REVUE DE STATISTIQUE APPLIQUÉE

A. HAYEK
J. P. LECOUTRE

Propriétés empiriques d'un prédicteur non paramétrique robuste

Revue de statistique appliquée, tome 46, n° 3 (1998), p. 77-88
<http://www.numdam.org/item?id=RSA_1998__46_3_77_0>

© Société française de statistique, 1998, 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/

Rev. Statistique Appliquée, 1998, **XLVI** (3), 77-88

PROPRIÉTÉS EMPIRIQUES D'UN PRÉDICTEUR NON PARAMÉTRIQUE ROBUSTE

A. Hayek*, J.P. Lecoutre**

* *Université libanaise. Faculté des Sciences. Hadath. Liban.*

** *Université Paris 6, U.R.A. 1321, 4 place Jussieu, 75252 Paris*

RÉSUMÉ

Les performances d'un prédicteur non paramétrique robuste sont comparées avec celles du prédicteur à noyau, du médianogramme et de la méthode de Box et Jenkins, pour des séries déjà étudiées où des données aberrantes sont introduites.

Mots-clés : *prédicteur non paramétrique, robustesse.*

ABSTRACT

For real or simulated series, we compare the efficiency of a nonparametric robust predictor with respect to the usual kernel predictor and the Box-Jenkins approach. Introduction of outliers in the series proves the superiority of the robust predictor.

Keywords : *nonparametric prediction, robustness.*

1. Introduction

À partir d'observations $x_1, \ldots, x_n$ d'une série chronologique, on souhaite prévoir la valeur future x_{n+h}, $h \in \mathbb{N}^*$. On suppose pour cela qu'il s'agit de réalisations d'un processus discret stationnaire et, selon les hypothèses introduites, on pourra généralement construire un prédicteur, optimal dans le cadre du modèle retenu. Si on choisit un modèle comportant un nombre fini de paramètres inconnus, on se situe dans le cadre de la statistique paramétrique où la classe des modèles ARMA est dominante, la procédure de prévision associée étant connue sous le nom de méthode de Box et Jenkins. Conscients des limites de cette modélisation, certains auteurs ont proposé récemment des extensions non linéaires comme par exemple les modèles ARCH.

Une autre voie que la complexification de modèles existants consiste à utiliser des méthodes non paramétriques où seules des hypothèses générales de régularité sur le processus observé sont nécessaires. Le problème de prédiction est alors un problème d'estimation d'un paramètre fonctionnel. Les propriétés de convergence de plusieurs classes de prédicteurs non paramétriques sont obtenues sous une hypothèse d'indépendance asymptotique du processus observé, très souvent une hypothèse de

mélangeance (cf Bosq, 1996), et même sous la condition minimale d'ergodicité (Delecroix, 1987).

Bien entendu, toutes les propriétés sont établies sous l'hypothèse que toutes les observations sont celles d'un même processus. Malheureusement, dans la pratique il peut arriver que parmi les observations utilisées figurent des données qui ont été perturbées par un phénomène parasite et que nous qualifierons de valeurs aberrantes. Même si la proportion de telles valeurs est faible, la qualité du prédicteur peut être gravement détériorée. En particulier, une procédure paramétrique, optimale dans le cadre strict du modèle, devient de très mauvaise qualité si l'on s'en écarte un peu. Pour pallier cet inconvénient on utilise des procédures statistiques robustes qui introduisent une certaine continuité dans les propriétés (Bosq-Lecoutre, 1992). Dans un cadre non paramétrique, on considère une famille beaucoup plus vaste de lois de probabilité, ce qui atténue l'effet de valeurs aberrantes éventuelles. Cependant, des perturbations importantes dans les données auront des effets importants sur la qualité des prédicteurs. Dans le cas de volumineux fichiers de données où il y a suspicion de valeurs aberrantes, il est donc nécessaire de recourir à une classe de prédicteurs non paramétriques robustes. Les propriétés de convergence d'une telle classe de prédicteurs à noyau ont été établies par Hayek-Lecoutre (1990) pour un processus fortement mélangeant. Nous proposons ici une comparaison empirique avec d'autres classes de prédicteurs, complétant l'article de Carbon-Delecroix (1993).

