

N^o d'ordre : 4257

THESE

présentée à

L'UNIVERSITE BORDEAUX 1

Ecole Doctorale Sciences et Environnements

Par **Mohamedou SOW**

Pour obtenir le grade de

DOCTEUR

Spécialité : Géochimie et écotoxicologie

Option : Statistique

**Développement de modèles non
paramétriques et robustes : application à
l'analyse du comportement de bivalves et
à l'analyse de liaison génétique**

Soutenue le 20 Mai 2011.

Devant la commission d'examen formée de :

M. Alain PAVÉ	PROF.	Université Lyon 1	Rapporteur
Mme. Brigitte MANGIN	DR INRA.	Toulouse	Rapporteur
M. Jérôme SARACCO	PROF.	Université Bordeaux 1	Examinateur
M. Jean-Charles MASSABUAU	DR CNRS.	Bordeaux	Examinateur
M. Laurent BRIOLLAIS	Assistant PROF.	Université Toronto	Directeur de thèse
M. Gilles DURRIEU	PROF.	Université Bretagne Sud	Directeur de thèse

EA Université Bordeaux 1 et UMR
CNRS 5805 EPOC
Place du Dr Peyneau
33120 Arcachon

Samuel Lunenfeld Research Insti-
tute Mont Sinai Hospital
60 Murray street M5G 1X5,
Toronto, ON

Je dédie cette thèse à la mémoire de mon grand père, Ardo Sélo.

Remerciements

Je tiens tout d'abord à remercier Jean-Charles MASSABUAU pour m'avoir accueilli chaleureusement au sein de son équipe et pour sa disponibilité, son accessibilité et pour tout ce qu'il m'a appris en biologie. En particulier, je lui suis très reconnaissant de s'être intéressé à mes travaux, tout le long de ma thèse, et d'avoir eu des discussions constructives avec lui.

C'est avec une profonde sincérité que j'adresse toute ma reconnaissance à Gilles Durrieu et Laurent Briollais pour m'avoir confié cette étude. Leurs grandes qualités humaines, leur disponibilité à toute épreuve et leurs remarquables compétences ont rendu ce travail particulièrement enrichissant. Je voudrais aussi les remercier pour la confiance qu'ils m'ont accordée tout au long de ces années. Je voudrais enfin les remercier pour m'avoir communiqué leur passion pour la recherche que ce soit en terme de résultats théoriques ou de leur mise en oeuvre concrète.

Je tiens bien entendu à exprimer ma sincère reconnaissance à Monsieur Alain Pavé, Professeur à l'Université Claude-Bernard, Lyon 1 et à Madame Brigitte Mangin, directrice de recherche à l'INRA de Toulouse, qui ont accepté, malgré toutes leurs occupations, d'être les rapporteurs de ce manuscrit de thèse. Je leur sais gré de l'intérêt qu'ils ont porté à mon travail. Je souhaite les remercier pour leur lecture attentive de ce mémoire ainsi que pour leurs remarques et commentaires qui m'ont permis d'en améliorer le contenu.

J'adresse mes plus vifs remerciements à Monsieur Jérôme Saraco, Professeur à l'Université de Bordeaux 1, pour l'intérêt qu'il a porté à cette étude et pour me faire l'honneur de présider mon jury de thèse.

Cette thèse a été réalisée grâce au support financier de la Région Aquitaine (Bourse doctorale région Aquitaine 2007-2009), le Samuel Lunenfeld Research Institute et le CIHR (grants 110053 et 43821). Je tiens à adresser mes remerciements à ces structures qui ont permis la réalisation de ce projet dans d'excellentes conditions.

Mes remerciements vont aussi à Pierre CIRET pour sa gentillesse, sa modestie et son efficacité pour le moindre problème.

Je tiens également à remercier chaleureusement tous les membres de l'équipe EA pour leur gentillesse et leur bonne humeur : Régine, Jean-Paul, Agnès, Nathalie, Magalie, Alexia, Patrice, Bruno et Henri. Je fais un clin d'oeil tout particulier à Damien et Fabien pour les nombreuses discussions toujours très intéressantes et souvent fructueuses.

Un grand merci également à Flo et Cathy, les deux petites mamans du labo pour leur bonne humeur et leur aide quotidienne.

J'adresse toutes mes amitiés aux jeunes docteurs, doctorants et ingénieurs avec qui j'ai partagé de très bons moments : Ika, Sebastian, Jade, Nicolas, Adeline, Audrey, Hugues, Amina, Dalia, Sophie, Simona et Cerise. J'ai également une pensée chaleureuse pour les étudiants du Master 2 : Piwi, Cecile, Vai et Nico.

J'adresse mes plus sincères remerciements à mes collègues de Toronto : Isabel, Taraneh, Mathieu (aka mister shawarma) et Rafiq. Merci aussi à Teresa pour son aide quotidienne pour les démarches administratives.

Je ne manquerais pas d'envoyer toute ma sympathie aux membres de l'équipe ECOBIOC : Benoit, Nico, Sabrina, Aurelie chaalali, Aurelie Garcia, Aurelie Ciutat, Roger, Guillaume, Flo et Alicia.

Merci à l'équipe de LMAM pour son accueil chaleureux.

Je tiens à remercier Bernard Bercu pour sa disponibilité pour répondre à certaines de mes questions concernant l'estimation fonctionnelle.

Je tiens à remercier mes amis, qui ont participé de près ou de loin à la bonne réalisation de cette thèse. Un clin d'oeil tout particulier à Inthy Diarra (aka Pape) et Kane Mamadou pour les temps de détente que nous avons pu partager.

Sur un plan plus personnel, je voudrais souligner le soutien inconditionnel de toute ma famille qui m'a accompagnée tout au long de mes études. Je souhaite donc remercier mes parents, ma soeur, mes frères et mes cousins qui m'ont toujours encouragé dans mes études et m'ont appuyé tant dans les bons que dans les moins bons moments.

Enfin, j'adresse mes remerciements à tous ceux qui ont participé de près ou de loin à la bonne réalisation de la soutenance de cette thèse.

Résumé

Le développement des approches robustes et non paramétriques pour l'analyse et le traitement statistique de gros volumes de données présentant une forte variabilité, comme dans les domaines de l'environnement et de la génétique, est fondamental. Nous modélisons ici des données complexes de biologie appliquées à l'étude du comportement de bivalves et à l'analyse de liaison génétique. L'application des mathématiques à l'analyse du comportement de mollusques bivalves nous a permis d'aller vers une quantification et une traduction mathématique de comportements d'animaux *in-situ*, en milieu proche ou lointain. Nous avons proposé un modèle de régression non paramétrique et comparé 3 estimateurs non paramétriques, récursifs ou non, de la fonction de régression pour optimiser le meilleur estimateur. Nous avons ensuite caractérisé des rythmes biologiques, formalisé l'évolution d'états d'ouvertures, proposé des méthodes de discrimination de comportements, utilisé la méthode des shot-noises pour caractériser différents états d'ouverture-fermetures transitoires et développé une méthode originale de mesure de croissance en ligne.

En génétique, nous avons abordé un cadre plus général de statistiques robustes pour l'analyse de liaison génétique. Nous avons développé des estimateurs robustes aux hypothèses de normalités et à la présence de valeurs aberrantes, nous avons aussi utilisé une approche statistique, où nous avons abordé la dépendance entre variables aléatoires via la théorie des copules. Nos principaux résultats ont montré l'intérêt pratique de ces estimateurs sur des données réelles de QTL et eQTL.

Mots-clefs

Biomonitoring, régression non paramétrique, estimation à noyau, valvométrie HFNI, bivalve, estimateur robuste, eQTL, méthode de paires de germains.

Abstract

The development of robust and nonparametric approaches for the analysis and statistical treatment of high-dimensional data sets exhibiting high variability, as seen in the environmental and genetic fields, is instrumental. Here, we model complex biological data with application to the analysis of bivalves' behavior and to linkage analysis. The application of mathematics to the analysis of mollusk bivalves' behavior gave us the possibility to quantify and translate mathematically the animals' behavior *in situ*, in close or far field. We proposed a nonparametric regression model and compared three nonparametric estimators (recursive or not) of the regression function to optimize the best estimator. We then characterized the biological rhythms, formalized the states of opening, proposed methods able to discriminate the behaviors, used shot-noise analysis to characterize various opening/closing transitory states and developed an original approach for measuring online growth. In genetics, we proposed a more general framework of robust statistics for linkage analysis. We developed estimators robust to distribution assumptions and the presence of outlier observations. We also used a statistical approach where the dependence between random variables is specified through copula theory. Our main results showed the practical interest of these estimators on real data for QTL and eQTL analysis.

Keywords

Biomonitoring, nonparametric regression, kernel estimation, valvometrie HFNI, bivalve, robust estimator, eQTL, sib-pair method.

Table des matières

Introduction générale	11
0.1 Développement de méthodes statistiques pour l'analyse du comportement de bivalves	12
0.2 Développement de méthodes statistiques pour l'analyse de liaison génétique	15
PARTIE 1. Développement de méthodes statistiques pour l'analyse du comportement de bivalves	19
1 Modélisation non paramétrique basée sur des estimateurs de type noyau	25
1.1 Introduction	25
1.2 Modèle et estimateurs	26
1.2.1 Estimation non paramétrique de la densité de probabilité . . .	26
1.2.2 Estimation non paramétrique de la fonction de régression . . .	27
1.2.3 Estimation des vitesses de fermetures et d'ouvertures	28
1.2.4 Choix du paramètre de lissage	28
1.3 Application de l'estimation fonctionnelle	29
1.3.1 Densité de probabilité des états d'ouvertures et fermetures . .	29
1.3.2 Rythme biologique et corrélation entre l'activité valvaire et la marée	29
1.4 Article 1 : Water quality assessment by means of HFNI valvometry and high-frequency data modeling. Sow, M., Durrieu, G., Briollais, L., Ciret, P., Massabuau, J.C. (2011). Environ Monit Assess. DOI 10.1007/s10661-010-1866-9.	34
1.5 Détection de la ponte par l'estimation de la vitesse	51
2 Article 2 : Comparaison d'estimateurs de régression non paramétriques : application en valvométrie	55
2.1 Introduction	56
2.2 Modèle et estimateurs	57
2.3 Propriétés asymptotiques	58
2.4 Choix de la fenêtre	59
2.5 Application en valvométrie	60

3	Quantiles de régression : application à l'estimation de la croissance	63
3.1	Introduction	63
3.2	Modèle et estimateurs	64
3.3	Application	67
3.4	Article 3 : In situ giant clam growth rate behavior in relation to temperature : a one year coupled study of high-frequency non-invasive valvometry and sclerochronology. Schwartzmann, C., Durrieu, G., Sow, M., Ciret, P., Lazareth, C.E., Massabuau, J.C. (2011). Limnology and Oceanography. (en révision favorable).	70
4	Estimation par la théorie des shot noise	111
4.1	Introduction	111
4.2	Méthodes	114
4.2.1	Densités de probabilité et extrêmes	114
4.2.2	Densité spectrale	115
4.2.3	Shot noise	116
4.3	Article 4 : Statistical study of bivalve high frequency microclosing behavior : scaling properties and shot noise analysis. Schmitt, F. G., De Rosa, M., Durrieu, G., Sow, M., Ciret, P., Tran, D., Massabuau, J.C. (2011). International Journal of Bifurcation and Chaos. (sous presse).	118
	PARTIE 2. Développement de méthodes statistiques pour l'analyse de liaison génétique	133
5	Approches robustes pour l'analyse de paires de germains : application à l'analyse de QTL et eQTL	135
5.1	Introduction	135
5.2	Article 5 : Robust Alternatives to Sib-Pair Methods for QTL and eQTL analyses. Sow, M., Durrieu, G., Gagnon, F., Briollais, L. (en préparation).	137
6	Copules et régression quantile : application à l'analyse de eQTL	161
6.1	Introduction	161
6.2	Modèles et estimateurs	162
6.3	Application	163
6.4	Article 6 : A Robust Statistical Framework for eQTL Analysis. Sow, M., Durrieu, G., Briollais, L. (2010). Joint Statistical Meeting proceedings.	164
7	Discussion générale	181
	Table des figures	189
	Bibliographie	191

Introduction générale

La recherche biologique actuelle dispose d'outils d'analyse de plus en plus performants, capables de générer des quantités de données de plus en plus grandes. Ces techniques ouvrent des champs d'analyse complètement nouveaux, souvent largement vierges. La protection de l'environnement et la découverte de gènes responsables de maladies sont parmi les plus grands des enjeux scientifiques, thérapeutiques et économiques de ces dernières années.

Les récentes innovations réalisées sur les méthodes d'acquisitions et les appareils de mesures ainsi que l'utilisation intensive de moyens informatiques permettent souvent de récolter des volumes de données qui étaient inimaginables il y a peu d'années. De nouvelles questions, demandant de nouvelles solutions mathématiques, sont donc apparues. Par exemple, mais de façon très générale, les données récoltées présentent beaucoup de sources de problèmes (distributions non gaussiennes, gros volumes de données, valeurs manquantes, observations dépendantes, valeurs aberrantes, non linéarité, etc). En fait, dès lors que l'on se confronte aux données réelles générées par ces technologies (en particulier dans les domaines de l'environnement et de la génétique), la modélisation devient complexe.

Cependant, il est merveilleux de constater, dans les découvertes de la génétique et de l'environnement, que le calcul des probabilités fournit des modèles mathématiques efficaces et la statistique un cadre rigoureux d'analyse et de modélisation des résultats expérimentaux. La modélisation est devenue une méthodologie incontournable. Associée à la simulation numérique, elle montre une redoutable efficacité. L'évolution des moyens de calcul de ces dernières années a permis non seulement d'augmenter considérablement le volume des simulations mais aussi la résolution numérique de modèles complexes pouvant prendre en compte un grand nombre de variables.

L'objet de ce travail est d'apporter des solutions aux problèmes statistiques soulevés par l'analyse et le traitement de gros volumes de données dans les domaines de l'environnement et de la génétique quantitative en développant des méthodes statistiques (robustes et non paramétriques) pour l'analyse de données :

- environnementales, acquises sur des sites multiples (actuellement bassin d'Arcachon, Nouvelle Calédonie et Espagne mais prochainement Bretagne, Norvège, Russie, ...), afin de développer un outil de surveillance de la qualité de l'eau basé sur l'analyse en continue du comportement de bivalves (bio-indicateur de pollution sur le terrain).
- de génétique quantitative afin de contribuer au développement de méthodes robustes en liaison génétique et traitement des puces à ADN.

Dans la première partie de cette introduction, nous présentons la problématique environnementale et dans la seconde la problématique génétique. Enfin nous donnons le plan de cette thèse.

0.1 Développement de méthodes statistiques pour l'analyse du comportement de bivalves

Nous savons tous que l'activité humaine a généré et continue de générer de multiples contaminants. Une prise de conscience se développe dans de très nombreux pays à travers le monde mais il est clair que le problème va s'aggraver à court et moyen terme si les sociétés ne réagissent pas vigoureusement. Dans le domaine de l'environnement, des réglementations et des contrôles sur la qualité des eaux ont été mis en place. De nouvelles techniques et de nouveaux outils doivent être imaginées et développées. 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 traces (concentrations très faibles) de contaminants via par exemple l'accumulation dans des tissus animaux ou végétaux spécifiques (organotropisme), des modifications sur des structures populationnelles (phytosociologie), etc.

Nous avons travaillé ici sur le développement d'un outil d'analyse du comportement de mollusques bivalves, qui permet d'aborder leur éthologie, mais qui est aussi un formidable moyen de surveillance de la qualité de l'eau. Le cas des mollusques bivalves est particulièrement intéressant en tant qu'espèce bioindicatrice car se sont des animaux sédentaires qui peuvent être témoins de pollutions locales éventuelles et on les trouve partout dans le monde, des tropiques aux pôles (possibilité d'application très large). Nous avons développé un grand nombre de travaux sur l'huître pour différentes raisons. D'une part, tout simplement parce que l'équipe EA (Ecotoxicologie Aquatique) de l'UMR CNRS 5805 est situé au sein de la Station Marine d'Arcachon, et que l'huître y est un matériel expérimental commun. D'autre part parce que le Bassin d'Arcachon est le premier centre de reproduction d'huîtres creuses en France. Les éleveurs bretons, normands et méditerranéens s'approvisionnent majoritairement en naissain d'origine arcachonnaise. Jusqu'à ces dernières années, et le problème dramatique de mortalités inexplicables que traverse l'ostréiculture arcachonnaise, on estimait que 60 à 70% des huîtres creuse commercialisées en France étaient natives du bassin d'Arcachon. La commercialisation du naissain constitue (ou constituait si les problèmes continuaient ?), une part significative de la ressource des professionnels arcachonnais. Une enquête menée par les services des Affaires Maritimes d'Arcachon montrait qu'une quarantaine d'entreprises ostréicoles du bassin vit essentiellement du commerce du naissain (Auby et Maurer, 2004). Une autre raison est que la compréhension des mécanismes fondamentaux du réglage de l'activité ventilatoire des huîtres et des bivalves en général reste un problème largement non résolu de physiologie respiratoire chez les animaux aquatiques et d'écophysiologie. Comprendre ces questions est aussi un problème d'écologie qui vise à mieux comprendre de nouveaux aspects de leur comportement trophique et de leur croissance ce qui permet de mieux gérer leurs populations, ainsi que de savoir comment

elles peuvent influencer leur milieu et comment elles peuvent être influencées par ce milieu. 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'activités des valves de mollusques). Les premiers enregistrements d'activité de bivalves ont été publiés en 1909 (Marceau, 1909) mais depuis, la technique a été souvent reprise pour évoluer régulièrement. En très bref il s'agit de fixer sur les valves des animaux un système d'enregistrement qui permet de suivre en permanence l'état d'ouverture/fermeture des animaux ainsi que les divers mouvements associés. Tous les systèmes utilisés à l'heure actuelle étaient basés sur la fixation par une valve de l'animal sur le dispositif expérimental. La seconde valve était libre et portait un seul capteur. L'imagerie vidéo est apparue récemment comme alternative intéressante mais avec les limitations qu'on imagine pour une optique de caméra devant éventuellement rester sous l'eau pour de longues périodes. Les deux applications les plus connues dans le domaine, sans doute parce qu'elles ont trouvé une application commerciale, sont le Mosselmonitor (Kramer et Foekema, 2001) et le Dreissena monitor (Borcherding, 2006). Ifremer a aussi développé un système avec la société Micrel NKE.

Au sein de l'UMR 5805 EPOC, nous avons développé notre système parce que nous ne voulions pas travailler avec des animaux collés (cas des systèmes précédents) mais sur des animaux libres de tous mouvements pour de longues, voir très longues (plusieurs mois) périodes de temps. Des animaux stressés par des contraintes expérimentales peuvent en effet présenter une perte de sensibilité très importante. La valvométrie a d'abord été développée en laboratoire pour la recherche fondamentale. Les buts étaient de mieux comprendre certains aspects de physiologie respiratoire et d'écophysiologie en respectant au mieux le comportement naturel spontané de petits bivalves d'eau douce, les corbicules (*Corbicula fluminea*). Elle a aussi permis d'aborder diverses questions en écotoxicologie (thèse de Damien Tran, 2001). A partir de 2006, la technologie a évolué pour permettre le travail sur le terrain afin de caractériser et traiter statistiquement le signal obtenu sur de très longues séries. Un des buts est de faire du suivi à très long terme afin de comprendre le fonctionnement et la trajectoire temporelle des systèmes environnementaux, leurs réponses à des événements perturbateurs et leur dynamique d'ajustement.

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 1-2 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. Comme ils ne disposent la plupart du temps pas de moyens de fuite bien efficace, leur protection principale est basée sur la fermeture de leurs valves. De nombreuses anomalies telles que le contact d'un prédateur ou l'arrivée d'un contaminant dans la colonne d'eau détectés (grâce aux différents types de mécano- et chimiorécepteurs) 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 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 du milieu (Chambon et al., 2007 ; Tran et al., 2003-2004 ; Sow et al., 2011 ; Schmitt et al., 2011).

Un des projets récents à l'UMR 5805 EPOC est la mise en place de dispositifs expérimentaux permettant de suivre 24h/24 le comportement *in situ* de lots de 16 bivalves. Le premier dispositif qui a été installé, a été placé sous la jetée d'Eyrac sur le bassin d'Arcachon, devant la Station Marine d'Arcachon. Le dispositif qui a été installé le plus loin utilise des lots de 16 bénitiers, qui ont été mis à 17 000 km d'Arcachon, dans le lagon sud de Nouvelle Calédonie. Divers sites sur la côte Atlantique française et espagnole (Port de Santander) ont été équipés pour quelques semaines ou quelques mois. Aujourd'hui, les projets à courts termes sont en Bretagne Sud (Vannes), en Norvège (Longyearbyen au Svalbard) et en Russie (région de Murmansk, en mer de Barents). Une caractéristique particulière de ce projet est son approche, totalement pluridisciplinaire, alliant biologie, électronique et mathématiques appliquées (statistiques).

