Quantiles extrêmes conditionnels et application à la surveillance en temps réel d'un système aquatique

Gilles Durrieu & Ion Grama & Quang-Khoai Pham & Jean-Marie Tricot

Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud et UMR CNRS 6205
Campus de Tohannic, 56017 Vannes.
{gilles.durrieu, ion.grama, quang-khoai.pham,jean-marie.tricot}@univ-ubs.fr

Résumé. Nous présentons une méthode statistique pour estimer les quantiles extrêmes de la distribution de la vitesse de fermeture des valves en valvométrie. L'idée de notre approche est d'ajuster la queue de la fonction de distribution de cette vitesse par une distribution de Pareto de paramètre $\theta_{t,\tau}$ au delà d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimé en utilisant un estimateur à noyau non paramétrique de taille de fenêtre h à partir des observations plus grandes que τ. Nous donnons sous des hypothèses de régularités les vitesses de convergence des estimateurs des quantiles extrêmes et de $\theta_{t,\tau}$. Notre modélisation fournit en temps réel une analyse du comportement d'un bioindicateur du milieu marin et apparaît comme un moyen efficace pour la surveillance de la qualité des eaux d'un système aquatique.

Mots-clés. Estimateur non paramétrique, valeurs extrêmes, valvometrie, environnement.

Abstract. We present a statistical method to estimate the extreme quantiles of the distribution of valve closing velocity in valvometry. The idea of our approach is to adjust the tail of the distribution function of valve closing velocity with a Pareto distribution with parameter $\theta_{t,\tau}$, starting from a threshold τ. The parameter $\theta_{t,\tau}$, is estimated using a nonparametric kernel estimator of bandwidth h based on the observations larger than τ. Under some regularity assumptions, the rates of convergence of the proposed estimators are given. Our modeling approach provides real time in situ analysis of a bioindicator behavior of marine ecosystem and appears as an effective early warning tool in ecological risk assessment and marine environment monitoring.

Keywords. Nonparametric estimator, extreme values, valvometry, environment.

1 Introduction

Les activités humaines génèrent et continuent de générer de multiples pollutions. Il est établi que les phénomènes météorologiques extrêmes enregistrés sur toute la planète sont étroitement liées aux pollutions et au réchauffement climatique induit par les activités humaines. Dans le domaine de l'environnement, des réglementations et des contrôles sur la qualité des eaux ont été mis en place. Parmi ces techniques de contrôles, les bioindicateurs sont de plus en plus utilisés et sont très efficaces par leurs capacités à révéler la présence de concentrations très faibles de contaminants via par exemple leur accumulation dans des tissus animaux ou végétaux spécifiques, ou via des modifications sur des structures populationnelles, etc.

Dans ce travail, nous nous intéressons particulièrement au développement d'une méthode statistique d'analyse du comportement extrême de mollusques bivalves, qui est un moyen de surveillance de la qualité de l'eau. Le cas des mollusques bivalves est particulièrement intéressant en tant qu'espèce bioindicatrice car ce sont des animaux sédentaires qui peuvent être témoins de pollutions locales éventuelles et on les trouve partout dans le monde, des tropiques aux pôles. 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. Nous utilisons ici comme technique de surveillance du milieu aquatique la valvométrie (mesure de l'activité des valves de mollusques). Aujourd'hui la valvométrie HFNI (Haute Fréquence Non Invasive) permet d'enregistrer, pratiquement en ligne, les mouvements des valves à haute fréquence sans interférer avec le comportement normal des animaux, n'importe où dans le monde pendant des périodes de 1 à 2 ans sans intervention humaine *in-situ*. La valvométrie est ainsi devenue une technique qui permet d'enregistrer les réactions de bivalves sur le terrain, proche ou distant, face aux changements de la qualité de l'eau dans laquelle ils vivent. L'idée de base est que les mollusques bivalves ventilent tout au long de la journée pour se nourrir et respirer. De nombreuses anomalies telles que le contact d'un prédateur ou l'arrivée d'un contaminant dans la colonne d'eau, détectées comme dangereuses pour l'animal, se traduisent par des comportements de protection allant de la fermeture brève à la fermeture prolongée en passant par des modifications de la fréquence et des vitesses des séquences d'ouverture et de fermeture des valves. Lorsque ces comportements affectent non pas uniquement un seul individu mais le groupe, ces modifications sont le reflet de perturbations du milieu proche des animaux et donc le comportement valvaire peut être alors un excellent outil d'aide à la gestion et à la surveillance du milieu (Azaïs et al. (2014), Coudret et al. (2015), Schmitt et al (2011), Sow et al. (2011), Tran et al. 2004).