2. Définition de la classe de prédicteurs

Soit $(X_t; t \in \mathbb{N})$ un processus réel stationnaire et r une fonction d'autorégression de X_{s+h}, $s \in \mathbb{N}^*$, sur $(X_1, \ldots, X_s)$. À partir des observations $X_1, \ldots, X_T$, $T > s + h$, on construit un estimateur r_T de r et on en déduit le prédicteur $r_T(X_{T-s+1}, \ldots, X_T)$. L'estimateur $r_T(x)$ de $r(x)$, $x \in \mathbb{R}^s$, est défini ici comme solution de l'équation en θ :

$$\sum_{i=1}^{n} w_{ni}(x)\psi(X_{i+s+h} - \theta) = 0$$

où ψ est une application réelle strictement monotone et continûment différentiable, $n = T - s - h$, les $w_{ni}(x)$, $1 \leqslant i \leqslant n$, étant des poids qui dépendent de la distance entre x et $(X_{i+1}, \ldots, X_{i+s})$.

Si ψ est la fonction identité, on retrouve la forme usuelle des prédicteurs non paramétriques s'exprimant comme une moyenne pondérée des observations du passé. Le choix de $\psi(x) = \text{sign}(x)$ conduit à un prédicteur basé sur l'estimation de la médiane conditionnelle, étudié par Gannoun (1990). Plus généralement, le choix d'une fonction ψ bornée permet d'obtenir un prédicteur robuste, c'est-à-dire peu affecté par la présence de valeurs aberrantes. La classe de prédicteurs étudiée dans Hayek-Lecoutre (1990) correspond au choix :

$$w_{ni}(x) = \frac{1}{nb_n^s} K[b_n^{-1}(x - (X_{i+1}, \ldots, X_{i+s}))]$$

où K est un noyau de Parzen-Rosenblatt (cf. Bosq-Lecoutre, 1987) positif et pair, (b_n) une suite de nombres positifs qui tend vers 0 avec $1/n$.

La comparaison théorique des performances de ces prédicteurs, suivant le choix de ψ, avec le prédicteur paramétrique de Box et Jenkins est délicate et nécessite de définir un modèle de contamination. On peut envisager le modèle général suivant étudié par Bustos (1982) où le processus observé (Y_t) s'écrit :

$$Y_t = V_t Z_t + (1 - V_t) X_t \qquad t \in \mathbb{N}$$

(Z_t) étant le processus de contamination, indépendant de (X_t), et (V_t) une suite de v.a. indépendantes et de même loi caractérisant le type de contamination.

En prenant $P(V_t = 1/2) = 1$ on obtient le modèle AO (additive outlier) introduit par Denby-Martin (1979), où les valeurs aberrantes sont obtenues par addition à la valeur observée $(P(Z_t = 0) > 0)$.

Si $P(V_t = 0) = 1$ on retrouve le modèle IO étudié par Fox (1972) si (X_t) est un processus autorégressif dont l'innovation suit une loi normale contaminée.

Le cas général où V_t suit une loi de Bernoulli, de paramètre $\gamma > 0$ faible, correspond au modèle SO (substitutive outlier).

Étudions le modèle simple AO où $Y_t = X_t + Z_t$, les v.a. Z_t étant indépendantes, de même loi

$$(1 - \gamma)\delta_0 + \gamma N(0, \sigma^2) \qquad 0 < \gamma < 1$$

c'est-à-dire mélange d'une loi normale de variance élevée σ^2 et d'une masse de Dirac à l'origine (γ de valeur faible), le processus (X_t) étant un AR(1) :

$$X_t = \phi X_{t-1} + \varepsilon_t$$

l'innovation ε_t suivant une loi $N(0, \sigma_\varepsilon^2)$.

Dans un tel modèle la fonction de prévision définie par $\widehat{X}_t(h) = \widehat{\phi}^h X_t$ sera de mauvaise qualité en raison d'un biais asymptotique important dans l'estimation de ϕ, d'expression :

$$-\frac{\gamma\sigma^2}{\sigma_X^2 + \gamma\sigma^2}\phi$$

où $\sigma_X^2 = \sigma_\varepsilon^2/(1 - \phi^2)$.