Le premier site équipé, dans le bassin d'Arcachon au pied de la jetée d'Eyrac face à la station marine, est à ce jour celui qui a généré la plus grande quantité de données (1 728 000 mesures par jour) depuis février 2006. Les écartements valvaires de 16 huîtres ont été et sont mesurées 24h/24. Les données sont accumulées et stockées pendant 24h sur site (chaque fichier représente environ 18M octets pour une journée) puis transmises au laboratoire par protocole de transfert de données GPRS (General Packet Radio Service) en utilisant le réseau de la téléphonie mobile et Internet (FTP). Cet important volume de données enregistrées à haute fréquence avec un nombre de variables obtenues en parallèle qui pourra être important dans l'avenir (couplage avec d'autres capteurs comme des sondes multi-paramètres), nécessite le développement de modèles statistiques performants afin de bien décrire le comportement des animaux *in situ* rapidement. En effet, un des buts est de ne pas laisser s'accumuler des données non traitées et d'obtenir immédiatement des métadonnées évolutives, présentées sous forme de graphiques lisibles instantanément. Une des approches en particulier repose sur l'analyse du fonctionnement et du dysfonctionnement des rythmes biologiques (circadien et circatidaux).

Le but des méthodes statistiques développées et appliquées a été entre autres d'extraire des rythmes biologiques et de les présenter sous forme graphique simple pour permettre, par la caractérisation de perturbations de ces rythmes, de détecter *in fine* une éventuelle pollution du milieu. Notre hypothèse de travail est que les organismes aquatiques doivent présenter des perturbations de leur comportement de base et de leurs rythmes biologiques face aux contraintes du milieu. Dans ce raisonnement, on pense que les caractéristiques typiques de leurs activités physiologiques ou comportementales seront modifiées et seront le témoin sensible et immédiat de cette perturbation du milieu. Nous avons utilisé plusieurs approches afin d'aborder divers aspects du comportement décrivant au niveau du jour, de la semaine, du mois ou sur une plus longue période de temps des traits de vie caractéristiques d'un mollusque bivalve.

Actuellement et en partie grâce aux résultats de nos travaux, l'acquisition, le transfert, et le traitement des données fonctionnent de manière automatique quelque soit le site où un système est positionné dans le monde. C'est ce qui est fait en routine sous la jetée d'Eyrac sur le bassin d'Arcachon, en Nouvelle Calédonie et en Espagne (projet européen Portonovo). Les enregistrements et les résultats du traitement statistique sont accessibles sur le site web "L'oeil du mollusque" (http://www.domino.u-bordeaux.fr/molluscan_eye).

0.2 Développement de méthodes statistiques pour l'analyse de liaison génétique

Dans le contexte de la génétique humaine, la reconnaissance des gènes à l'origine de maladies multifactorielles complexes (cancer du sein, diabète, hypertension, etc ...) suscite des enjeux scientifiques et médicaux considérables (compréhension du fonctionnement, de l'évolution des organismes, diagnostique génétique et thérapie génétique, nouveaux médicaments, ...). Mais au-delà de ces retombées, la découverte de nouveaux gènes a engendré d'importants développements techniques et technologiques. Le traitement, l'interprétation, et l'automatisation des procédés ont conduit à des expériences multiples, générant simultanément un grand nombre d'informations. Enfin, les statistiques et l'informatique sont dès lors devenus indispensables pour automatiser les expériences comme pour sauvegarder et analyser les résultats.

En général, beaucoup de maladies humaines peuvent être décrites par des traits quantitatifs (exemple pression sanguine, index d'obésité, etc ...). Un caractère quantitatif d'intérêt (par exemple le poids ou la taille), peut être plus particulièrement influencé par certains gènes ou expressions de gènes sur le génome.

L'objectif des généticiens est de détecter et localiser les gènes de susceptibilités à des pathologies. Certains de ces gènes ont un effet suffisamment important pour être détecté en utilisant simplement les observations faites sur des individus : ces gènes sont appelés gènes majeurs. D'autres ne peuvent pas être détectés avec ces seules observations, et il est nécessaire d'utiliser l'information fournie par des cartes génétiques. Ils sont alors appelés Quantitative Trait Loci (QTL).

Il existe aussi l'approche eQTL (QTL d'expression) pour l'identification de gènes responsables de la variabilité de caractères quantitatifs. Cette approche consiste, par l'analyse de liaison génétique réalisée non plus sur des traits quantitatifs associés aux maladies (comme cela est fait pour une analyse QTL), mais sur des données d'expression génique, pour identifier des régions dites eQTL (Expression Quantitative Trait Loci). Une région eQTL contrôle ainsi la variabilité d'expression d'un gène ou d'un groupe de gènes qui lui sont liés. Si certaines régions QTL et eQTL, identifiées dans un seul et même dispositif, sont co-localisées, il est possible que le gène responsable des variations phénotypiques pour le caractère d'intérêt soit aussi responsable de la variation d'expression des gènes reliés à la région eQTL.

L'émergence de nouvelles technologies comme la SAGE (Serial Analysis of Gene Expression) permet d'identifier et de quantifier, simultanément, le niveau d'expression de plusieurs milliers de gènes, dans un type cellulaire donné (Velculescu et al., 1995). On peut citer aussi la technologie des puces à ADN (Schena et al., 1995 ;

Lockhart et al., 1996), moins coûteuse et surtout plus évolutive en terme d'applications. En effet, les puces a ADN permettent non seulement de visualiser, simultanément, le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier ; mais aussi d'étudier la séquence des gènes dans un échantillon, les mutations ou le polymorphisme (Mantripragada et al., 2004). Elles sont donc rapidement devenues les outils privilégiés pour l'analyse de l'expression des gènes. La masse de données générées par ces nouvelles technologies présente des défis plus considérables car la distribution de nombreuses expressions de gènes s'écarte souvent d'une distribution gaussienne et peut être affectée par des observations aberrantes notamment par des événements extrêmes (Durrieu et al., 1999 ; Briollais et al., 2007 ; Durrieu et Briollais, 2009 ; Sow et al., 2010, Schmitt et al., 2011), plus que ce qui est observé dans une analyse classique de QTL.

Une comparaison rigoureuse des performances de différentes méthodes d'estimation nécessite généralement la simulation de données afin d'évaluer le comportement des estimateurs et ainsi appliquer correctement les méthodes sur des données réelles. Les propriétés asymptotiques et les performances des estimateurs ont été étudiées sur des données réelles de QTL et eQTL.

Comme nous l'avons déjà dit plus haut, la thèse se divise en 2 parties pour des raisons de facilité d'écriture et de lecture bien que les outils mathématiques développés présentent de forts points communs ou sont carrément identiques parfois. Nous développons donc des modèles statistiques pour l'étude de systèmes biologiques complexes dans le domaine de l'environnement (partie 1) et de la génétique (partie 2) que nous allons présenter dans les différents chapitres de ce mémoire. A ce jour, quatre articles sont tirés de ce travail, un article est soumis et 2 sont en préparation.

La partie 1 se divise en quatre chapitres. Le chapitre 1 s'intéresse à la mise en place d'un modèle de régression non paramétrique basée sur des estimateurs de type noyau pour l'analyse du comportement de bivalves. Nous montrons la présence de rythmes biologiques liés aux marées chez l'huître *Crassostrea gigas*. Ces résultats permettent de caractériser une perturbation de ces rythmes qui peuvent être liées à la perturbation du milieu comme par exemple une pollution. Ce travail est décrit dans un article publié récemment dans *Environmental Monitoring and Assessment*. A partir des estimateurs développés, l'estimateur de la vitesse d'ouverture/fermeture permet de détecter automatiquement un changement de comportement de l'animale comme par exemple une ponte. Dans le chapitre 2, nous comparons au sens du critère de la variance asymptotique l'estimateur classique de Nadaraya-Watson, sa version récursive, et l'estimateur récursif de Révész sur des données acquises en valvométrie. Ce travail fait l'objet d'un article publié au 41^{èmes} Journées de la société Française de Statistique et a été réalisé en collaboration avec Thi Mong Ngoc Nguyen de l'Institut de Mathématiques de Bordeaux. Le chapitre 3 s'intéresse à la croissance des bénitiers en utilisant l'estimation non paramétrique de type quantiles conditionnels. Ce travail est actuellement soumis dans *Limnology and Oceanography* et a été réalisé en collaboration avec Claire Lazareth de l'IRD (Nouméa et Bondy). Enfin, le chapitre 4 est consacré à l'utilisation de la théorie des shot noise pour décrire les

micro-fermetures de bivalves par l'intermédiaire de quelques paramètres, contôlant principalement la densité de probabilité des inter-temps et celle de l'amplitude des micro-fermetures, ce chapitre fait l'objet d'un article accepté dans le journal "International Journal of Bifurcation and Chaos".

La partie 2 se divise en deux chapitres. Le chapitre 5 est consacré au développement et à la comparaison d'estimateurs robustes sur des données réelles de QTL et eQTL pour l'analyse de paire de germains. Enfin, le chapitre 6 s'intéresse aux quantiles conditionnelles de régression ou la distribution conditionnelle est modélisée en utilisant des copules spécifiant les dépendances entre variables. Une application sur des données eQTL montre l'intérêt pratique de cette approche. Ce chapitre fait l'objet d'un article publié dans le proceeding de l'American Statistical Association JSM (joint statistical meeting). Nous terminons ce mémoire par une conclusion générale et nos perspectives de recherche future.

PARTIE 1. Développement de méthodes statistiques pour l'analyse du comportement de bivalves

Introduction

Dans cette partie, nous développons des modèles statistiques pour l'étude de systèmes biologiques complexes dans le domaine de l'environnement. Le but est de caractériser l'activité valvaire de bivalves (qui résume à elle seule l'ensemble du comportement locomoteur de ce type d'animal) en fonction des paramètres du milieu afin de décrire des comportements normaux en fonction des variations naturelles du milieu (éthologie marine) puis de discriminer ensuite des comportements atypiques liés à une perturbation du milieu. Notre hypothèse de travail est que les organismes aquatiques doivent présenter des perturbations de leur comportement de base et de leurs rythmes biologiques face aux contraintes du milieu. Les caractéristiques typiques de leurs activités physiologiques ou comportementales seront modifiées et seront le témoin de cette perturbation du milieu.

Dans le domaine de l'environnement, il est courant d'obtenir de grands ensembles de données. Dans ce domaine, nous avons utilisé plusieurs approches afin d'aborder divers aspects du comportement décrivant au niveau du jour, de la semaine, du mois ou sur une plus longue période de temps des traits de vie caractéristiques d'un mollusque bivalve.

Après une présentation des données environnementales, nous présentons quatre articles dont 3 sont publiés, et un est actuellement soumis. Le chapitre 1 s'intéresse à la mise en place d'un modèle de régression non paramétrique basé sur des estimateurs de type noyau pour l'analyse du comportement individuel et moyen des individus afin de déterminer un fonctionnement caractéristique de bivalves. A partir de nos estimations, nous avons pu extraire des rythmes biologiques liés aux marées chez l'huître *Crassostrea gigas* et les présenter sous forme graphique simple pour permettre, par la caractérisation de perturbations de ces rythmes, de détecter *in fine* une éventuelle perturbation du milieu comme par exemple une pollution. Ce travail est décrit dans un article publié dans Environmental Monitoring and Assessment. Une extension de ce travail est consacrée à l'estimation de la vitesse d'ouverture/fermeture en vue de détecter automatiquement un changement de comportement de l'animal comme par exemple une ponte. Dans le chapitre 2, nous comparons au sens du critère de la variance asymptotique l'estimateur classique de Nadaraya-Watson, sa version récursive et l'estimateur récursif de Révész sur des données acquises en valvométrie. Ce travail fait l'objet d'un article en collaboration avec Thi Mong Ngoc Nguyen de l'Institut de Mathématiques de Bordeaux. Dans le chapitre 3, nous abordons la croissance des bénitiers en utilisant l'estimation non paramétrique de type quantiles conditionnels. Ce travail est actuellement soumis dans Limnology and Oceanography et a été réalisé en collaboration avec Claire Lazareth et Caroline Schwartzmann de l'IRD (Nouméa et Bondy). Enfin, le chapitre 4 est consacré à l'utilisation de la théorie des shot noise pour décrire les micro-fermetures de bivalves par l'intermédiaire de quelques paramètres, contrôlant principalement la densité de probabilité des inter-temps et celle de l'amplitude des micro-fermetures. Ce chapitre fait l'objet d'un article accepté dans le journal International Journal of Bifurcation and Chaos.

En ce qui concerne le dispositif expérimental, nous travaillons dans un esprit de collecte de données à haute fréquence (10 Hz) et sur le long terme (l'unité de temps

est l'année) pour aller vers une analyse fine *in situ*. Un fait important est que de plus en plus de collaborations se développent. Le résultat est donc une augmentation rapide du nombre de données à traiter.

Données

Le principe du biocapteur valvométrique développé est le suivant : chaque huître, bémittier ou autre type de bivalve est muni de deux électrodes (une émettrice et l'autre réceptrice) très légère (56 mg hors enrobage) qui sont deux bobines électromagnétiques enrobées dans une résine et collées sur chaque valve (Fig. 1). Le montage est non invasif (l'animal est libre de se positionner et se déplacer). La mesure de la distance entre les deux bobines est effectuée grâce à un principe magnétique insensible aux variations de la physicochimie de l'environnement. L'intensité du champ électrique produit entre ces deux bobines décroît avec la distance selon la loi $1/D$, D étant la distance entre le point de mesure et le centre de la bobine. Cette mesure est effectuée par le valvomètre qui est relié aux électrodes émettrices et réceptrices par un câble fin. Le signal mesuré (D_{mv} en millivolt) est converti en millimètre en utilisant l'équation (1). Les paramètres du modèle ont été estimés par la méthode des moindres carrés et le modèle est très significatif ($R^2 = 0.99$, P-value < 0.001). L'électronique de mesure utilise le principe de détection synchrone qui est sensible et immunisé contre les perturbations. Les électrodes sont gérées par une carte analogique immergée à côté des animaux en boîtier étanche. Le tout est relié par un câble ombilical à une seconde carte électronique en surface pilotée sous Linux (Fedora 12) et équipée du matériel de transmission et d'une antenne. Pour la version terrain, le tout consomme 2 watts et est alimentable par des batteries avec ou sans panneaux solaires.

$$\hat{D}_{mv} = \frac{151}{(\hat{D}_{mm})^{0.35}} - 1.48. \quad (1)$$

La fréquence d'échantillonnage est fixée à une mesure toutes les 0.1 s mais cette fréquence peut être changée. Pour avoir un nombre d'animaux représentatif, nous travaillons sur des groupes de 16 animaux. Nous disposons donc d'un total de 864000 triplets de points par jour (toutes les 0.1 sec, on a une distance au niveau d'une paire d'électrodes, l'heure de la mesure, le numéro de l'animal). Une fois les données brutes arrivées au laboratoire elles sont modélisées et traitées statistiquement. Une trentaine de minutes après le transfert depuis le terrain, les résultats peuvent être mis en accès public et/ou professionnel sur le site web. Le tableau 1 présente un extrait de l'enregistrement des données brutes. La première colonne correspond aux huîtres (numérotées de 1 à 16), la deuxième à l'amplitude d'ouverture en mm, et la troisième au inter temps en ms. A noter que dans cette configuration, avec 16 huîtres chaque animal est interrogé toute les 1.6 secondes puisque la fréquence d'échantillonnage est de 10 Hz. Les mesures de l'activité valvaire commencent à partir de minuit tous les jours.

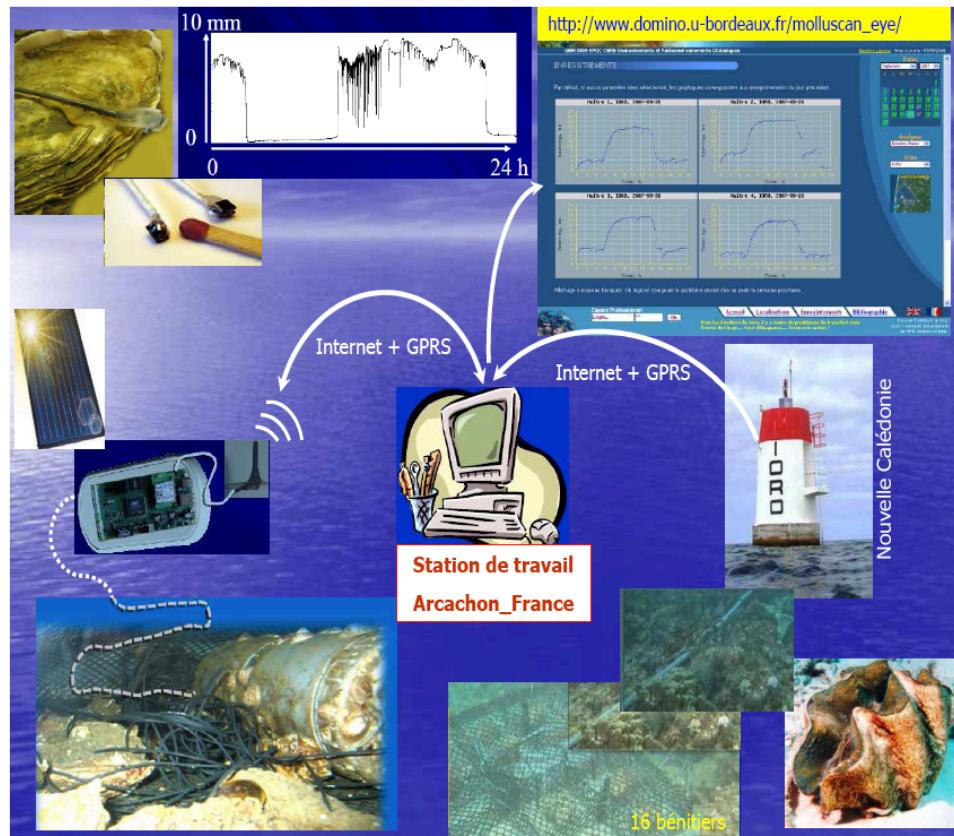


FIGURE 1 – Description du protocole expérimental de 2 sites : Eyrac (Bassin d'Arcachon, France) et récif Ioro (lagon sud de Nouvelle Calédonie, Pacifique sud). Le principe est le même pour les autres sites.

Les estimations de hauteur d'eau sont disponibles par le Service Hydrographique et Océanographique de la Marine (SHOM, <http://www.shom.fr>). Pour déterminer le rythme circadien " (jour/nuit) " sur l'activité valvaire, nous avons implémenté un code PHP qui permet de récupérer de manière automatique les horaires du lever et du coucher du soleil sur la station de travail en fonction des coordonnées GPS du site.

TABLE 1 – Extrait de l'enregistrement des données bruts (N=16 huîtres)

Huître	Ouverture (mm)	Inter temps (ms)
1	8.34	100
2	9.54	100
3	8.53	100
4	9.10	100
5	9.85	100
6	10.1	100
7	7.84	100
8	10.51	100
9	9.3	100
10	7.31	100
11	9.5	100
12	10.67	100
13	10.43	100
14	8.35	100
15	7.93	100
16	10.21	100
...	...	...

Ci-dessous un exemple de paramètres que nous pouvons estimer à partir du tableau 1 :

- Les périodes d'ouvertures et fermetures
- Durée d'ouverture et fermeture journalières
- Nombre d'ouvertures et fermetures journalières
- Nombre de micor-fermetures avec seuillage d'amplitude
- Amplitude d'ouverture minimum et maximum
- ...

Le principe de l'acquisition des données décrit ci-dessus est le même pour les différents chapitres de la première partie.

Chapitre 1

Modélisation non paramétrique basée sur des estimateurs de type noyau

1.1 Introduction

La théorie de l'estimation non paramétrique s'est développée considérablement ces deux dernières décennies. Un des problèmes souvent rencontrés en statistique est celui de l'estimation fonctionnelle associée à la loi des observations telles que par exemple la fonction de densité ou la fonction de régression.

Dans le domaine de l'environnement, on peut citer les travaux portant sur la prédiction de pics d'ozone (Damon et Guillas, 2002 ; Cardot et al., 2006), de la pollution par certains gaz à effet de serre (Febrero et al., 2007) et de la qualité de l'eau (Henderson, 2006 ; Nerini et Ghattas, 2007). Dans le domaine de la génétique plusieurs approches ont été proposées pour l'analyse d'expression de gènes (Durrieu et al., 1999 ; Leng et Müller, 2006 ; Opgen-Rhein et Strimmer, 2006 ; Song et al., 2007 ; Durrieu et Briollais, 2009 ; Sow et al., 2010). On peut citer aussi Briollais et al., (2007) qui ont proposés une approche pour le génome scan meta-analysis (GSMA). Dans le domaine de la médecine, on peut citer les études de l'évolution de certains cancers (Ramsay et Silverman, 2005 ; Cao et Ramsay, 2007) et de l'activité cardiaque (Clot, 2002 ; Ratcliffe et al., 2002a ; 2002b). Cette littérature illustre le nombre conséquent de travaux et la diversité des données de nature fonctionnelle à laquelle le statisticien peut être confronté. Dans ce chapitre, nous nous intéressons à la mise en place d'un modèle de régression non paramétrique basée sur des estimateurs de type noyau pour l'analyse du comportement de bivalves. Nous présentons d'abord l'estimateur à noyau classique d'une densité de probabilité (Rosenblatt, 1956 ; Parzen, 1962 ; Silverman, 1986), l'estimateur de la fonction de régression (Nadaraya, 1964 et Watson, 1964 ; Härdle, 1990), puis nous rappelons les propriétés asymptotiques de ces estimateurs. Ensuite nous établissons les rythmes biologiques liés aux marées chez l'huître *Crassostrea gigas*, puis nous proposons une approche statistique pour prouver cette corrélation. Une extension de cet article est consacrée à la détection de la ponte dans le bassin d'Arcachon en utilisant la dérivée première

par rapport au temps de l'estimateur de Nadaraya-Watson pour estimer les vitesses de fermetures et d'ouvertures.

