Estimation non paramétrique de la densité et de la régression - Prévision non paramétrique

Michel CARBON et Christian FRANCQ

Laboratoire de Probabilités et Statistique
Bâtiment M2
59650 Villeneuve d'Ascq Cedex
FRANCE

Résumé

Après un aperçu sur l'estimation non paramétrique de la densité et de la régression, nous détaillons et interprétons une méthode de prévision, dite prévision non paramétrique. Nous en montrons les différents aspects aussi bien techniques que pratiques, et la comparons sur quelques exemples à la méthodologie de Box et Jenkins.

1 Introduction

Considérons une série chronologique $x_1, x_2, \cdots, x_n$. A partir de ces n observations, on voudrait prévoir x_{n+h} (h est l'horizon de prévision). La première question qui se pose est « quelle méthode choisir » ? Bien que la question soit simple, la réponse générale est hélas décevante. Il n'existe pas de méthode meilleure que les autres, dans tous les cas de figures. Depuis quelques temps apparaissent de nouvelles méthodes, appelées non paramétriques, qui semblent apporter un regard nouveau sur la question. Ces nouvelles techniques présentent l'avantage d'une mise en œuvre très aisée, d'une interprétation plus intuitive, et d'une certaine robustesse, vis à vis des méthodes dites de Box et Jenkins.

Dans ce papier, nous allons faire un petit panorama de certaines approches non paramétriques, en vue surtout de montrer comment s'y insère la prévision non paramétrique. En particulier, nous commençons par l'estimation non paramétrique de la densité. Le lecteur intéressé est renvoyé aux travaux de Bosq et Lecoutre (1987) ou Praksa Rao (1983). Nous poursuivons par une présentation assez rapide de la régression non paramétrique, dont la prévision non paramétrique est un cas particulier. Nous tentons de montrer que, finalement, l'interprétation est tout à fait naturelle. Nous terminons enfin par des exemples montrant une certaine supériorité de la méthode non paramétrique vis à vis de la méthodologie de Box et Jenkins.

2 Estimation non paramétrique de la densité

2.1 Histogramme

L'estimateur le plus rudimentaire pour estimer une densité est l'histogramme des fréquences. Supposons que l'on ait $x_1, \ldots, x_n$, n observations issues d'une même loi de probabilité de densité f, où f est à support borné $[a, b[$. Pour estimer cette densité f par la méthode de l'histogramme, ce qui revient à approcher f par une fonction en escaliers, on découpe $[a, b[$ en k classes $[\alpha_i, \alpha_{i+1}[$ où $i = 1, \ldots, k$, avec $a = \alpha_1$ et $b = \alpha_{k+1}$. L'estimateur histogramme s'écrit alors : $\forall t \in [a, b[, \quad \exists i = 1, \ldots, k$ tel que $t \in [\alpha_i, \alpha_{i+1}[$ et

$$\hat{f}_n(t) = \frac{f_i}{\alpha_{i+1} - \alpha_i},$$

où f_i est la fréquence du nombre de points de la classe correspondante. Ce que l'on peut encore écrire plus concisément : $\forall t \in [a, b[$,

$$\hat{f}_n(t) = \sum_{i=1}^{k} \frac{f_i}{\alpha_{i+1} - \alpha_i} \cdot \mathbf{1}_{[\alpha_i, \alpha_{i+1}[}(t),$$

où

$$f_i = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{[\alpha_i, \alpha_{i+1}[}(x_j).$$

Soit encore :

$$\hat{f}_n(t) = \sum_{i=1}^{k} \frac{1}{n(\alpha_{i+1} - \alpha_i)} \sum_{j=1}^{n} \mathbf{1}_{[\alpha_i, \alpha_{i+1}[}(x_j). \tag{1}$$

Pour simplifier les notations, on supposera maintenant les classes de même largeur, c'est-à-dire que pour tout $i = 1, \ldots, k$; $\alpha_{i+1} - \alpha_i = b(n)$. Il est aisé de remarquer que $\hat{f}_n$ est une densité de probabilité. Si on pense à la convergence de cet estimateur, il est clair que $\hat{f}_n$ sera d'autant plus proche de la vraie densité f que les largeurs de classe seront plus étroites, d'où la nécessité d'imposer que $b(n) \to 0$ quand $n \to \infty$. En revanche, il ne faut pas que $b(n)$ tende trop vite vers 0, sinon on pourrait avoir des classes ne contenant aucun point, et donc une fonction en escalier $\hat{f}_n$ avec des marches d'ordonnée nulle, très éloignée de la réalité. Il faut donc que, certes $b(n)$ tende vers 0 avec n, et que, malgré cela, il « tombe » de plus en plus de points dans chaque classe, ce que l'on peut résumer dans la condition :

$$nb(n) \to \infty \quad \text{quand} \quad n \to \infty.$$

L'erreur quadratique moyenne $E[\hat{f}_n(x) - f(x)]^2$ est de l'ordre de $n^{-2/3}$, pour un choix optimal de $b(n)$.

2.2 Histogramme mobile

Nous allons tenter d'améliorer cet estimateur histogramme $\hat{f}_n$. Considérons la classe $C_i = [\alpha_i, \alpha_{i+1}[$, et imaginons que le point t de C_i où l'on veut estimer $f(t)$ par $\hat{f}_n(t)$ se situe près de l'extrémité α_i (voir figure 1). Alors tous les points de la classe

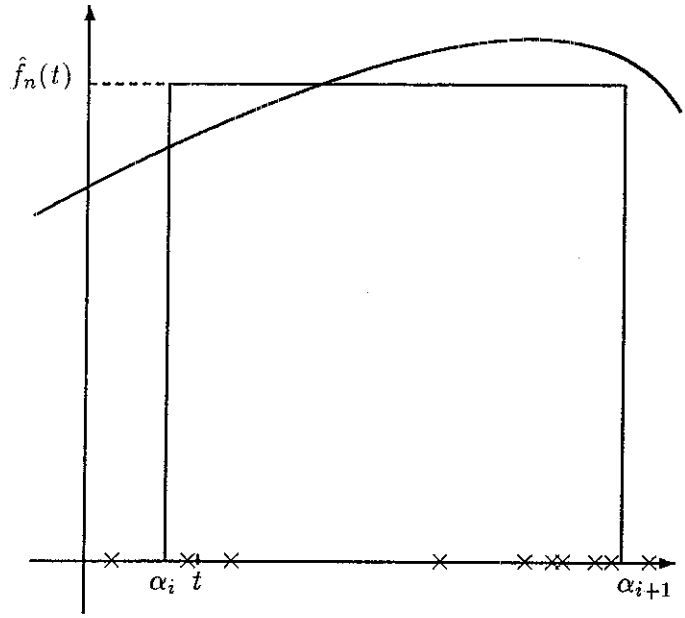


FIG. 1: Histogramme des fréquences

C_i interviennent dans le calcul de $\hat{f}_n$, mais on se rend compte qu'un point situé près de α_{i+1} sera pris en compte, alors qu'il est assez éloigné de t, et qu'un point situé tout près de t dans la classe C_{i-1} n'entre pas en ligne de compte dans le calcul de $\hat{f}_n$.

Pour remédier à cet inconvénient, on peut alors utiliser l'histogramme mobile, qui est un translaté de l'histogramme de manière à ce que le point t où l'on estime, se retrouve au centre d'une classe, plus précisément au centre de la classe $[t-h(n), t+h(n)[$ où $h(n)$ désigne la demi-largeur d'une classe. L'estimateur histogramme mobile s'écrit alors :

$$\hat{f}_n(t) = \frac{1}{2nh(n)} \sum_{j=1}^{n} \mathbf{1}_{[t-h(n),t+h(n)[}(x_j). \tag{2}$$

Remarquons que :

$$t - h(n) \leq x_j < t + h(n) \quad \Longleftrightarrow \quad -1 \leq \frac{x_j - t}{h(n)} < 1. \tag{3}$$

D'où :

$$\hat{f}_n(t) = \frac{1}{2nh(n)} \sum_{j=1}^{n} \mathbf{1}_{[-1,1[}\left(\frac{x_j - t}{h(n)}\right).$$

L'estimateur s'écrit alors :

$$\hat{f}_n(t) = \frac{1}{nh(n)} \sum_{j=1}^{n} \mathbf{K}\left(\frac{x_j - t}{h(n)}\right),$$

où

$$\mathbf{K}(x) = \frac{1}{2}\mathbf{1}_{[-1,1[}(x).$$

2.3 Estimation par la méthode du noyau

L'estimateur ainsi construit peut encore être amélioré. En effet, maintenant que la classe est centrée en t, on peut tout de même remarquer que tous les points de cette classe ont le même rôle quant au calcul de $\hat{f}_n(t)$. Il serait plus judicieux de penser que plus un point est proche de t, plus il doit intervenir dans le calcul de $\hat{f}_n(t)$. L'idée alors la plus naturelle est de pondérer les observations en mettant d'autant plus de poids qu'on se trouve proche de t, et d'autant moins qu'on s'en trouve éloigné.

On a déjà vu un exemple de fonction de poids, notée $\mathbf{K}$ au paragraphe précédent. C'était une densité de probabilité (la loi uniforme sur $[-1/2, 1/2[$). Cette fonction de poids est trop brutale et ne répond pas à nos préoccupations. On choisira donc des fonctions de poids dans des classes plus larges de densités, comprenant notamment des densités à support non borné, et ayant un seul mode à l'origine (par exemple la loi normale centrée réduite).

On notera que, puisque $\mathbf{K}$ est une densité de probabilité, $\hat{f}_n(t)$ est aussi une densité de probabilité. L'estimateur à noyau s'écrit :

$$\hat{f}_n(t) = \frac{1}{nh(n)} \sum_{j=1}^{n} \mathbf{K}\left(\frac{x_j - t}{h(n)}\right) \tag{4}$$

et $\mathbf{K}$ s'appelle un noyau. Quant aux propriétés de convergence, on montre que $\hat{f}_n$ est asymptotiquement sans biais. On peut aussi montrer que, pour que $Var[\hat{f}_n(t)] \to 0$, il faut que $h(n) \to 0$ et $nh(n) \to +\infty$ quand $n \to \infty$. La vitesse de convergence, cette fois, pour le choix le meilleur de $h(n)$, et au sens de l'erreur quadratique moyenne est de l'ordre de $n^{-4/5}$.

Il nous reste à parler du choix du noyau et du paramètre $h(n)$. Ces choix ne peuvent s'effectuer que par utilisation de certains critères. Sans entrer dans tous les détails, il s'avère que le choix du noyau n'a pas d'influence majeure s'il est choisi dans une classe raisonnable d'estimateurs. Le choix d'un noyau gaussien, par exemple, est tout à fait recommandé. Le choix du $h(n)$ est crucial par contre. On recommande le choix suivant :

$$h(n) = S_n n^{-1/5},$$

où S_n désigne l'écart-type estimé des observations.

Un point essentiel plaidant en la faveur de cette technique d'estimation est la formule (4) qui reste quasi inchangée dans le cas multidimensionnel :

$$\hat{f}_n(t) = \frac{1}{nh^s(n)} \sum_{j=1}^{n} \mathbf{K}\left(\frac{x_j - t}{h(n)}\right) \tag{5}$$

où t, les x_j sont dans $\mathbf{R}^s$, et où $\mathbf{K}$ est une densité définie sur $\mathbf{R}^s$. $\mathbf{K}$ est en général choisi comme produit de noyaux de $\mathbb{R}$ dans $\mathbb{R}$. $h(n)$ est obtenu par des techniques de validation. Pour plus de détails voir Devroye et Györfi (1985).

2.4 Régression non paramétrique

Supposons que le comportement d'une variable aléatoire Y soit lié à une autre variable X. Il est classique d'essayer d'exprimer Y linéairement en fonction de X. C'est le problème bien connu de la régression linéaire. Il serait peut-être mieux de

tenter d'exprimer Y par une fonction R non nécessairement linéaire de X, c'est-à-dire de trouver une expression de la forme :

$$Y = R(X) + \varepsilon \quad \text{où} \quad \varepsilon \quad \text{est le résidu.}$$

On peut alors chercher à déterminer R comme solution du problème de minimisation :

$$\min E[Y - R(X)]^2.$$

On obtient, sous certaines conditions de régularité (essentiellement d'intégrabilité de Y) :

$$R(X) = E(Y|X),$$

c'est-à-dire la régression de Y en X.