Si on utilise un prédicteur non paramétrique de la classe générale précédente, la variance asymptotique de $\sqrt{nb_n^s}[r_n(x) - r(x)]$ en un point x de $\mathbb{R}^s$ est (cf. Robinson, 1984) :

$$\sigma^2(x) = \frac{\psi_1[r(x), x] \int K^2(u)du}{f(x)\psi_2^2[r(x), x]}$$

où f est la densité de $(X_{i+1}, \ldots, X_{i+s})$ et où on a posé :

$$\psi_1(\theta, x) = \int \psi^2(y - \theta) f(y|x) dy \qquad \psi_2(\theta, x) = \frac{\partial}{\partial\theta} \int \psi(y - \theta) f(y|x) dy$$

$f(.|x)$ étant la densité conditionnelle de X_{i+s+h} sachant que $(X_{i+1}, \ldots, X_{i+s}) = x$. Le choix $\psi(x) = x$ permet de retrouver pour un processus α-mélangeant :

$$\sigma^2(x) = \frac{\nu(x) \int K^2(u)du}{f(x)}$$

où $\nu(x)$ est la variance conditionnelle de X_{i+s+h} sachant que $(X_{i+1}, \ldots, X_{i+s}) = x$. Si le processus X_t est contaminé, ce terme peut devenir très élevé et conduire ainsi à un prédicteur de mauvaise qualité. Par contre, en prenant ψ bornée le terme ψ_1 aura toujours une valeur finie, permettant ainsi d'obtenir un prédicteur robuste. Pour le choix particulier $\psi(x) = \text{sign}(x)$, on obtient :

$$\sigma^2(x) = \frac{\int K^2(u)du}{4f^2(Md|x)f(x)}$$

où Md est la médiane de la loi conditionnelle de X_{i+s+h} sachant que $(X_{i+1}, \ldots, X_{i+s}) = x$.

3. Comparaison empirique de prédicteurs

Nous allons effectuer la comparaison de quatre méthodes de prédiction sur les neuf premières séries utilisées dans l'article de Carbon-Delecroix (1993) et reprises par Gannoun (1990). Il s'agit de cinq séries simulées, de trois issues du livre de Pankratz (1983) et une de celui de Box-Jenkins (1976). La méthode paramétrique de Box et Jenkins (BJ) est comparée avec trois méthodes non paramétriques, celle classique du noyau gaussien (NG) et deux méthodes robustes, le médianogramme (M) associé à $\psi(x) = \text{sign}(x)$ et le prédicteur de Tukey (TK) associé à $\psi(x) = x(1-x^2)^2\mathbf{1}_{[-1,1]}(x)$. Nous effectuons la prévision $\widehat{X}_j$, pour l'horizon $h = 1$, des m dernières observations X_j de la série et retenons comme critère d'erreur la quantité :

$$E = \frac{1}{m} \sum_{j=N-m+1}^{m} \frac{|\widehat{X}_j - X_j|}{X_j}$$

Pour chaque méthode nous faisons figurer la valeur de l'erreur la plus faible, parmi celles associées aux différents choix de paramètres pour cette méthode, notamment le choix des ordres p et q pour un modèle ARMA ou le choix du nombre s d'observations précédentes retenues pour la prédiction, pour une méthode non paramétrique. Enfin, nous retenons le même choix de fenêtre b_n que dans Carbon-Delecroix (1993), c'est-à-dire proportionnel à $n^{-1/(s+4)}$ et à l'écart-type empirique de la série.

Les v.a. ε_t des modèles simulés suivent une loi $N(0, 5)$ et u_t une loi exponentielle de paramètre 1/300.

Série 1 : AR1

$X_t = 0,9X_{t-1} + 1000 + \varepsilon_t \quad T = 100 \quad m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 1, q = 0$	0,089
NG	$s = 19$	0,062
M	$s = 1$	0,076
TK	$s = 2$	0,035

Série 2 : MA6

$X_t = \varepsilon_t - 2,848\varepsilon_{t-1} + 2,6885\varepsilon_{t-2} - 1,64645\varepsilon_{t-3}$
$+2,972\varepsilon_{t-4} - 2,1492\varepsilon_{t-5} + 0,67716\varepsilon_{t-6} \quad T = 100 \quad m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 0, q = 7$	2,77
NG	$s = 2$	2,77
M	$s = 2$	3,18
TK	$s = 2$	2,60

Série 3 : AR2

$X_t = 0,7X_{t-1} + 0,2X_{t-2} + 1000 + \varepsilon_t \quad T = 100 \quad m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 2, q = 0$	0,012
NG	$s = 2$	0,014
M	$s = 1$	0,085
TK	$s = 2$	0,010

