en «profil»

variable individu	X_{105}	...	X_t	...	X_{156}	total de la ligne
ω_1						1
\vdots			\vdots			
ω_j		...	$\frac{x_t(\omega_j)}{\sum_{t=105} x_t(\omega_j)}$...		1
\vdots						
ω_{1577}						1

L'analyse des correspondances [Benzécri, 1979] utilise une distance entre les lignes du tableau, dite distance du Khi-2, qui répond très précisément à ces critères. La distance entre grossistes est une distance entre profils, donc indépendante des volumes, et qui fait jouer à chaque date (colonne) un rôle indépendant du volume global des ventes de la date.

On donne ci-dessous l'histogramme de décroissance des valeurs propres de l'analyse. La première valeur peut paraître exagérément forte pour un tableau de cette taille (75% de la variance totale). En réalité, elle est essentiellement la traduction de la très forte corrélation temporelle des données.

Les trois premiers facteurs de l'analyse rendent compte de près de 88% de la dispersion des valeurs observées des ventes sur les 52 semaines de l'année. Les 12% restant se répartissent sur les 10 axes suivant. On admettra que l'essentiel des liaisons structurelles entre les données est contenu dans cet espace des trois premières dimensions (ce qui revient à considérer comme *perturbations aléatoires* ce qui est ajouté aux données par les dimensions ultérieures). En ignorant les dimensions ultérieures, on réalise une sorte de «lissage» des données ne retenant que la partie la plus facile à comparer ou à modéliser.

EDITION DES VALEURS PROPRES : AFC sur 1577 grossistes et 52 numeros (1992)

TRACE DE LA MATRICE : .0329

HISTOGRAMME DES 13 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE
1	.0249	75.61	75.61
2	.0028	8.44	84.05
3	.0012	3.68	87.73
4	.0006	1.77	89.50
5	.0004	1.36	90.86
6	.0003	.98	91.83
7	.0003	.81	92.64
8	.0002	.60	93.24
9	.0002	.49	93.73
10	.0001	.45	94.17
11	.0001	.39	94.57
12	.0001	.35	94.91
13	.0001	.28	95.18

2.2. Typologie des grossistes

La matrice initiale de dimension 1577 par 52 est ramenée à une matrice 1577 par 3 à la suite de l'analyse des correspondances. C'est sur ce tableau des données «lissées» que la classification sera réalisée. En utilisant la distance euclidienne usuelle sur les coordonnées factorielles de l'analyse des correspondances, on évalue entre les grossistes la distance entre profils de ventes telle qu'on l'a souhaitée et définie plus haut (distance du Khi-2).

La méthode de classification choisie est la classification hiérarchique utilisant le critère d'agrégation de Ward [Lebart *et al.*, 1984]. L'intérêt de ce critère est ici sa «compatibilité» avec l'analyse factorielle (dans la mesure où il s'agit d'un critère basé sur la variance) et sa tendance à construire des classes plutôt «sphériques».

Le dendrogramme des niveaux supérieurs de la hiérarchie (figure 1) est dessiné pour guider le choix du nombre de classes à retenir. On cherche à couper l'arbre à un niveau qui met en évidence des branches bien séparées, chaque branche ne comportant que des éléments très proches (notion de palier des indices de niveau).

A défaut d'un critère formel d'optimalité (qui de toute façon ne manquerait pas d'être arbitraire) on peut choisir plusieurs partitions pour ne retenir que celle qui semblera convenir le mieux à l'application. Dans notre application, après plusieurs essais, on a finalement retenu la partition en 6 classes.

Noter que, du point de vue de l'homogénéité des classes, la partition créée par coupure d'un arbre ne saurait être optimale. Elle est en effet construite en respectant la contrainte d'emboîtement des partitions successives de la hiérarchie. On peut donc améliorer cette partition par une procédure d'itérations autour de centres mobiles (dite procédure de *consolidation*). Sur la figure 1, les effectifs entre parenthèses donnent les effectifs des classes obtenues directement par la coupure; les effectifs sans parenthèses sont ceux des classes optimisées obtenues après consolidation.

Remarque. La stratégie de classification en plusieurs phases (passant par une analyse factorielle pour en supprimer les derniers axes, suivie d'une classification hiérarchique utilisant le critère de Ward, avec coupure de l'arbre et consolidation de

la partition) constitue l'ossature des typologies dans [SPAD.N, 1993]. Les arguments pour cette stratégie, essentiellement empiriques, ne suffiraient pas à justifier les choix si une assez longue pratique n'avait pas déjà fait la preuve de son efficacité dans de nombreuses applications. En réalité, l'intérêt et la force de cette stratégie est de «donner la main» à tout instant à l'utilisateur qui doit faire des choix et vérifier leur pertinence à toutes les étapes des calculs.