L'objectif de ce travail est de développer des modèles statistiques originaux basés sur la théorie des valeurs extrêmes sur des volumes de données importants afin de fournir une meilleure connaissance du fonctionnement et du comportement de l'huître dans son environnement naturel. Nous cherchons à caractériser/à discriminer automatiquement des perturbations extrêmes de l'activité des animaux par l'étude d'un bivalve filtreur (l'huître dans le golfe du Morbihan), de la nature d'un contaminant ou d'un facteur du milieu.

1.1 Modélisation

Une nouvelle méthode statistique est proposée pour étudier les vitesses de fermeture extrêmes des valves d'huîtres $(X_t)_{t\in[0,T_{\max}]}$ en fonction du temps dans le cadre de la théorie des valeurs extrêmes. Cette théorie a souvent été utilisée en environnement et biologie. Nous nous intéressons à une version fonctionnelle de la méthode publiée par Grama et Spokoiny en 2008 et développée dans Grama et al. (2014) et Durrieu et al. (2014).

L'objectif principal est de fournir une estimation ponctuelle de la fonction de survie $1 - F_t(x)$ où $F_t(x) = P(X_t \leq x)$ est la fonction de répartition de la vitesse de fermeture des valves X_t. Nous proposons une estimation des p-quantile extrêmes du processus $F_t^{-1}(p)$ sur $[0, T_{\max}]$, où ce quantile est d'ordre p élevé. La variable aléatoire X_t est supposée de densité f_t strictement positive. Nous supposons que les distributions F_t sont dans le domaine d'attraction de la distribution de Fréchet. Par le théorème de Fisher-Tippet-Gnedenko (Beirlant (2004)) nous

avons, pour tout $t \in [0, T_{\max}]$,

$$\lim_{\tau \to \infty} \frac{1 - F_t(\lambda\tau)}{1 - F_t(\tau)} = \lambda^{-1/\theta_t}, \qquad \text{pour tout } \lambda \geq 1. \tag{1}$$

Pour un seuil fixé $\tau \geq x_0$, nous désignons par

$$F_{t,\tau}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\tau)}, \quad x \in [\tau, \infty)$$

la fonction de distribution d'excès au delà de τ.

La condition (1) précise que si x est assez grand, $F_{t,\tau}(x)$ peut être approchée par la distribution de Pareto

$$G_{\tau,\theta_t}(x) = 1 - \left(\frac{x}{\tau}\right)^{-1/\theta_t}, \quad x \geq \tau \text{ quand } \tau \to \infty, \tag{2}$$

où le paramètre θ_t est l'indice de queue conditionnel.

D'après (1) et (2), quand le seuil τ tend vers l'infini, F_t peut être approchée par le modèle semi-paramétrique :

$$F_{t,\tau,\theta_t}(x) = \begin{cases} F_t(x) & \text{si } x \in [x_0, \tau], \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta_t}(x)) & \text{si } x > \tau. \end{cases} \tag{3}$$

Ainsi, d'après le modèle (3), F_t est estimée par la fonction de répartition empirique pondérée sur l'interval $[0, \tau]$, tandis que sur l'intervalle $[\tau, \infty)$ F_t is ajustée par la distribution de Pareto de paramètre θ_t.

Nous considérons le noyau Gaussien tronqué :

$$K(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) & \text{si } x \in [-1, 1] \\ 0 & \text{sinon,} \end{cases}$$

et les poids $W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right)$, où $h > 0$ est une taille de fenêtre inconnue. Pour tout $t \in [0, T_{\max}]$, la log-vraisemblance pondérée s'écrit

$$\begin{aligned} \mathcal{L}_{t,h}(\tau, \theta) &= \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}) \\ &= \sum_{i=1}^{n} 1_{\{X_{t_i} \leq \tau\}} W_{t,h}(t_i) \log f_t(X_{t_i}) \\ &\quad + \sum_{i=1}^{n} 1_{\{X_{t_i} > \tau\}} W_{t,h}(t_i) \log \left(\frac{(1 - F_t(\tau))}{\tau\theta} \left(\frac{X_{t_i}}{\tau}\right)^{-\frac{1}{\theta}-1} \right), \end{aligned}$$

où 1_A prend la valeur 1 quand la condition A est vérifiée et 0 sinon. En maximisant $\mathcal{L}_{t,h}(\tau, \theta)$ par rapport à θ, nous obtenons l'estimateur du paramètre θ_t :