Série 4 : ARMA(1,1)
$X_t = 0,8X_{t-1} + 1000 + \varepsilon_t + 0,2\varepsilon_{t-1} \quad T = 100 \quad m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 1, q = 1$	0,123
NG	$s = 30$	0,074
M	$s = 30$	0,105
TK	$s = 1$	0,183

Série 5 : Marges bénéficiaires (SARIMA)
Pankratz $T = 80, m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 1, d = 0, q = 0$ $P = 2, D = 1, Q = 4, S = 4$	4,85
NG	$s = 24$	1,17
M	$s = 1$	0,076
TK	$s = 2$	0,035

Série 6 : Consommation de cigares (SARIMA)
Pankratz $T = 96, m = 6$

Méthode	Paramètres	Erreur E
BJ	$p = 2, q = 0, d = 1$ $P = 1, Q = 0, D = 1, S = 12$	8,758
NG	$s = 12$	5,70
M	$s = 24$	34,281
TK	$s = 12$	4,123

Série 7 : Sinusoïde perturbée
$X_t = 3000\sin(\pi t/15) + u_1, T = 200, m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 2, q = 0, d = 1$ $P = 2, Q = 0, D = 1, S = 30$	21,88
NG	$s = 15$	7,81
M	$s = 30$	136,00
TK	$s = 2$	23,00

Série 8 : Charbon
Pankratz $T = 90, m = 10$

Méthode	Paramètres	Erreur E
BJ	$p = 1, q = 3$	3,11
NG	$s = 1$	2,94
M	$s = 5$	4,59
TK	$s = 4$	3,35

Série 9 : Données d'un processus chimique
Box-Jenkins $T = 70, m = 5$

Méthode	Paramètres	Erreur E
BJ	$p = 0, q = 2$	25,70
NG	$s = 2$	17,88
M	$s = 5$	35,78
TK	$s = 1$	15,12

L'examen des tableaux précédents montre les très bonnes performances du prédicteur de Tukey, même pour ces séries qui ne justifient pas l'emploi d'un prédicteur robuste. La valeur d'erreur est la plus faible pour six des neuf séries, le médianogramme n'étant meilleur que pour la série 4. Les méthodes de Box et Jenkins et du noyau l'emportent pour les séries 4, 7 et 8, l'écart n'étant significatif qu'avec le noyau pour la série 7 (sinusoïde perturbée).

Bien entendu, c'est la présence de valeurs aberrantes qui peut justifier pleinement l'emploi d'une telle méthode. Compte tenu des résultats précédents, nous ne retiendrons pour la comparaison dans ce cas que la méthode non paramétrique classique du noyau gaussien et le prédicteur robuste de Tukey. Nous allons donc introduire de une à trois valeurs aberrantes dans les séries 2, 3, 7 et 9 en multipliant par 10 la valeur observée. La série de longueur $T - m$ utilisée pour la prévision est découpée en 1, 2 ou 3 parties suivant le nombre de points aberrants retenus et les valeurs modifiées sont tirées au hasard dans chacune de ces sous-séries. Les tableaux ci-après indiquent la fragilité de la méthode du noyau, l'introduction de chaque valeur aberrante supplémentaire augmentant beaucoup la valeur de l'erreur. Par contre, la valeur d'erreur augmente peu avec ce facteur 10 pour la méthode de Tukey. Pour une altération d'un facteur 20, 50 ou même 100, l'erreur augmente sensiblement tout en restant très inférieure à celle du prédicteur non robuste. Notons cependant le comportement étrange de la série 7 où l'erreur diminue avec l'introduction de nouvelles valeurs aberrantes, les ordres de grandeur étant les mêmes pour les deux méthodes.

La première colonne précise le nombre de valeurs altérées et la seconde le(s) facteur(s) multiplicatif(s) utilisés. Les deux dernières colonnes indiquent la valeur de l'erreur pour la méthode du noyau gaussien et celle de Tukey.