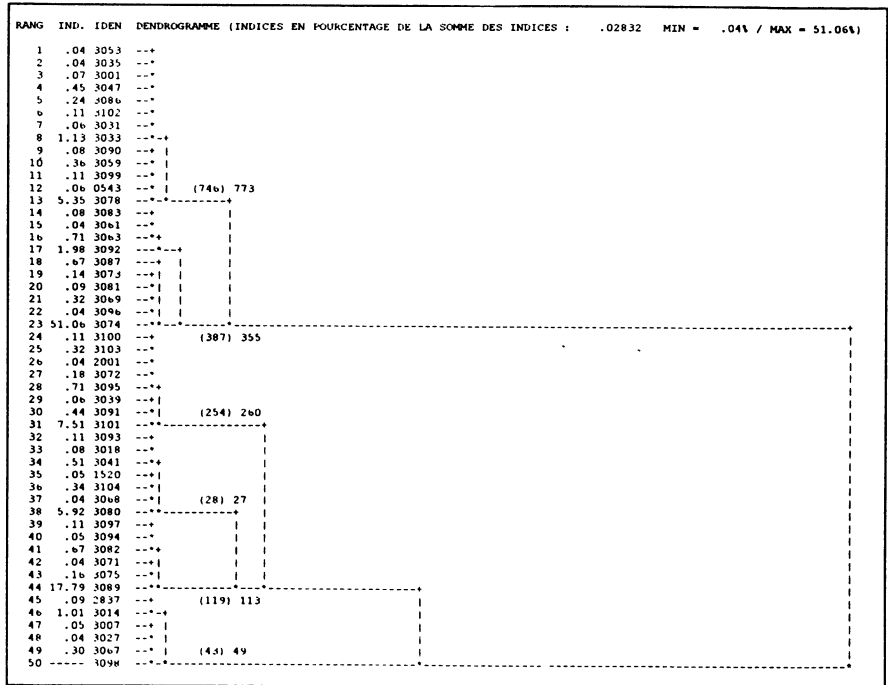


FIGURE 1

Dendrogramme des derniers niveaux de la hiérarchie

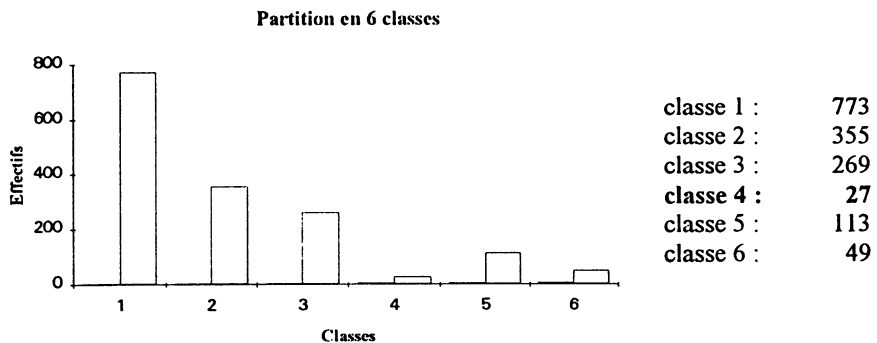


FIGURE 2

Les 6 classes de grossistes

Les fréquences (volumes de ventes) illustratives de la classe 4

Du point de vue de l'utilisateur, la qualité d'une partition doit être jugée de façon ultime d'une part par la possibilité de justifier *a posteriori* les regroupements obtenus, d'autre part par les qualités opératoires de ces classes pour répondre au problème (ici, le problème de prévision). On retient dans cette application la partition en 6 classes (cf. figure 2) et à titre d'illustration on étudie plus particulièrement la classe 4 (petite classe de 27 grossistes dont la liste complète est fournie dans le tableau 5).

Le tableau 3 fournit une aide à l'interprétation du regroupement des grossistes dans la classe 4. On y trouve, classés par ordre d'importance, les numéros (ou semaines) où l'hebdomadaire est plus vendu par la classe 4 que par le reste des grossistes. Voici un exemple de lecture des résultats :

- la colonne «POIDS» donne le volume des ventes; les autres chiffres sont des pourcentages; la classe 4 des grossistes réalise 1% seulement des ventes de l'année;
- le numéro du 22 février représente 4,14% (FRE/CLA) des ventes réalisées par les grossistes de la classe 4, alors qu'il représente seulement 1,82% (GLOBAL) des ventes réalisées par l'ensemble des grossistes;
- 2,28% (CLA/FRE) des ventes concernant ce même numéro ont été réalisées par les grossistes de la classe 4 alors que ces mêmes grossistes réalisent seulement 1% (GLOBAL) du total des ventes (tous numéros confondus).

Ainsi on lit au fil de ce tableau que les grossistes de la classe 4 réalisent mieux que les autres les ventes des numéros de février, juillet et août. Ceci constitue une première indication sur la nature du regroupement qui a été réalisé par la classification.

TABLEAU 3

*Caractérisation de la classe 4
par les volumes de ventes hebdomadaires*

DATES	CLA/FRE	FRE/CLA	GLOBAL	POIDS
CLASSE 4 / 6			1.00	602458
22 février	2.28	4.14	1.82	11094552
8 août	2.10	4.67	2.23	11337531
29 février	2.20	4.07	1.85	11114442
1er août	1.97	4.48	2.28	11371112
15 août	1.90	4.08	2.15	11292016
7 mars	1.76	3.23	1.84	11104618
26 décembre	1.68	3.13	1.87	11122936
15 février	1.68	3.04	1.81	11090658
18 juillet	1.63	3.20	1.96	11180897
25 juillet	1.61	3.30	2.05	11234760
11 juillet	1.54	3.00	1.95	11172083
14 mars	1.49	2.84	1.91	11150548
22 août	1.42	2.94	2.08	11248234
8 février	1.34	2.62	1.96	11176385
1er février	1.29	2.36	1.83	11101402
19 décembre	1.25	2.20	1.77	11064585
21 mars	1.13	2.02	1.80	11080080
25 janvier	1.11	2.07	1.86	11120323
4 juillet	1.10	2.00	1.82	11094106

Les variables qualitatives illustratives de la classe 4

On peut chercher à caractériser la classe 4 par les autres informations disponibles et qui n'ont pas participé à la construction des classes. Il s'agit des colonnes illustratives du tableau : type d'agglomération où se trouve le grossiste, de saisonnalité des ventes, type de secteur géographique, etc.