2.5 Régressogramme

D'un point de vue pratique, à partir de n observations (x_i, y_i), on va estimer $R(t) = E(Y|X = t)$, en fabriquant des classes C_i où se situent ou non les x_j. Pour la classe C_i où se trouve le point t, on effectue la moyenne des y_j correspondants aux x_j de cette classe C_i.

En notant k le nombre de points x_j de la classe C_i, pour tout t de C_i, on estime $R(t)$ par :

$$\hat{R}_n(t) = \frac{1}{k}\sum_{j=1}^{k} y_j \quad \text{avec} \quad k = \sum_{j=1}^{n} \mathbf{1}_{C_i}(x_j).$$

D'où

$$\hat{R}_n(t) = \frac{\sum_{j=1}^{n} \mathbf{1}_{C_i}(x_j) y_j}{\sum_{j=1}^{n} \mathbf{1}_{C_i}(x_j)}.$$

Bien sûr, comme pour l'histogramme, le régressogramme $\hat{r}_n$ est constant sur chaque classe C_i. Pour l'améliorer, on va suivre alors la même démarche que pour l'histogramme.

2.6 Régressogramme mobile

Comme pour l'histogramme mobile, la première amélioration consiste à centrer la classe en le point t où l'on estime la régression. Ce régressogramme mobile s'écrit :

$$\hat{R}_n(t) = \frac{\sum_{j=1}^{n} \mathbf{1}_{[t-h(n),t+h(n)[}(x_j) y_j}{\sum_{j=1}^{n} \mathbf{1}_{[t-h(n),t+h(n)[}(x_j)},$$

ce qui, grâce à la remarque (3), donne :

$$\hat{R}_n(t) = \frac{\sum_{j=1}^{n} \mathbf{1}_{[-1,1[}[(x_j - t)/h(n)] y_j}{\sum_{j=1}^{n} \mathbf{1}_{[-1,1[}[(x_j - t)/h(n)]}.$$

2.7 Régression par la méthode du noyau

Par analogie au cas de l'histogramme, la dernière étape consiste à remplacer la densité de probabilité $\frac{1}{2}\mathbf{1}_{[-1,1[}$ par un noyau quelconque du type décrit dans le cas de l'histogramme. La régression par la méthode du noyau s'écrit alors :

$$\hat{r}_n(t) = \frac{\sum_{j=1}^{n} \mathbf{K}[(x_j - t)/h(n)]y_j}{\sum_{j=1}^{n} \mathbf{K}[(x_j - t)/h(n)]} \tag{6}$$

La vitesse de convergence, au sens de l'erreur quadratique moyenne, est également de l'ordre de $n^{-4/5}$.

Il est aussi à remarquer que la formule (6) est inchangée quand les observations sont dans $\mathbf{R}^s$ (*i.e.* (X, Y) est dans $\mathbf{R}^s \times \mathbf{R}$), à la seule différence que t, et les x_i sont dans $\mathbf{R}^s$. Pour plus de détails voir Härdle (1989).

3 Prévision non paramétrique

3.1 Lien entre régression et prévision

On considère un processus stationnaire $(X_t)_{t\in\mathbf{Z}}$. On suppose avoir observé $X_1, \ldots, X_T$. Et on cherche à prédire X_{T+k} (avec $k \in \mathbf{N}^*$) à partir des variables observées. Un prédicteur naturel de X_{T+k} basé sur $X_1, \ldots, X_T$ est alors donné par $E(X_{T+k}|X_1, \ldots, X_T)$. Cette espérance conditionnelle est hélas impossible à estimer correctement si on ne fait pas d'hypothèses supplémentaires sur le processus $(X_t)_{t\in\mathbf{Z}}$. Si par contre on suppose a priori que le processus est r-Markovien, alors :

$$E(X_{T+k}|X_T, \ldots, X_1) = E(X_{T+k}|X_T, \ldots, X_{T-r+1}).$$

On cherchera donc à régresser X_{T+k} sur son propre « passé proche » $X_T, \ldots, X_{T-r+1}$. Une telle démarche permet souvent d'obtenir une prévision « raisonnable » de X_{T+k}, même si le processus $(X_t)_{t\in\mathbf{Z}}$ n'est pas r-Markovien. On peut noter que dans les méthodes de lissage, comme dans la méthodologie de Box et Jenkins, une même troncature est usuellement réalisée. Par exemple, quand le prédicteur de X_{T+k} est basé sur une représentation autorégressive infinie du type $\left(X_t = \sum_{i=1}^{\infty} a_i X_{t-i}\right)$, il est effectivement calculé comme une somme finie $\left(\sum_{i=1}^{r} \hat{a}_i X_{t-i}\right)$ avec les $\hat{a}_i$ estimés à partir des données $X_1, \ldots, X_T$. Nous indiquerons plus loin comment identifier le « meilleur » r possible.

3.2 Écriture du prédicteur à noyau

Il est à remarquer que le problème précédent de régression a pour cas particulier celui de la prévision, où on essaie d'expliquer X_{T+k} en fonction de $X_T, \ldots, X_{T-r+1}$. Le prédicteur à noyau s'écrit alors :

$$\hat{L}_{T,k}^{(r)} = \frac{\sum_{t=r}^{T-k} \mathbf{K}[(X_T^{(r)} - X_t^{(r)})/h(T)] X_{t+k}}{\sum_{t=r}^{T-k} \mathbf{K}[(X_T^{(r)} - X_t^{(r)})/h(T)]} \tag{7}$$

avec $X_T^{(r)} = (X_T, \ldots, X_{T-r+1})$ et $X_t^{(r)} = (X_t, \ldots, X_{t-r+1})$. Ce qui s'écrit encore :

$$\hat{L}_{T,k}^{(r)} = \sum_{t=r}^{T-k} \hat{\alpha}_{t,T}^{(r)} X_{t+k}, \tag{8}$$

où

$$\hat{\alpha}_{t,T}^{(r)} = \frac{\mathbf{K}[(X_T^{(r)} - X_t^{(r)})/h(T)]}{\sum_{t=r}^{T-k} \mathbf{K}[(X_T^{(r)} - X_t^{(r)})/h(T)]}. \tag{9}$$

L'écriture (8) montre que le prédicteur non paramétrique est typiquement un lisseur. En effet, (8) est une moyenne pondérée des valeurs passées. La différence fondamentale étant qu'ici $\hat{\alpha}_{t,T}^{(r)}$ est aléatoire, et non déterministe comme dans le cas des lissages exponentiels.

3.3 Interprétation

Si on examine la figure 2, où le paramètre r a été fixé à 3, on peut remarquer que les formules (8) et (9) utiles au calcul des prévisions servent à calculer en chaque point t du passé du processus un poids (voir (9)) $\hat{\alpha}_{t,T}^{(r)}$. La prévision à l'horizon 1 se calcule alors en faisant la moyenne des X_{t+1} avec les pondérations $\hat{\alpha}_{t,T}^{(r)}$. La prévision à l'horizon 2 se calcule de manière similaire en faisant la moyenne des X_{t+2} avec les mêmes pondérations $\hat{\alpha}_{t,T}^{(r)}$. Cela revient donc à chercher dans le passé propre du processus les séquences qui ressemblent le plus à la dernière $X_T^{(r)} = (X_T, \ldots, X_{T-r+1})$. Toutes les séquences $X_t^{(r)} = (X_t, \ldots, X_{t-r+1})$ jouent un rôle bien sûr, mais d'autant plus important qu'il y a similarité entre $X_t^{(r)} = (X_t, \ldots, X_{t-r+1})$ et $X_T^{(r)} = (X_T, \ldots, X_{T-r+1})$.

3.4 Remarques

On peut ainsi remarquer que la prévision à l'horizon 2 ne fait pas intervenir la prévision à l'horizon 1. De manière générale, le calcul des prévisions successives ne fait jamais intervenir les prévisions aux horizons antérieurs, contrairement à ce qui se passe par exemple dans la méthodologie de Box et Jenkins. On en déduit bien entendu une robustesse plus grande de la méthode, par non accumulation des erreurs précédentes.

Nous avons fait au début de ce paragraphe l'hypothèse de stationnarité. Si le processus n'était pas stationnaire, il y aurait lieu de d'abord éliminer la tendance par une différenciation ad hoc. Cette stationnarité est bien sûr nécessaire, sauf en ce qui concerne une saisonnalité éventuelle. Une série présentant une saisonnalité apporte

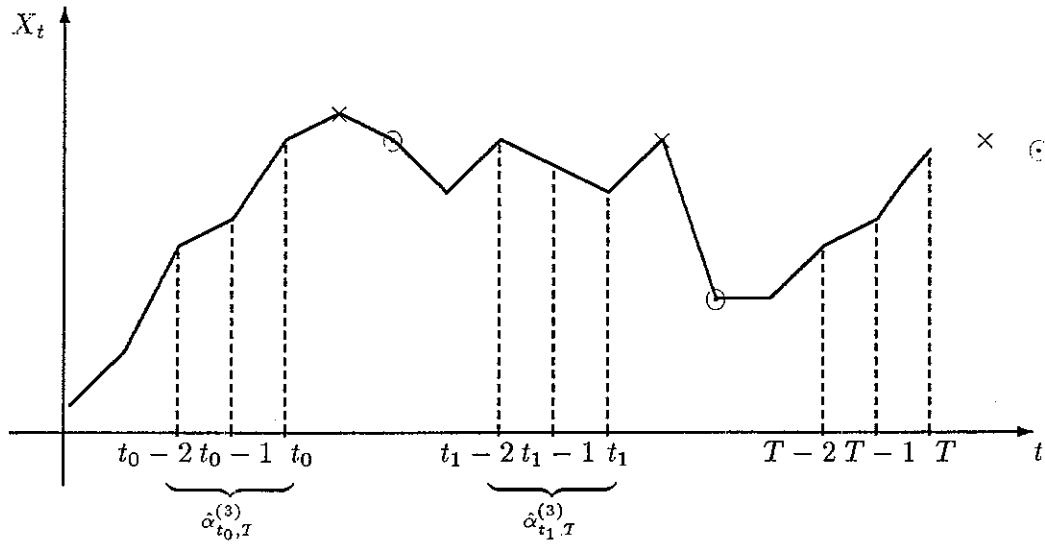


FIG. 2: Interprétation de la méthode de prévision non paramétrique : Le poids $\hat{\alpha}^{(3)}_{t_1,T}$ sera plus important que le poids $\hat{\alpha}^{(3)}_{t_0,T}$ parce que la séquence $(X_{t_0-2}, X_{t_0-1}, X_{t_0})$ « ressemble plus » que $(X_{t_1-2}, X_{t_1-1}, X_{t_0})$ à (X_{T-2}, X_{T-1}, X_T). Le calcul de la prévision à l'horizon 1 (respectivement l'horizon 2), qui est matérialisée par le symbole $\times$ (resp. $\odot$), fait intervenir les valeurs de X_{t_0+1} et de X_{t_1+1} (resp. X_{t_0+2} et X_{t_1+2}) (également indiquées par le symbole $\times$ (resp. $\odot$)) pondérées par les poids $\hat{\alpha}^{(3)}_{t_0,T}$ et $\hat{\alpha}^{(3)}_{t_1,T}$.

une information supplémentaire importante et qu'on a intérêt à exploiter pour affiner nos prévisions. Intuitivement, si l'on pense en termes de recherche de similarité du passé au présent, on voit bien que, s'il y a saisonnalité, on a intérêt à choisir un r qui soit la longueur d'une saisonnalité ou d'une demi-saison, par exemple.

Maintenant, imaginons dans la série un point « aberrant » en t_0. Alors, toute séquence du passé contenant t_0 aura une similarité avec $X_T^{(r)} = (X_T, \ldots, X_{T-r+1})$ quasi nulle, et ainsi n'interviendra nullement dans le calcul des prévisions. Il n'y a même pas le problème de savoir si un point est ou non aberrant, ou au nom de quel critère il l'est ou non, il est éliminé de manière automatique dans les calculs. C'est aussi un avantage notable de cette technique.