1.2 Modèle et estimateurs

Un des modèles le plus fréquemment rencontré en statistique paramétrique ou non-paramétrique est le modèle de régression dont nous donnons ci-dessous une description sommaire.

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

$$Y_i = m(T_i) + \varepsilon_i. \quad (1.1)$$

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 .

Dans le cas classique, l'erreur commise est modélisé par une variable aléatoire gaussienne et la relation entre la variable dépendante Y et la ou les variables explicatives T est linéaire.

Il arrive fréquemment que ces postulats ne soient pas respectés, souvent de façon plus évidente lorsque que l'on possède un nombre important de données. Dans ce cas, le chercheur désire habituellement obtenir un modèle plus complexe, qui reflète mieux la relation entre Y et T . Il y a différentes façons d'y arriver, nous considérons ici des estimateurs non paramétriques de type noyau dans le cas du dispositif expérimental à effets fixes (la distance entre les électrodes est mesurée de manière équidistante).

Le principal avantage de la régression non paramétrique est qu'elle ne suppose aucune forme spécifique pour l'estimateur, ce qui lui donne beaucoup plus de flexibilité.

1.2.1 Estimation non paramétrique de la densité de probabilité

L'estimateur de type histogramme est classiquement utilisé en estimation de loi mais il reste très sensible au choix du nombre de classes. Nous avons donc choisi d'utiliser l'estimateur à noyau de la fonction de densité de probabilité (Rosenblatt, 1956 ; Parzen, 1962 ; Silverman, 1986 ; Wand et Jones 1995). Cet estimateur de la densité de probabilité a été utilisé pour estimer la densité de probabilité et les durées des états d'ouvertures et fermetures de chaque animal au cours d'une journée. Plus précisément, si on considère un échantillon de n variables aléatoires indépendantes et de même loi (f inconnue), $X_1, X_2, \dots, X_n$, l'estimateur de type noyau de f s'écrit :

$$\hat{f}_n(t) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right), \quad (1.2)$$

ou dans sa forme récursive :

$$\tilde{f}_n(t) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i} K\left(\frac{t - T_i}{h_i}\right). \quad (1.3)$$

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

1.2.2 Estimation non paramétrique de la fonction de régression

Nous avons considéré l'estimateur de Nadaraya-Watson (Nadaraya, 1964 et Watson, 1964 ; Härdle, 1990 ; Briollais et al., 2007 ; Durrieu et Briollais, 2009) pour la fonction de régression, c'est à dire la relation entre le temps au cours d'une journée et l'amplitude d'ouverture. Il est construit à partir d'une fonction noyau $K(\cdot)$ et d'un paramètre de lissage h_n , de manière analogue à l'estimateur à noyau de la fonction de densité de probabilité. L'estimateur de Nadaraya-Watson **NW** s'écrit sous la forme d'une moyenne pondérée des valeurs $(Y_1, \dots, Y_n)$. Il est donné par :

$$\hat{m}_{h_n}(t) = \begin{cases} \frac{\frac{1}{n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) Y_i}{\frac{1}{n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right)} & \text{si } f_T(t) \neq 0, \\ \frac{1}{n} \sum_{i=1}^n Y_i & \text{sinon.} \end{cases} \quad (1.4)$$

Cet estimateur nous a permis de décrire les mouvements des valves au cours de la journée.

Propriétés asymptotiques

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

Pour l'estimateur de f , nous avons sous les conditions C1–C3, le biais tend vers 0 et l'estimateur de type noyau $\hat{f}_{h_n}$ est un estimateur consistant de f dont la loi asymptotique est normalement distribuée quand $n \rightarrow \infty$:

$$\sqrt{n h_n} \left(\hat{f}_{h_n}(t) - f(t) \right) \xrightarrow{D} N \left(0, \tau^2 f(t) \right). \quad (1.5)$$

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 pour tout $t \in R$ tel que $f(t) > 0$, nous avons quand $n \rightarrow \infty$:

1.

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

2.

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

Il est possible d'estimer la variance de la loi normale limite. Pour cela, la densité marginale f est estimée par (1.2) et nous avons :

$$\hat{\sigma}^2(t) = \frac{1}{\hat{f}_n(t)} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) (Y_i - \hat{m}_n(t))^2.$$

1.2.3 Estimation des vitesses de fermetures et d'ouvertures

Nous avons estimé la dérivée première par rapport au temps de l'estimateur de Nadaraya-Watson pour caractériser les vitesses de fermetures et d'ouvertures afin d'identifier les changements d'états (ouvert, fermé) et des séries de contractions rapides comme lors du phénomène de pontage ou en présence de contaminants. A l'exception de la dérivée de l'estimateur de Nadaraya-Watson, le comportement asymptotique des autres estimateurs a été étudié par de nombreux auteurs. Nous avons établi la convergence en loi de cet estimateur sur des données acquises au niveau de la jetée d'Eyrac à Arcachon et les premières simulations montrent un bon comportement de cet estimateur.

Nous considérons ici la vitesse associée à l'estimateur **NW** :

$$\frac{d\hat{m}_{h_n}(t)}{dt} = \frac{1}{nh_n} \frac{\sum_i^n K'\left(\frac{t - T_i}{h_n}\right) Y_i}{f_T(t)}. \quad (1.6)$$

1.2.4 Choix du paramètre de lissage

L'utilisation pratique de ces méthodes requiert la fonction noyau et le paramètre de lissage. Si le choix du noyau est réputé comme peu influent sur l'estimateur, il n'en est pas de même pour le paramètre de lissage. Le paramètre de lissage est un facteur important dans l'estimation par la méthode des noyaux. Il représente en quelque sorte une fenêtre qui, centrée sur chaque observation, détermine le degré de lissage de l'estimation de la fonction. Un faible paramètre de lissage implique un faible degré de lissage et résulte en une fonction irrégulière. A l'opposé, une large valeur du paramètre de lissage conduit à une estimation lisse. En pratique, ce paramètre est choisi comme un compromis entre le biais et la variance. Une importante littérature est consacrée à ce sujet, et en particulier aux méthodes de sélection automatique par minimisation d'un critère (Silverman, 1986 ; Sheather et Jones, 1991).

Nous utilisons comme critère la méthode de la validation croisée (Härdle et Marron, 1985 ; Härdle, 1990) qui consiste à minimiser par rapport à h_n la fonction

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

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

1.3 Application de l'estimation fonctionnelle

1.3.1 Densité de probabilité des états d'ouvertures et fermetures

Sur des graphiques aisément lisible l'estimateur de la densité de probabilité permet de voir immédiatement si l'huître reste ouverte ou fermée toute la journée (distribution unimodale) ou si elle alterne entre ouverture et fermeture (distribution bimodale). On observe sur la figure 1.1b, un seul pic qui correspond à une ouverture autour de 6mm (99% de temps ouvert). Sur la figure 1.1c on observe une amplitude majoritaire autour de 1 mm (12% de temps ouvert) et sur la figure 1.1a, l'huître alterne entre 1mm et 7 mm (74% de temps ouvert). Ce type d'analyse peut être bien sur être utilisé pour comparer des huîtres dans différentes conditions, normales et en présence de contaminants. Des expériences en laboratoire (Tran et al., 2010) ont par exemple montré qu'en présence de l'algue toxique *Alexandrium minutum*, l'amplitude d'ouverture diminuait fortement.

1.3.2 Rythme biologique et corrélation entre l'activité valvaire et la marée

Un des objectifs, à coté de l'analyse du comportement journalier, est de se servir des résultats de nos modèles afin d'extraire des rythmes biologiques circadiens, circatidaux, saisonniers, annuels, ... Dans un premier temps, nous représentons graphiquement les périodes d'ouvertures et fermetures des bivalves sur l'ensemble des données disponibles depuis mars 2006 au niveau de la jetée d'EYRAC et depuis septembre 2007 en Nouvelle Calédonie. Ces mois d'enregistrements (Fig. 1.2) nous ont permis de mettre en évidence des rythmes biologiques liés aux rythmes des marées chez l'huître *Crassostrea gigas*. Une de nos hypothèses de travail est que la modification de ces rythmes biologiques pourrait alors être symptomatique d'un problème au niveau de la qualité de l'eau. C'est en effet ce que nous pouvons voir en présence de fermetures "anormales" à des moments de la marée où la probabilité de fermeture est faible ou nulle en situation de référence.

La figure 1.2 nous montre que les activités de fermetures sont corrélées à la marée et plus précisément nous pouvons observer que les huîtres ferment leurs valves à l'étale de basse mer. Nous proposons une approche statistique pour démontrer cette corrélation. Nous déterminons les moments de la journée où les huîtres ferment leurs

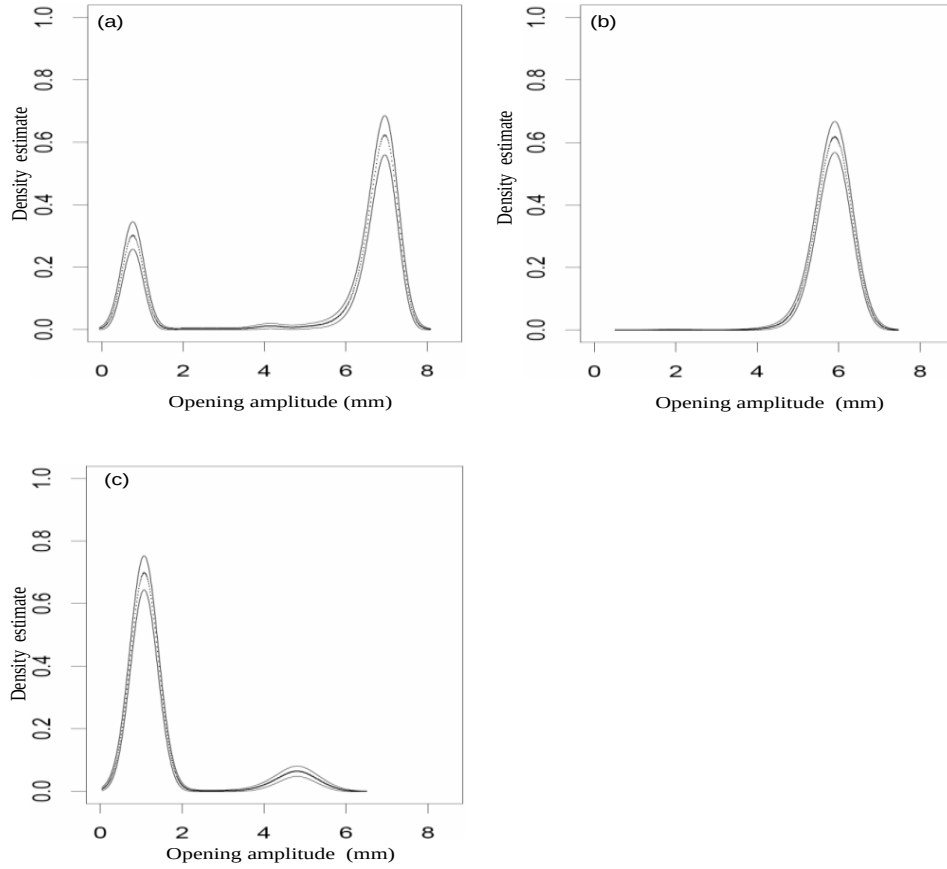


FIGURE 1.1 – Description de la distribution des états de fermetures/ouvertures au cours de la journée, avec en abscisse l’amplitude d’ouverture et en ordonnée la densité. Avec les bornes de l’intervalle de confiance à 95%.

valves et nous corrélons ces moments avec les paramètres de marée. Les périodes de fermeture des huîtres sont calculées en utilisant un seuil qui dépend du pourcentage de l’amplitude d’ouverture pour chaque animal et chaque jour. Le vecteur de ces estimations de seuils est

$$\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_{16}) \quad (1.7)$$

où pour tout $i = 1, \dots, 16$

$$\hat{s}_i = \min(\hat{Y}_i) + (\max(\hat{Y}_i) - \min(\hat{Y}_i)) \times 5\%.$$

Le seuil de 5% a été choisi après représentation de différents enregistrements mais ce choix de seuil peut se modifier. En utilisant les données marée du SHOM (Service Hydrographique et Océanographique de la Marine), nous définissons une classe notée C qui caractérise l’intervalle de temps entre les deux pleines mer pour un jour lunaire. En utilisant $\hat{\mathbf{s}}$, nous estimons chaque jour k , le vecteur $\boldsymbol{\tau}$ qui correspond au temps entre 0 et 24 h où la première fermeture des 16 animaux arrivent dans les 2 classes.

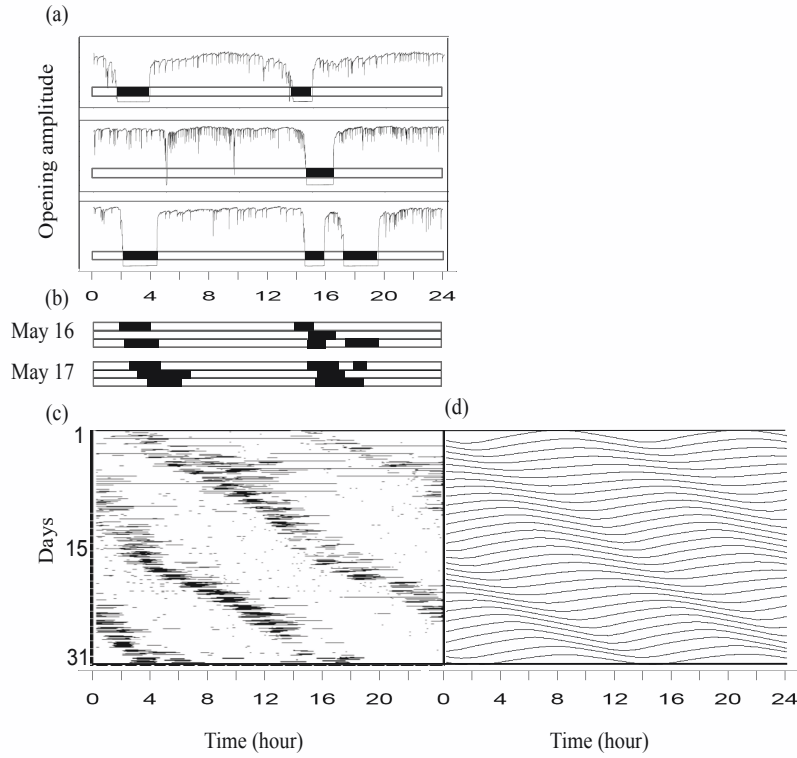


FIGURE 1.2 – Principe du rythme biologique lié à la marée : a) activité de fermetures/ouvertures pour 3 différentes huîtres (trait noir fermeture et trait blanc ouverture) ; b) superposition de fermetures/ouvertures de 3 huîtres différentes (16 et 17 Mai) ; c) représentation de l'activité de fermeture/ouverture sur 31 jours (pour chaque jour, il y a 16 lignes représentant les périodes de fermetures/ouvertures pour 16 huîtres) ; d) Représentation de l'évolution de la hauteur d'eau par heure. (Mai 2006)

$$\hat{\tau}^k = (\hat{\tau}_1^k, \dots, \hat{\tau}_{16}^k) \quad (1.8)$$

où pour tout $i = 1, \dots, 16$

$$(\hat{\tau})_i^k = \min \{t_j \in C^k \mid \hat{m}_i^k(t_j) \leq s_i^k\}.$$

Pour avoir une idée du comportement moyen des 16 animaux, nous estimons la moyenne et l'erreur type du temps où la première fermeture des huîtres a lieu dans la

classe C . La représentation graphique de ce résultat (Fig. 1.3) montre une corrélation très significative entre le moment où les animaux se ferment pour la première fois et l'heure de la marée basse. L'estimation par les moindres carrés de la pente et de l'ordonnée à l'origine sont respectivement égales à 0.995 et -0.36 ; $R^2 = 0.94$, $p\text{-value} < 0.001$.

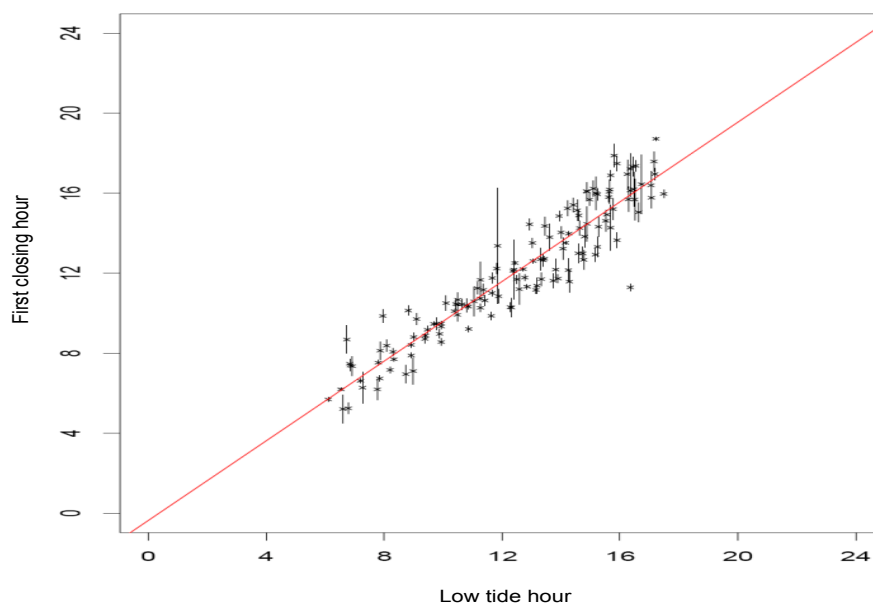


FIGURE 1.3 – Relation entre l'heure de la première fermeture et l'heure de la marée basse.

Ainsi, nous prouvons qu'en moyenne les huîtres se ferment au minimum de marée basse. Comme nous pouvons le remarquer dans la figure 1.2, les huîtres étudiées sous la jetée d'Eyrac du Bassin d'Arcachon ont tendance à ne pas se fermer entre les 2 marée basses. Nous avons proposés une méthode qui permet de visualiser les probabilités de fermeture au cours des jours. Ainsi, nous pouvons déceler des fermetures atypiques sur un graphe simple. Cette analyse statistique est publiée dans le journal "Environmental Monitoring and Assessment".

- 1.4 Article 1 : Water quality assessment by means of HFNI valvometry and high-frequency data modeling. Sow, M., Durrieu, G., Briollais, L., Ciret, P., Massabuau, J.C. (2011). Environ Monit Assess. DOI 10.1007/s10661-010-1866-9.**

Water quality assessment by means of HFNI valvometry and high-frequency data modeling

Mohamedou Sow · Gilles Durrieu · Laurent Briollais ·
Pierre Ciret · Jean-Charles Massabuau

Received: 19 July 2010 / Accepted: 21 December 2010
© Springer Science+Business Media B.V. 2011

Abstract The high-frequency measurements of valve activity in bivalves (e.g., valvometry) over a long period of time and in various environmental conditions allow a very accurate study of their behaviors as well as a global analysis of possible perturbations due to the environment. Valvometry uses the bivalve's ability to close its shell when exposed to a contaminant or other abnormal environmental conditions as an alarm to indicate possible perturbations in the environment. The modeling of such high-frequency serial valvometry data is statistically challenging, and here, a

nonparametric approach based on kernel estimation is proposed. This method has the advantage of summarizing complex data into a simple density profile obtained from each animal at every 24-h period to ultimately make inference about time effect and external conditions on this profile. The statistical properties of the estimator are presented. Through an application to a sample of 16 oysters living in the Bay of Arcachon (France), we demonstrate that this method can be used to first estimate the normal biological rhythms of permanently immersed oysters and second to detect perturbations of these rhythms due to changes in their environment. We anticipate that this approach could have an important contribution to the survey of aquatic systems.