Le tableau 4 présente, par ordre d'importance décroissante, les éléments qui expliquent le regroupement des grossistes de la classe. Cet ordre est défini par le critère situé dans la colonne «valeurs-tests» [cf. Alevizos *et al*, 1992]. Le critère, assimilable à un nombre d'écarts-types d'une loi normale, est positif si la modalité est sur-représentée dans la classe, négatif si elle est sous-représentée. Les valeurs absolues sont d'autant plus importantes que l'attribut est plus caractéristique de la classe.

La classe 4 est une classe homogène formée de grossistes caractérisés par une saisonnalité «Hiver-Eté» (la «valeur-test» relative au code «saison» est très supérieure en valeur absolue à 2, seuil usuel des tests au risque de 5%). Il lui est lié aussi, à un moindre degré, le secteur géographique «Rhône-Alpes».

TABLEAU 4

Autres variables illustratives de la classe 4 des grossistes

V.TEST	PROBA	MODALITES	CARACTERISTIQUES DES VARIABLES---	POURCENTAGES ----			POIDS	
				CLA/MOD	MOD/CLA	GLOBAL		
		CLASSE 4 / e					1.71	27
13.42	.000	Hiver-Eté	Saison	50.98	96.30	3.23	51	
6.07	.000	Rhône Alpes	Secteur géographique	7.20	70.37	16.74	264	
-2.35	.009	Villes qui se vident	Saison	.00	.00	15.73	248	
-2.57	.005	Stables	Saison	.26	3.70	24.22	382	
-2.98	.001	Est	Secteur géographique	.00	.00	21.37	337	
-3.24	.001	Ouest	Secteur géographique	.00	.00	23.84	376	
-4.26	.000	Estival	Saison	.00	.00	34.62	546	

Les parangons de la classe 4

Le logiciel fournit également des «parangons» pour chaque classe. Ce sont des éléments centraux de la classe, les plus proches du centre. Ils constituent, d'un certain point de vue, des représentants typiques de la classe.

La figure 3 fournit la représentation de l'évolution des deux premiers parangons de la classe 4 (Chamonix et Combloux) ainsi que l'évolution moyenne de la classe. Ce graphique illustre comment les parangons représentent la moyenne de la classe par leurs profils (les extrema coïncidant sur les trois courbes) et non par le volume des ventes (qui peut être très différent).

TABLEAU 5

Liste complète des grossistes de la classe 4

65	Cauterets	74	Taninges
66	Font Romeu	38	Autrans
73	Aigueblanche	38	Bourg d'Oisans
73	Aime	38	Villars de Lans
73	Bourg St Maurice	25	Les Hôpitaux Neufs
73	Modane	05	Briançon
73	St Avre	05	Vars
73	Val d'Isère	05	Veynes
74	Les Carroz	04	St André Les Alpes
74	Chamonix	04	Barcelonnette
74	La Clusaz	64	Eaux Bonnes
74	Combloux	39	Les Bois d'Amont
74	Megève	39	Les Rousses
74	St Gervais les Bains		

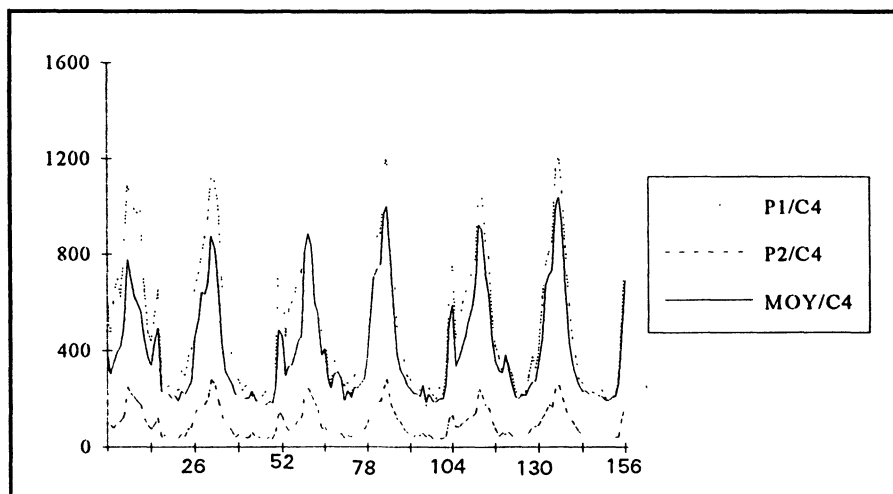


FIGURE 3

Série des ventes moyennes de la classe 4 et séries des ventes du 1^{er} paragon «Megève» et du 2^{ème} paragon «Combloux»

3. Modèle de prévision pour une classe de grossistes

3.1. Analyse en composantes principales et modélisation

3.1.1. La théorie

Des applications de l'analyse factorielle de processus *continus* ont déjà été menées notamment dans le domaine démographique [Deville, 1974] et dans le domaine médical [Pardoux, 1989]. On étudie ici des processus *discrets* puisqu'il s'agit de séries de flux hebdomadaires.

On va réaliser une analyse en composantes principales non normée [Lebart *et al*, 1979] sur les données des ventes des trois années relatives à la classe 4 (*cf.* tableau 6).

Disposant d'observations de T variables (ici 157 variables) sur n individus (ici 27 individus), on cherche un petit nombre de *variables synthétiques* qui représentent au mieux les T variables initiales.

TABLEAU 6

$x_t(\omega_i)$ représente le nombre d'hebdomadaires de la semaine t vendus par le grossiste ω_i

variable	X_0	...	X_t	...	X_{156}
individu					
grossiste ω_1					
grossiste ω_i	$x_t(\omega_i)$
grossiste ω_{27}					

Chacune des T variables X_t est mesurée sur n individus et peut être caractérisée par sa valeur moyenne $\left\{ \bar{x}_t = \frac{1}{n} \sum_{i=1}^n x_t(\omega_i) \right\}$ et la variation autour de cette valeur moyenne mesurée par la variance.

