Comparaison d'estimateurs de régression non paramétriques : application en valvométrie

Gilles DURRIEU[2], Thi Mong Ngoc NGUYEN[1] & Mohamedou SOW[2]

[1] *Université Bordeaux 1, Institut de Mathématiques de Bordeaux, UMR 5251, 351 cours de la libération, 33405 Talence.*

[2] *Université Bordeaux 1 et UMR CNRS 5805 Place du Dr Peyneau - 33120 Arcachon.*

Résumé La mesure de l'activité de mollusques bivalves est un moyen d'enregistrer le comportement de bivalves *in situ* et donc d'évaluer des changements de la qualité de l'eau. Nous proposons un modèle de régression non paramétrique et comparons trois estimateurs non paramétriques, récursifs ou non, de la fonction de lien sur les données acquises en Nouvelle Calédonie.
Mots clés estimateur à noyau, estimateur récursif, validation croisée.

Abstract Measurement of mollusks bivalves activity is a way to record the animal behaviour and so to evaluate possible changes in the water quality. The huge volume of data collected necessitates the development of statistical models. We propose a nonparametric regression model and we compare three non parametric estimators (recursive or not) of the link function, on the data collected in New Caledonia.
Key words kernel estimator, recursive estimator, cross validation.

1 Introduction

Les activités humaines sont responsables d'importants rejets d'agents polluants dans le milieu naturel. Ces polluants entraînent la dégradation de nombreux biotopes, perturbant les écosystèmes et posant également des problèmes en termes de santé publique. Des réglementations et des contrôles sur la qualité des eaux ont été mis en place. Parmi ces contrôles, les bioindicateurs sont de plus en plus utilisés et sont très efficaces par leurs capacités à révéler la présence de traces (concentrations très faibles) de contaminant. Nous utilisons ici comme moyen de surveillance du milieu la valvométrie. La valvométrie (mesure de l'activité des valves de mollusques) est une technique qui permet d'enregistrer les réactions de bivalves, face aux changements de la qualité de l'eau dans laquelle ils vivent (Tran et al., 2003). Les mollusques bivalves ventilent tout au long de la journée pour se nourrir et respirer. Ils sont équipés de récepteurs qui leur permettent en permanence d'estimer la qualité de l'eau dans laquelle ils baignent, de façon à pouvoir réagir immédiatement face à une eau qu'ils jugent nocive pour leur intégrité. Le cas des mollusques bivalves est particulièrement intéressant en tant qu'espèce bioindicatrice car ce sont des animaux sédentaires qui peuvent être témoins de changement locaux de la qualité de l'eau. Le suivi du comportement de bivalves permet donc de rendre compte jour après jour de leur état de santé et au-delà, de l'évolution de la qualité de l'eau. Actuellement, l'aquisition, le transfert et le traitement des données fonctionnent de manière automatique sous la jetée d'Eyrac sur le bassin

d'Arcachon et en Nouvelle Calédonie. L'important volume de données enregistrées à haute fréquence avec un nombre de variables pouvant être important, nécessitent le développement de modèles statistiques performants afin de bien décrire le comportement des animaux *in situ* dans le but d'extraire des rythmes biologiques qui permettraient par la caractérisation de perturbations de ces rythmes de détecter une pollution du milieu. Dans cette communication, nous proposons un modèle de régression non paramétrique et comparons trois estimateurs non paramétriques, récursifs ou non, de la fonction de lien sur les données acquises depuis septembre 2007 au niveau du récif IORO en Nouvelle Calédonie. L'objectif final est d'utiliser la valvométrie comme un système de biosurveillance de la qualité du milieu pour suivre l'impact potentiel d'une nouvelle mine de Nickel et de cobalt. Les enregistrements et les résulats du traitement statistique sont accessibles sur le site web "L'oeil du mollusque" (http://www.domino.u-bordeaux.fr/molluscan_eye).

2 Modèle et estimateurs

Nous disposons d'un échantillon composé de n couples indépendants de variables aléatoires $(T_1, Y_1), \ldots, (T_n, Y_n)$ et nous considérons le modèle de régression non paramétrique donné, pour $i = 1, \ldots, n$, par

$$Y_i = m(T_i) + \varepsilon_i. \tag{1}$$

Dans ce modèle intervient une fonction m inconnue à estimer qui exprime la valeur moyenne de l'écartement valvaire de nos bivalves en fonction du temps T et un terme aléatoire d'erreur ε de loi inconnue et indépendant de T. Nous proposons trois estimateurs non paramétriques de la fonction m. Le premier estimateur est l'estimateur de Nadaraya-Watson (Nadaraya, 1964 et Watson, 1964), noté **NW**. Il est construit à partir d'une fonction noyau K et d'une fenêtre h_n, de manière similaire à l'estimateur à noyau de la fonction de densité de probabilité (Silverman, 1986). Cet estimateur de la densité f de T s'écrit :

$$\hat{f}_n(t) = \frac{1}{n\,h_n} \sum_{i=1}^{n} K(\frac{t - T_i}{h_n}), \tag{2}$$

ou dans sa forme récursive :

$$\tilde{f}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K(\frac{t - T_i}{h_i}). \tag{3}$$