Keywords Environmental monitoring · Nonparametric regression · High-frequency data · Valvometry · Water quality · Bivalves

M. Sow · G. Durrieu (✉) · P. Ciret · J.-C. Massabuau
CNRS UMR 5805 EPOC, University of Bordeaux 1,
Place du Dr Peyneau, 33120 Arcachon, France
e-mail: gilles.durrieu@univ-ubs.fr,
g.durrieu@epoc.u-bordeaux1.fr

M. Sow · L. Briollais
Division of Epidemiology and Biostatistics,
Samuel Lunenfeld Research Institute,
Mount Sinai Hospital,
Toronto, ON, M5G 1X5, Canada

Present Address:
G. Durrieu
LMAM, University of South Brittany,
Campus de Tohannic, 56017 Vannes, France

Introduction

Protection of the aquatic environment is a top priority for marine managers, policy makers, and the general public. Due to an increasing interest in the health of aquatic systems, there is a compelling need to using remote online sensors to instantly and widely distribute information on a daily basis, specifically in marine environment

(Kröger and Law 2005). Among these sensors, bio-indicators are increasingly used and are highly effective in their ability to reveal the presence of trace amounts (very low concentrations) of contaminants through accumulation in tissues (see the US Mussel Watch, <http://ccma.nos.noaa.gov/about/coast/nsandt/welcome.html>). The ability of mollusk bivalves to taste their environment is one of the possible ways to monitor water quality. Monitoring their opening and closing activities over time is yet another way to evaluate the behavior of the bivalves in reaction to their environmental exposure. The interest in investigating the bivalve's activities by recording their valve movements (e.g., valvometry) is not recent. Marceau was certainly pioneer in this method and performed the first recordings of molluscan valve movements (Marceau 1909). This technique has been explored in ecotoxicology for more than 20 years, first under laboratory conditions. The basic idea of valvometry is to use the bivalve's ability to close its shell when exposed to a contaminant as an alarm signal (Doherty et al. 1987; Shumway and Cucci 1987; Byrne et al. 1990; Kramer and Botterweg 1991; Borcharding 1992; Sluyts et al. 1996; Curtis et al. 2000; Kadar et al. 2001; Nagai et al. 2006). Nowadays, valvometers are available on the market and use the principle of electromagnetic induction (Sloff et al. 1983; Jenner et al. 1989) such as the Mossel Monitor (Kramer et al. 1989) or the Dreissena Monitor (Borcharding 1992). There has been a clear research interest in recent years to measuring the bivalves' behaviors directly in real conditions (Riisgard et al. 2006; Robson et al. 2007; Garcia-March et al. 2008). This is also our case where our technical innovations include the development of an online data system (no delayed analysis and graph production; near-real time data production), a remote control, a strong resistance under bad sea conditions, and no limit of distance between the sampling sites and the laboratory. Thus, our newly developed technique and analytical tools are able to monitor and analyze quantitatively the bivalve's behaviors very accurately. More specifically, we want to describe statistically the normal bivalve's activities under laboratory and/or field conditions in order to provide a reference from different seas, biotopes, and seasons

and then to further determine abnormal patterns. For this purpose, since 2003, we have developed a new approach using a noninvasive valvometric technique (high frequency, noninvasive (HFNI) valvometry) coupled with (a) the integration of the bivalve's respiratory physiology and ethology and (b) an in-depth statistical analysis that can handle very large amount of data, a key to fast online analysis of molluscan bivalve's behavior at a large scale. Three original key features have been introduced (Tran et al. 2003). First, a new valvometric system was built using lightweight electrodes (approximately 100 mg) linked by thin flexible wires to high-performance electronic units. This system allows the bivalves to be studied in their natural positions with minimal experimental constraints. Second, variations in the probability of spontaneous closures over the animals' biological rhythms were first taken into account under laboratory conditions by working during specific periods, for example, when the probability of spontaneous closing events was the lowest (Tran et al. 2003) and by controlling the animals' respiratory physiology and behavior (Tran et al. 2000, 2001; Massabuau 2001). Third, a statistical approach was developed, and it demonstrated its ability to use bivalves' closure response as a method for detecting trace amounts of metals (Tran et al. 2003, 2004, 2007; Fournier et al. 2004). It is noteworthy that from our results and observations, Liao et al. (2005, 2006) and Jou et al. (2009) developed mathematical models that describe the interaction between spontaneous valve daily rhythms and various valve response characteristics. In this study, we established two field stations, the first in the Bay of Arcachon, France, in February 2006 and the second one in the Havannah channel, New Caledonia, in August 2007. These field stations use the light electromagnetic and electronic principles as described in Chambon et al. (2007). Two major challenges we faced were (a) one data point is generated every 0.1 s at each site, 24 h a day, to optimize the description of the animals' behaviors and (b) the data sampled have a high variability. The objective of this paper is to apply and develop appropriate statistical tools to successfully summarize and analyze the high-frequency valvometry data. Although our statistical modeling was conducted

on the oysters from the Bay of Arcachon, our statistical approach was also validated on the giant clams from the site in New Caledonia. Here, we illustrate some applications using valvometry data on a group of 16 oysters from the Bay of Arcachon with valve activity measured over a period of 183 days. On a daily basis, the data and their statistical analysis are presented graphically and made available to public in the MolluSCAN Eye website http://www.domino.u-bordeaux.fr/molluscan_eye.

Materials and methods

Experimental sites and animal species

The monitoring site is located in the Bay of Arcachon, France, at the Eyrac pier (latitude 44°40 N, longitude 1°10 W). Since February 2006, groups of 16 Pacific oysters, *Crassostrea gigas*, measuring from 8 to 10 cm in length are permanently installed on this site. These oysters are all from the same age group (1.5 years old) and come from the same local supplier. They also all grew in the Bay of Arcachon. They were placed in a traditional oyster farmer bag on a bench in the bottom and permanently immersed at 3 to 7 m deep in the water, depending on the tide activity. We present here a study covering the period from 1st April 2006 to 31 September 2006.

Measurements of valve activity

The first electronic principle of monitoring was described by Tran et al. (2003) and further modified by Chambon et al. in 2007. The modification was necessary for field application in the ocean and was purely mechanical. It involved the hardening of the monitoring system to allow its constant immersion in seawater even in strong water current conditions and even during storms. Some information about these specific aspects can be found on http://www.domino.u-bordeaux.fr/molluscan_eye. The main challenge was to ensure the complete autonomy of the system without in situ human intervention for at least one full year. In brief, each animal is equipped with two light coils (sensors), ≈ 53 mg each (unembedded), fixed on the edge of each valve. These coils mea-

sured $2.5 \times 2.5 \times 2$ mm and were coated with a resin-sealing before fixation on the valves. One of the coils emits a high-frequency, sinusoidal signal which is received by the other coil. In each group of animals, one measurement was received every 0.1 s (10 Hz) on one of the 16 animals. This means that each animal's behavior was measured every 1.6 s. Every day, a data set with 864,000 pairs of values (one distance value, one stamped time value) was generated representing a total of 1,728,000 data points. Each oyster's behavior was described by a total of 108,000 measurements per day. The strength of the electric field produced between the two coils decreases with the distance according to the transformation $1/D$, where D is the distance between the point of measurement and the center of the transmitting coil. The measured signal (D_{mv} in millivolts) is converted into distance in millimeters (D_{mm}) using the following calibration model

$$\hat{D}_{mv} = \frac{151}{(\hat{D}_{mm})^{0.35}} - 1.48. \quad (1)$$

The model parameters were estimated using least-squares estimation, and the model was very significant ($R^2 = 0.99$, P value < 0.001).

Data collection and data transmission

On the shore (i.e., by a pier or lighthouse), a second electronic unit handles the data acquisition and the programmed emission. The unit is equipped with a GSM/GPRS modem and uses Linux operating system for driving the first control unit, managing the data storage with a time stamp, accessing the internet, and transferring the data. A self-developed software module runs on mobile phone technology. After each 24-h period (or any other programmed period of time), the data collected are transmitted to a central workstation server located at the Marine Biological Station in Arcachon, France. The valve activity data are stored in a central database in a merged data set. Access to this database is provided via an Intranet or Internet connection (Fig. 1).

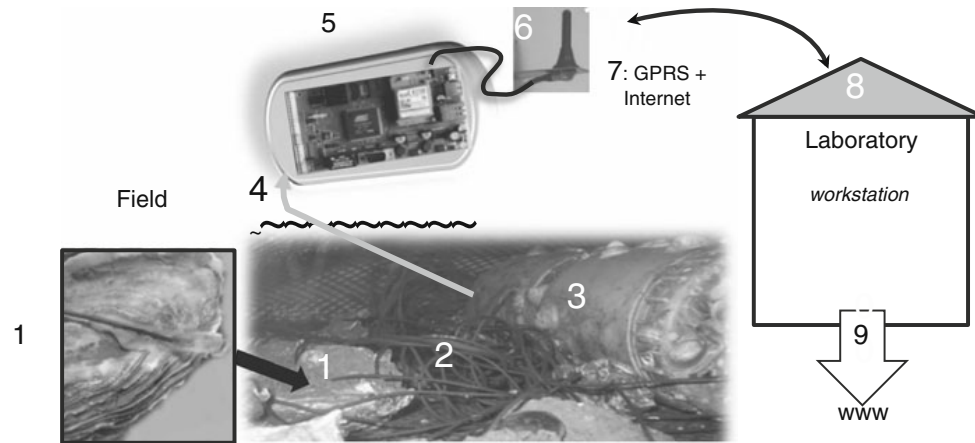


Fig. 1 Synoptic representation of the system, from field to laboratory. 1 Oyster equipped with two electrodes, 2 connecting cables from oysters to first-level electronic card, 3 first-level electronic card in a waterproof case (immersed), 4 electrical connection between the first and the

second electronic card, 5 second electronic card (emersed), 6 GPRS antennae, 7 GPRS and internet connection, 8 Marine Biological Station of Arcachon, 9 daily update on Internet

Tide and diurnal/nocturnal data

Tidal data were collected including a measurement of the height of the water column every hour and the times of the low and high tides, provided by the hydrographic and oceanographic service of the marine (<http://www.shom.fr/>). To characterize the diurnal/nocturnal rhythms of the bivalve, we estimated the time of sunrise and sunset at the sampling site.

Estimators and models

We propose a kernel-based approach for the modeling and a statistical treatment of the high-frequency valvometry data. The main novelty is the use of a nonparametric regression, which is particularly advantageous in the context of environmental studies that often exhibit high variability. The principle of our approach is first to produce a density estimation of each animal's opening and closing valve activities over a 24-h period, second to make inference about these individual profiles using kernel regression, and third to correlate them with some measured environmental data. A graphical display is also provided to represent the daily and annual profiles of each animal, which can be easily read and interpreted by specialists and nonspecialists. Another possible choice of estimator is a local polynomial estimator

(Fan and Gijbels 1996), but Einmahl and Mason (2000) and Quian and Mammitzsch (2000) emphasized the advantage of kernel regression and the use of the Nadaraya–Watson approach for the variance estimator (Nadaraya 1964; Watson 1964). One of the strengths of this estimator consists in its automatic adaptation to designs where the local polynomial estimator may not perform reliably overall since its variance may fail to exist. Also, the Nadaraya–Watson estimator retains some optimality properties as exposed in Härdle and Marron (1985). Moreover, we recently applied successfully kernel estimation procedure to high-frequency data in genetics as described in Briollais et al. (2007) and Durrieu and Briollais (2009).

Estimation of the probability density function of valve activities

To determine a general profile of valve activities during a given period of time (24 h in general), we estimated the density of probability of the theoretical states of opening and closing which are unknown in our sample. There are a number of statistical problems for which it is necessary to obtain estimates of the density. For instance, it is of great interest to reveal features of the data such as skewness and multimodality which make possible the identification of different behaviors in

the whole population (a mixture of several distributions and clusters). A histogram-type estimator is the most commonly used estimator, but it could be strongly affected by the number of classes chosen. Therefore, we decided to use a nonparametric kernel density estimator that behaves better statistically (consistency, asymptotic normality, and numerical behavior). A kernel estimator of a function f is defined by Silverman (1986), Wand and Jones (1995):

$$\hat{f}_{h_n}(t) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) \quad (2)$$

where n is the sample size and h_n is the bandwidth (the smoothing parameter) satisfying (C1) $h_n \rightarrow 0$ and $n h_n \rightarrow \infty$, as $n \rightarrow \infty$. The kernel K is a symmetric function satisfying (C2) $\int_{-\infty}^{+\infty} K(t) dt = 1$, $\int_{-\infty}^{+\infty} t K(t) dt = 0$, $\int_{-\infty}^{+\infty} |t| K(t) dt < +\infty$, and the distribution function f is twice differentiable (C3). We introduce the notation $\int_{-\infty}^{+\infty} K^2(t) dt = \tau^2$ to simplify the following formulas, and here, we considered the Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right). \quad (3)$$

Under conditions (C1)–(C3), the bias and variance of the estimator $\hat{f}_h(\cdot)$ can be approximated when $n \rightarrow \infty$ by:

$$\text{bias}(\hat{f}_{h_n}(t)) = \frac{h_n^2}{2} f''(t) \int t^2 K(t) dt + o(h_n^2), \quad (4)$$

and

$$\text{var}(\hat{f}_{h_n}(t)) = \frac{1}{n h_n} f(t) \tau^2 + o\left(\frac{1}{n h_n}\right). \quad (5)$$

Under conditions (C1)–(C3), the bias tends to zero and the kernel density estimator $\hat{f}_{h_n}$ is a consistent estimate of f asymptotically normal distributed as $n \rightarrow \infty$:

$$\sqrt{n h_n}(\hat{f}_{h_n}(t) - f(t)) \xrightarrow{D} N(0, \tau^2 f(t)). \quad (6)$$

Using Eq. 6, we derive the $(1 - \alpha)\%$ confidence interval of f when $n \rightarrow \infty$

$$\left[\hat{f}_{h_n}(t) - z_{1-\alpha/2} \sqrt{\frac{\hat{f}_{h_n}(t) \tau^2}{n h_n}}, \right. \\ \left. \hat{f}_{h_n}(t) + z_{1-\alpha/2} \sqrt{\frac{\hat{f}_{h_n}(t) \tau^2}{n h_n}} \right] \quad (7)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the standard normal distribution.

Nonparametric kernel regression model

The relationship between the distance between the two electrodes (Y) and the time of the measurement (T) is modeled using a nonparametric regression model given by:

$$Y_i = m(T_i) + \epsilon_i, \quad \text{for } i = 1, \dots, n \quad (8)$$

where n , m , and ϵ denote the sample size (total number of couples of values), the regression function to be estimated, and the model error term, respectively. The stochastic distribution f of ϵ is typically unknown and is unlikely to follow any familiar distribution such as a normal distribution and is independent of T . We therefore decided to use nonparametric statistics. Here, we consider the case of a fixed equidistant design. Also, the T_i 's have a uniform known design density f_T . However, the limiting bias and variance depend on unknown quantities which have to be estimated consistently in order to construct the asymptotic confidence intervals. The variation ϵ_i is a random variable with a mean value equal to 0 and the stochastic distribution f of these random variations enable us to characterize the variation of the random variable Y around m :

$$m(t) = E(Y/T = t) = \frac{\int y f(y, t) dy}{f(t)}. \quad (9)$$

The regression function m thus depends on the joint density which is unknown. The joint density

was estimated using a multiplicative kernel and calculations show that

$$\int y \hat{f}(y, t) dy = \frac{1}{n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) Y_i. \quad (10)$$

Using the kernel estimator of the joint densities function, we deduced the regression estimator of m given by

$$\hat{m}_{h_n}(t) = \begin{cases} \frac{1}{n} \frac{\sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) Y_i}{f_T(t)} & \text{if } f_T(t) \neq 0, \\ \frac{1}{n} \sum_{i=1}^n Y_i & \text{otherwise.} \end{cases} \quad (11)$$

Under conditions (C1)–(C3), the bias and variance of the estimator $\hat{m}_{h_n}$ are given when $n \rightarrow \infty$ by:

$$\text{bias}(\hat{m}_{h_n}(t)) = \frac{1}{2} h_n^2 m''(t) \int t^2 K(t) dt + o(h_n^2), \quad (12)$$

and

$$\text{Var}(\hat{m}_{h_n}(t)) = \frac{1}{nh_n} \sigma^2(t) \tau^2 + o\left(\frac{1}{nh_n}\right), \quad (13)$$

where $\sigma^2(t)$ is the conditional variance $\text{Var}(Y/T=t)$. We chose $h_n = n^{-\eta}$ and replaced $\sigma^2(t)$ by its estimator

$$\hat{\sigma}^2(t) = \frac{1}{n f(t)} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) (Y_i - \hat{m}_{h_n}(t))^2. \quad (14)$$

We added a new condition (C4) stating that m is twice differentiable, $E(Y^2) < \infty$ and $\eta \in [1/5, 1[$. Under conditions (C1)–(C4), $\hat{m}_{h_n}(t)$ is a consistent estimate of $m(t)$ and is asymptotically normally distributed (Härdle 1990) when $n \rightarrow \infty$:

$$\sqrt{nh_n} (\hat{m}_{h_n}(t) - m(t)) \xrightarrow{D} N(B, \sigma^2(t) \tau^2), \quad (15)$$

where B denotes the bias of $\hat{m}_{h_n}(t)$. Our analyses show that the bias is negligible compared to the variance estimator (see “[Checking the model behavior](#)”). Then we can set the bias of $\hat{m}_{h_n}(t)$ equal to zero. Equation 15 allows to compute an

$(1 - \alpha)\%$ confidence interval for $m(t)$, and it is given by

$$\left[\hat{m}_{h_n}(t) - z_{1-\alpha/2} \sqrt{\frac{\sigma^2(t) \tau^2}{n h_n}}, \hat{m}_{h_n}(t) + z_{1-\alpha/2} \sqrt{\frac{\sigma^2(t) \tau^2}{n h_n}} \right] \quad (16)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the standard normal distribution. From the expression of $\hat{m}_{h_n}(t)$ (Eq. 11), we can estimate by derivation with respect to time the speed of opening and closing by using the derivative over time:

$$\frac{d\hat{m}_{h_n}(t)}{dt} = \frac{1}{nh_n} \frac{\sum_{i=1}^n K'\left(\frac{t - T_i}{h_n}\right) Y_i}{f_T(t)}. \quad (17)$$

Bandwidth selection

Selecting a proper bandwidth parameter h_n is a critical step in estimating the density and the regression functions. Although in practice one can choose the bandwidth subjectively, this can lead to inaccurate estimation of the density and its variance. The choice of h_n is much more important for the behavior of the estimators than the choice of the kernel function K . Small values of h_n make the estimate look wiggly and show wrong features, whereas large values of h_n will lead to an estimate which is too smooth in the sense that is heavily biased and may not reveal important structural features, a bimodality for example. To optimize h_n for the density, we used a plug-in procedure developed by Sheather and Jones (1991). For the regression function estimator, we used a cross-validation (CV) criterion which leads to the score function

$$\text{CV}(h_n) = \sum_{i=1}^n (Y_i - \hat{m}_{(-i)}(T_i; h_n))^2 \quad (18)$$

with respect to h_n where $\hat{m}_{(-i)}(T_i; h_n)$ is the regression function estimator without considering (T_i, Y_i) . This bandwidth is asymptotically optimal and yields the same speed of convergence than other techniques (Härdle 1991).

All our analyses were performed on a DELL workstation using both Bash Linux (Fedora 12) and mathematical codes written in R (Venables and Ripley 1999, <http://www.R-project.org/>).

Results and discussion

Analysis of the oysters' behaviors

The kernel density estimator (Eq. 2) provides an estimation of the distribution and probability of closing and opening states and is computed each day of our study period. Three examples are given in Fig. 2. When an oyster has several sequences of opening and closing states during the day, the distribution function is bimodal. In Fig. 2a, the density has two peaks, the first one corresponding to a closing state and the second one to an opening state. The probability of the closing state for this particular oyster is equal to 0.252, and the corresponding 95% confidence interval (Eq. 7) is [0.196, 0.323], so the total duration of the closing state during the studied day is 6.048 h and the corresponding 95% confidence interval is [4.704, 7.752] with $\hat{h} = 0.42$. In contrast, when the oyster remains completely closed or open during the day, the distribution function is unimodal. In Fig. 2b, it is observed that the oyster has spent more time open than closed (with an opening amplitude >4 mm). The probability of closing for this oyster and this day is equal to 0.00208, and the corresponding 95% confidence interval is [0, 0.0206], so the total duration of closure is 0.0672 hours and the corresponding 95% confidence interval is [0, 0.494] with $\hat{h} = 0.053$. The closing probability for the oyster in Fig. 2c is 0.883, and the 95% confidence interval is [0.779, 0.986], so the total duration of closure is 21.16 h and the corresponding 95% confidence interval is [18.696, 23.664] with $\hat{h} = 0.41$. This means that this oyster spent more time closed than open during the day.

The application of the estimator in Eq. 11 gives an estimate of the function m , which models the opening amplitude as a function of time. Then, by using asymptotic normality and consistency properties of $\hat{m}_{h_n}$, we derived the following 95% confidence interval based on (Eq. 16). The bandwidth estimator has been determined by CV where the CV functions are convex and have a clear minimum. This bandwidth provides a good fit of our data (see Fig. 3).

Then, from the estimator in Eq. 11, we estimated the speed of each closing and opening

moves for each oyster (Eq. 17). In Fig. 4, we notice that the positive speeds correspond to the opening moves and the negative speeds to the closing moves of the valves. Table 1 gives the values corresponding to this activity. It shows that oyster number 1 (Fig. 3a) had three closing states and their total duration was 6.07 h. The second closing state is much quicker with a speed equals to 17.96 mm/s and opening speed equals to 2.4 mm/s. In Fig. 3b, we observe that the oyster had two micro-closures during the day that were shorter than 1 min. In Figs. 3c and 4c, this oyster has one opening move with speed equal to 7.93 mm/s and a total duration of closing state equal to 21.03 h. Therefore, the density estimation and the kernel regression both provide very useful information to establish the normal bivalve's activity for a given period of time and then can be used as a reference to determine abnormal responses to the environment (for example, due to water contamination), as illustrated below.