La première première variable synthétique ou première composante principale ξ_1 est définie par le *vecteur propre unitaire* f_1 associé à la plus grande valeur propre

λ_1 de la matrice des covariances V :

$$\xi_1(\omega_i) = \sum_{t=1}^T f_{1,t} (x_t(\omega_i) - \bar{x}_t)$$

Le vecteur f_1 a T composantes. La coordonnée $f_{1,t}$ ($t = 1, \dots, T$) est associée à la $t^{\text{ème}}$ semaine. La suite des $f_{1,t}$ constitue ainsi une série chronologique de longueur T .

Si l'inertie expliquée par le premier axe factoriel est élevée, on va considérer les séries chronologiques $\{\bar{x}_t, t = 1, \dots, T\}$ et $\{f_{1,t}, t = 1, \dots, T\}$ comme caractéristiques de l'évolution des ventes de l'ensemble des grossistes de la classe. Si de plus, le second vecteur propre peut être considéré comme réalisation d'un « bruit blanc » ou possède un faible pourcentage d'inertie, alors on pourra écrire pour tout t ($1 \leq t \leq T$) et tout ω_i la formule de reconstitution des données :

$$x_t(\omega_i) \approx \bar{x}_t + \xi_1(\omega_i) \cdot f_{1t} \quad (1)$$

Pour une reconstitution correcte des données de certaines classes, on peut être amené à prendre en compte le second vecteur propre. Dans ce cas, la formule s'écrit pour tout t ($1 \leq t \leq T$) :

$$x_t(\omega_i) \approx \bar{x}_t + \xi_1(\omega_i) \cdot f_{1t} + \xi_2(\omega_i) \cdot f_{2t}$$

3.1.2. Reconstitution des données de la classe 4 et modélisation

Pour cette classe, une analyse en composantes principales faite sur la matrice des covariances (dimension 157×157) associée au tableau 6 fournit un premier axe extrêmement dominant. Cela tient à la très forte dépendance structurelle des données temporelles, augmenté du fait que nous avons choisi de travailler sur les covariances plutôt que sur les corrélations. La réussite de l'entreprise repose en grande partie sur cette remarque.

HISTOGRAMME DES 39 PREMIERES VALEURS PROPRES			
NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE
1	156.4081	88.38	88.38
2	7.5792	4.28	92.67
3	5.8844	3.33	96.00
4	3.1728	1.79	97.78
5	1.1052	.62	98.41
6	.6535	.37	98.78
7	.3898	.22	99.00
8	.3240	.18	99.18

Etant donnée la valeur très élevée du pourcentage d'inertie expliqué par le premier axe factoriel, on va pouvoir reconstituer très correctement les données en utilisant la série moyenne $\{\bar{x}_t, t = 0, \dots, 156\}$ et la seule série $\{f_{1,t}, t = 0, \dots, 156\}$ (formule 1).

Pour quelques éléments particuliers de la classe (notamment les grossistes à faibles ventes), on verra qu'il est utile, pour améliorer la reconstitution, de tenir compte de la seconde série $\{f_{2,t}, t = 0, \dots, 156\}$. Dans un premier temps, nous ne retiendrons que les séries $\{\bar{x}_t, t = 0, \dots, 156\}$ et $\{f_{1,t}, t = 0, \dots, 156\}$.

La représentation graphique de ces deux séries caractéristiques (cf. figure 4) montre que la chronique $\{f_{1,t}, t = 0, \dots, 156\}$ est en quelque sorte le «reflet» de la chronique $\{\bar{x}_t, t = 0, \dots, 156\}$. En se référant à la formule (1) de reconstitution des données, on lit que l'évolution du $i^{\text{ème}}$ grossiste se déduit de l'évolution moyenne en lui additionnant la série $\{f_{1,t}, t = 0, \dots, 156\}$ avec une pondération $\xi_1(\omega_i)$ propre à ce grossiste. Si cette pondération est positive, on diminue l'effet de la série moyenne, et si elle est négative, on l'augmente.

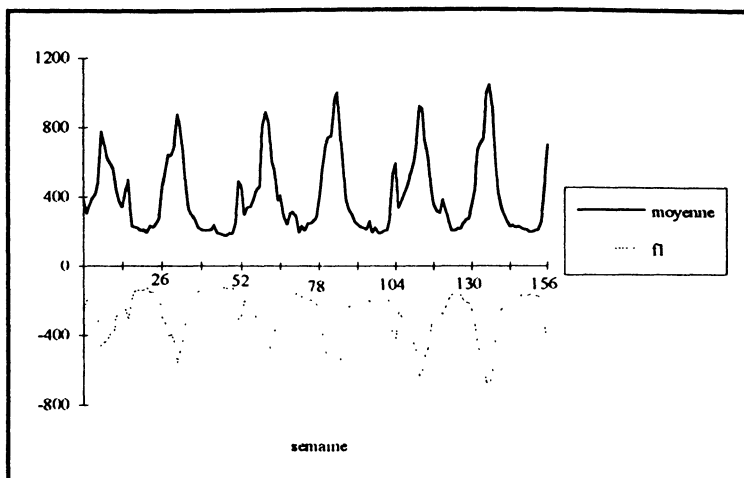


FIGURE 4

Les deux séries représentatives de la classe 4

L'examen graphique de la reconstitution des évolutions de quelques éléments montre que le modèle fonctionne correctement. Sur les figures 5, 6, 7, 8 on a représenté quelques séries observées (trait plein) et les séries calculées par le modèle (en pointillé). Les figures 5, 6 et 7 correspondent aux trois premiers parangons alors que la figure 8 correspond au deuxième grossiste le plus éloigné du centre de la classe.