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\{X_{t_i} > \tau\}} \log \left(\frac{X_{t_i}}{\tau}\right)$$

où

$$\widehat{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\{X_{t_i} > \tau\}}$$

est le nombre pondéré d'observations au delà du seuil τ. La fonction $F_t(x)$ au temps t est alors estimée par

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x) & \text{si } \ x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}(\tau)\right) \left(\frac{x}{\tau}\right)^{-1/\widehat{\theta}_{t,h,\tau}} & \text{si } \ x > \tau, \end{cases}$$

qui combine le fonction de répartition empirique pondérée

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\{X_{t_i} \leq x\}}$$

et la distribution de Pareto ajustée. Pour tout $p \in (0,1)$, l'estimateur du p-quantile de X_t est défini par

$$\widehat{q}_p(t) = \widehat{q}_p(t,h,\tau) = \begin{cases} \widehat{F}_{t,h}^{-1}(p) & \text{si} \quad p < \widehat{p}_\tau, \\ \tau \left(\frac{1-\widehat{p}_\tau}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{sinon,} \end{cases}$$

où $\widehat{p}_\tau = \widehat{F}_{t,h}(\tau)$.

Les estimateurs $\widehat{\theta}_{t,h,\tau}$, $\widehat{F}_{t,h,\tau}(x)$ et $\widehat{q}_p(t,h,\tau)$ dépendent des paramètres inconnus τ et h. Le bon choix de ces paramètres est crucial pour ces estimateurs. Durrieu et al. (2014) et Pham (2015), ont proposé de choisir τ par une méthode adaptative basée sur une procédure de tests multiples. La sélection de la taille de la fenêtre h est également obtenue par deux méthodes : une procédure de type validation croisée et une procédure adaptative. Le choix simultané de τ et de h est proposé par Pham 2015.

Sous des conditions de régularité et pour des suites convenables $\{\tau_n\}$, $\{\hat{\tau}_n\}$ et $\{h_n\}$, les estimateurs adaptatifs sont consistants et leurs vitesses de convergence sont données par :

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\widehat{\tau}_n}, \theta_{t,\tau_n}\right) = O_P\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}}\right) \qquad \text{quand } n \to \infty,$$

et

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_n}}\right) = O_P\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}}\right) \qquad \text{quand } n \to \infty,$$

où

$$\overline{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i)(1 - F_{t_i}(\tau))$$

et $\mathcal{K}(P,Q) = \int \log \frac{dP}{dQ} dP$ désigne l'entropie de Kullback-Leibler entre deux mesures équivalentes P et Q, et $\mathcal{K}(\theta,\theta') = \frac{\theta}{\theta'} - 1 - \log\left(\frac{\theta}{\theta'}\right)$, $\theta, \theta' > 0$.

2 Application en environnement

Nous appliquons la méthode statistique sur les données recueillies depuis le 3 mars 2011 au niveau du site de Locmariaquer en Bretagne Sud. Les mesures sont collectées par deux cartes électroniques qui gèrent à la fois l'acquisition (toutes les 0.1 seconde) sur un groupe de 16 huîtres et le transfert automatique des données sur une station de travail. En complément de ces mesures de valvométrie, des dosages de l'acétylcholinestérase (AChE, biomarqueur de neurotoxicité) et de la Glutathion S-Transférase GST(biomarqueur du stress oxydatif) ont été aussi réalisés sur des huîtres installées à proximité. La Figure 1 montre un bon comportement de l'estimateur des quantiles extrêmes des vitesses de fermeture sur les données observées sur une journée.

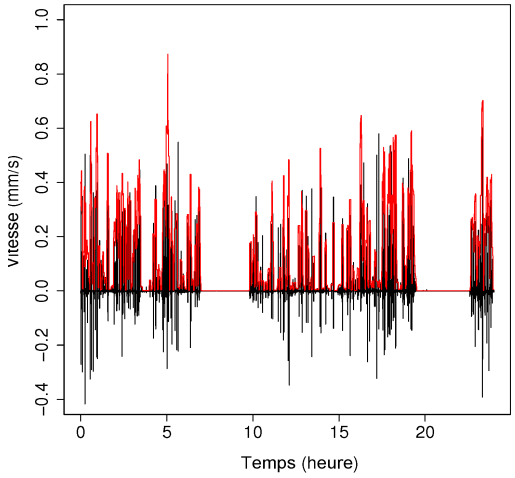

FIGURE 1 – Représentation des vitesses d'ouverture des huîtres en noir et de l'estimateur du 0.999-quantile extrême des vitesses de fermeture en rouge.