Association of oysters' mean behavior with nycthemeral and tidal biological rhythms

Concerning the nycthemeral rhythms in the Bay of Arcachon, the difference between diurnal and nocturnal mean duration of closing was not significant in our study period (i.e., April to September 2006) with a Mann and Whitney U test giving a P value = 0.17. The mean duration of the closing state for an oyster was 5.06 h per day with a confidence interval of [1.98, 8.16] (mean $\pm$ 1 SE) decomposed into 2.63 h ([0.89, 4.39]) for the night and 2.43 h ([0.69, 4.17]) for the day. Using the model described in Eqs. 8 and 11, we can also model the biological rhythms of the permanently immersed bivalves. To extract a biological rhythm, we used a graphical representation of the period of closing estimates. The principle of this approach is represented in Fig. 5a, b, which shows a series of opening and closing states for each oyster during a 24-h period. The length of the closing period is proportional to the length of the solid line. Figure 5c shows this representation during a full month (May 2006), the x -axis represents the time in a 24-h time period, and the y -axis represents the number of days during this month. For each day, there are 16 lines which

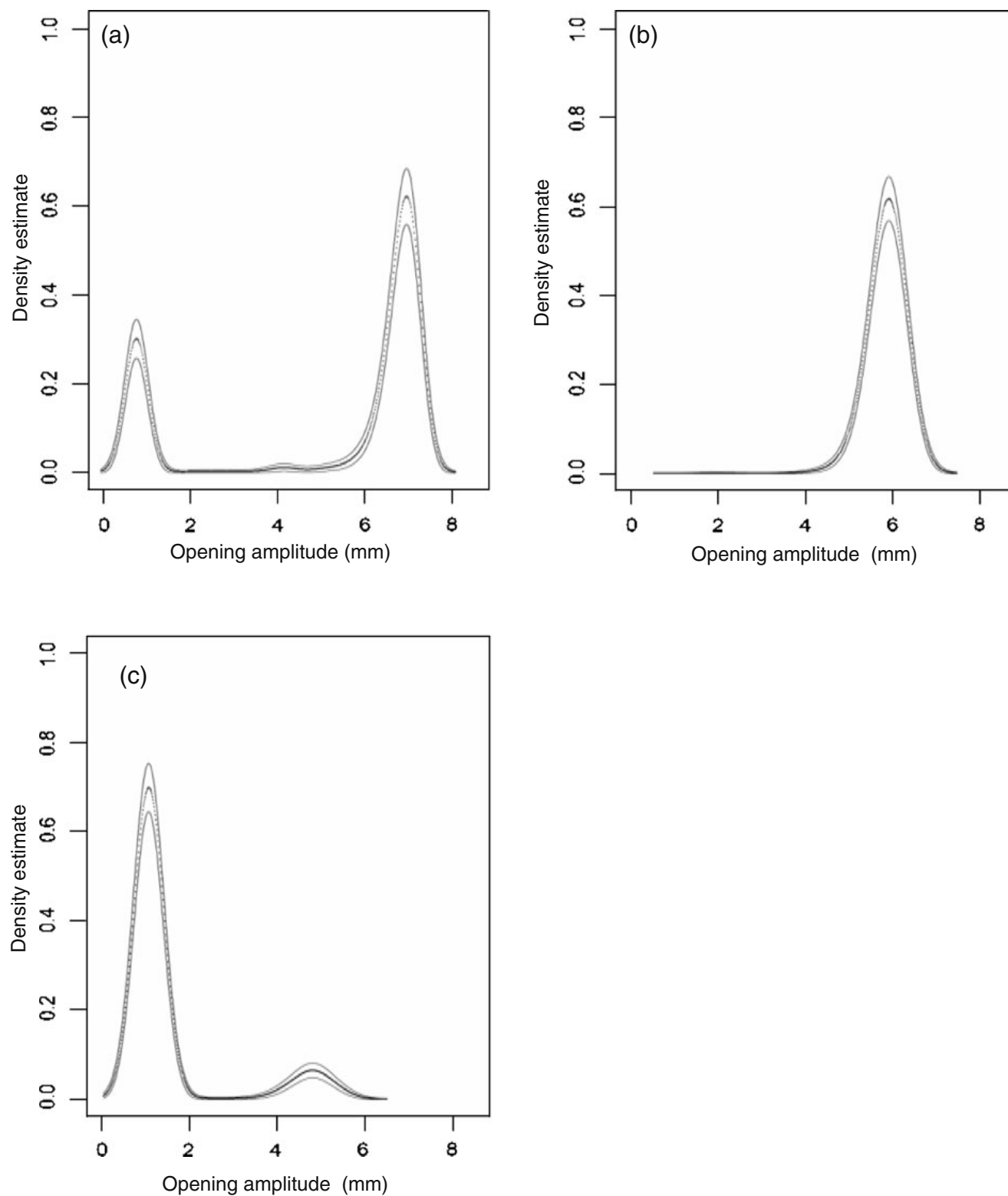


Fig. 2 a–c Plot of kernel density estimator of opening amplitude (*dotted lines*) and 95% confidence bands for the underlying distribution function (*solid line*), for three different oysters (**a**), (**b**), and (**c**)

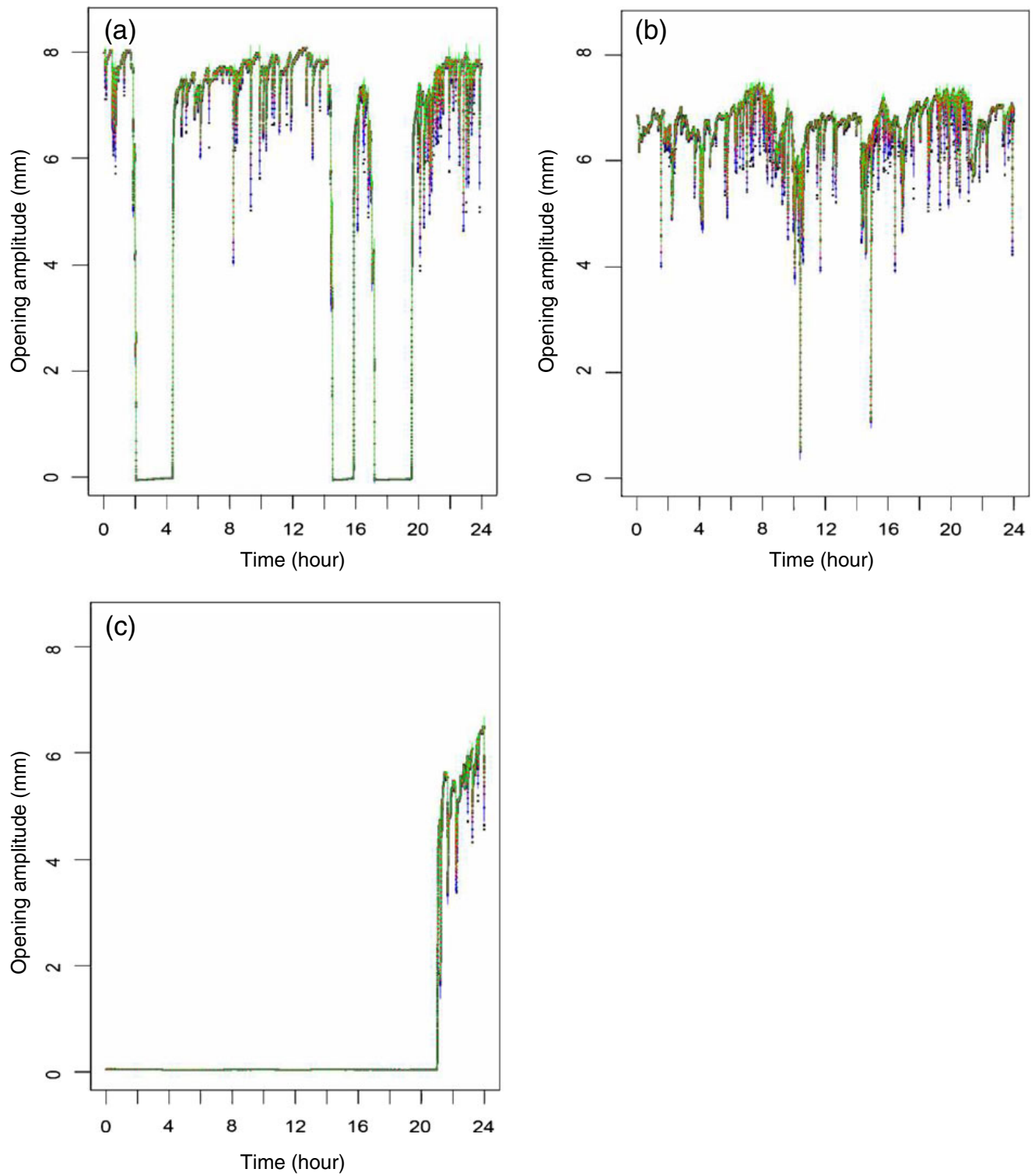


Fig. 3 a–c Relationship between the opening amplitude (millimeters) and the time of the experiment (in hour; represented by *points*), estimator of link regression estimator

(*dashed lines*), and 95% confidence interval of the link regression (*solid line*) for three different oysters (**a**), (**b**), and (**c**)

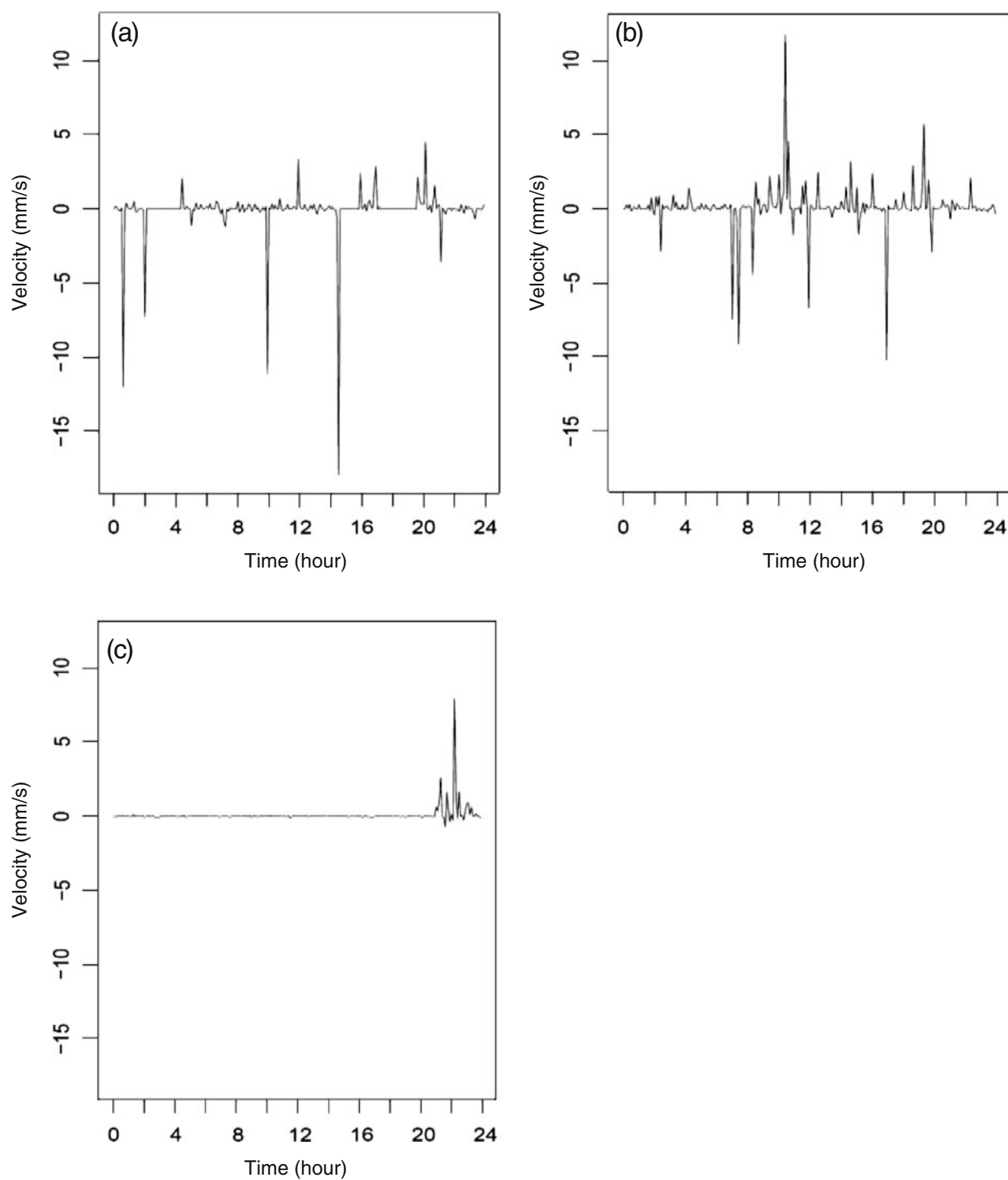


Fig. 4 a–c Estimator of the opening and closing velocity (millimeters per second) according to time (in hour) for three different oysters (a), (b), and (c)

Table 1 Example of estimated parameters for three different oysters (a), (b), and (c)

Oysters	Closing (h)	Opening (h)	Duration (h)	Sunrise (h)	Sunset (h)	Day/night	Tide amplitude (cm)	Closing velocity (mm/s)	Opening velocity (mm/s)
a	2.034	4.382	2.347	6.650	19.328	Night	424	7.29	2.07
a	14.522	15.869	1.346	6.650	19.328	Day	424	17.96	2.4
a	17.191	19.565	2.373	6.650	19.328	Day/night	424	0.011	2.15
b	10.388	10.403	0.014	6.650	19.328	Day	424	0.36	11.77
b	14.915	14.924	0.009	6.650	19.328	Day	424	0.24	1.44
c	0.001	21.026	21.026	6.650	19.328	Day/night	424	0.071	7.93

represent the periods of different time intervals of each oyster's closing states (solid lines). The advantage of this representation is to give the time duration and number of closing and opening states for each of the 16 oysters for a given time period in one single graph. We also notice in Figs. 5c, d and 6 (April to August) that the closing and opening time periods do not occur randomly. In particular, this is very clear in Fig. 5d which depicts the change of the tide levels beginning at noon every day. More specifically, these figures show that the closing activity is highly correlated with the tidal amplitude and that the closing state

is synchronized with the low tide period. Based on these observations, we proposed a statistical method to evaluate the relationship between the oysters' closing events and the tide amplitude. This is the first time that the bio-rhythms of the permanently immersed oysters is studied.

Correlation between tide amplitude and valve closing states

We determined the moment of the day where the oysters close their valves and correlated this event with the tide parameters in our entire study

Fig. 5 Principle of biological rhythm related to tide (May 2006): **a** closing periods of three different oysters (filled and open bars denote respectively the closing and opening periods); **b** superposing of the three oysters closing/opening periods for 2 days (May 16 and 17); **c** superposing of the oyster closing/opening periods for 31 days (for each day, there are 16 lines which represent the periods of different time intervals of the closing/opening states for the 16 oysters); **d** representation of the tide levels per hour. The values of the x-axis represent the hours per day

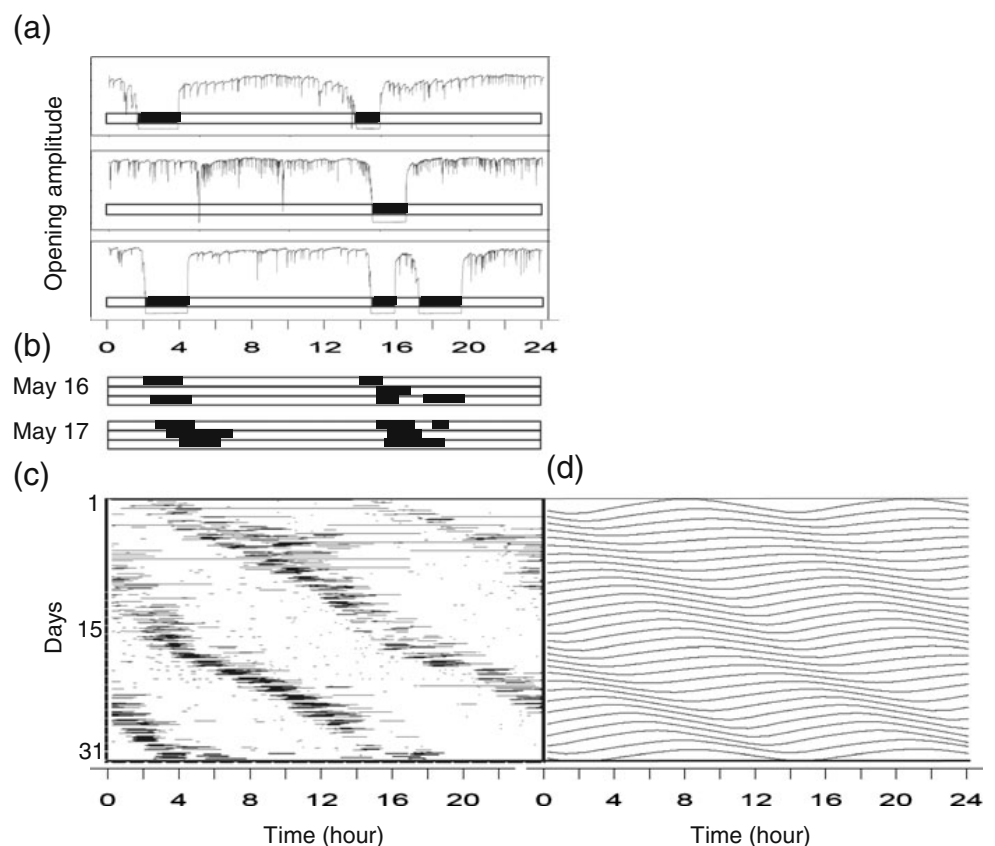
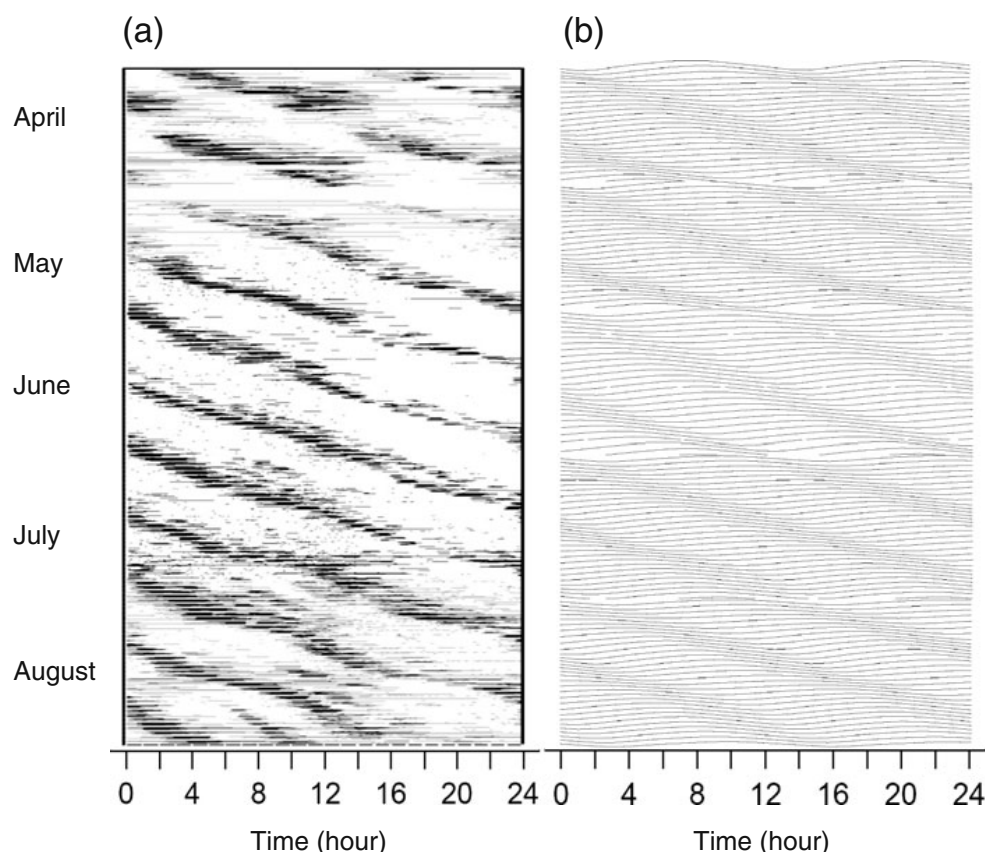


Fig. 6 Biological rhythm related to tide:

a superposing of the oyster closing/opening periods from April 1 to August 31 following the same graph principle as in Fig. 5 (filled and open bars denote respectively the closing and opening periods), **b** the tide levels per hour from April to August 2006



period. For any particular day, there are 16 broken lines representing the time intervals of the closing states for the 16 oysters. Figure 5b represents for example this information for 3 oysters and during a 2-day period. Figure 5c compiles this information for the 16 oysters for a 1-month period. The boundaries of the closing state interval were calculated using a threshold value that takes into account the distance between the two valves. We denote the estimated vector of threshold values for the 16 oysters by

$$\hat{s} = (\hat{s}_1, \dots, \hat{s}_{16})$$

where for $i = 1, \dots, 16$

$$\hat{s}_i = \min(\hat{Y}_i) + (\max(\hat{Y}_i) - \min(\hat{Y}_i)) \times 5\%. \quad (19)$$

Using the tide data (<http://www.shom.fr/>), we defined the interval C as the time period between two high tides occurring every lunar day. Using the vector of thresholds, we estimated for each day k , the vector $\hat{\tau}$ corresponding to the time at which the first closing event occur for an oyster i in the interval C :

$$\hat{\tau}^k = (\hat{\tau}_1^k, \dots, \hat{\tau}_i^k, \dots, \hat{\tau}_{16}^k)$$

where for $i = 1, \dots, 16$

$$(\hat{\tau}_i^k = \min \{t_j \in C^k | \hat{m}_{h_n}^k(t_j) \leq s_i^k\}. \quad (20)$$

Then, we estimated the mean of each oyster's time at the first closing period in the interval C denoted by $(\bar{\tau}_i^k)$ and the standard error for the 16 oysters. In Fig. 7, we plotted $(\bar{\tau}_i^k)$ as a function of time of the low tide for each day of the study period. We found evidence for a very significant linear relationship (least square estimators of the slope and the intercept being respectively equal to 0.995 and -0.36 ; $R^2 = 0.94$, P value < 0.001). So, we have showed that on average, the oyster's first closing hour is significantly correlated with the time of low tide.