La fenêtre h_n désigne une suite de nombres réels strictement positifs vérifiant (**C1**) $h_n \to 0$ et $n\,h_n \to \infty$ lorsque $n \to \infty$. Le noyau est une fonction mesurable, positive et bornée satisfaisant (**C2**) $\int_R K(x)\,dx = 1$, $\int x\,K(x)\,dx = 0$, $\int_R |x|\,K(x)\,dx < +\infty$ et $\int_R K^2(x)\,dx = \tau^2$.

L'estimateur **NW** s'écrit sous la forme d'une moyenne pondérée des valeurs $(Y_1, \ldots, Y_n)$.

Il est donné par :

$$\hat{m}_n(t) = \begin{cases} \dfrac{\sum_{i=1}^{n} K(\dfrac{t-T_i}{h_n}) Y_i}{\sum_{i=1}^{n} K(\dfrac{t-T_i}{h_n})} & \text{si } \sum_{i=1}^{n} K(\dfrac{t-T_i}{h_n}) \neq 0, \\ \dfrac{1}{n}\sum_{i=1}^{n} Y_i & \text{sinon.} \end{cases} \tag{4}$$

On propose également d'utiliser l'estimateur de Nadaraya-Watson récursif (Duflo, 1997), noté **NWR**, défini par :

$$\tilde{m}_n(t) = \begin{cases} \dfrac{\sum_{i=1}^{n} \dfrac{1}{h_i} K(\dfrac{t-T_i}{h_i}) Y_i}{\sum_{i=1}^{n} \dfrac{1}{h_i} K(\dfrac{t-T_i}{h_i})} & \text{si } \sum_{i=1}^{n} \dfrac{1}{h_i} K(\dfrac{t-T_i}{h_i}) \neq 0, \\ \dfrac{1}{n}\sum_{i=1}^{n} Y_i & \text{sinon.} \end{cases} \tag{5}$$

Enfin, nous considérons l'estimateur récursif de Révész (Révész, 1977 et Mokkadem et al., 2008), noté **R**, défini par :

$$\check{m}_n(t) = \check{m}_{n-1}(t) + \frac{1}{nh_n} K(\frac{t-T_n}{h_n})(Y_n - \check{m}_{n-1}(t)). \tag{6}$$

Ces trois estimateurs de m sont donc dépendants du choix de la fenêtre et du noyau. Le noyau K détermine "la forme du voisinage" autour du point t et la fenêtre h_n contrôle "la taille de ce voisinage", c'est-à-dire grossièrement le poids des observations pris pour effectuer le calcul de l'estimateur en t. Le choix du paramètre h_n est par conséquent un point crucial pour la qualité de l'estimation. Cependant, le choix du noyau permet aussi de réduire le biais des estimateurs en se basant sur les propriétés de régularité de la fonction de lien.

3 Propriétés asymptotiques

Nous rappelons ici les principales propriétés asymptotiques des estimateurs **NW**, **NWR** et **R**. Nous introduisons tout d'abord les notations $h_n = n^{-\alpha}$ et $\sigma^2(t) = \text{var}\,(Y \mid T = t)$. Nous ajoutons deux conditions de régularité : (**C3**) la fonction de lien m et la densité f sont bornées et deux fois continûment dérivables sur R et (**C4**) $E(Y^2) < \infty$.

Théorème 1 (NW) *Sous les conditions* **C1** –**C4** *et pour tout* $\alpha \in [1/5, 1[$, *à chaque point de continuité de* $\sigma^2(t)$ *et pout tout* $t \in R$ *tel que* $f(t) > 0$, *nous avons quand* $n \to \infty$:

1.
$$\hat{m}_n(t) \xrightarrow{ps} m(t).$$

2.
$$\sqrt{nh_n}\,(\hat{m}_n(t) - m(t)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2(t)\,\tau^2}{f(t)}\right).$$

Théorème 2 (NWR) *Sous les conditions* **C1** –**C4** *et pour tout* $\alpha \in]1/3, 1[$*, à chaque point de continuité de* $\sigma^2(t)$ *et pout tout* $t \in R$ *tel que* $f(t) > 0$*, nous avons quand* $n \to \infty$ *:*

1.
$$\tilde{m}_n(t) \xrightarrow{ps} m(t).$$

2.
$$\sqrt{nh_n}\,(\tilde{m}_n(t) - m(t)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2(t)\,\tau^2}{f(t)(1+\alpha)}\right).$$