L'ajustement correspondant à *Combloux*, où le volume des ventes est très faible, pourrait être notablement amélioré en utilisant la série $\{f_{2,t}, t = 0, \dots, 156\}$ associée au second vecteur propre.

Considérons maintenant le problème de la prévision en nous plaçant dans le cas le plus simple suivant :

$$x_t(\omega_i) \approx \bar{x}_t + \xi_1(\omega_i) \cdot f_{1t}$$

Nous sommes amenés à rechercher un modèle de prévision pour les deux séries chronologiques $\{\bar{x}_t, t = 0, \dots, 156\}$ et $\{f_{1,t}, t = 0, \dots, 156\}$ au lieu de réaliser une prévision pour chacune des 27 séries initiales.

3.2. Analyse d'une série chronologique

L'étude d'une chronique commence par l'examen de son allure générale à l'aide de sa représentation graphique. Nous avons sur la figure 4 les représentations graphiques des séries $\{\bar{x}_t, t = 0, \dots, 156\}$ et $\{f_{1,t}, t = 0, \dots, 156\}$.

Ces chroniques sont sans tendance apparente et possèdent à l'évidence une composante saisonnière de période 52. On observe en effet des *extrema* tous les ans en février et août, et des *extrema* locaux en décembre et avril-mai. Ceci confirme d'une certaine façon la caractérisation trouvée en §2.2 pour la classe 4 (saisonnalité hiver-été).

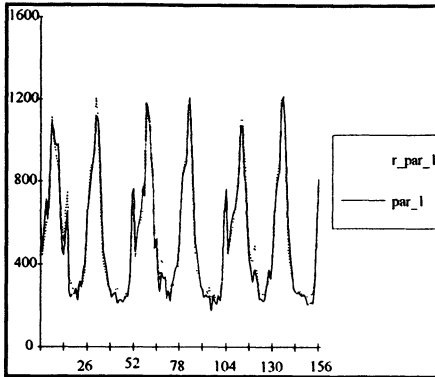


FIGURE 5
Reconstitution du 1^{er} paragon
Megève

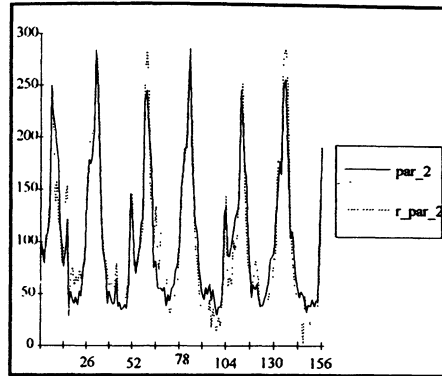


FIGURE 6
Reconstitution du 2^{ème} paragon
Combloux

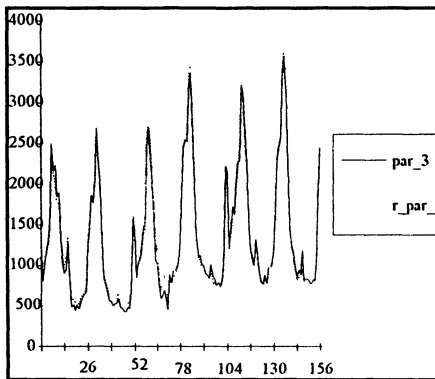


FIGURE 7
Reconstitution du 1^{er} paragon
Taninges

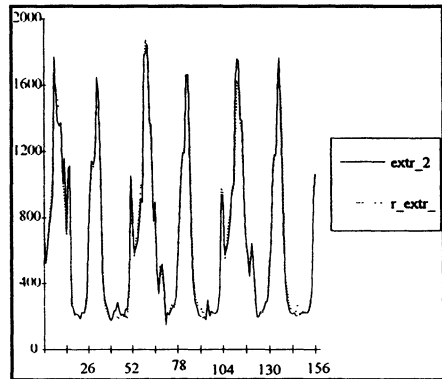


FIGURE 8
Reconstitution du 2^{ème} paragon
Bourg d'Oisans

Avant d'établir le modèle de prévision, il est important d'examiner si les maxima correspondent bien aux mêmes dates tous les ans. Ceci peut être vérifié par

l'examen des courbes annuelles superposées (cf. figure 9); la procédure correspond à la représentation graphique de la table de Buys-Ballot (tableau 7).

Au préalable, on élimine la première observation qui correspond à une date exceptionnelle : c'est la semaine notée 0 qui précède les 3 années d'observation. Pour les trois années étudiées, on constate des décalages pour les maxima en février et en avril. Ces décalages sont certainement dus à la variabilité du calendrier scolaire.

Il est donc nécessaire de repérer les semaines de vacances de février, mars, avril et mai non plus par leur numéro de semaine dans l'année, mais par le type de vacances associé : «semaine vacances février des zones A et B», «semaine vacances février des zones B et C», etc. Avant l'analyse des séries $\{\bar{x}_t, t = 0, \dots, 156\}$ et $\{f_{1,t}, t = 0, \dots, 156\}$, on prend soin de permuter des colonnes de la table de Buys-Ballot afin que les semaines de vacances se correspondent. L'examen des courbes annuelles superposées associées à cette nouvelle table permet de juger si on a bien une série périodique de période 52.

C'est sur les séries associées à cette nouvelle table que l'on va s'appuyer pour établir les modèles des prévisions.