Detection of unusual behaviors

Animals' reactions to changes in their environment can be very variable. One way to assess unusual oysters' activity due to environmental changes is to estimate the proportion of oysters which close their valves at a moment of the day where they usually are open. For example, the

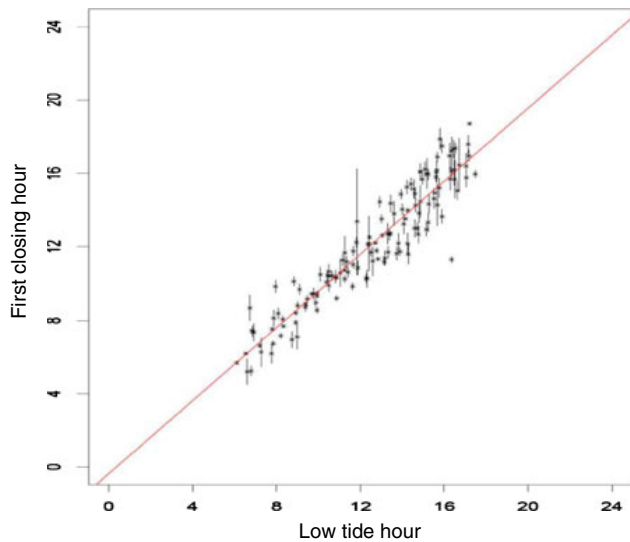
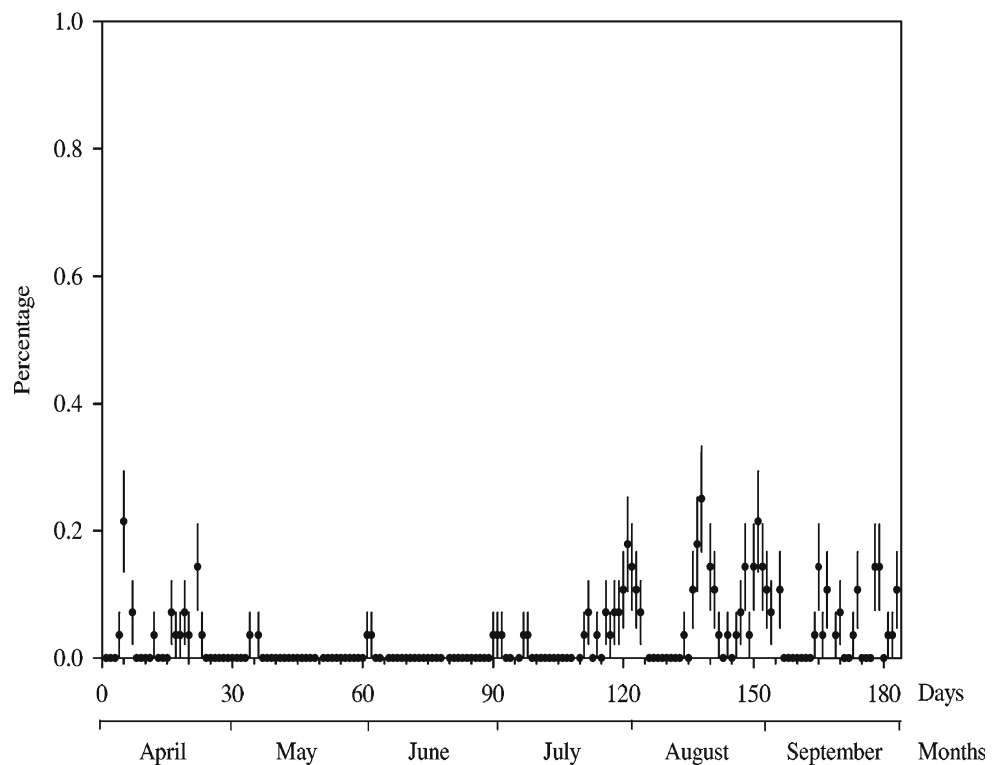


Fig. 7 Relationship between the time of the low tide and the average of first closing period (16 oysters) for each day (April and September). Error bars denote the standard errors from the mean denoted by asterisk

proportion of oysters closing their valves during the high tide period ± 1 h should be very small. This proportion estimated for our study period was 0.064 ± 0.0068 (mean ± 1 SE) whereas in the low tide period ± 1 h, this proportion was equal

to 0.45 ± 0.015 (mean ± 1 SE), SE denoting the standard error. Figure 8 shows some results for the proportion of oysters closing their valves during the high tide period over our study period. We can easily detect the days when this rhythm was the most modified: April 5, 2006 ($p = 41\%$), July 30, 2006 ($p = 36\%$), August 16, 2006 ($p = 51\%$), and August 29, 2006 ($p = 42\%$). To better characterize the animals' behaviors during these periods, we defined two time periods, one corresponding to a reference activity (46 days between June 1 and July 15, 2006) and another one corresponding to the modified activity during the consumption ban (i.e., 46 days between August 1 and September 15, 2006; Ifremer 2006). The proportion of oysters closing their valves during the high tide interval ± 1 was significantly different between the two periods (Mann and Whitney U test, P value = $5.508E - 06$). An estimation of the probability that this proportion be superior or inferior to a fixed threshold, for example 25%, was 0.048 and 0.29 for the normal and perturbed period, respectively. It suggests that changes in the bio-rhythms might be associated with field perturbations. The reasons for these changes remain to be determined although there is some

Fig. 8 Percentage of oysters that closed their valves in the time interval high tide ± 1 h for each day (April to September 2006)



evidence that the oyster *C. gigas* can modified its normal behavior in presence of the harmful algae *Alexandrium minutum* (Tran et al. 2010). Similar results were also provided by Basti et al. (2009) in the manila clam *Ruditapes philippinarum*.

Checking the model behavior

We numerically checked the asymptotic distribution of $m(x)$ and its first derivative. We selected one oyster from our sample and chose two time points randomly between 0 and 24 h over a period of $N = 450$ days. Figure 9 shows that the distribution of the bias of $\hat{m}_{h_n}$ is normally distributed and centered around 0. The mean of the bias is 0.001 with a confidence interval of $[-0.01, 0.014]$, we employed a Mann and Whitney U test to compare the mean of the bias to 0, and the test yielded a

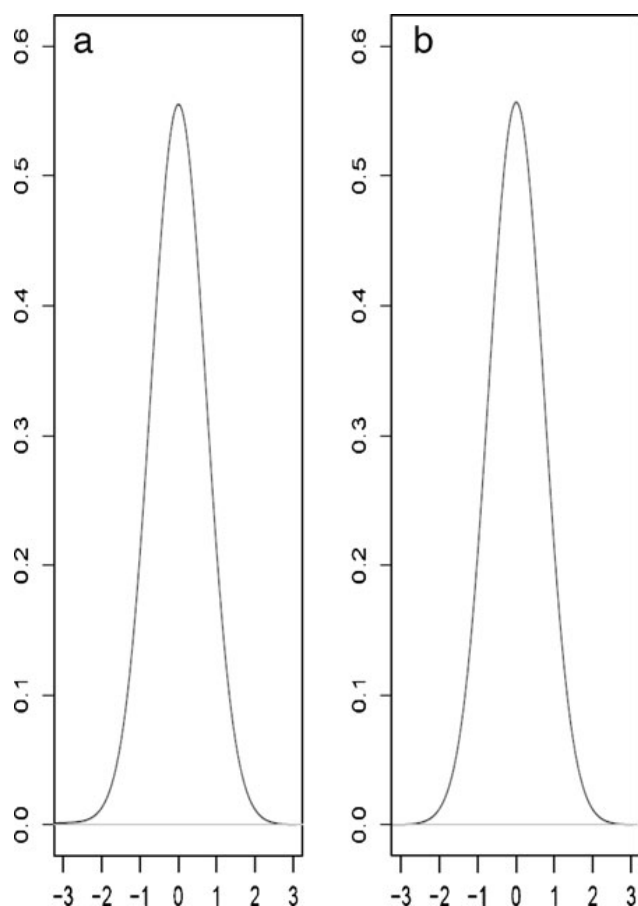


Fig. 9 Convergence in distribution of estimator of the fixed design regression estimator on $N = 450$ days at 1.5 h (a) and 10 h (b)

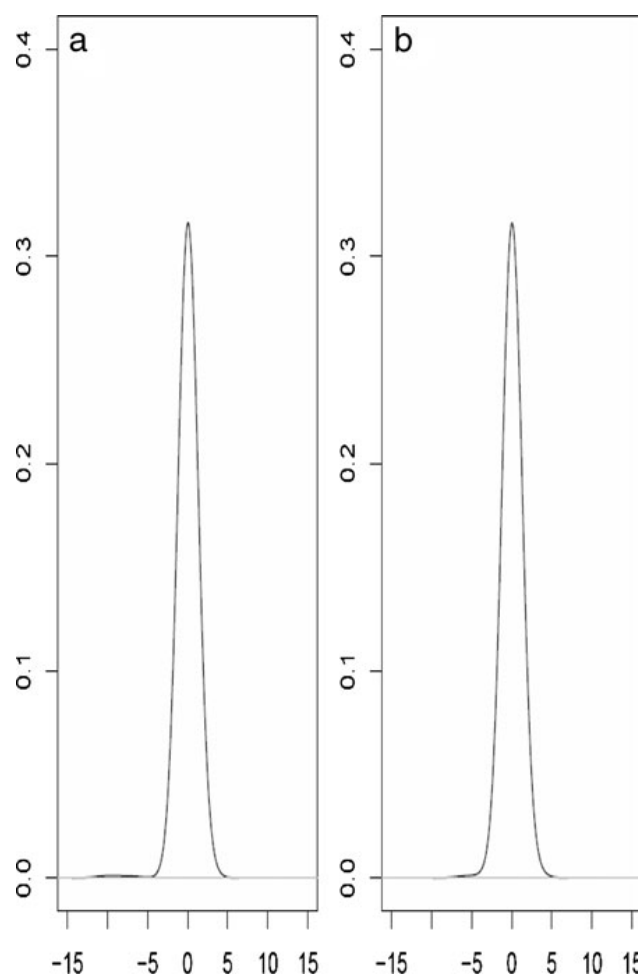


Fig. 10 Convergence in distribution of fixed design derivative estimator on $N = 450$ days at 1.5 h (a) and 10 h (b)

P value = 0.887. Figure 10 also shows that the first derivative of $\hat{m}_{h_n}$ is normally distributed.

Conclusion

This work demonstrates the possibility to develop a general mathematical and statistical approach for analyzing high-frequency serial valvometry data in the context of an environmental study. The main interest of this work is to provide efficient analytical tools to synthesizing and analyzing the bivalves' behaviors directly exposed to their natural environment. We present an application to a group of 16 oysters living in the Bay of Arcachon, studied in 2006. Our statistical modeling based on density estimation and nonparametric kernel regression allows a significant reduction

in the calculations needed for the estimation of the oysters' activity over a long period of time and provides a very informative summary from a huge flow of data produced at high frequency. This methodology also enables us to make some statistical inference and correlate the serial valvometry data with serial environmental data in order to explain unusual behaviors of the animals studied. For example, we can identify several days in our study period at which the oysters exhibited abnormal activities, corresponding to periods of oyster banning from human consumption. We think that our statistical approach provides a very useful tool for measuring the oysters' behaviors and also assessing changes in these behaviors due to perturbations in their environment. Therefore, we anticipate that this approach could have an important contribution to the survey of aquatic systems.

Acknowledgements M. SOW was supported by a grant from the Aquitaine Region and University of Toronto. All experiments presented in this article complied with the current laws of France, where they were performed. The authors thank Lindsay Hiscop for the improvement of the narrative style of the manuscript.

References

- Basti, L., Nagai, K., Shimasaki, Y., Oshima, Y., Honjo, T., & Segawa, S. (2009). Effects of the toxic dinoflagellate *Heterocapsa circularisquama* on the valve movement behaviour of the Manila clam *Ruditapes philippinarum*. *Aquaculture*, 291, 41–47.
- Borcherding, J. (1992). Another early warning system for the detection of toxic discharges in the aquatic environment based on valve movements of the freshwater mussel *Dreissena polymorpha*. In D. Neumann & H. A. Jenner (Eds.), *The zebra mussel Dreissena polymorpha: Ecology, biological monitoring and first applications in the water quality management* (pp. 127–146). New York: Gustav Fischer.
- Briollais, L., Durrieu, G., & Upathilake, R. (2007). Novel approach for genome scan meta-analysis of rheumatoid arthritis: A kernel-based estimation procedure. *BMC Genetics*, 1, 1–6.
- Byrne, R., Gnaiger, E., McMahon, R., & Dietz, T. (1990). Behavioral and metabolic responses to emersion and subsequent reimmersion in the freshwater bivalve *Corbicula fluminea*. *The Biological Bulletin*, 178, 251–259.
- Chambon, C., Legeay, A., Durrieu, G., Gonzalez, P., Ciret, P., & Massabuau, J. C. (2007). Influence of the parasite worm *Polydora* sp. on the behaviour of the oyster *Crassostrea gigas*: A study of the respiratory impact and associated oxidative stress. *Marine Biology*, 152, 329–338.
- Curtis, T. M., Williamson, R., & Depledge, M. H. (2000). Simultaneous, long-term monitoring of valve and cardiac activity in the blue mussel *Mytilus edulis* exposed to copper. *Marine Biology*, 136, 837–846.
- Doherty, F.G., Cherry, D.S., & Cairns, J. (1987). Valve closure responses of the Asiatic clam *Corbicula fluminea* exposed to cadmium and zinc. *Hydrobiologia*, 153, 159–167.
- Durrieu, G., & Briollais, L. (2009). Sequential design for microarray experiments. *Journal of the American Statistical Association*, 486, 650–660.
- Einmahl, U., & Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel type function estimators. *Journal of Theoretical Probability*, 13, 1–37.
- Fan J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.
- Fournier, E., Tran, D., Denison, F., Massabuau, J. C., & Garnier-Laplace J. (2004). Valve closure response to uranium exposure for a freshwater bivalve *Corbicula fluminea*: Quantification of the influence of pH. *Environmental Toxicology and Chemistry*, 23, 1108–1111.
- Garcia-March, J. R., Sanchis Solsona, M. A., & Garcia-Carrascosa, A. M. (2008). Shell gaping behavior of *Pinna nobilis* L., 1758: Circadian and circalunar rhythms revealed by in situ monitoring. *Marine Biology*, 153, 689–698.
- Härdle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.
- Härdle, W. (1991). *Smoothing techniques with implementation in S*. New York: Springer.
- Härdle, W., & Marron, J. S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika*, 72, 481–484.
- Ifremer (2006). *Bilan de la crise phycotoxines sur le Bassin d'Arcachon en 2006* (35 pp.). Direction des opérations/Laboratoires Environnement Ressources d'Arcachon.
- Jenner, H. A., Noppert, F., & Sikking, T. (1989). A new system for the detection of valve movement response of bivalves. *Kema Scientific and Technical Reports*, 7(2), 91–98.
- Jou, L. J., Chen, W. Y., & Liao, C. M. (2009). Online detection of waterborne bioavailable copper by valve daily rhythm in freshwater clam *Corbicula fluminea*. *Environmental Monitoring Assess.* doi:10.1007/s10661-008-0433-0.
- Kadar, E., Salanki, J., Jugdaohsingh, R., Powell, J. J., McCrohan, C. R., & White, K. N. (2001). Avoidance responses to aluminium in the freshwater bivalve *Anodonta cygnea*. *Aquatic Toxicology*, 55, 137–148.
- Kramer, K. J. M., & Botterweg, J. (1991). *Aquatic biological early warning systems: An overview*. In D. W. Jeffrey & B. Madden (Eds.), *Bioindicators and environmental management* (pp. 95–126). London: Academic.
- Kramer, K. J. M., Jenner, H. A., & De Zwart, D. (1989). The valve movement response of mussels: A tool in

- biological monitoring. *Hydrobiologia*, 188(189), 433–443.
- Kröger, S., & Law R. J. (2005). Sensing the sea. *TRENDS in Biotechnology*, 23, 250–256.
- Liao, C. M., Jou, L. J., & Chen, B. C. (2005). Risk-based approach to appraise valve closure in the clam *Corbicula fluminea* in response to waterborne metals. *Environmental Pollution*, 135, 41–52.
- Liao, C. M., Lin, C. M., Jou, L. J., & Chiang, K. C. (2006). Linking valve closure behavior and sodium transport mechanism in freshwater clam *Corbicula fluminea* in response to copper. *Environmental Pollution*, 147, 656–667.
- Marceau, F. (1909). Recherches sur la morphologie, et l'histologie, et la physiologie compars des muscles adducteurs des mollusques acphales. *Archives de Zoologie expérimentale et générale*, 2, 295–469.
- Massabuau, J. C. (2001). From a low blood- to low tissue-oxygenation strategy, an evolutionary theory. *Respiratory Physiology and Neurobiology*, 128, 249–262.
- Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Application*, 9, 141–142.
- Nagai, K., Honjo, T., Go, J., Yamashita, H., & Oh, S. J. (2006). Detecting the shellfish killer *Heterocapsa circularisquama* (Dinophyceae) by measuring the bivalve valve activity with Hall element sensor. *Aquaculture*, 255, 395–401.
- Quian, W., & Mammitzsch, V. (2000). Strong uniform convergence for the estimator of the regression function under Φ mixing conditions. *Metrika*, 52, 45–61.
- Riisgard, H. U., Lassen, J., & Kittner, C. (2006). Valve-gape response times in mussels (*Mytilus edulis*): Effects of laboratory preceding feeding conditions and in situ tidally induced variation in phytoplankton biomass. *The Journal of Shellfish Research*, 25, 901–911.
- Robson, A., Wilson, R., & Garcia de Leaniz, C. (2007). Mussels flexing their muscles: A new method for quantifying bivalve behavior. *Marine Biology*, 151, 1195–1204.
- Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B*, 53, 683–690.
- Shumway, S. E., & Cucci, T. L. (1987). The effects of the toxic dinoflagellate *Protogonyaulax tamarensis* on the feeding and behavior of bivalve mollusks. *Aquatic Toxicology*, 10, 9–27.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.
- Sloff, W., De Zwart, D., & Marquenie, J. M. (1983). Detection limits of a biological monitoring system for chemical water pollution based on mussel activity. *Bulletin of Environmental Contamination and Toxicology*, 30(4), 400–405.
- Sluys, H., Van Hoof, F., Cornet, A., & Paulussen, J. (1996). A dynamic new alarm system for use in biological early warning systems. *Environmental Toxicology and Chemistry*, 15, 1317–1323.
- Tran, D., Boudou, A., & Massabuau, J. C. (2000). Mechanism of oxygen consumption maintenance under varying levels of oxygenation in the freshwater clam *Corbicula fluminea*. *Canadian Journal of Zoology*, 78, 2027–2036.
- Tran, D., Boudou, A., & Massabuau, J. C. (2001). How water oxygenation level influences cadmium accumulation pattern in the Asiatic clam *Corbicula fluminea*: A laboratory and field study. *Environmental Toxicology and Chemistry*, 20(9), 2073–2080.
- Tran, D., Ciret, P., Ciutat, A., Durrieu, G., & Massabuau, J. C. (2003). Estimation of potential and limits of bivalve closure response to detect contaminants: application to cadmium. *Environmental Toxicology and Chemistry*, 22, 116–122.
- Tran, D., Fournier, E., Durrieu, G., & Massabuau, J. C. (2004). Copper detection in the Asiatic clam *Corbicula fluminea*: Optimum valve closure response. *Aquatic Toxicology*, 65(3), 317–327.
- Tran, D., Fournier, E., Durrieu, G., & Massabuau, J. C. (2007). *Corbicula fluminea* valve closure response as sensitivity threshold to inorganic mercury contamination. *Environmental Toxicology and Chemistry*, 26(7), 1545–1551.
- Tran, D., Haberkorn, H., Soudant, P., Ciret P., & Massabuau, J. C. (2010). Behavioral responses of *Crassostrea gigas* exposed to the harmful algae *Alexandrium minutum*. *Aquaculture*, 298, 338–345.
- Venables, W. N., & Ripley, B. D. (1999). *Modern applied statistics with S-PLUS*. New York: Springer.
- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, 26, 359–372.