Théorème 3 (R) *Sous les conditions* **C1** –**C4** *et pour tout* $\alpha \in]1/2, 1[$*, à chaque point de continuité de* $\sigma^2(t)$ *et pout tout* $t \in R$ *tel que* $2\,f(t) > 1 - \alpha$*, nous avons quand* $n \to \infty$ *:*

1.
$$\check{m}_n(t) \xrightarrow{ps} m(t).$$

2.
$$\sqrt{nh_n}\,(\check{m}_n(t) - m(t)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2(t)\,\tau^2\,f(t)}{2f(t) - (1-\alpha)}\right).$$

Il est possible d'estimer la variance de la loi normale limite. Pour cela, la densité marginale f est estimée par (2) pour **NW** et par (3) pour **NWR** et **R**. La variance conditionnelle $\sigma^2(t)$ est estimée respectivement pour **NW**, **NWR** et **R** par :

$$\hat{\sigma}^2(t) = \frac{1}{\hat{f}_n(t)} \sum_{i=1}^{n} K(\frac{t - T_i}{h_n})(Y_i - \hat{m}_n(t))^2 \quad \text{et} \quad \tilde{\sigma}^2(t) = \frac{1}{\tilde{f}_n(t)} \sum_{i=1}^{n} K(\frac{t - T_i}{h_i})(Y_i - \hat{B}_n(t))^2,$$

avec $\hat{B}_n(t) = \tilde{m}_n(t)$ et $\hat{B}_n(t) = \check{m}_n(t)$.

4 Choix de la fenêtre

Le choix de ce paramètre est crucial pour nos trois estimateurs. En pratique, ce paramètre est choisi comme un compromis entre la variance et le biais de l'estimation. Une importante littérature est consacrée à ce sujet, et en particulier aux méthodes de sélection automatique par minimisation d'un critère. Nous utilisons comme critère la méthode de la validation croisée (Härdle et Marron, 1985 et Härdle, 1990) qui consiste à minimiser par rapport à h la fonction

$$CV(h) = \sum_{i=1}^{n} \left(Y_i - \hat{m}_{(-i)}(T_i; h)\right)^2$$

où $\hat{m}_{(-i)}(T_i; h)$ désigne un estimateur (**NW**, **NWR** ou **R**) de la fonction de lien au point T_i calculé sur l'échantillon privé du couple (T_i, Y_i).

5 Application en valvométrie

Nous évaluons ici les trois estimateurs non paramétriques sur les données recueillies depuis septembre 2007 au niveau du site IORO en Nouvelle Calédonie. Les mesures sont collectées par deux cartes électroniques qui gèrent à la fois l'acquisition (toutes les $1,6$ secondes) sur un groupe de 16 bénitiers et le transfert des données. Ainsi ce dispositif génère tous les jours, pour chaque bénitier, 54000 couples de valeurs (T_i, Y_i) qui sont le temps en heure et l'écartement valvaire en mm.

Nous avons choisi le noyau Gaussien pour les estimateurs **NW** et **NWR**. Pour l'estimateur **R**, l'article de Révész (1977) recommande le choix du noyau uniforme. Les largeurs des fenêtres sont déterminées en utilisant la méthode de validation croisée. Les fonction CV obtenues numériquement sont toutes convexes pour nos trois estimateurs et les fenêtres optimales ainsi obtenues sont respectivement égales à $n^{-0,47}$ pour **NW**, $n^{-0,50}$ pour **NWR** et $n^{-0,99}$ pour **R**. Avec ces choix de noyaux et de fenêtres, nous obtenons un très bon ajustement des modèles de régression aux données (Figure 1).

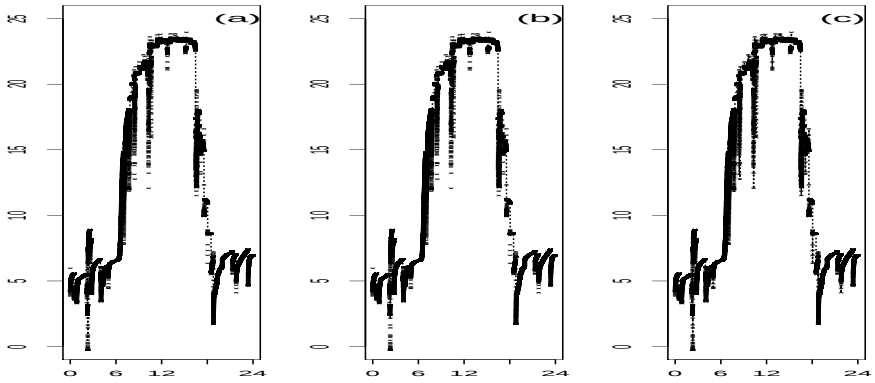

Figure 1: Représentation des estimateurs (traits en pointillés) **NW** (a), **NWR** (b) et **R** (c) sur les données brutes du bénitier 7 avec en abscisse le temps (h) et en ordonnée l'écartement valvaire (mm).