La Figure 2 révèle la présence de 3 clusters, chacune incluant une des classes de chaque variable. Dans le premier cluster au milieu à droite, les valeurs élevées de GST et faibles de AChE caractérisant une possible pollution sont associées aux plus faibles vitesses des fermetures des valves. Le second cluster en bas à gauche inclut des valeurs faibles de GST et élevées de AChE liées à une absence de pollution et associées à des vitesses élevées. Le dernier cluster correspond aux classes intermédiaires des 3 variables. Ces résultats montrent une forte dépendance entre les activités extrêmes des huîtres et une pollution du milieu. Ainsi, cette étude montre que la valvométrie fournit en instantané un outil de surveillance des risques de pollution environnementale.

Bibliographie

[1] Azaïs, R., Coudret, G. and Durrieu, G. (2014), A hidden renewal model for monitoring aquatic systems biosensors, *Environmetrics*, 25, 189-199.
[2] Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004), *Statistics of Extremes : Theory and Applications*. Wiley, Chichester, 514 pages.

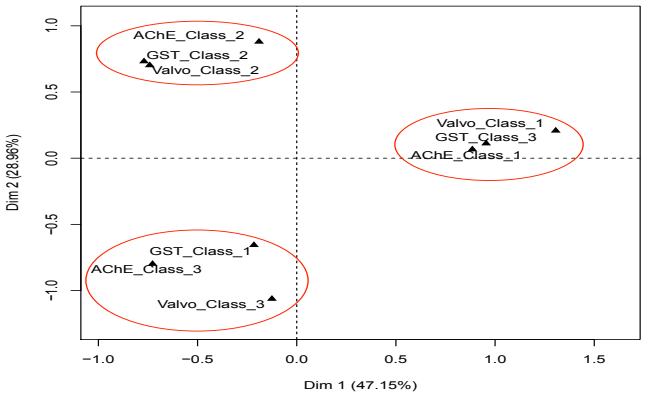

FIGURE 2 – Projection dans le premier plan principal des variables GST, AChE et des 0.999-quantiles extrêmes des vitesses de fermeture discrétisées chacune en 3 classes.

[3] Coudret, R., Durrieu, G. and Saracco, J., (2015), Comparison of kernel density estimators with assumption on number of modes, *Communication in Statistics - Simulation and Computation*, 44(1), 196-216.
[4] Durrieu, G., Grama, I., Pham, Q.K., and Tricot, J.M. Nonparametric adaptive estimation of conditional probabilities of rares events and extreme quantiles, in revision.
[5] Grama, I. and Spokoiny, V. (2008), Statistics of extremes by oracle estimation, *Annals of Statistics*, 36(4), 1619-1648..
[6] Grama, I., Tricot, J.M. and Petiot, J.F. (2014), Estimation of the extreme survival probabilities from censored data, *Buletinul Academiei de Stiinte a Republicii Moldova. Mathematica*, 74(1), 33-62.
[7] Pham, Q.K. (2015), Estimation non paramétrique adaptative dans la théorie des valeurs extrêmes : application en environnement. Phd thesis, Université de Bretagne-Sud, 138 pages.
[8] Schmitt, F. G., De Rosa, M., Durrieu, G., Sow, M., Ciret, P., Tran, D. and Massabuau, J. C. (2011), Statistical analysis of bivalve high frequency microclosing behavior : scaling properties and shot noise modeling, *International Journal of Bifurcation and Chaos*, 21(12), 3565-3576.
[9] Sow, M., Durrieu, G. and Briollais, L. (2011), Water quality assessment by means of HFNI valvometry and high-frequency data modeling, *Environmental Modeling and Assessment*, 182, 155-170.
[10] Tran, D., Fournier, E., Durrieu, G. and Massabuau, J.-C., (2004), Copper detection in the Asiatic clam Corbicula fluminea : Optimum valve closure response, *Aquatic Toxicology*, 66, 333-343.