1.5 Détection de la ponte par l'estimation de la vitesse

L'émission des gamètes est un événement biologique très bref mais aussi très important pour l'huître. Les conditions d'émission et notamment les facteurs déclenchant de la ponte restent mal connus faute de pouvoir dater précisément, à l'heure près, l'émission des gamètes. Sur les traces de Galtsoff (1938), nous avons tenté de mettre en évidence *in situ* les comportements valvaires associés à l'émission des ovules chez l'huître femelle décrits par Galtsoff (Fig. 1.4).

La mesure de la vitesse des mouvements valvaires donne des informations particulièrement importantes sur la physiologie de l'animal, la vitesse de contraction de son muscle adducteur et l'élasticité de sa charnière. Elle permet de caractériser des épisodes particuliers comme les réactions en présence de contaminants ou lors de la ponte.

En collaboration avec Ifremer Argenton (Ismael Bernard et Stephane Pouvreau) nous avons travaillé sur un outil de détection automatique de la ponte. Nous avons estimé la dérivée première par rapport au temps de l'estimateur de Nadaraya-Watson pour caractériser les vitesses de fermetures et d'ouvertures afin d'identifier les changements d'états (ouvert, fermé) et des séries de contractions rapides comme le phénomène de la ponte, qui est un événement bref. Une ponte dure en moyenne 31 min $\pm$ 8 durant lesquelles on a autour de 51 ± 9 micro-fermetures et arrive 1 fois/an. La figure 1.5, présente un exemple de ponte présentant une suite de contractions rapides caractéristiques (le 15 juillet 2007) et la figure 1.6, un index de la vitesse permettant une visualisation immédiate (le 15 juillet 2007). Notons que cet index de la vitesse a permis d'identifier en tout 35 pontes dont 22 sur le bassin d'Arcachon (5 en 2006, 8 en 2007, 2 en 2008, et 7 en 2010) et 13 sur le bassin de Marennes-Oléron (6 en 2008 et 7 en 2009).

La fréquence d'échantillonnage actuelle est de 10 Hz sur un groupe de 16 animaux soit 0.6 Hz sur un animal. Le stage de Jon Haritschelhar (master 1) a permis de mettre en évidence les limites de l'appareillage et du modèle mathématique utilisés. Pour cela, il a comparé en laboratoire des mesures valvométriques électroniques à des mesures faites grâce à l'imagerie vidéo. Il a montré, et caractérisé la nécessité de diminuer le pas d'échantillonnage à 10 ms pour le calcul des vitesses maximales de fermetures-ouvertures valvaire. On voit malgré tout qu'ici, malgré une fréquence d'échantillonnage de 1/1.6 sec, grâce à la modélisation et au calcul de dérivée fait sur l'équation caractérisant chaque animal on détecte remarquablement bien l'événement de ponte.

PHYSIOLOGY OF REPRODUCTION OF OSTREA VIRGINICA

I. SPAWNING REACTIONS OF THE FEMALE AND MALE

PAUL S. GALTSOFF

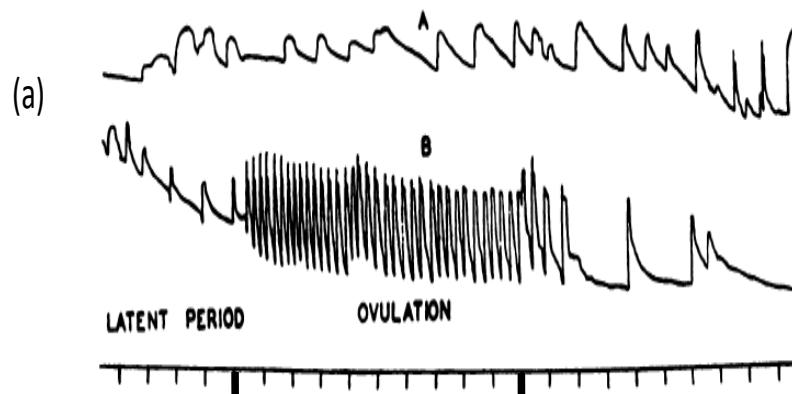


FIG. 10. *Ostrea commercialis*, spawning reaction of a female. A, muscular behavior during feeding. B, spawning reaction. Time interval, 1 minute (bottom line).

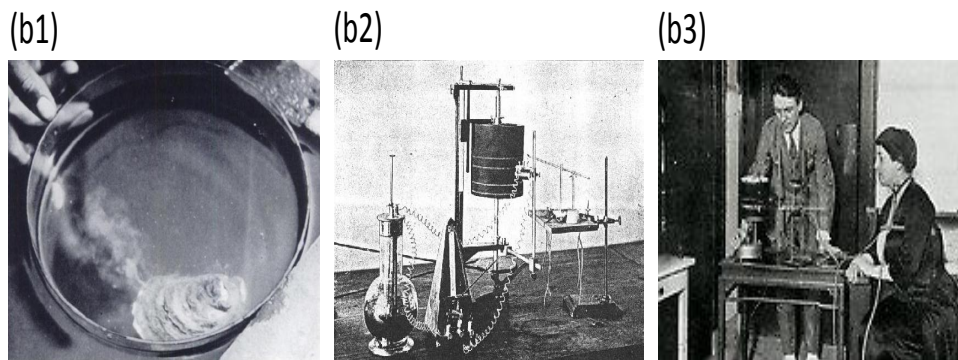


FIGURE 1.4 – Ponte chez l'huître *Ostrea virginica*. a) premier enregistrement publié (Galtsoff 1938). b1) ponte d'une huître femelle. b2, b3) Enregistreur papier.

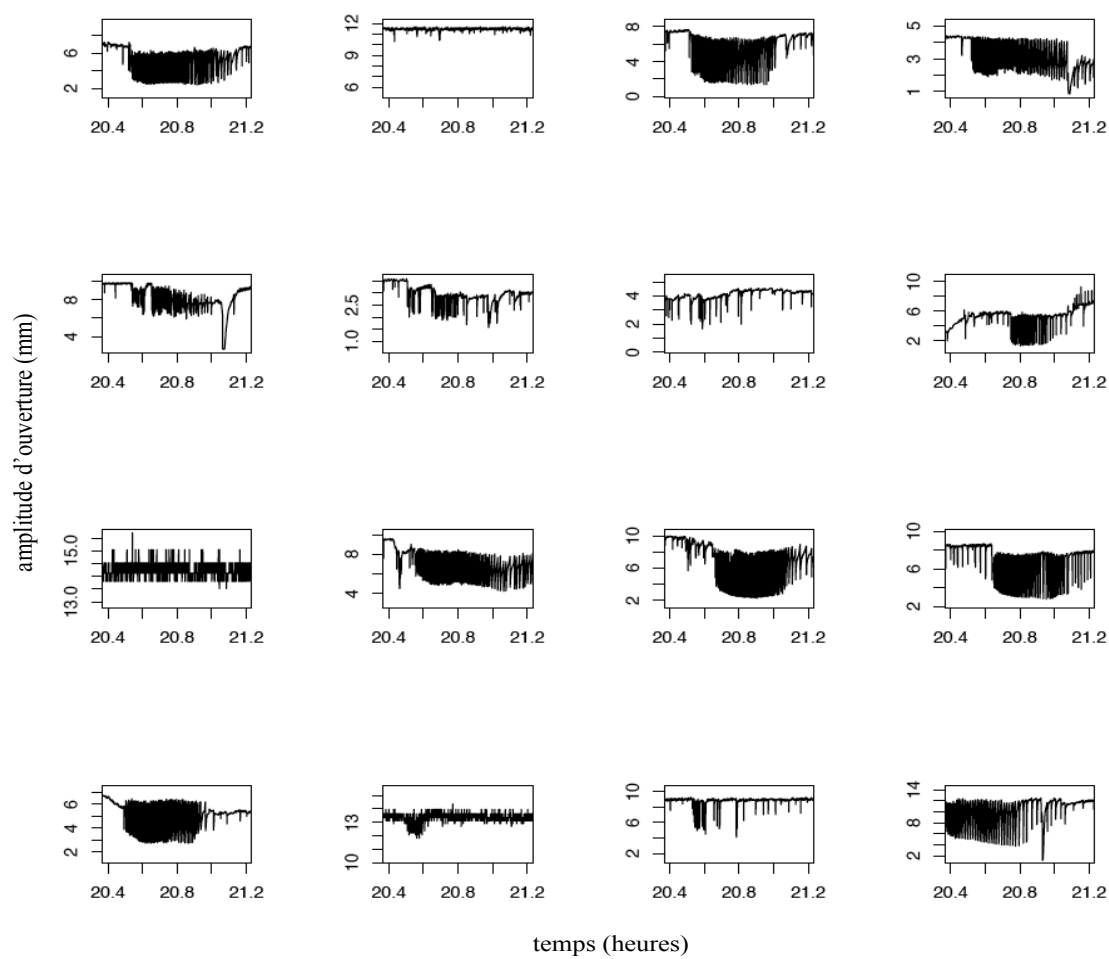


FIGURE 1.5 – Extrait de l'enregistrement des données brutes correspondant à la ponte de 8 huîtres femelles, le 15 juillet 2007, en abscisse le temps (heures décimales) et en ordonné l'amplitude d'ouverture (mm). Jetée d'Eyrac, Arcachon, France.

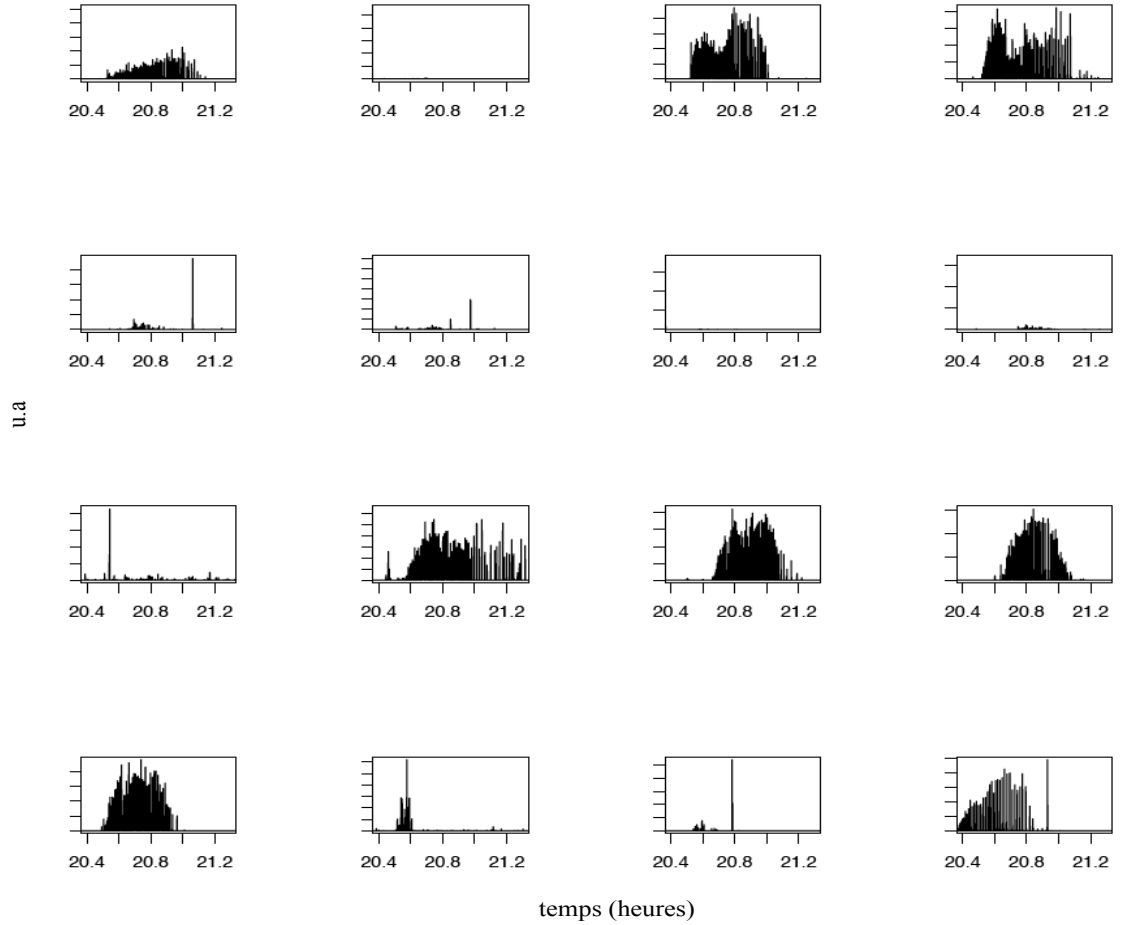


FIGURE 1.6 – Index de la vitesse correspondant aux enregistrements des données brutes, le 15 juillet 2007, en abscisse le temps (heures décimales) et en ordonné l'index de vitesse (u.a, unité arbitraire). L'index visuel est obtenu en calculant la valeur absolue à la puissance 5 de la vitesse réelle.

Le but était de détecter l'événement et d'avoir une connaissance plus précise de l'heure et de la date des pontes dans le bassin, afin de caractériser si des facteurs environnementaux particuliers synchronisent les pontes pour une meilleure efficacité du groupe en terme de reproduction. La vitesse permet aussi de caractériser des fermetures particulières (par exemple en présence de contaminant).

Afin d'améliorer la précision, tout en optimisant les temps de calcul, nous avons comparé l'estimateur de Nadaya-Watson avec l'estimateur de Nadara-Watson récursif (Duflo, 1997) et l'estimateur de Révész (Révész, 1977) au sens du critère de la variance asymptotique. Ce travail sera présenté dans le chapitre suivant.

Chapitre 2

Article 2 : Comparaison d'estimateurs de régression non paramétriques : application en valvométrie

Dans ce chapitre, nous comparons au sens du critère de la variance asymptotique l'estimateur classique de Nadaraya-Watson, sa version récursive, et l'estimateur récursif de Révész sur des données acquises depuis septembre 2007 au niveau du récif IORO en Nouvelle Calédonie (bénitiers *Hippopus hippopus*). Ce travail fait l'objet d'un article publié au 41^{èmes} Journées de la société Française de Statistique et a été réalisé en collaboration avec Thi Mong Ngoc Nguyen de l'Institut de Mathématiques de Bordeaux.

COMPARAISON D'ESTIMATEURS DE RÉGRESSION NON PARAMÉTRIQUES : APPLICATION EN VALVOMÉTRIE

Gilles DURRIEU², Thi Mong Ngoc NGUYEN¹ & Mohamedou SOW²

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.

2.1 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'acquisition, 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 résultats du traitement statistique sont accessibles sur le site web "L'oeil du mollusque" (http://www.domino.u-bordeaux.fr/molluscan_eye).

2.2 Modèle et estimateurs

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

$$Y_i = m(T_i) + \varepsilon_i. \quad (2.1)$$

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 :

$$\hat{f}_n(t) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right), \quad (2.2)$$

ou dans sa forme récursive :

$$\tilde{f}_n(t) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i} K\left(\frac{t - T_i}{h_i}\right). \quad (2.3)$$

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

L'estimateur **NW** s'écrit sous la forme d'une moyenne pondérée des valeurs

$(Y_1, \dots, Y_n)$. Il est donné par :

$$\hat{m}_n(t) = \begin{cases} \frac{\sum_{i=1}^n K(\frac{t-T_i}{h_n})Y_i}{\sum_{i=1}^n K(\frac{t-T_i}{h_n})} & \text{si } \sum_{i=1}^n K(\frac{t-T_i}{h_n}) \neq 0, \\ \frac{1}{n} \sum_{i=1}^n Y_i & \text{sinon.} \end{cases} \quad (2.4)$$

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

$$\tilde{m}_n(t) = \begin{cases} \frac{\sum_{i=1}^n \frac{1}{h_i} K(\frac{t-T_i}{h_i})Y_i}{\sum_{i=1}^n \frac{1}{h_i} K(\frac{t-T_i}{h_i})} & \text{si } \sum_{i=1}^n \frac{1}{h_i} K(\frac{t-T_i}{h_i}) \neq 0, \\ \frac{1}{n} \sum_{i=1}^n Y_i & \text{sinon.} \end{cases} \quad (2.5)$$

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

$$\check{m}_n(t) = \check{m}_{n-1}(t) + \frac{1}{nh_n} K(\frac{t-T_n}{h_n})(Y_n - \check{m}_{n-1}(t)). \quad (2.6)$$

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.

2.3 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 $\sigma^2(t) = \text{var}(Y | T = t)$. 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 2 (NW). Sous les conditions C1–C4 et pour tout $\alpha \in [1/5, 1]$, à chaque point de continuité de $\sigma^2(t)$ et pour tout $t \in R$ tel que $f(t) > 0$, nous avons quand $n \rightarrow \infty$:

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

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

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 (2.2) pour **NW** et par (2.3) pour **NWR** et **R**. La variance conditionnelle $\sigma^2(t)$ est estimée respectivement pour **NW**, **NWR** et **R** par :

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

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

2.4 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) .

2.5 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).

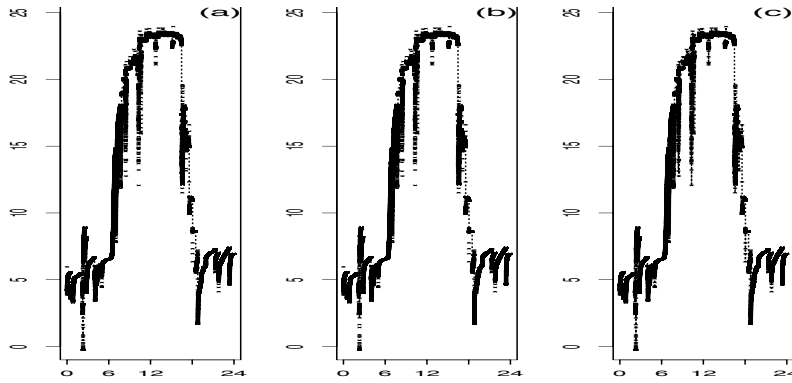


FIGURE 2.1 – Représentation des estimateurs (traits en pointillés) **NW** (a), **NWR** (b) et **R** (c) sur les données brutes du bénitier 7 avec en abscisse le temps (h) et en ordonnée l'écartement valvaire (mm).

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 $\text{var}(\mathbf{R}) < \text{var}(\mathbf{NWR}) < \text{var}(\mathbf{NW})$ ($p < 0.05$). 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.

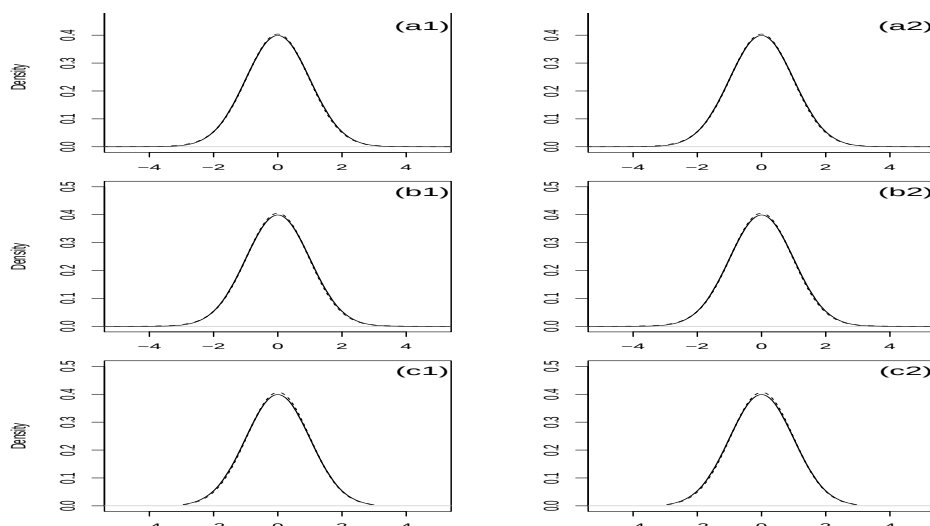


FIGURE 2.2 – Illustration de la convergence en distribution sur $N = 475$ jours : (a1, b1, c1) et (a2, b2, c2) correspondent respectivement aux estimateurs de la densité de **NW**, **NWR** et **R** ($h = 0.96$) pour le temps 1 (5 h) et le temps 2 (23 h).

Bibliographie

- [1] Duflo, Marie. (1997) Random Iterative Models. *Collection mathématiques et applications*, Springer, 385 pages.
- [2] Härdle, Wolfgang. (1990) Applied nonparametric 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 nonparametric 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.

Chapitre 3

Quantiles de régression : application à l'estimation de la croissance

Dans ce chapitre, nous abordons la croissance des bécards en utilisant l'estimation non paramétrique de type quantiles conditionnels.

3.1 Introduction

Les mollusques bivalves ont comme leur nom l'indique, une coquille constituée de deux valves qui sont reliées entre elles au niveau de la charnière grâce à un ligament et qui peuvent s'ouvrir et se fermer grâce à la contraction d'un ou deux muscles attachés aux deux valves (muscles adducteurs). La fabrication de la coquille, tout au long de la vie du mollusque, est rythmée par le métabolisme de l'animal et la température mais aussi par les paramètres environnementaux qui agissent sur "l'horloge biologique" des organismes. La croissance de la coquille se fait essentiellement de façon incrémentale, par ajouts journaliers successifs de couches de croissance à l'intérieur de la coquille. L'étude de ces incréments de croissance est susceptible d'apporter des informations sur les variations des conditions environnementales de vie des mollusques tout en permettant de dater les événements dans la coquille. La coquille des mollusques s'agrandit presque tous les jours lorsque l'animal n'est pas perturbé, ni par son environnement (température de l'eau trop basse par exemple comme chez la coquille St Jacques où elle s'arrête en dessous de 7°C), ni par ses propres rythmes biologiques.