Il faut enfin noter la facilité de mise en œuvre informatique, ce qui n'est pas un mince avantage.

3.5 Choix du paramètre de troncation r

Pour un r fixé, on peut, pour chaque t $(r+k \leq t \leq T-k)$, calculer une prévision comme on l'a fait précédemment par :

$$\hat{X}_{t+k}^{(r)} = \sum_{i=r}^{t-k} \hat{\alpha}_{i,t}^{(r)} X_{i+k} \tag{10}$$

ce qui donne une erreur de prévision estimée :

$$e_t^k = |\hat{X}_{t+k}^{(r)} - X_{t+k}| \tag{11}$$

Il semble alors très classique d'utiliser un critère d'erreur quadratique pour estimer le r adéquat, c'est-à-dire minimiser :

$$(T-2k-r+1)^{-1} \sum_{t=r+k}^{T-k} (e_t^k)^2 \tag{12}$$

Bien sûr, on se fixe une valeur r_0 maximale et on utilise alors (12) pour estimer r, avec $0 < r \leq r_0$. Il faut bien entendu remarquer que les premières prévisions calculées par (10) sont pauvres en information car elles ne dépendent que de peu d'observations. Il y a alors lieu de ne commencer les prévisions qu'à partir d'une date raisonnable t_0. (voir Carbon-Delecroix 1993 pour des précisions sur cette validation).

3.6 Intervalle de prévision

On va réutiliser les erreurs de prévision (voir (11)) e_t^k pour $t = r+k, \ldots, T-k$. On définit $\hat{q}^k$ le quantile empirique à 95 % (raisonnement identique pour d'autres niveaux) basé sur les e_t^k. Pour cela notons N_t^k le nombre de e_i^k supérieurs à e_t^k et t_0 un indice tel que $N_{t_0}^k$ soit le plus grand entier vérifiant

$$\frac{1}{T-2k-r+1} N_{t_0}^k \leq 0,05.$$

On a alors

$$\hat{q}^k = e_{t_0}^k,$$

ce qui donne un estimé de l'intervalle de prévision (à 95 %) pour X_{T+k} :

$$[\hat{X}_{T+k} - \hat{q}^k, \hat{X}_{T+k} + \hat{q}^k]$$

On pourra remarquer que cette manière d'estimer un intervalle de prévision ne fait aucune hypothèse sur la forme des lois sous-jacentes (Box et Jenkins font l'hypothèse de normalité). De plus, s'il y a des points aberrants, ils se retrouvent dans les 5 % restants et n'interviennent nullement ici non plus dans le calcul de l'intervalle de prévision.

4 Exemples et comparaisons

Nous étudions ici quelques séries temporelles. Pour chacune d'entre elles, nous allons comparer les efficacités de la méthode non paramétrique vis à vis de la méthode de Box et Jenkins. Ces diverses séries sont ou bien simulées ou bien extraites de la littérature. Pour les séries simulées, qui sont des modèles ARMA, nous avons bien sûr retenu l'identification optimale pour la méthode de Box et Jenkins. Pour les autres séries, prises dans la littérature, nous avons retenu les identifications indiquées par les auteurs, pour éviter toute ambiguïté. Pour chacune d'entre elles, nous avons tronqué la série de manière à avoir une comparaison des prévisions par rapport aux vraies observations. Dans les figures 3 à 14 le graphique (a) représente les données utilisées pour faire l'ensemble des prévisions. Le graphique (b) permet de comparer les prévisions paramétriques (représentées par les symboles $\times$ joints par un trait en pointillés), les prévisions non paramétriques (représentées par les symboles + joints par des tirets), et les vraies valeurs (représentées comme dans (a)). Les intervalles de confiance pour les prévisions paramétriques et non paramétriques sont respectivement délimités par des tirets et par des pointillés.

Nous avons défini deux critères de comparaison. L'un appelé EMO (erreur moyenne observée) défini par :

$$EMO = K^{-1} \sum_{i=T-K+1}^{I} (|X_i - \hat{X}_i|/|X_i|)$$

où K désigne le nombre maximum de prévisions calculées.

Si on désigne par $\hat{q}_i$ le quantile estimé du quantile théorique q_i défini par :

$$P(|\hat{X}_i - X_i| < q_i) = 0,95,$$

alors, on définit le critère EMP (erreur moyenne de prévision) par :

$$EMP = K^{-1} \sum_{i=T-K+1}^{T} (\hat{q}_i/|\hat{X}_i|)$$

Pour les deux méthodes (non paramétrique et Box-Jenkins), nous avons préalablement stationnariser la série de la même manière. Bien entendu, dans les cas où il y a saisonnalité, on ne traite pas préliminairement celle-ci pour la méthode non paramétrique.

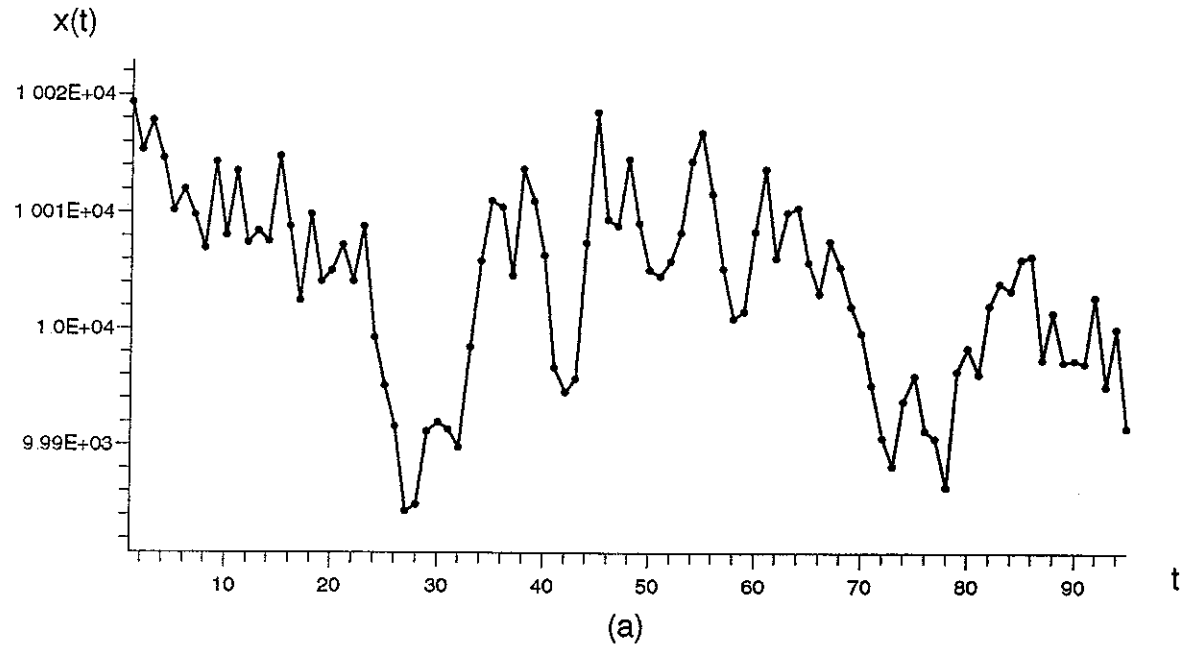
Sur les différentes figures qui suivent nous constatons que :

1. Dans la plupart des simulations de modèles ARMA (figures 3, 6, 7, 10 et 14), modèles bien adaptés à la méthodologie Box et Jenkins, les résultats obtenus par la méthode non paramétrique sont très proches de ceux obtenus par la méthode de Box et Jenkins.

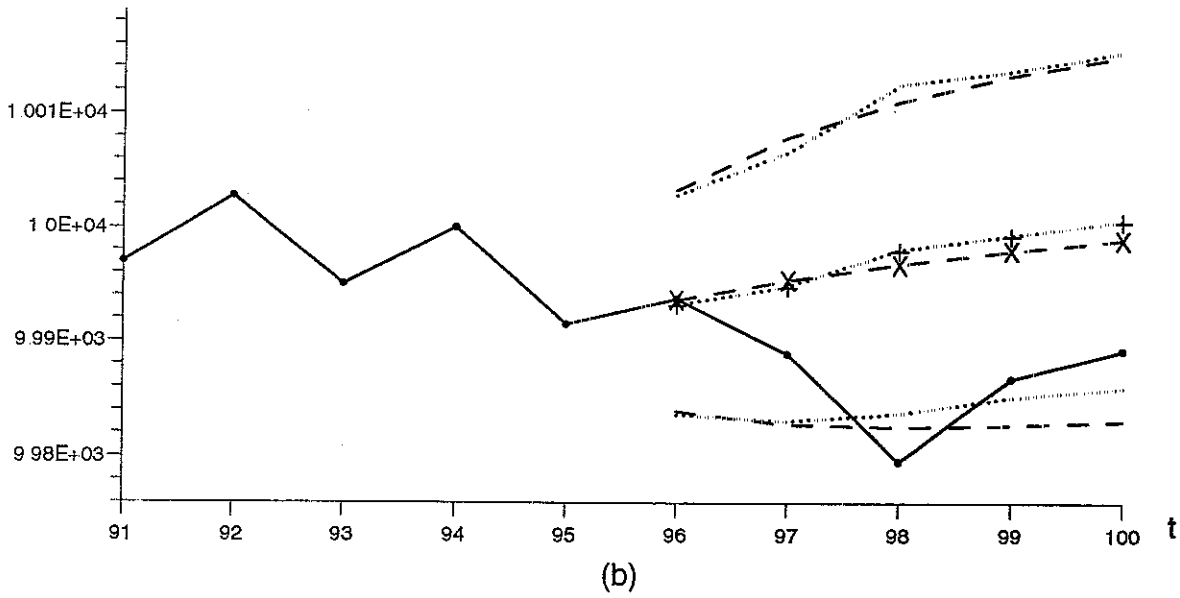
2. Pour certaines séries on remarque parfois un meilleur comportement des prévisions non paramétriques (voir par exemple la figure 11).

3. L'intervalle de prévision dans le cas non paramétrique s'élargit moins vite que dans le cas Box et Jenkins, attestant de la robustesse plus grande de la méthode (voir par exemple la figure 9).

Pour d'autres exemples, voir Carbon et Delecroix (1993).