TABLEAU 7

Table de Buys-Ballot de la série moyenne

	1 ^{ère} semaine	...	j ^{ème} semaine	...	52 ^{ème} semaine	moyenne annuelle
1990	\bar{x}_1	...	\bar{x}_j	...	\bar{x}_{52}	$\frac{1}{52} \sum_{j=1}^{52} \bar{x}_j$
1991	\bar{x}_{53}	...	\bar{x}_{j+52}	...	\bar{x}_{104}	$\frac{1}{52} \sum_{j=1}^{52} \bar{x}_{j+52}$
1992	\bar{x}_{105}	...	\bar{x}_{j+104}	...	\bar{x}_{156}	$\frac{1}{52} \sum_{j=1}^{52} \bar{x}_{j+104}$
moyenne hebdomadaire		...	$\frac{1}{3} \sum_{k=0}^2 \bar{x}_{j+52k}$...		moyenne générale \bar{x}

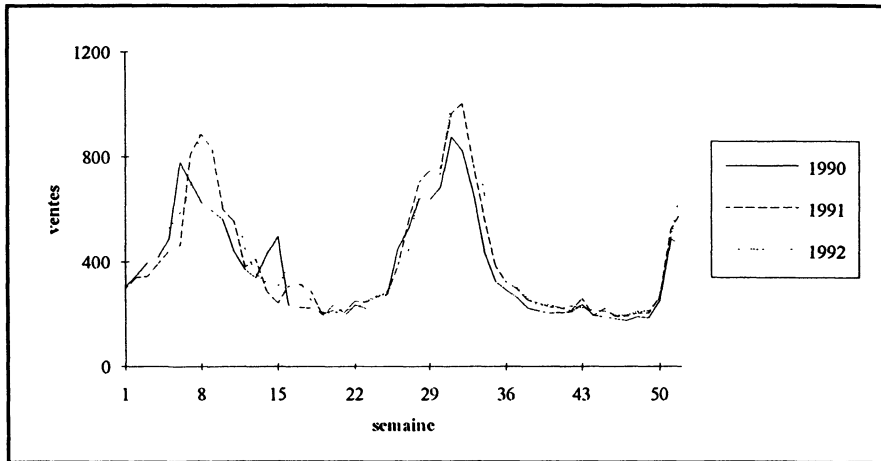


FIGURE 9
*Courbes annuelles superposées de la série moyenne
 avant le décalage permettant de faire coïncider les extrema*

3.3. Prédiction à l'aide d'un modèle SARIMA

A partir des prévisions à court terme réalisées pour chacune des deux séries temporelles $\{\bar{x}_t, t = 0, \dots, T\}$ et $\{f_{1,t}, t = 0, \dots, T\}$, nous pouvons calculer les prévisions pour les ventes $\hat{x}_{T+h}(\omega_i)$ du grossiste ω_i au cours de la semaine $T + h$ ultérieure à la dernière observation ($h > 0$) :

$$\hat{x}_{T+h}(\omega_i) \approx \widehat{\bar{x}_{T+h}} + \xi_1(\omega_i) \cdot \hat{f}_{1,T+h}$$

On peut compléter la prévision par l'estimation de la borne supérieure d'un intervalle de confiance d'un niveau donné $(1 - \alpha)$.

Les prévisions pour les séries temporelles saisonnières peuvent être obtenues à partir d'un modèle SARIMA [Box et Jenkins, 1976, Gouriéroux et Monfort, 1990]. Après estimation, on contrôle le modèle par comparaison avec la réalisation effective de la série.

3.3.1. Les modèles SARIMA

Un processus $\{X_t, t \in \mathbb{Z}\}$ est faiblement stationnaire ou stationnaire du second ordre si pour tout $t \in \mathbb{Z}$:

$$E(X_t) = \mu, E(X_t^2) < \infty \quad \text{et} \quad \text{cov}(X_t, X_{t+h}) = \gamma(h) \quad \text{pour tout } h \in \mathbb{Z},$$

$\gamma(h)$ porte le nom de fonction d'autocovariance du processus. Si $\mu \neq 0$, on associe le processus centré $\{Y_t = X_t - \mu, t \in \mathbb{Z}\}$. Dans la suite de cet article, le terme «stationnaire» signifiera «stationnaire du second ordre».

D'autre part, nous appelons *bruit blanc* une suite d'aléas $\{a_t, t \in Z\}$ centrés, indépendants et de même variance :

$$E(a_t) = 0, \quad E(a_t a_{t'}) = \begin{cases} 0 & \text{si } t \neq t' \\ \sigma^2 & \text{si } t = t' \end{cases}$$

Un processus stationnaire centré $\{X_t, t \in Z\}$ est dit *autorégressif-moyenne mobile d'ordres p et q* ou ARMA(p, q), s'il satisfait à une relation du type :

$$X_t - \sum_{i=1}^p \varphi_i X_{t-i} = a_t - \sum_{j=1}^q \theta_j a_{t-j}$$

que l'on peut écrire sous la forme :

$$\varphi_p(B)X_t = \theta_q(B)a_t$$

avec :

$$\varphi_p(B) = 1 - \sum_{i=1}^p \varphi_i B^i, \quad \theta_q(B) = 1 - \sum_{j=1}^q \theta_j B^j$$

où a_t est un bruit blanc qui s'interprète comme la suite des erreurs de prévision, B est l'opérateur «retard» tel que $B^k X_t = X_{t-k}$ pour tout entier k , les racines des polynômes $\varphi(z)$ et $\theta(z)$ étant à l'extérieur du disque unité.

Un modèle ARMA vise donc à représenter un processus stationnaire centré par une combinaison de valeurs passées et d'erreurs passées. Lorsqu'un processus n'est pas stationnaire et a une tendance représentée par un polynôme de degré n , l'opérateur des différences $(1 - B)^n$ permet de faire disparaître la tendance.