Série 2 : MA6

$$X_t = \varepsilon_t - 2,848\varepsilon_{t-1} + 2,6885\varepsilon_{t-2} - 1,64645\varepsilon_{t-3} + 2,972\varepsilon_{t-4} - 2,1492\varepsilon_{t-5} + 0,67716\varepsilon_{t-6} \quad T = 100 \quad m = 5$$

Points aberrants	Coefficients multiplicateurs	NG	TK
0		2,77	2,60
1	10	12,92	3,30
2	10,10	23,59	3,73
3	10, 50, 100	183,46	46,40

Série 3 : AR2

$$X_t = 0,7\,X_{t-1} + 0,2\,X_{t-2} + 1000 + \varepsilon_t \quad T = 100 \quad m = 5$$

Points aberrants	Coefficients multiplicateurs	NG	TK
0		0,014	0,010
1	10	9,64	0,243
2	10,10	19,85	0,64
3	10, 10, 10	30,53	1,08
3	10, 20, 50	87,37	25,95

Série 7 : Sinusoïde perturbée
$X_t = 3000 \sin(\pi t/15) + u_t,\ T = 200,\ m = 5$

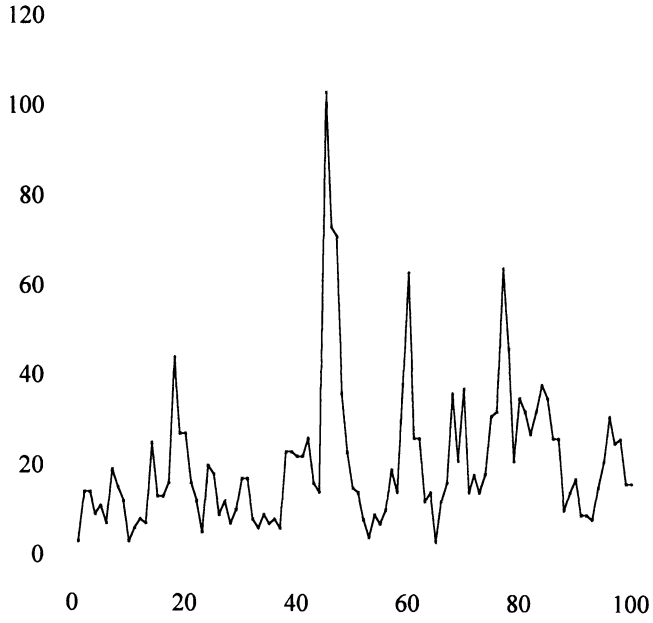
Points aberrants	Coefficients multiplicateurs	NG	TK
0		7,81	23,00
1	10	65,34	65,16
2	10, 10	62,37	62,44
3	10, 20, 50	56,36	44,97

Série 9 : Données d'un processus chimique
Box-Jenkins $T = 70, m = 5$

Points aberrants	Coefficients multiplicateurs	NG	TK
0		17,88	15,12
1	10	36,26	28,47
1	20	64,93	33,15
1	50	128,04	35,35
2	10, 10	52,48	32,79
2	10, 100	239,29	37,10
2	100, 100	328,51	85,55
3	10, 10, 10	74,93	33,89

La pollution de l'air en milieu urbain étant un sujet sensible actuellement et la présence de pics de pollution correspondant tout à fait à cette notion de valeurs aberrantes, nous avons retenu deux séries réelles récentes liées à la pollution en région parisienne pour conclure cette comparaison. Ces deux séries nous ont été fournies par la société AIRPARIF, chargée de la surveillance de la qualité de l'air en Ile-de-France. La première (série 10) est une série de moyennes journalières de dioxyde de soufre, exprimées en PPB (nombre de molécules par milliard), dans le 14ème arrondissement de Paris, pour les 100 derniers jours de l'année 1995. La seconde (série 11) est constituée de 222 relevés de sulfure d'hydrogène en PPB, toutes les 30 mn de 20h30 le 12/12/96 à 11h30 le 17/12/96, dans une station proche de Saint-Germain-en-Laye.