(a)

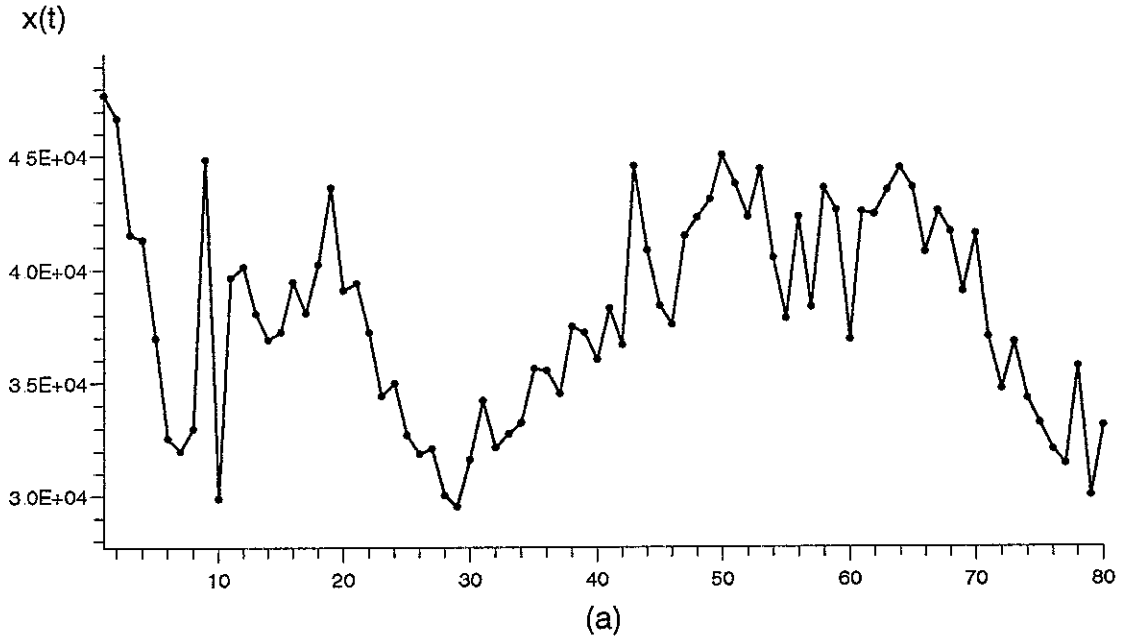


(b)

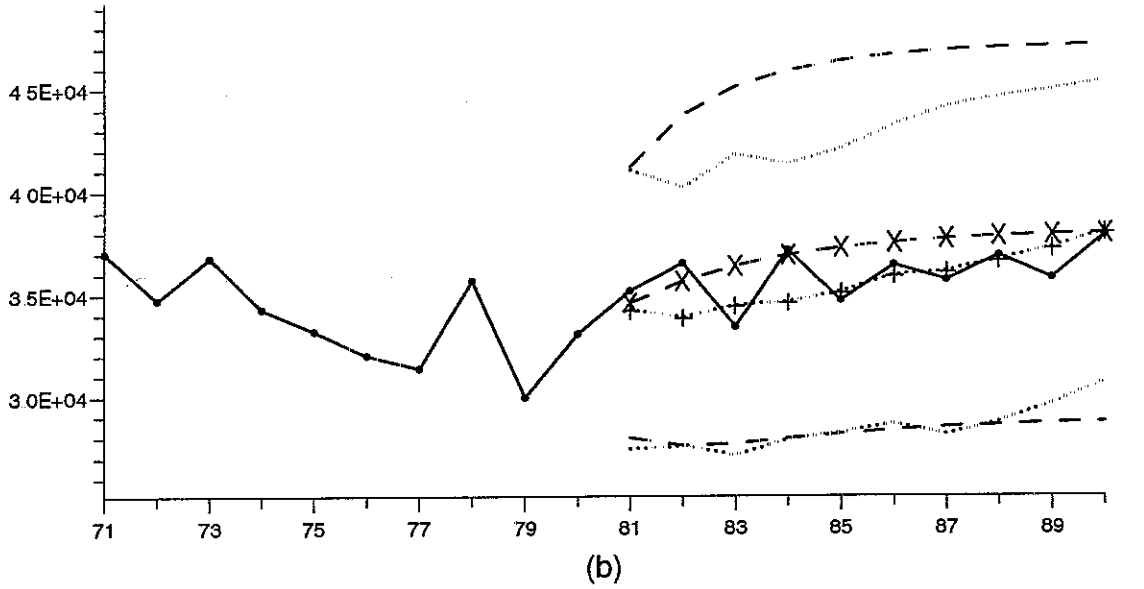
	EMO	*EMP*
Box et Jenkins	0.090%	0.136 %
Non paramétrique	0.098%	0.130%

(c)

FIG. 3: Comparaison des prévisions paramétriques ($\times$) et non paramétriques ($+$) pour l'AR(1) $X_t = 0.9X_{t-1} + 10000 + \epsilon_t$, où (ϵ_t) est un bruit blanc fort $\mathcal{N}(0,5)$ (voir la section 4 pour la description des graphiques)



(a)

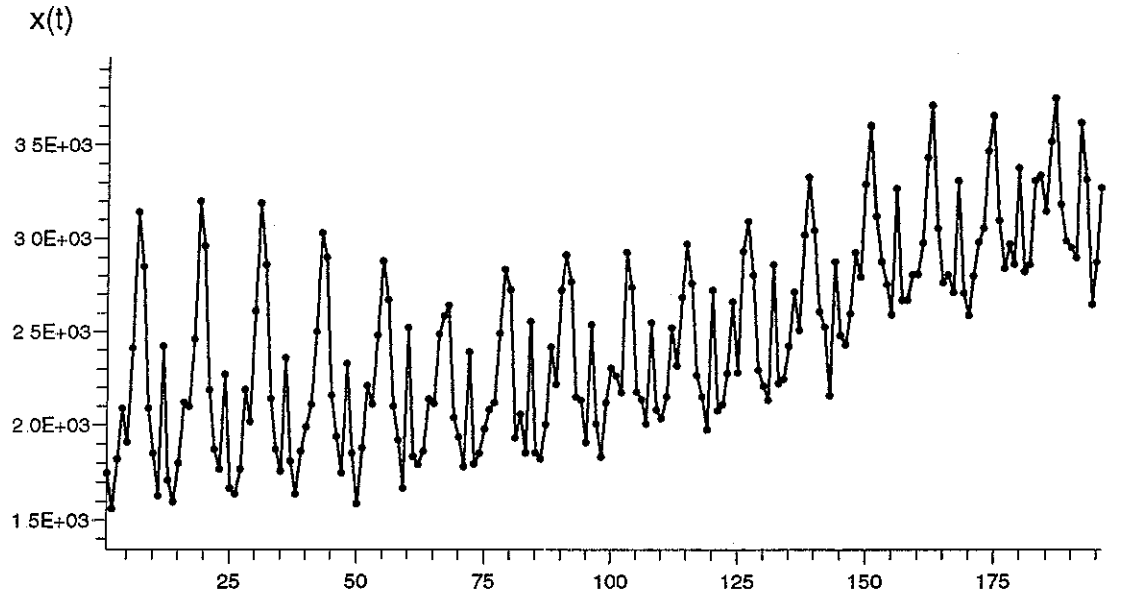


(b)

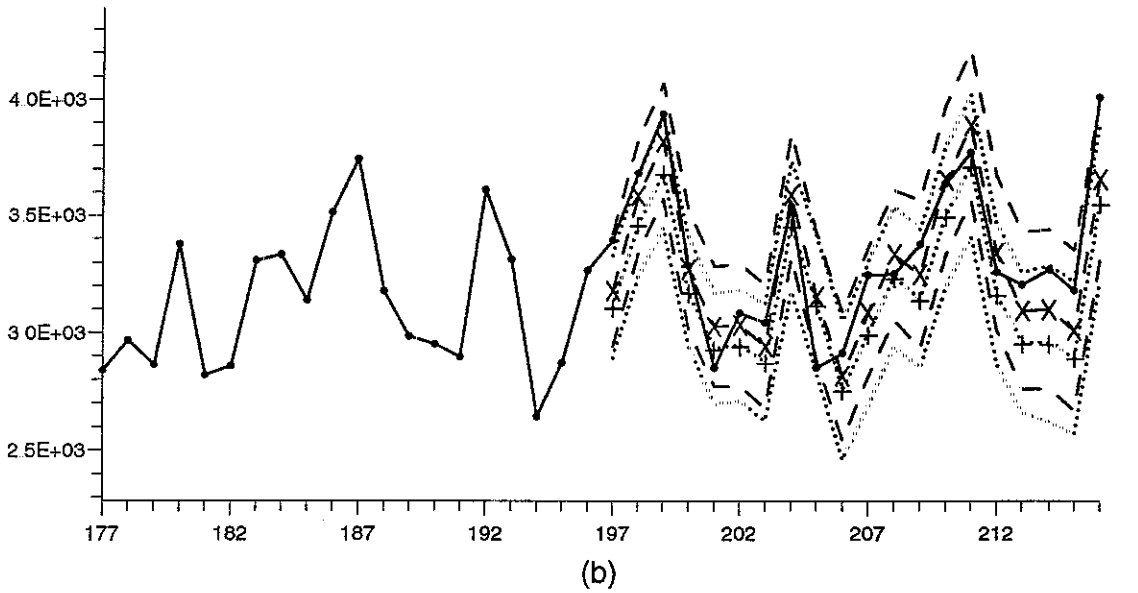
	EMO	*EMP*
Box et Jenkins	3.83%	23.63%
Non paramétrique	2.86%	20.29%

(c)

FIG. 4: Comparaison des prévisions paramétriques et non paramétriques pour la production de charbon (cf. Pankratz, 1983)



(a)

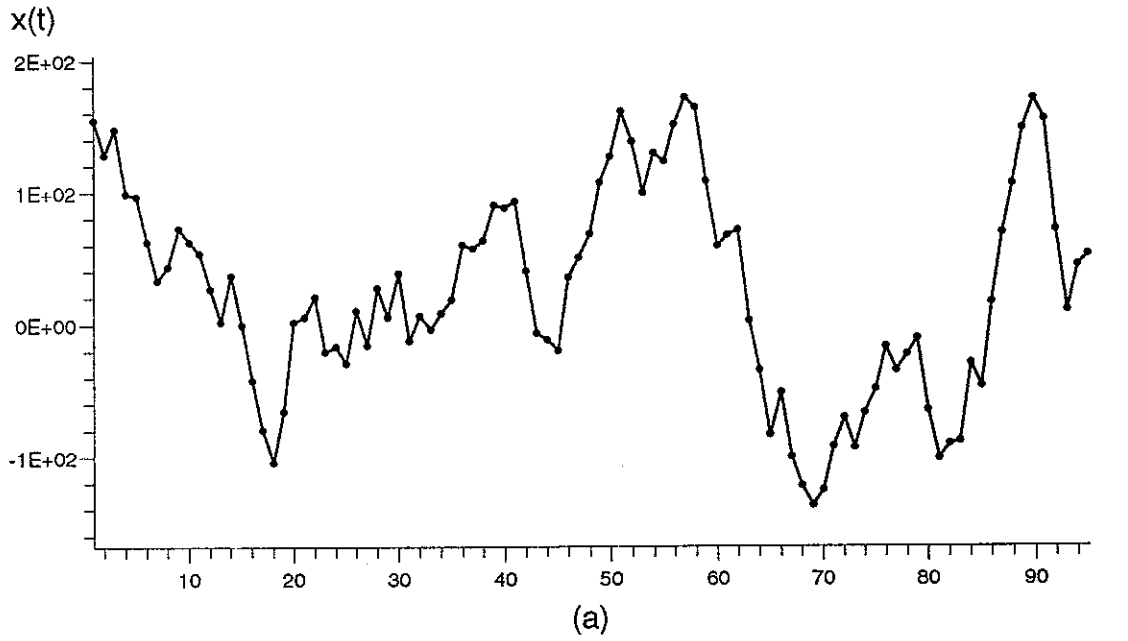


(b)

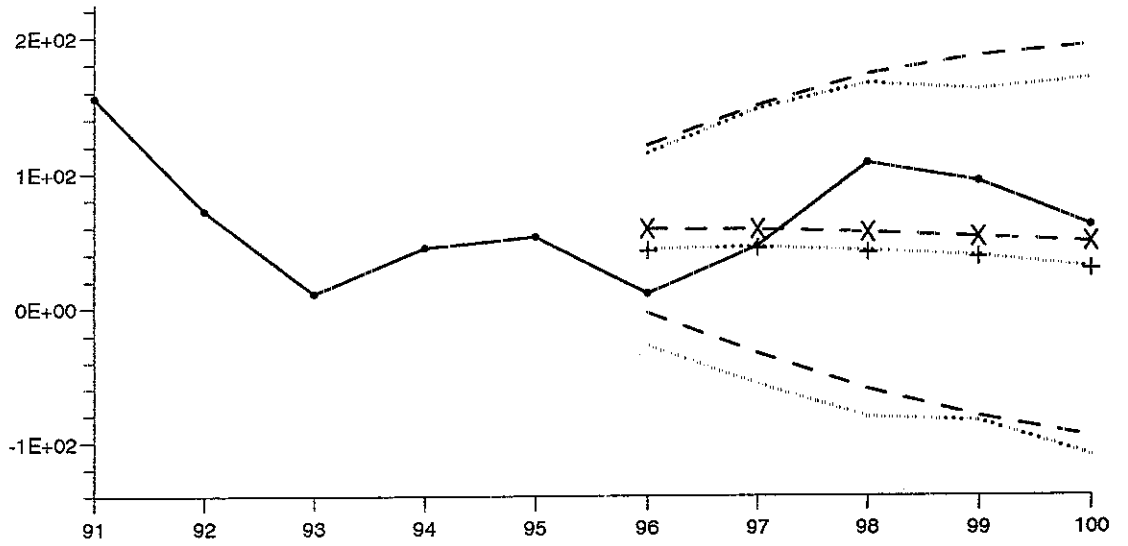
	EMO	*EMP*
Box et Jenkins	3.95%	8.85%
Non paramétrique	5.77%	8.85%