Un processus non stationnaire $\{X_t, t \in Z\}$ sera dit ARIMA(p, d, q) si le processus $\{(1 - B)^d X_t, t \in Z\}$ peut être représenté par un ARMA(p, q) - d étant le degré d'intégration de la série (I comme «Integrated») :

$$\varphi_p(B)(1 - B)^d X_t = \theta_q(B)a_t \quad (1)$$

Si le processus $\{X_t, t \in Z\}$ présente également une composante saisonnière de période s et si on lui ajuste un modèle ARIMA, alors la suite des aléas $\{a_t, t \in Z\}$ ne sera pas un bruit blanc et sera périodique de période s . Lorsque la tendance du processus $\{a_t, t \in Z\}$ peut être ajustée par un polynôme de degré D , le processus $\{(1 - B^s)^D a_t, t \in Z\}$ est stationnaire et on peut le représenter par un modèle ARMA :

$$\Phi_P(B^s)(1 - B^s)^D a_t = \Theta_Q(B^s)\varepsilon_t \quad (2)$$

où ε_t est un bruit blanc, les racines des polynômes $\Phi(z)$ et $\Theta(z)$ étant à l'extérieur du disque unité.

En combinant les modèles (1) et (2), on obtient un modèle SARIMA ou *Seasonal AutoRegressive Integrated Moving Average* :

$$\varphi_P(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D\tilde{X}_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (3)$$

où \tilde{X}_t est défini à partir de la variable X_t :

$$\tilde{X}_t = \begin{cases} X_t - \mu & \text{si } d = D = 0 \\ X_t & \text{sinon} \end{cases}$$

où ε_t est un bruit blanc, les racines des polynômes $\varphi(z)$, $\theta(z)$, $\Phi(z)$ et $\Theta(z)$ étant à l'extérieur du disque unité.

Ce modèle est noté SARIMA_s[(p, d, q), (P, D, Q)].

A partir des observations x_1, \dots, x_T , on estime les ordres p, d, q, P, D et Q du modèle SARIMA qui les représente le mieux; c'est la phase d'identification du modèle, qui est suivie de la phase d'estimation des paramètres. Le logiciel [A• TEMP, 1993] procède à l'identification complète du modèle en estimant ses ordres par minimisation du critère d'Akaike et ses paramètres selon la méthode de Marquardt. Le modèle est ensuite testé à l'aide du test «portmanteau» de Box-Pierce [Gouriéroux et Monfort, 1990, O'Donovan, 1983, Pankratz, 1983]. Si le modèle est accepté, on passe à la prévision.

3.3.2. Utilisation d'un modèle SARIMA pour la classe 4

Pour chacune des deux séries représentatives de la classe 4 $\{\bar{x}_t, t = 0, \dots, 156\}$ et $\{f_{1,t}, t = 0, \dots, 156\}$ qui sont saisonnières de période 52, nous utilisons un modèle SARIMA₅₂[(2,0,0),(0,1,0)]. C'est un modèle qui prend en compte, pour chaque valeur prévue, les deux dernières valeurs observées ainsi que l'observation correspondant à la même semaine de l'année précédente. En remplaçant les paramètres par leurs valeurs dans l'équation (3), ce modèle s'écrit :

$$(1 - \varphi_1 B - \varphi_2 B^2)(X_t - X_{t-52}) = \varepsilon_t \\ \Rightarrow X_t = X_{t-52} + \varphi_1(X_{t-1} - X_{t-53}) + \varphi_2(X_{t-2} - X_{t-54}) + \varepsilon_t$$

Utilisons ce modèle pour les prévisions des deux séries représentatives de la classe 4 :

$$\widehat{x_{T+h}} = \overline{x_{T+h-52}} + \widehat{\varphi}_1(\overline{x_T} - \overline{x_{T-52}}) + \widehat{\varphi}_2(\overline{x_{T-1}} - \overline{x_{T-53}}) \\ \widehat{f_{1,T+h}} = f_{1,T+h-52} + \widehat{\varphi}_1^*(f_{1,T} - f_{1,T-52}) + \widehat{\varphi}_2^*(f_{1,T-1} - f_{1,T-53})$$

Ces prévisions à la date T pour l'horizon h prennent en compte la valeur observée de la même semaine de l'année précédente et les différences entre les deux dernières observations et celles des mêmes semaines de l'année précédente. Ces deux différences

sont interprétées, le cas échéant, comme étant dues à des modifications du niveau des ventes. Voici les estimations des deux séries pour les horizons 1, 2 et 3 :

	\widehat{x}_{156+h}	$\widehat{f}_{1,156+h}$
$h = 1$	403,2	-280,2
$h = 2$	400,9	-314,5
$h = 3$	483,5	-364,5

En affectant la prévision $\widehat{f}_{1,156+h}$ de la pondération $\xi_1(\omega_i)$, on en déduit une prévision de ventes pour le grossiste ω_i :

$$\widehat{x}_{156+h}(\omega_i) \approx \widehat{x}_{156+h} + \xi_1(\omega_i) \cdot \widehat{f}_{1,156+h}$$

Dans le tableau 8, nous donnons en exemple le nombre d'exemplaires vendus par quelques grossistes pour la première semaine à prévoir (horizon 1). La colonne *estimation automatique* donne la valeur prévue par le modèle; cette valeur est comparée à la valeur réelle observée. L'erreur de prévision est notée dans la colonne *% erreur automatique*. La dernière colonne indique la prévision traditionnelle fournie par le service de prévision de l'entreprise. La comparaison des deux dernières colonnes montre la performance du modèle.