Les propriétés asymptotiques des estimateurs sont illustrées numériquement sur un bénitier en se fixant arbitrairement deux temps entre 0 et 24 h. La Figure 2 illustre bien les résultats théoriques concernant la normalité asymptotique. En effet, ces représentations nous montrent un très bon ajustement de **NW**, **NWR** et **R** avec une loi normale centrée pour le temps $t_1 = 5$ h (Figure 2 a1-b1-c1) et le temps $t_2 = 23$ h (Figure 2 a2-b2-c2) en considérant $N = 475$ jours.

En comparant nos trois estimateurs au sens du critère de la variance, nous montrons que significativement $\mathrm{var}(\mathbf{R}) < \mathrm{var}(\mathbf{NWR}) < \mathrm{var}(\mathbf{NW})$ ($p < 0.05$). Ainsi, ce résultat nous suggère d'utiliser plutôt un estimateur récursif de type **NWR** ou **R** au sens de la variance asymptotique minimum, mais le temps calcul de **R** est plus important.

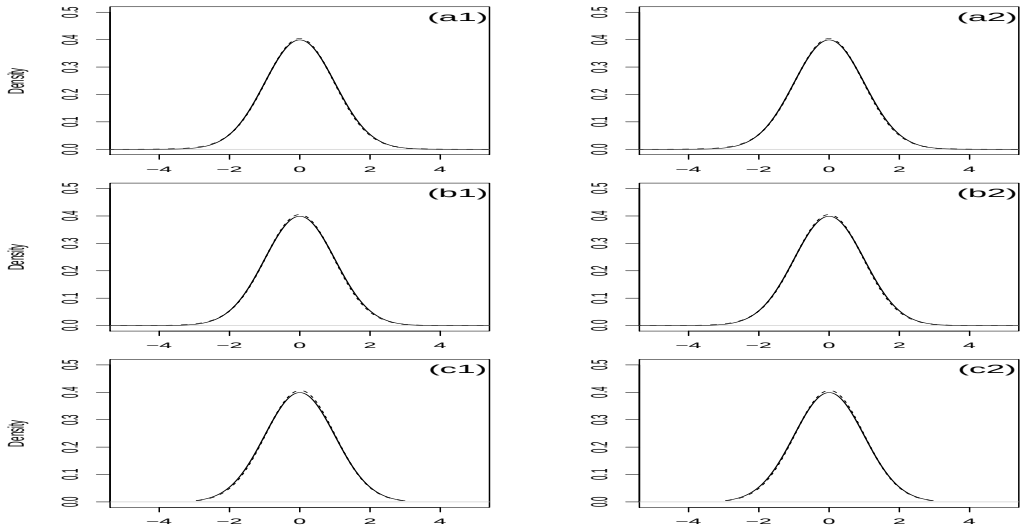

Figure 2: Illustration de la convergence en distribution sur $N = 475$ jours: (a1, b1, c1) et (a2, b2, c2) correspondent respectivement aux estimateurs de la densité de **NW**, **NWR** et **R** ($h = 0.96$) pour le temps 1 (5 h) et le temps 2 (23 h).

Bibliographie

[1] Duflo, Marie. (1997) Random Iterative Models. *Collection mathématiques et applications, Springer*, 385 pages.

[2] Härdle, Wolgang. (1990) Applied nonparametrice regression. *Econometric Society Monographs*, 333 pages.

[3] Härdle, W. and Marron, J.S (1985) Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, **13**, 4, 1465-1481.

[4] Mokkadem, A., Pelletier, M. and Slaoui, Y (2008) Revisiting Révész stochastic approximation method for the estimation of a regression function. *math.ST* arXiv : 0812.3973v1.

[5] Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and its Applications* **10**, 186-190.

[6] Révész, P. (1977). How to apply the method of stochastic approximation in the non-parametric estimation of a regression function. *Math.Operationsforsch. Statist., Ser. Statistics*, **8**, 119-126.

[7] Silverman, B. W (1986). Density estimation for statistics and data analysis, Chapman & Hall, 175 pages.

[8] Tran, D., Ciret, P., Ciutat, A., Durrieu, G. et Massabuau, J.C. (2003). Estimation of potential and limits of bivalve closure response to detect contaminants: application to cadmium. *Environmental Toxicology and Chemistry*, **22(4)**, 914-920.

[9] Watson, G.S. (1964) Smooth regression analysis. *Sankhya.* **26**, 359-372.