(c)

FIG. 5: Comparaison des prévisions paramétriques et non paramétriques pour le traffic voyageur (cf. Gouriéroux, 1990)



(a)

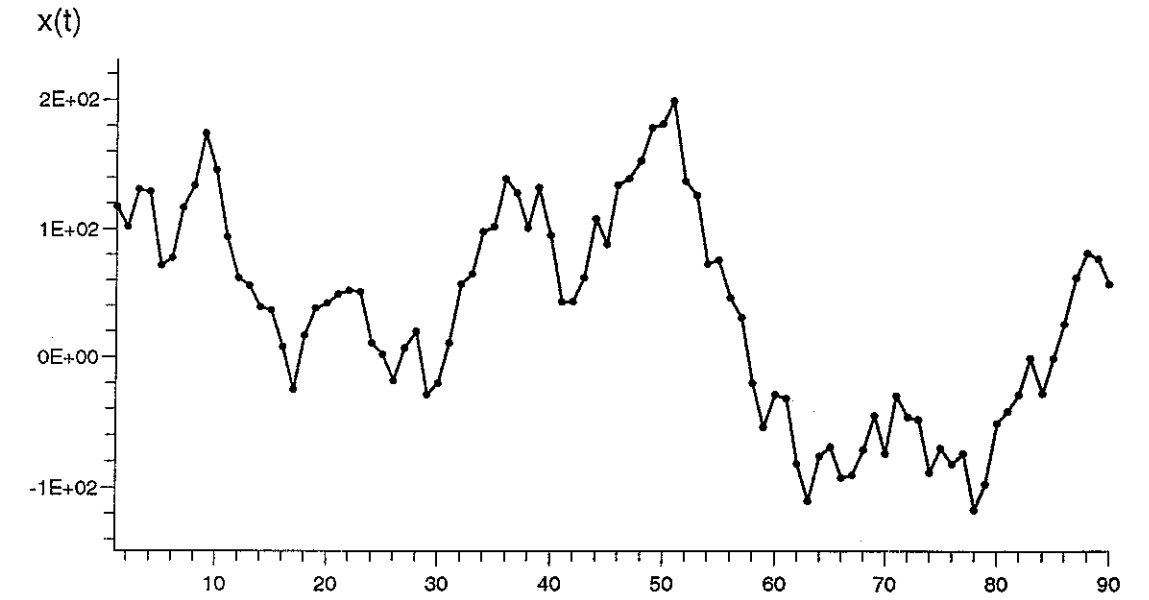


(b)

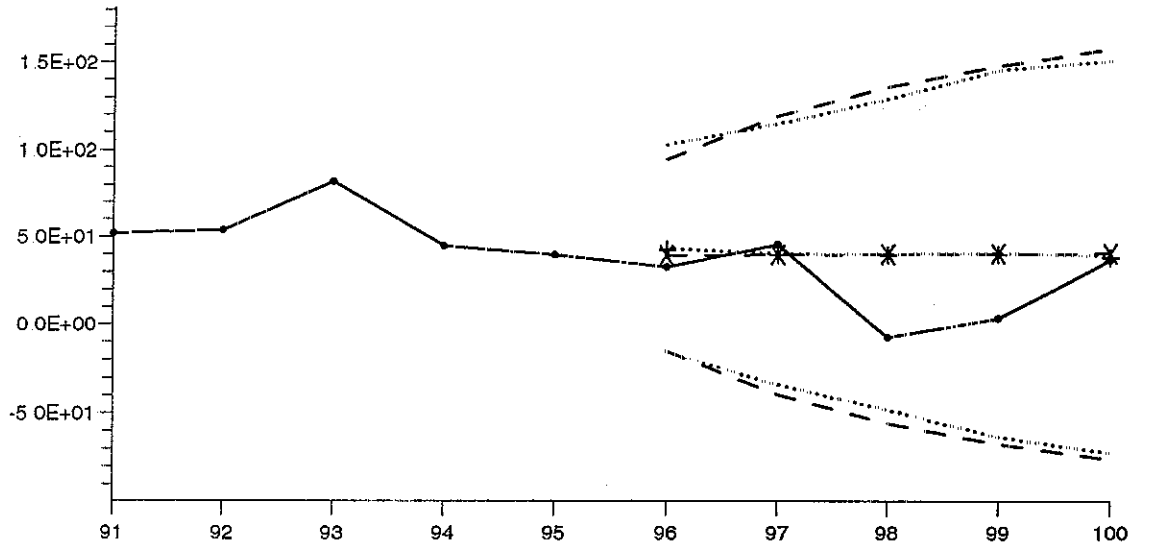
	EMO	*EMP*
Box et Jenkins	1.15	2.04
Non paramétrique	0.94	2.93

(c)

FIG. 6: Comparaison des prévisions paramétriques et non paramétriques sur une simulation de la série $X_t = \epsilon_t + 0.2X_{t-1} + X_{t-2} - 0.3X_{t-3}$, où (ϵ_t) est une suite i.i.d. de loi uniforme sur $[-49, 49]$



(a)

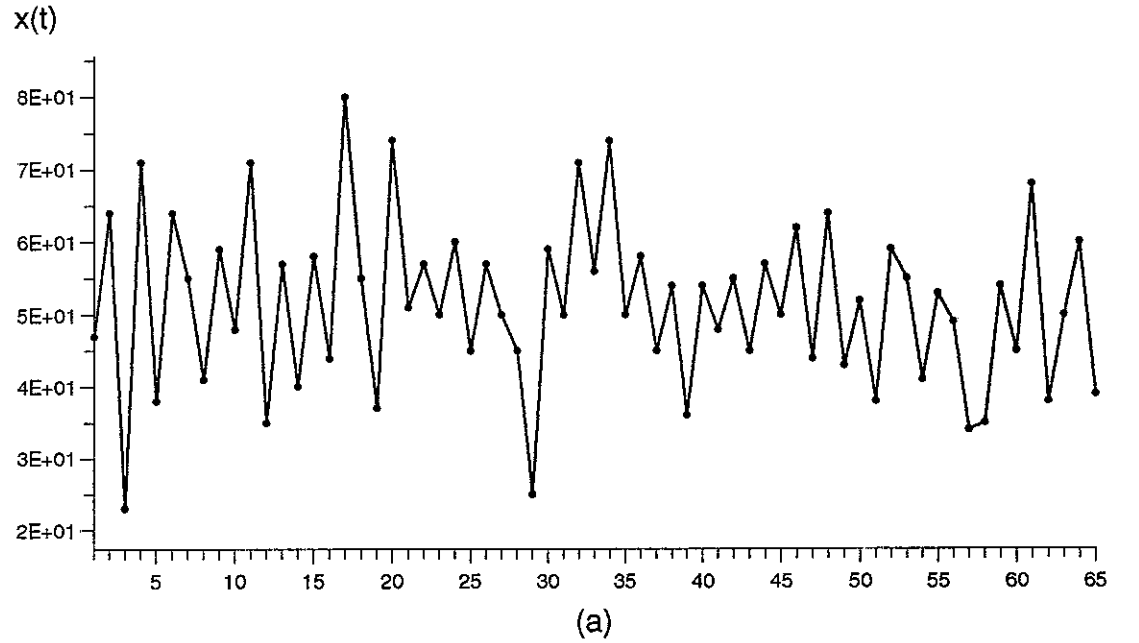


(b)

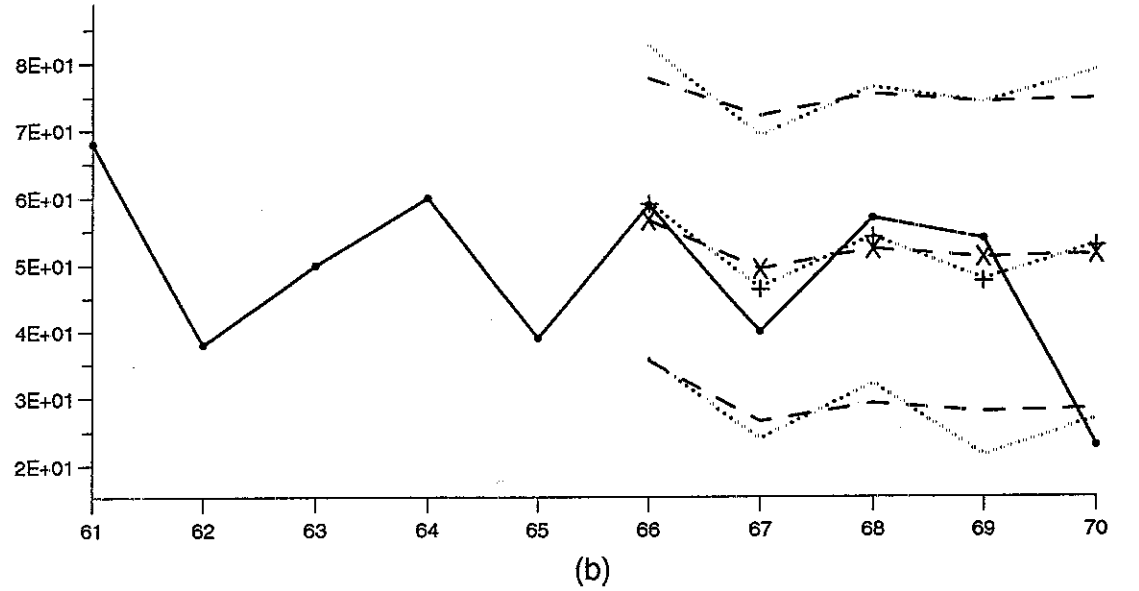
	EMO	*EMP*
Box et Jenkins	3.27	2.25
Non paramétrique	3.33	2.14

(c)

FIG. 7: Comparaison des prévisions paramétriques et non paramétriques sur une simulation de la série $X_t = \epsilon_t + X_{t-1} - 0.1X_{t-2}$, où (ϵ_t) est une suite i.i.d. de loi uniforme sur $[-49, 49]$



(a)

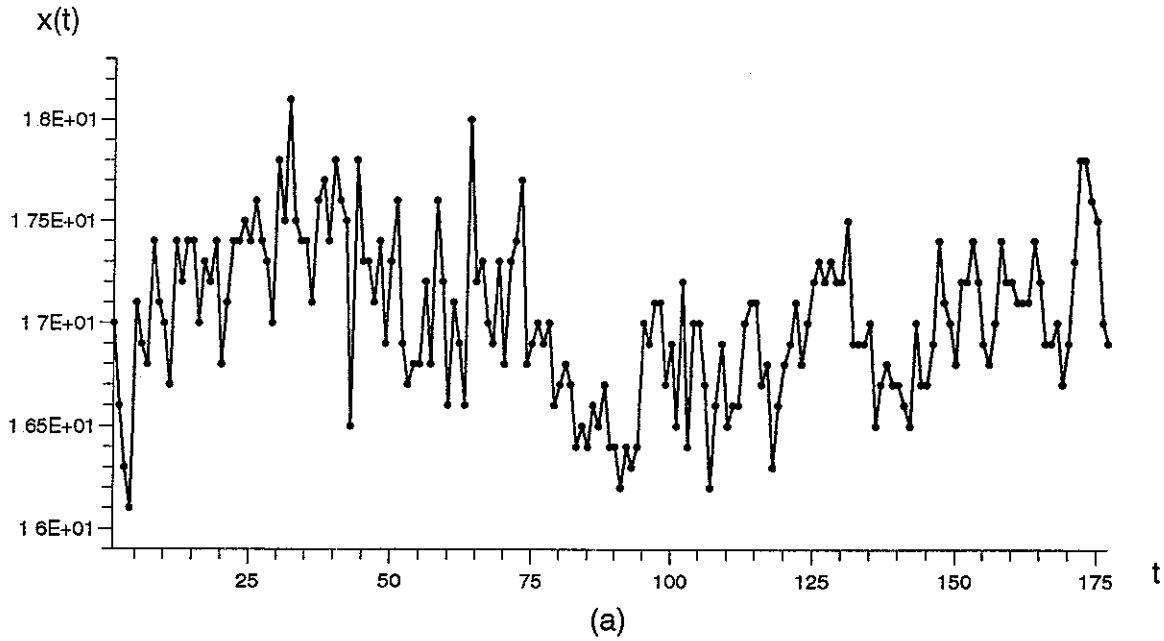


(b)