TABLEAU 8
Quelques exemples de prévision à l'horizon 1

Grossiste	estimation «automatique»	quantité vendue	% erreur «automatique»	% erreur «manuelle»
Megève	520	475	8,70	13,64
Taninges	1345	1301	3,30	16,06
Briançon	788	731	7,27	18,78
Val d'Isère	331	267	19,27	33,25

Nous présentons dans le tableau 9 deux exemples de prévisions pour les horizons 1, 2 et 3. A l'horizon 1 on peut comparer directement les «erreurs» de prévision. Pour les horizons 2 et 3, les chiffres fournis par le modèle sont bien des anticipations à 2 et 3 semaines alors que la colonne «erreur manuelle» donne les prévisions fournies par le service de prévision en tenant compte des observations jusqu'à la semaine précédente.

TABLEAU 9
Deux exemples de prévisions pour les horizons 1, 2 et 3

Grossiste	horizon	estimation «automatique»	quantité vendue	% erreur «automatique»	% erreur «manuelle»
St Gervais les B	1	493	464	5,79	15,64
	2	501	452	9,80	17,82
	3	556	511	8,15	14,98
Bourg d'Oisans	1	692	522	24,56	25,43
	2	725	694	4,26	7,47
	3	816	768	5,85	9,65

4. Bilan et perspectives

4.1. Bilan

Si le modèle de reconstitution des ventes est de bonne qualité, les méthodes de prévision donnent de bons résultats à condition de faire l'hypothèse d'une stabilité de l'environnement. Dans le cas d'un changement du niveau des ventes (regroupement de grossistes, campagne promotionnelle, numéro spécial) ou de la saisonnalité (incidence du décalage du calendrier des congés scolaires d'une année à l'autre), il conviendrait d'utiliser ces informations externes au modèle. On a vu comment tenir compte des zones du calendrier scolaire (cf. §3.2). On ne peut pas envisager des prévisions pertinentes sans la prise en compte des changements dans l'environnement.

Pour cette première étude, en utilisant la seule information donnée par les historiques (les observations ayant été corrigées par le nombre de jours de ventes) sans la «connaissance experte», les erreurs relatives des estimations automatiques et des estimations par «suivi ou simulation manuelle» sont en moyenne du même ordre :

pour la 1^{ère} semaine 23,5% pour les estimations automatiques,

23% pour les estimations manuelles,

pour la 2^{ème} semaine, 21,5% pour les estimations automatiques,

21% pour les estimations manuelles,

pour la 3^{ème} semaine, 19% pour les estimations automatiques,

21,5% pour les estimations manuelles.

On peut d'ailleurs améliorer les scores de prévision par modèle en raffinant le découpage en classes et donc les modèles de prévision.

De plus, rappelons qu'un avantage de la prévision automatique est qu'elle peut s'effectuer à un horizon de 2 ou 3 semaines.

4.2. Perspectives

Cette étude a montré :

- l'importance des méthodes de l'analyse des données, avec l'utilisation de toute la richesse des recodages pour exhiber les structures et les nuances cachées par l'égale importance des corrélations brutes initiales,
- l'intérêt de l'alternance des méthodes de modélisation des séries temporelles avec celles de l'analyse des données pour réduire un volume de séries initiales trop importantes à manipuler et synthétiser leur évolution (pour la classe présentée en exemple, on synthétise 27 chroniques observées à 157 dates par 2 chroniques observées aux mêmes dates et un coefficient propre à chaque unité statistique),
- la possibilité du développement d'une aide automatique à la prévision des ventes.

La stratégie proposée pourrait être enrichie :

- en prenant en compte, pour les recodages, d'indices spécifiques au domaine d'application (taux de variation, indices économiques, etc.),
- en affinant la classification dans les classes importantes pour constituer des sous-classes selon le volume des ventes (cet effet taille ayant été jusqu'à présent écarté),
- en affinant le modèle dans les classes homogènes par la prise en compte du 2nd facteur et d'informations de type expert (congés scolaires, etc.).

Bibliographie

- ALEVIZOS P., MORINEAU A. (1992) *Tests et Valeurs-Tests. Application à l'Etude des Mastics Utilisés dans la Fabrication des Vitraux*. Revue de Statistique Appliquée, vol. 40, n°4, p. 27-43.
- A•TEMP (1993) *Analyses des séries Temporelles*, version 1.0, CISIA, Saint-Mandé, France.
- BENZÉCRI, J.P. (1973) *L'analyse des données*, tome I : *La taxinomie*, tome II : *L'analyse des correspondances*, Dunod.
- BOX, G.E.P., JENKINS, G.M. (1976) *Time Series Analysis. Forecasting and Control*. Second edition, Holden-Day, San Francisco.
- CELEUX, G., DIDAY E., RALAMBONDRAINY H., LECHEVALLIER Y. GOVAERT G. (1989) *Classification automatique des données*. Dunod.
- DEVILLE J.C. (1974) *Méthodes statistiques et numériques de l'analyse harmonique*, Annales de l'INSEE , n° 15, 3-97.
- GOURIÉROUX C., MONFORT A. (1990) *Séries temporelles et modèles dynamiques*, Economica.
- LEBART L., MORINEAU A., FÉNELON J.-P. (1979) *Traitement des données statistiques*, Dunod.

- LEBART L., MORINEAU A., WARWICK K. M. (1984) *Multivariate Descriptive Statistical Analysis*, John Wiley.
- O'DONOVAN, T.M. (1983) *Short Term Forecasting : An Introduction to the Box-Jenkins Approach*, Wiley.
- PANKRATZ A. (1983) *Forecasting With Univariate Box-Jenkins Models : Concepts and Cases*, John Wiley
- PARDOUX, C. (1989) *Apport de l'analyse factorielle à l'étude d'un processus*. Revue de Statistique Appliquée, vol. 37, n°4, p. 41-60.
- SPAD•N (1993) *Système Portable d'Analyse des Données Numériques. Version 2.5*, CISIA, Saint-Mandé, France.