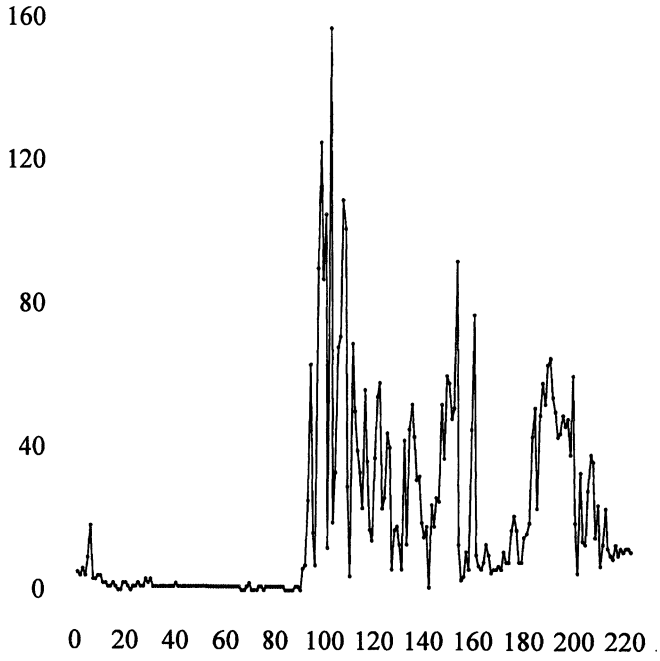
Dans la série 10, le principal pic de pollution correspond à la 45ème observation, deux autres pics moins importants étant situés dans la seconde moitié. Pour la prévision des 10 dernières observations nous avons ensuite retiré les 30 puis 40 premières valeurs, renforçant ainsi l'importance de ces points aberrants. On constate bien sûr une dégradation des performances des deux méthodes, mais avec un écart qui se creuse en faveur de la méthode robuste qui résiste beaucoup mieux.



Série 10 : Dioxyde de soufre à Paris 14ème
$m = 10$

T	NG	TK
100	0,395	0,345
70	0,492	0,396
60	0,590	0,457

Dans la série 11 on constate que les 90 premières observations sont des valeurs faibles, très peu dispersées, le principal pic de pollution correspondant à l'observation n° 102. Nous avons donc également tronqué cette série des premières valeurs, respectivement en nombre 57, 92 et 97, effectuant toujours la prédiction des mêmes dix dernières observations. On observe le même phénomène que pour la série précédente, mais on constate que si l'on supprime la totalité de la première partie plate de la série cela conduit à une légère amélioration des performances des prédicteurs (les cinq valeurs faibles conservées dans les 130 dernières observations prenant alors le caractère de valeurs aberrantes !).



Série 11 : Sulfure d'hydrogène (H_2S) à Saint-Germain-en-Laye
$m = 10$

T	NG	TK
222	0,241	0,190
165	0,566	0,340
130	1,046	0,775
125	0,937	0,690

Remerciements

Nous remercions les rapporteurs qui par leurs remarques pertinentes ont permis d'apporter certaines précisions améliorant la version initiale de cet article.

Bibliographie

BOSQ D. (1996) Nonparametric Statistics for Stochastic Processes. Estimation and Prediction. *Lecture Notes in Statist.* **110**. Springer Verlag.

BOSQ D., LECOUTRE J.P. (1987) Théorie de l'estimation fonctionnelle. Economica.

BOSQ D., LECOUTRE J.P. (1992) Analyse et prévision des séries chronologiques : méthodes paramétriques et non paramétriques. Masson.

BOX G.E.P., JENKINS G.M. (1976) Time Series Analysis : Forecasting and Control. Holden-Day.

BUSTOS O. (1982) General M-estimates for contamined p-th order autoregressive process : consistency and asymptotic normality. Z. Wahrsch. verw. Gebiete **59**, 491-504.

CARBON M., DELECROIX M. (1993) Non parametric vs parametric forecasting in time series : a computational point of view. Applied Stochastic Models and Data Analysis **9**, 215-229.

DELECROIX M. (1987) Sur l'estimation et la prévision non paramétrique des processus ergodiques. Thèse d'État, Univ. Lille 1.

DENBY L., MARTIN R.D. (1979) Robust estimation of the first-order autoregressive parameter. J. Amer. Statist. Assoc. **74**, 140-146.

FOX A.J. (1972) Outliers in time series. J. Roy. Statist. Soc. B **34**, 350-363.

GANNOUN A. (1990) Estimation non paramétrique de la médiane conditionnelle : médianogramme et méthode du noyau. Application à la prévision des processus. Pub. Inst. Stat. Univ. Paris , XXXV, 11-22.

HAYEK A., LECOUTRE J.P. (1990) Convergence uniforme presque sûre d'une classe de prédicteurs à noyau pour un processus fortement mélangeant. Pub. Inst. Stat. Univ. Paris, XXXV, 23-41.

PANKRATZ A. (1983) Forecasting with univariate Box-Jenkins models : concept and cases. Wiley.

ROBINSON P. (1984) Robust nonparametric autoregression. *Robust and nonlinear time series analysis. Lecture Notes in Statist.* **26**, 247-255.