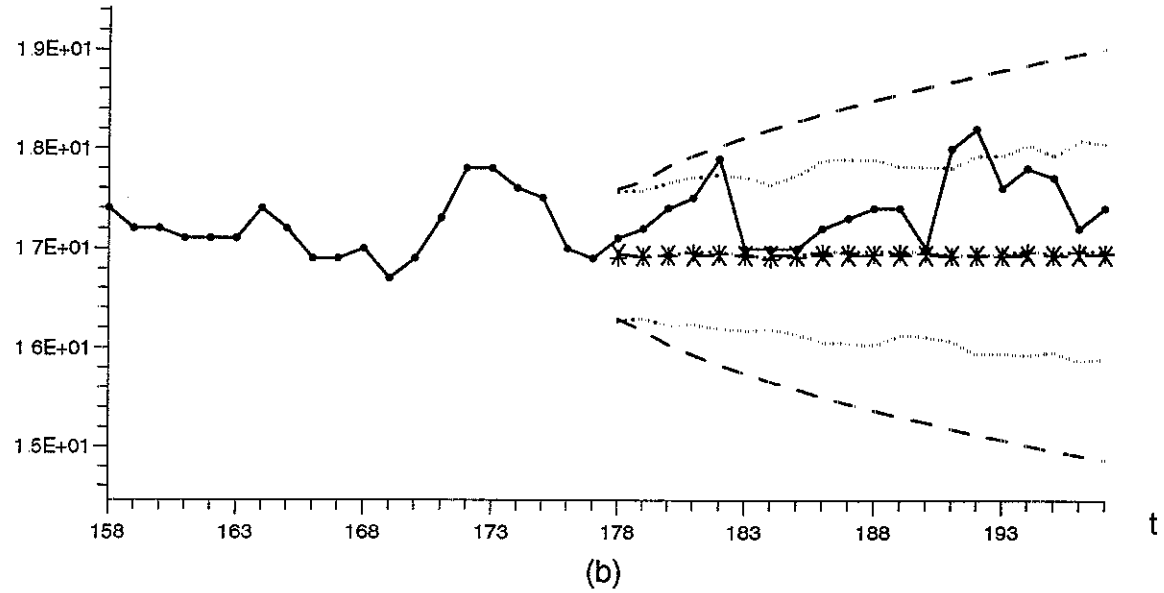
	EMO	*EMP*
Box et Jenkins	0.33	0.43
Non paramétrique	0.33	0.47

(c)

FIG. 8: Comparaison des prévisions paramétriques et non paramétriques sur des données d'un processus chimique (cf. Box et Jenkins, 1970)



(a)

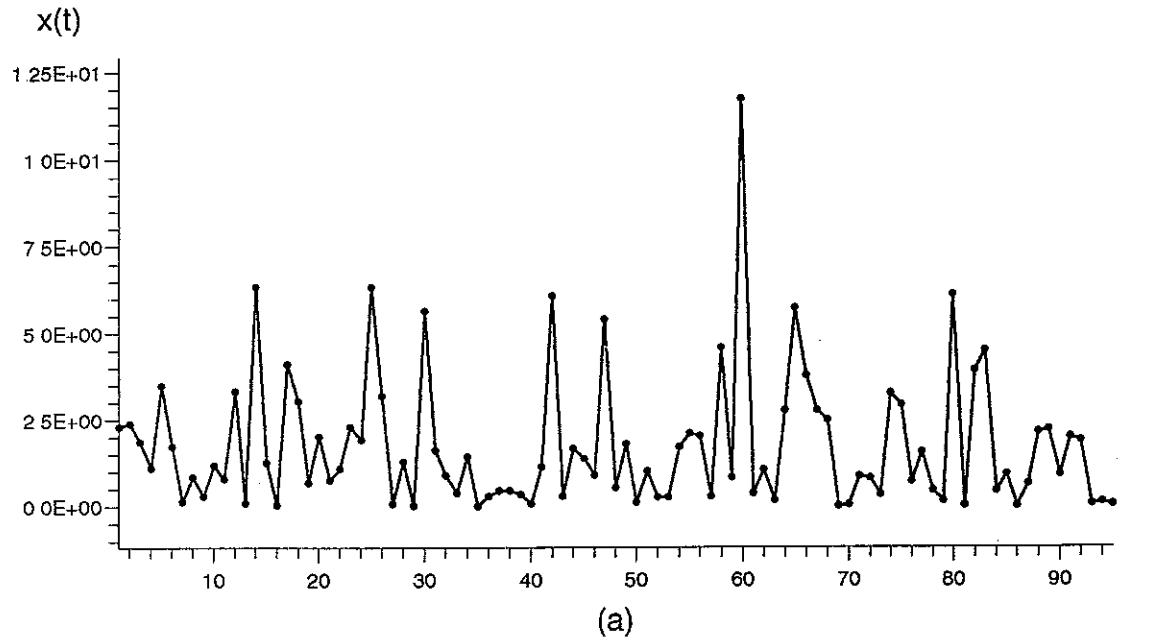


(b)

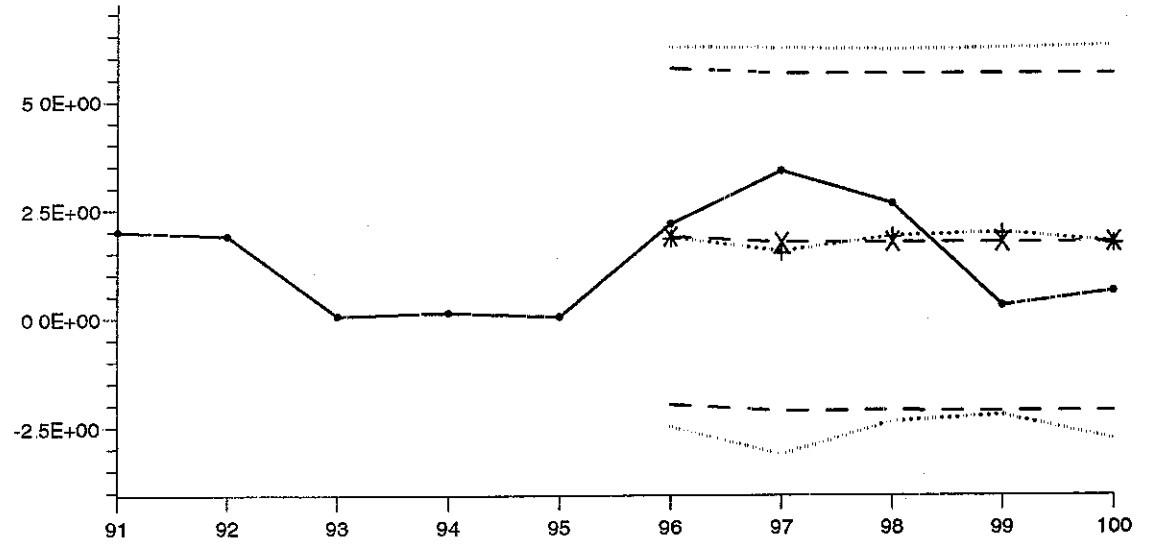
	EMO	*EMP*
Box et Jenkins	2.72%	8.67%
Non paramétrique	2.61%	5.11%

(c)

FIG. 9: Comparaison des prévisions paramétriques et non paramétriques sur des données d'un processus chimique de concentration (cf. Box et Jenkins, 1970)



(a)

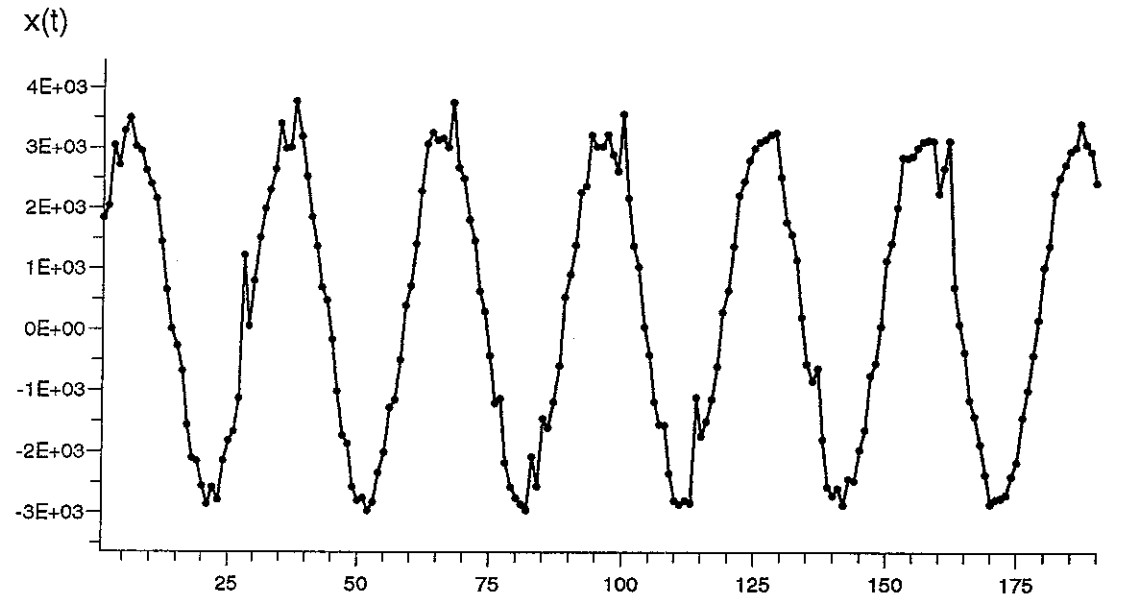


(b)

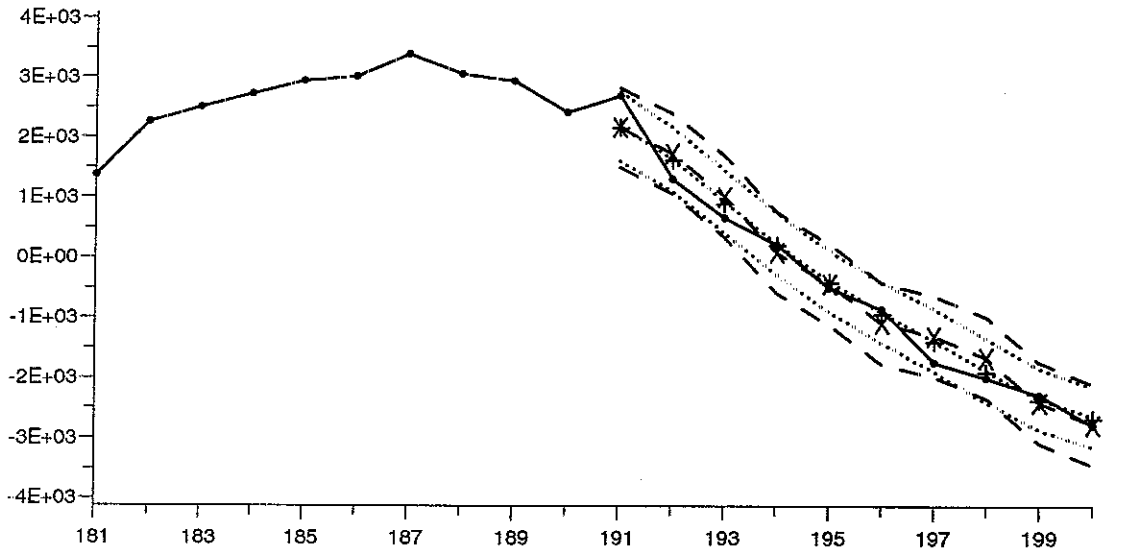
	EMO	*EMP*
Box et Jenkins	1.55	2.18
Non paramétrique	1.71	2.46

(c)

FIG. 10: Comparaison des prévisions paramétriques et non paramétriques sur une simulation de la série $X_t = \epsilon_t + 0.8X_{t-1}$, où (ϵ_t) est une suite i.i.d. de loi exponentielle de paramètre 2.5.



(a)

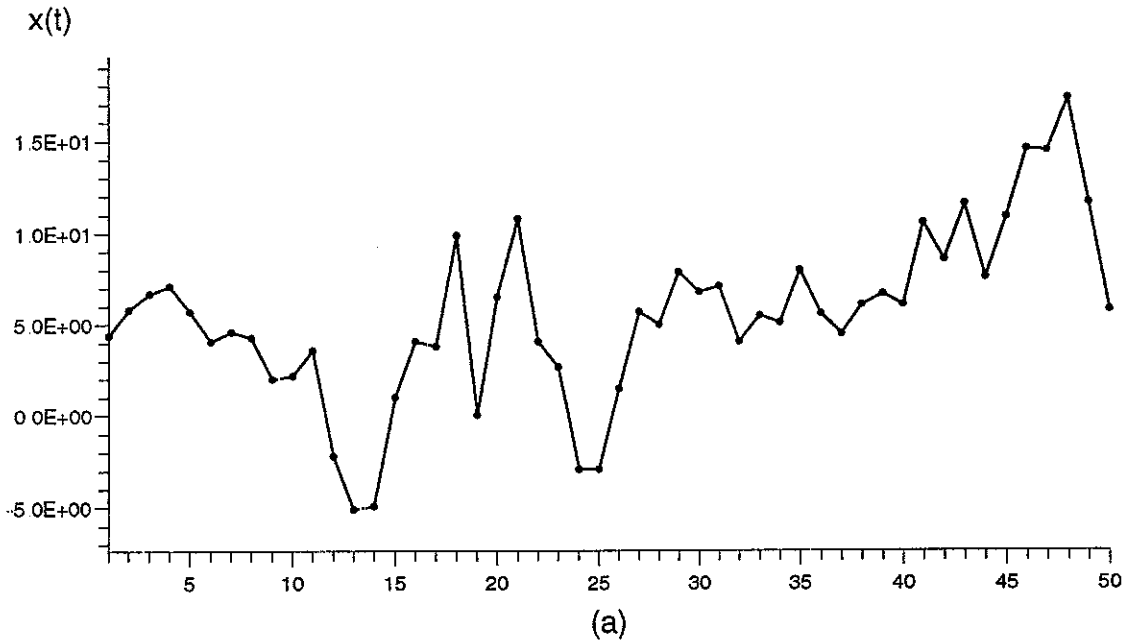


(b)

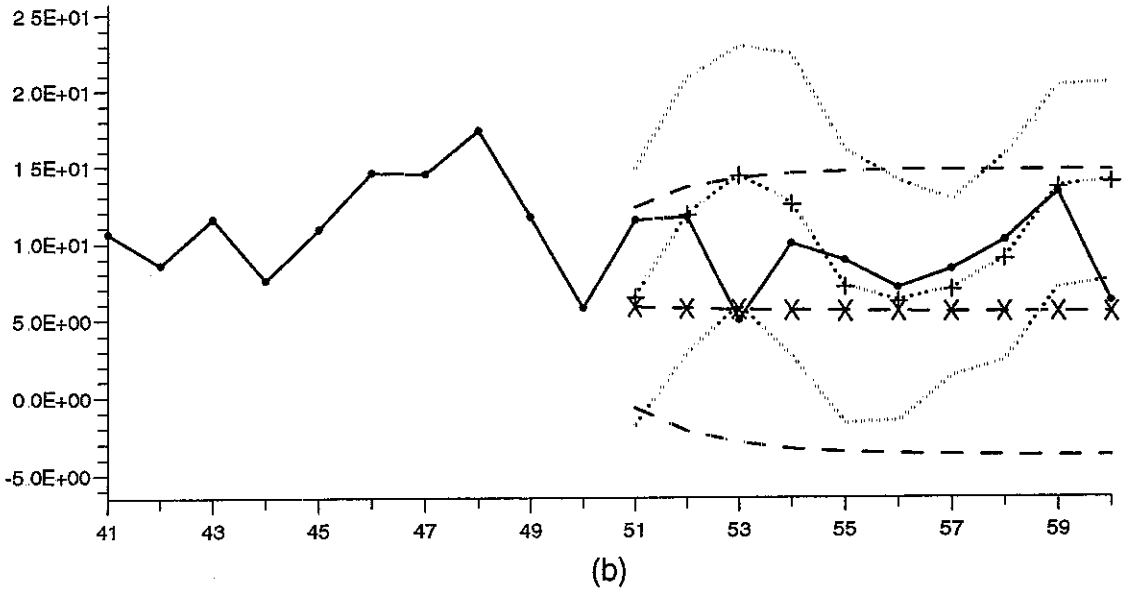
	EMO	*EMP*
Box et Jenkins	0.25	1.28
Non paramétrique	0.15	0.63

(c)

FIG. 11: Comparaison des prévisions paramétriques et non paramétriques sur une simulation de la série $X_t = \epsilon_t + 3000 \sin(\frac{\pi}{15} t)$, où (ϵ_t) est une suite i.i.d. de loi exponentielle de paramètre $\frac{1}{300}$.



(a)

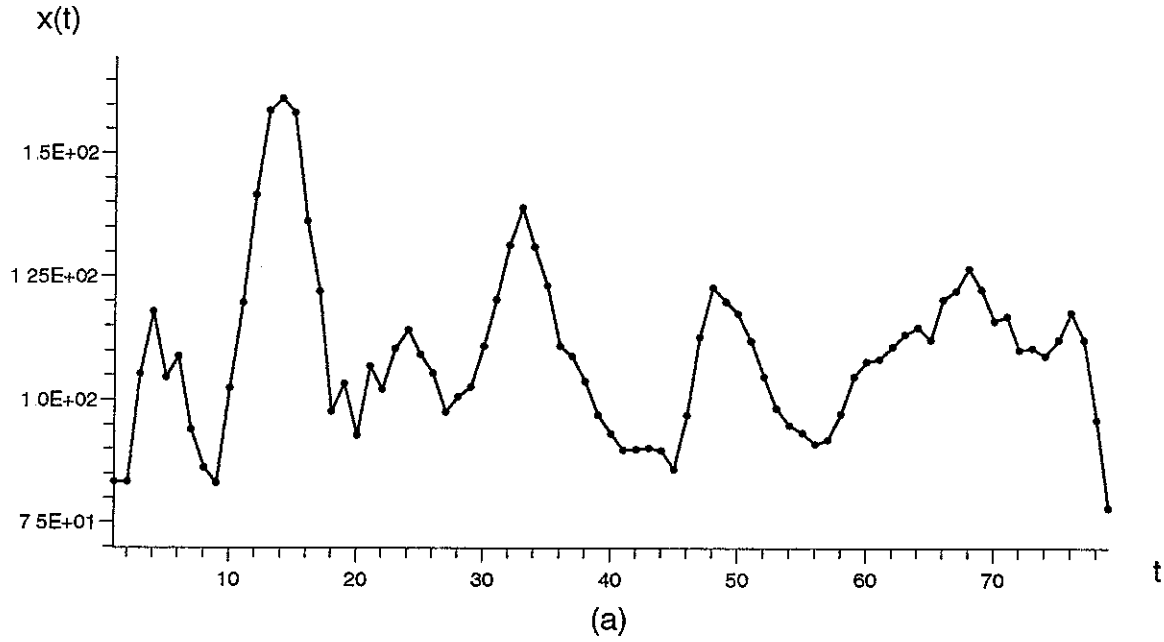


(b)

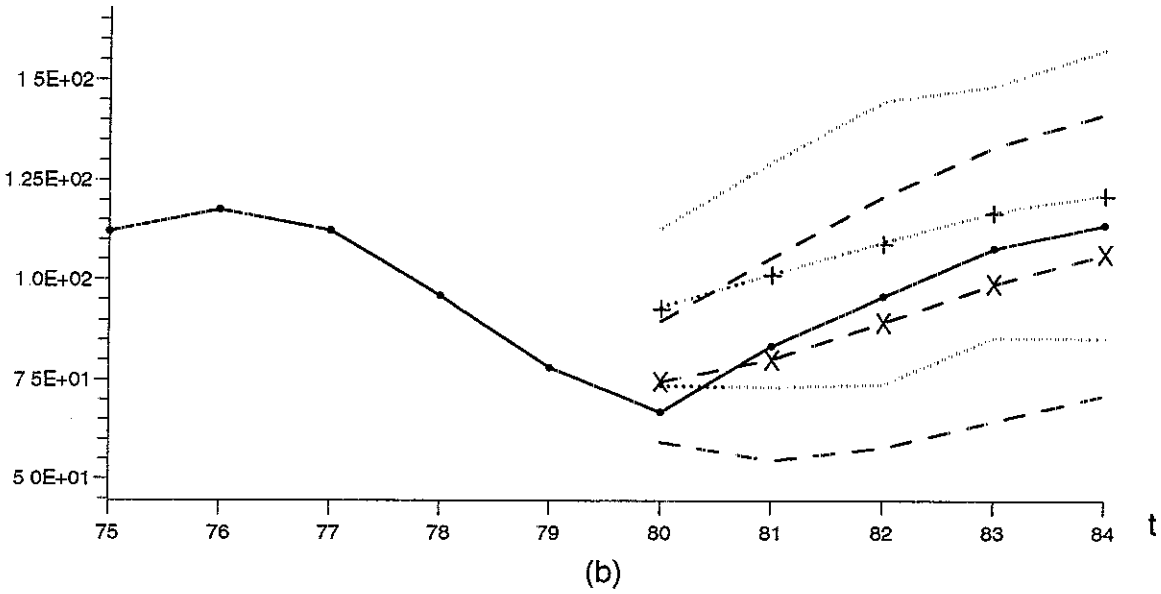
	EMO	*EMP*
Box et Jenkins	0.37	1.57
Non paramétrique	0.45	0.84

(c)

FIG. 12: Comparaison des prévisions paramétriques et non paramétriques sur la série « change in business inventories » (cf. Prankratz, 1983)



(a)

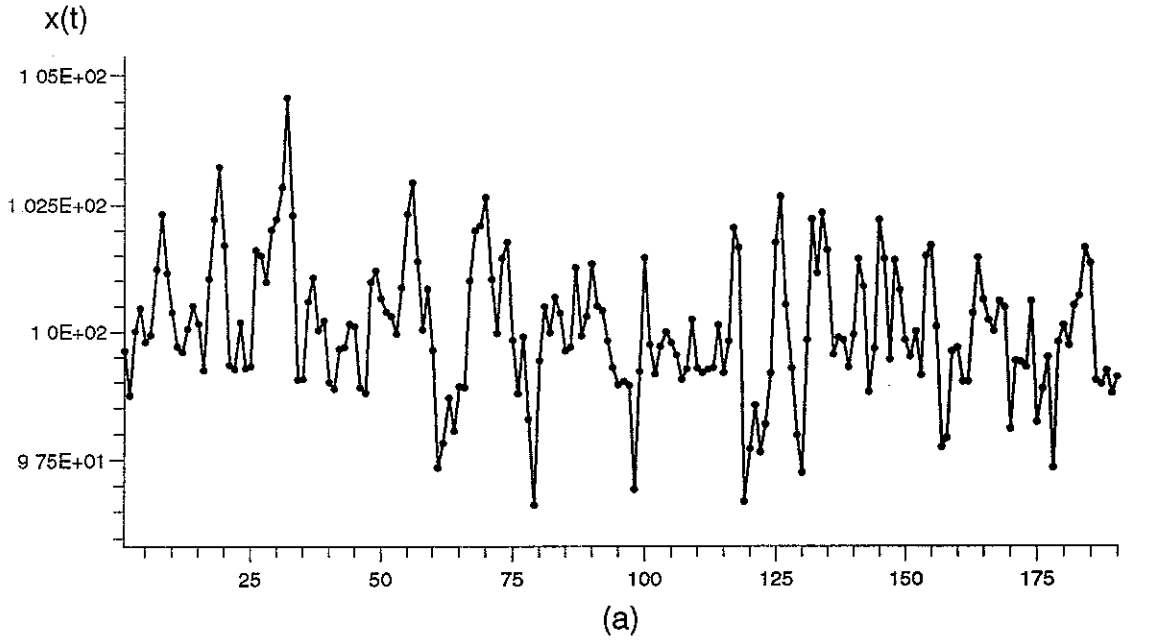


(b)

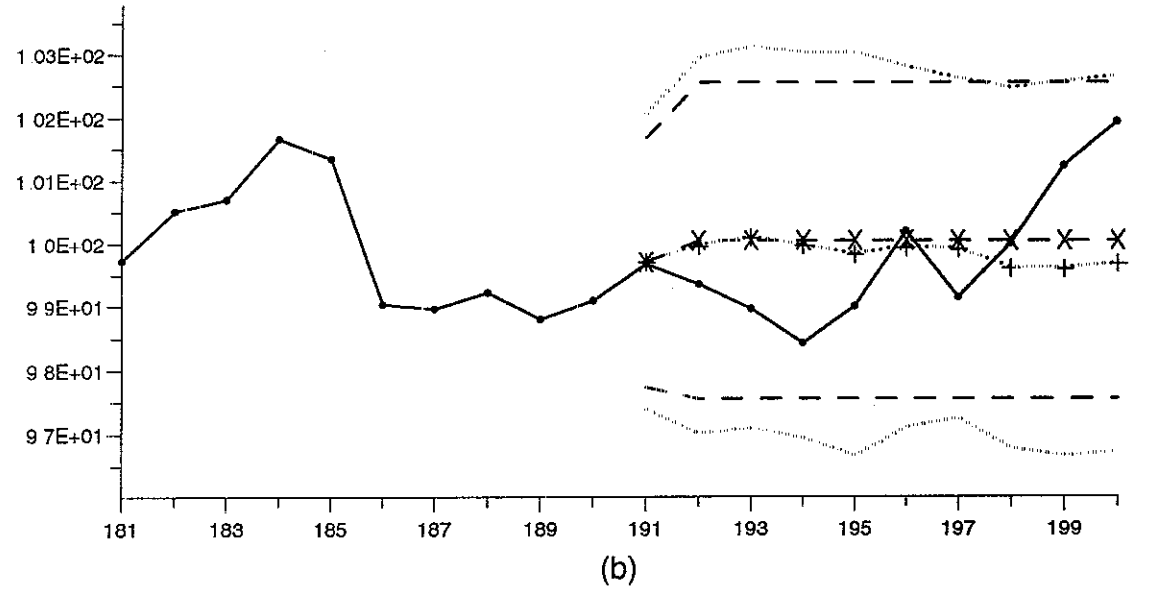
	EMO	*EMP*
Box et Jenkins	7.36%	31.00%
Non paramétrique	18.10%	27.58%

(c)

FIG. 13: Comparaison des prévisions paramétriques et non paramétriques sur les permis de construire (cf. Prankratz, 1983)



(a)

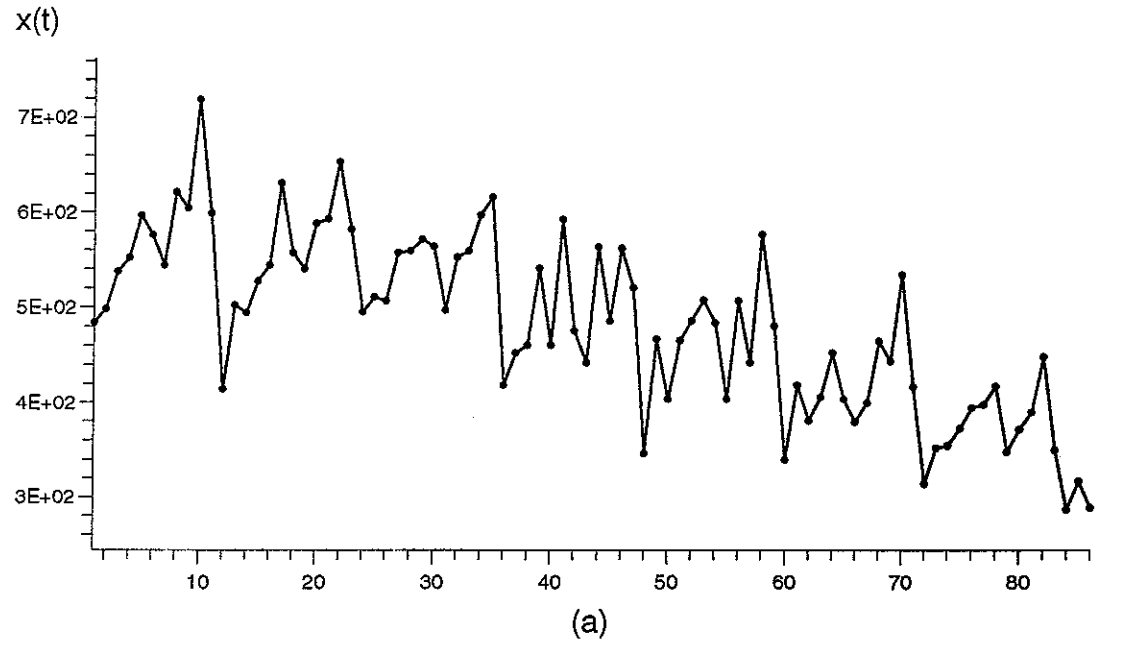


(b)

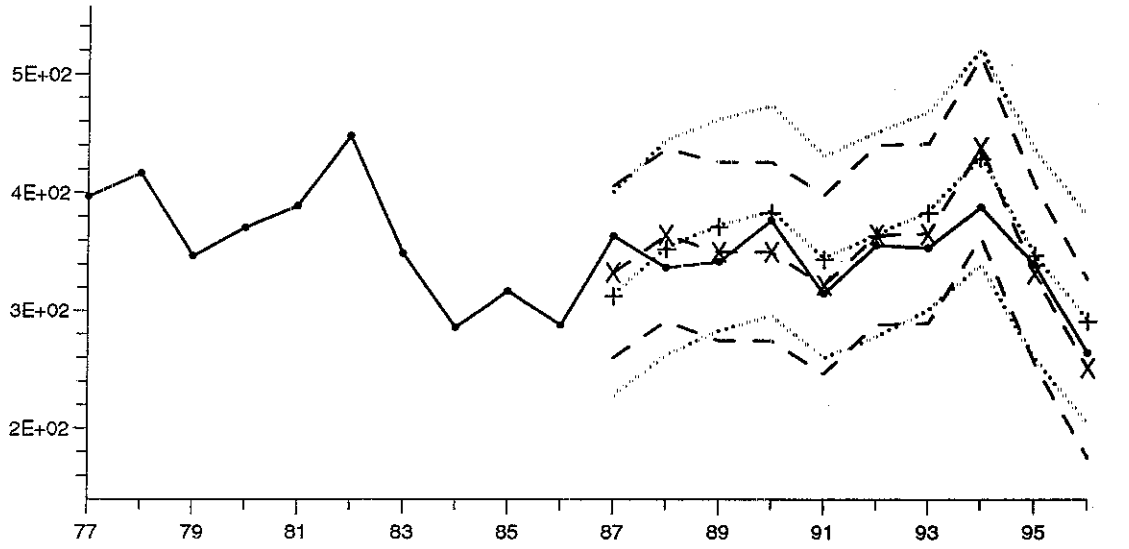
	EMO	*EMP*
Box et Jenkins	0.87%	2.45%
Non paramétrique	0.95%	2.89%

(c)

FIG. 14: Comparaison des prévisions paramétriques et non paramétriques sur la moyenne mobile $X_t = 100 + \epsilon_t + 0.8\epsilon_{t-1}$, où (ϵ_t) est un bruit blanc gaussien $\mathcal{N}(0,1)$



(a)



(b)

	EMO	*EMP*
Box et Jenkins	0.54%	22.12%
Non paramétrique	0.75%	24.72%

(c)

FIG. 15: Comparaison des prévisions paramétriques et non paramétriques sur la consommation de cigares (cf. Prankratz, 1983)

Références

Bosq, D. (1973) Sur l'estimation de la densité d'un processus stationnaire et mélangeant. C.R. Acad. Sci. Paris, 277, pp. 535-538.

Bosq, D. (1975) Inégalité de Bernstein pour les processus stationnaires et mélangeants. Applications. C.R. Acad. Sci. Paris, 281, pp. 1095-1098

Bosq, D. et Delecroix, M. (1985) Nonparametric prediction of a Hilbert space valued random variable. Stoch. Proc. and their Appl., 19, pp. 271-280.

Bosq, D. et Lecoutre, J.P. (1987) Théorie de l'estimation fonctionnelle. Economica. Paris.

Bosq, D. et Lecoutre, J.P. (1992) Analyse et prévision des séries chronologiques. Masson.

Box, G. et Jenkins, G. (1970) Time series analysis, forecasting and control. Holden Day.

Carbon, M. (1982) Sur l'estimation asymptotique d'une classe de paramètres fonctionnels pour un processus stationnaire. Thèse de l'Université de Lille.

Carbon, M (1983) Inégalité de Bernstein pour les processus fortement mélangeants non nécessairement stationnaires. C. R. Acad. Sci. Paris, 297, pp. 303-306.

Carbon, M. (1988) Inégalités de grandes déviations dans les processus. Applications à l'estimation fonctionnelle. Thèse de doctorat de l'Univ. de Paris 6.

Carbon, M. et Delecroix, M. (1993) Nonparametric vs. Parametric Forecasting in Time Series : A Computational Point of View. Appl. Stoch. Models and Data Anal., vol 9, pp. 215- 229.

Collomb, G. (1981) Estimation non paramétrique de la régression : revue bibliographique. Intern. Stat. Review, 49, pp. 75-93.

Collomb, G. (1983) From nonparametric regression to nonparametric prediction : Survey on the mean square error and original results on the predictogram. Lectures Notes in Statistics, 16, pp. 182-204.

Collomb, G. (1984) Propriétés de convergence presque complète du prédicteur à noyau. Z. Wahrsch. Verw. Gebiete 66,441-460.

Collomb, G. (1985a) Nonparametric regression : an up to date bibliography. Statistics, 2, pp. 309-324.

Collomb, G. (1985b) Nonparametric time series analysis and prediction : uniform almost sure convergence. Statistics, 2, pp. 297-307.

Deheuvels, P. (1977) Estimation non paramétrique de la densité par des histogrammes généralisés. Rev ; Stat. Appl., 35, pp. 5-42.

Devroye, L. et Györfi, L. (1985) Nonparametric density estimation : the L_1 view. Wiley.

Doukhan, P. (1991) Mixing. Properties and examples. Preprint. Univ. Paris, Orsay.

Gouriéroux, C. et Monfort, A. (1990) Séries temporelles et modèles dynamiques Economica.

Hall, P. (1983) Large sample optimality of least squared cross-validation in density estimation. Ann of Stat., 11, pp. 1156-1174.

Hall, P. (1984) Asymptotic properties of integrated squared errors and cross-validation for kernel estimation of a regression function. Z. Wahrsch. Verw. Gebiete, 67, pp. 175-196.

Härdle, W. (1989) Applied Nonparametric Regression. Economic Society Monograph Series, Cambridge University Press.

Härdle, W. et Marron, J.S. (1985) Optimal bandwith selection in nonparametric regression estimation. Annals of Statistics, 13, 4, pp. 1465-1481.

Marron, J.S. (1987) A comparison of cross-validation techniques in density estimation. Ann. of Statist., 15, pp 152-162.

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

Prakasa Rao, B.L.S. (1983) Nonparametric functional estimation. Academic Press.

Stone, C.J. (1982) Optimal global rates of convergence of non parametric regression. Ann. of Statist., 10, pp. 1040-1053.