Jusqu'à présent il n'existait qu'une méthode, la sclérochronologie (analogue à la dendrochronologie pour dater l'âge d'un arbre), pour quantifier cette croissance par analyse en microscopie de tranches de valves après sacrifice de l'animal. Nous avons mis au point, à partir des données de la valvométrie HFNI une technique pour faire la mesure en ligne d'un index de croissance. Au cours de notre travail, les mesures de croissance par sclérochronologie ont été effectuées par les collègues de l'IRD de Bondy sur 5 animaux d'un lot de bécards sur lequel nous avons fait nos mesures par valvométrie HFNI. Nous avons ensuite comparé les résultats

obtenus par les 2 techniques, la valvométrie (HFNI) avec la sclérochronologie (i.e. estimation du temps à partir des marques de croissance enregistrées et conservées dans les tissus durs, i.e. les valves). En ce qui concerne la sclérochronologie, le terme a d'abord été appliqué à une étude radiographique sur des squelettes des coraux par Buddemier et al. (1974) et Hudson et al. (1976), et sa signification a été récemment examinée par Gröcke et Gillikin (2008). L'objectif principal de la sclérochronologie est la caractérisation des modèles de croissance en fonction de contraintes de temps, des facteurs environnementaux, et les rythmes biologiques.

Pour obtenir l'historique de croissance à l'échelle quotidienne, nous avons utilisé la distance minimale entre les électrodes et le fait que chez les bivalves, la calcification se déroule dans la cavité du manteau. Si une couche de croissance est produite par jour, une conséquence est que, lorsque les valves se ferment tous les jours, la distance minimale entre les électrodes augmente également tous les jours. Pour obtenir un indice de croissance en ligne, nous avons représenté les valeurs minimum en fonction des jours (Fig. 3.1). Nous observons sur cette figure que l'évolution du minimum a une tendance linéaire et qu'il arrive des jours où l'animal ne se ferme pas complètement. On comprend donc immédiatement les limites d'une approche statistique classique et l'intérêt de développer un estimateur robuste qui passe par le bas du nuage des points.

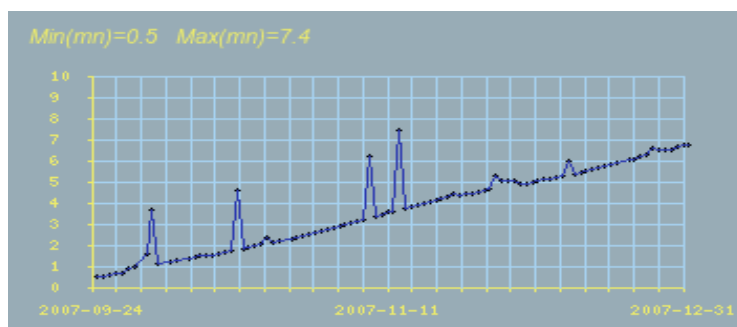


FIGURE 3.1 – Evolution du minimum d'ouverture par jour : bénitier numéro 7 du 24 septembre au 31 décembre 2007.

3.2 Modèle et estimateurs

Pour étudier la croissance des bivalves dans leur environnement naturel, nous avons proposé une méthode robuste basée sur les quantiles de régression (Koenker et Basset, 1978 ; Koenker et D'Orey, 1987 ; Dodge et Jurečková, 1995). Notre travail a d'abord été appliqué au cas des bénitiers en Nouvelle Calédonie.

Nous considérons le modèle de régression non paramétrique :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3.1)$$

où

- $\mathbf{Y} = (Y_1, \dots, Y_n)'$ est le vecteur des observations (ici la croissance mesurée en mm),
- $\mathbf{X}$ est une matrice connue de conditions expérimentales $n \times p$ ayant pour lignes $\mathbf{x}'_i$,
- $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \dots, \varepsilon_n)'$ est un vecteur d'erreurs indépendantes et identiquement distribué (i.i.d) de loi f inconnue,
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ est le vecteur inconnu des paramètres de régression à estimer.

L'application de la régression quantile dans le domaine de l'environnement et de la biologie a suscité un fort engouement ces dernières années (Terrell et al., 1996; Brown et Peet, 2003; Durrieu et Briollais 2009). Plusieurs auteurs ont étudié le comportement asymptotique des estimateurs basés sur la régression quantile sous certaines conditions de régularité (Koenker et Basset, 1978; Jurečková et Sen, 1996; Koenker, 2005). Les algorithmes de calcul ont été discutés dans Koenker et d'Orey (1987); et Narula et Wellington (1977).

L'intérêt de la régression quantile dans les domaines de l'environnement et de la génétique permet son application à des données où l'estimation d'une réponse moyenne n'est pas la plus adaptée. Se concentrer exclusivement sur les changements de la moyenne peut aboutir à sous-estimer, surestimer, ou ne pas parvenir à distinguer les changements réels dans les distributions hétérogènes (Terrell et al., 1996; Cade et al., 1999).

La régression quantile permet de modéliser de manière plus complète la distribution de la réponse, sans faire d'hypothèse sur la distribution du terme d'erreur.

Nous introduisons tout d'abord quelques notations et présentons les conditions de régularités pour la consistance et la normalité asymptotique de l'estimateur du quantile de régression.

Soit

- $\mathbf{D}_n = n^{-1} \mathbf{X}'_n \mathbf{X}_n$
- $\mathbb{I}(\mathcal{P})$ prend la valeur 1 ou 0 selon que la condition $\mathcal{P}$ est vérifiée ou non.
- θ désigne un nombre réel de l'intervalle $]0, 1[$.
- $\psi_\theta(x) = \theta - \mathbb{I}(x < 0)$ et $\rho_\theta(x) = x\psi_\theta(x)$.
- Q la *fonction quantile* de ε_n vérifiant $Q(1/2) = 0$.
- q la *fonction densité du quantile* définie par $q(\theta) = Q'(\theta)$.
- $\|\cdot\|$ dénote la norme euclidienne.

C1. F est de classe $\mathcal{C}^2$ et sa dérivée f est strictement positive sur $\{x : 0 < F(x) < 1\}$.

C2. $\lim_{n \rightarrow \infty} \mathbf{D}_n = \mathbf{D}$, matrice définie positive.

C3. $n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^4 = O(1)$ quand $n \rightarrow \infty$.

C4. $x_{i1} = 1$ pour $i = 1, \dots, n$.

Soit $(\nu_n)_{n \geq 1}$ une suite décroissante de nombres positifs (*tailles de fenêtre*). Les condi-

tions que l'on demandera à $(\nu_n)_{n \geq 1}$ appartiendront à la famille suivante :

$C5$. $n \nu_n \rightarrow \infty$ quand $n \rightarrow \infty$.

$C5'$. $n \nu_n^2 \rightarrow \infty$ quand $n \rightarrow \infty$.

$C6$. $n \nu_n^3 \rightarrow 0$ quand $n \rightarrow \infty$.

La dernière condition concerne la fonction noyau.

$C7$. k est une fonction continue à support compact telle que

$\int k(v) dv = 1$ et $\int v k(v) dv = 0$; on note aussi $K(x) = \int_{-\infty}^x k(y) dy$ et $\bar{K} = \int K^2(x) dx$.

Les quantiles de régression sont une extension des techniques de régressions linéaires aux quantiles. Plus précisément, on appelle θ -quantile de régression toute solution du problème de minimisation

$$\hat{\beta}(\theta) \equiv \hat{\beta}^n(\theta) = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\theta}(Y_i - \mathbf{x}_i' \beta). \quad (3.2)$$

Un cas particulier de cette classe d'estimateurs (obtenu pour $\theta = 1/2$) est l'estimateur L^1 qui s'obtient par résolution du problème de minimisation. La normalité asymptotique de $\hat{\beta}(\theta)$ a été établit par Koenker and Bassett (1978). Sous les conditions $C1$, $C2$ – $C4$, $C5$, $C6$ et $C7$, nous avons, quand $n \rightarrow \infty$,

$$\sqrt{n} (\hat{\beta}(\theta) - \beta(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma_{\theta}). \quad (3.3)$$

avec

$$\Sigma_{\theta} = (\theta(1 - \theta)/f^2(Q(\theta))) (\mathbf{X}' \mathbf{X})^{-1} \quad (3.4)$$

est la matrice de variance covariance asymptotique de $\hat{\beta}(\theta)$. La variance asymptotique de $\hat{\beta}^n(\theta)$ est inconnue car elle dépend de la fonction de densité du terme d'erreur dans le modèle (3.1), nous avons proposé un estimateur de cette variance en utilisant l'estimateur de la densité du quantile $q(\theta)$.

Estimateurs de $q(\theta)$:

La variance asymptotique dépendant de la densité de probabilité des erreurs (*inconnue*), nous avons besoin de “bons” estimateurs de la variance asymptotique (notée S_n) invariants par régression ($S_n(\mathbf{Y} + \mathbf{X}\beta) = S_n(\mathbf{Y})$, $\forall \mathbf{Y}, \beta$) et “invariants” par homothétie ($S_n(c\mathbf{Y}) = c S_n(\mathbf{Y})$, $\forall c > 0$). Nous allons décrire un estimateur de $q(\theta)$ dans le cas du modèle de régression (3.1) basés sur le concept de quantile de régression introduit par Koenker et Bassett (1978) qui sont invariants par régression et homothétie. Ceci est en quelque sorte une extension de l'approche de Falk (1986) pour la fonction quantile pour un échantillon $X_1, \dots, X_n$.

L'estimateur de type noyau dans le modèle de location a été considéré par Falk (1986) et Zeltermann (1990). Ces résultats ont été étendus aux modèles de régression linéaire par Dodge et Jurečková (1995) en particulier.

Cet estimateur de $q(\theta)$, $0 < \theta < 1$, est défini par

$$\hat{Z}_n(\theta) = \frac{1}{\nu_n^2} \int_0^1 \hat{\beta}_1^n(w) k\left(\frac{\theta - w}{\nu_n}\right) dw,$$

où les conditions $C5$, $C6$ sur la taille de la fenêtre sont satisfaites et k est une fonction (noyau) vérifiant $C7$. (cf. Chapitre 1 et 2).

Théorème 5 (Dodge and Jurečková, 1995). Sous les conditions $C1$, $C2$ – $C4$, $C5'$, $C6$ et $C7$, nous avons, quand $n \rightarrow \infty$,

1. $\sqrt{n\nu_n} \left(\hat{Z}_n(\theta) - q(\theta) \right) = O_p(1)$;
2. $\sqrt{n\nu_n} \left(\hat{Z}_n(\theta) - q(\theta) \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, q^2(\theta) \overline{K} \right)$.

Là aussi, cela implique la consistance de $\hat{Z}_n(\theta)$, c'est-à-dire

$$\hat{Z}_n(\theta) \xrightarrow{P} q(\theta) \text{ quand } n \rightarrow \infty.$$

Nous en déduisons pour estimateur de Σ_θ

$$\hat{\Sigma}_\theta = \theta(1 - \theta) \hat{Z}_n^2(\theta) (\mathbf{X}'\mathbf{X})^{-1}. \quad (3.5)$$

3.3 Application

Pour faire cette comparaison, nous avons sélectionné 5 bénitiers (sur une période comprise entre le 24 Septembre 2007 et le 18 Août 2008) représentatifs des croissances les plus fréquemment observées en posant deux postulats, validés par la valvométrie HFNI : (i) chaque strie correspond à un jour ; (ii) chaque strie est daté puisque le jour où les animaux ont été sortis de l'eau (19 août 2008) et le jour où ils ont été sacrifiés (20 août 2008) sont connus. Une étude sclérochronologique a donc été réalisée chez chacun de ces animaux pour être comparée aux données mesurées en valvométrie. Ainsi nous avons considéré l'estimateur (3.2) avec $\theta = 0.1$ sur les données évolution du minimum d'ouverture de l'activité valvaire. La figures 3.2 présente l'évolution de la croissance mesurée par sclérochronologie (a), l'estimée de la croissance par un estimateur de type quantile de régression (b) et la corrélation entre les deux croissances (c). On observe une corrélation très significative ($r = 0.975$, $p < 0.0001$) entre la croissance estimée par valvométrie et par la mesure directe via les stries de croissance.

Un des buts était aussi de pouvoir prédire cette tendance d'évolution du minimum d'ouverture et de pouvoir l'isoler pour éviter une dérive au niveau de la représentation des données en ligne. Pour chaque jour j , nous estimons l'évolution du minimum du jour suivant ($j+1$) par l'estimateur (3.2) avec $\theta = 0.1$ permettant ainsi de corriger le minimum du jour ($j + 1$).

Ainsi, notre outil de surveillance du milieu joint à la diffusion en ligne des résultats sur internet, nous permet de suivre la croissance des animaux et de visualiser ainsi facilement toute perturbation de cette croissance. Nous avons, par exemple, pu visualiser la mort d'un bénitier (le 6 Février à 2 h du matin, heure Nouvelle Calédonie) dont la croissance s'était arrêtée depuis un mois. On peut noter que cette analyse peut être faite avec différents quantiles conditionnels de la distribution de

la réponse. En particulier, nous avons aussi voulu tirer de l'information de l'analyse de l'évolution des états d'ouverture maximum journalier. L'idée de départ est que pour se protéger, une des réponses les plus courantes d'un bivalve est de fermer ses valves plus ou moins. L'analyse journalière des états d'ouverture et le suivi de tendances peut donc permettre de voir ce comportement particulier sur un animal ou un groupe d'animaux. Le suivi de la valeur maximale d'ouverture donne donc des informations complémentaires, particulièrement importantes, sur le bien-être de l'animal et son rapport aux variations de qualité de son environnement (état de stress).

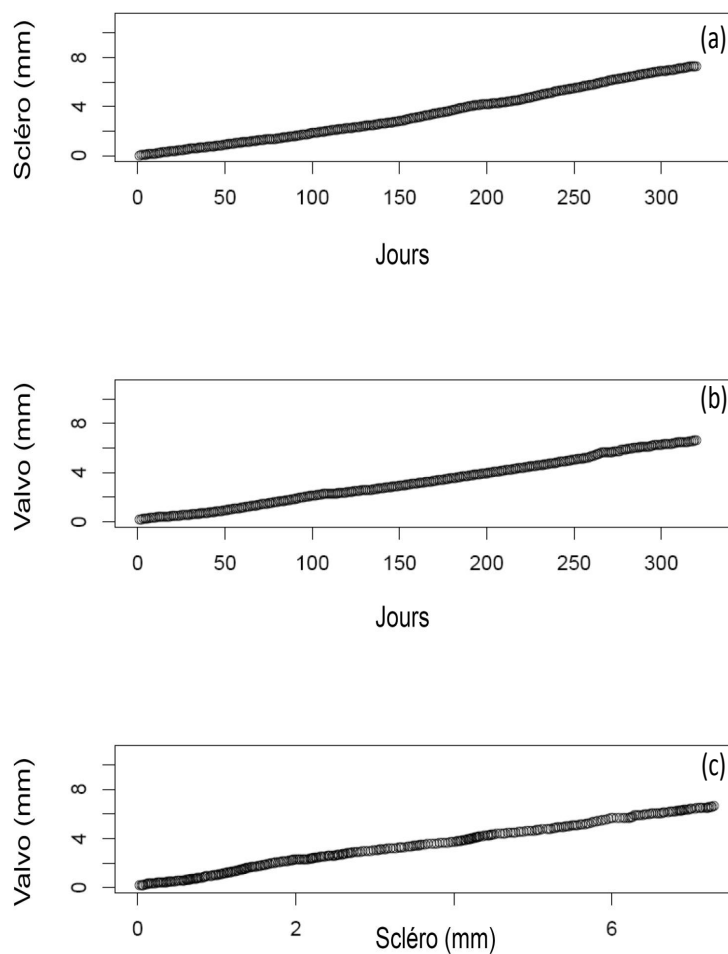


FIGURE 3.2 – (a) croissance (en mm) des bénitiers mesurée par la mesure des stries ; (b) croissance estimée par un estimateur de type quantile de régression et (c) corrélation entre les deux croissances.

- 3.4 Article 3 : In situ giant clam growth rate behavior in relation to temperature : a one year coupled study of high-frequency non-invasive valvometry and sclerochronology. Schwartzmann, C., Durrieu, G., Sow, M., Ciret, P., Lazareth, C.E., Massabuau, J.C. (2011). Limnology and Oceanography. (en révision favorable).**

In situ giant clam growth rate behavior in relation to temperature:
a one year coupled study of high-frequency non-invasive valvometry
and sclerochronology

Caroline Schwartzmann^{1, 2}, Gilles Durrieu¹, Mohamedou Sow¹, Pierre Ciret¹, Claire E.
Lazareth² and Jean-Charles Massabuau¹

1

Running head: In situ giant clam growth rate

¹ Corresponding author: E.mail: jc.massabuau@epoc.u-bordeaux1.fr

1 Station Marine, UMR 5805, CNRS and Université Bordeaux 1, Place du Dr. Peyneau, 33120, Arcachon,
France.

2 IPSL/LOCEAN, UPMC/CNRS/IRD/MNHN, Centre IRD France Nord, 32, avenue Henri Varagnat, 93143
Bondy Cedex, France

Acknowledgments

This work was granted by the Province Sud of New Caledonia and the Région Aquitaine, France. The authors thank all people in New Caledonia who helped us during the various field trips and more specifically the people from Province Sud and DENV, Phares et Balises of Nouvelle-Calédonie (Erwan Guivarch), IRD Nouméa (Fabrice Colin, Eric Folcher, Jean-Louis Menou, Jean-Michel Fernandez, Michel Menezo, Pascal Douillet, Olivier Pringault), the people from IFREMER St Vincent, Liet Chim, Dominique Pham, Nelly Wabete, Pierrette Lemaire and the people from the OPT. Special thanks to Pierre Brun and the Brun family, and Météo France New Caledonia for kindly providing environmental data.

Abstract

The life history of fifteen giant clams, *Hippopus hippopus*, was studied in situ in the southern lagoon of New Caledonia. For the first time, the growth rate and animal behavior were studied both by sclerochronology and HFNI valvometry. Electrodes glued on each valve of each specimen, recorded the shell gaping behavior at a 0.6 Hz frequency. A nonparametric regression model was used to model clam behavior. After etching, the daily increment thickness in the inner layer of 5 representative clams was measured. *H. hippopus* has its valves opened during the day and almost closed during the night all year round. Their shell growth is continuous. The cumulated growth using both techniques was not different, as was the mean daily increment thickness. The occurrence of one increment per day in *H. hippopus* shell is thus empirically proven. The 5 sclerochronological profiles were highly similar. If shell growth is significantly correlated to sea surface temperature, SST, while it is rising to reach summer values, this relationship vanishes when SST reaches 27°C in association with maximal solar irradiance. At this time, erratic gaping behavior and increment thickness changes were observed. This evidence of SST- and solar irradiance-related stress could be related to physiological oxidative stress triggered by zooxanthellae symbionts. In the present global warming context, our full data set support original evidence indicating that the giant clams *Hippopus hippopus* are already living out of their thermal comfort limits in summer in New Caledonia.

Key words: giant clam, *Hippopus hippopus*, valvometry, sclerochronology, global warming, reef ethology

Introduction

The coral reefs around the world represent a very diverse and vulnerable ecosystem and many factors are a threat to their survival. Increased water temperature and/or solar irradiance, sedimentation, salinity changes, pollution, tourism, diseases, fishing practices and high energy storms are among them. Coral bleaching is one of the most often reported problems which occurs when the fragile balance between the animal and its symbiotic algae, the zooxanthellae, collapses and this has been quite often discussed in terms of maximum sea surface temperature (SST) and solar irradiance during the austral summer (McClanahan et al. 2007; Suggett et al. 2011). Two major points of discussion are the adaptive capabilities of corals to rapid changes and the assumption that they could already live close to their thermal limits (Jokiel and Coles 1990; Maynard et al. 2008; Hoegh-Guldberg 2009). There are predictions stating that increasing seawater temperature by only a very few degrees (Celsius) above summer values could have major impacts (Brown 1997; Toscano et al. 2000; Hoegh-Guldberg 2009).

Although it is often referred to as the whitening of corals, temperature-induced bleaching has been reported in all marine invertebrates that host zooxanthellae, including giant clams (Blidberg et al. 2000; Przeslawski et al. 2008). Their mantle bleaches after exposure to high temperatures and high light intensities (Addessi 2001; Buck et al. 2002; Leggat et al. 2003). Together with corals, giant clams are said to rely heavily on zooxanthellae for the production of energy and metabolic processes. In coral, more than 50-90% of energy requirements are claimed to be provided through photosynthesis, although phyto- and zoo-plankton feeding are certainly also quite important (Goreau et al. 1971; Muscatine 1990; Grottoli et al. 2006). In the reef, we hypothesized that giant clams could be living in a situation very close to that of corals.

The aim of the present work was to get more insights into this possibility. We present an in situ study of growth rate pattern and behavior performed in giant clams *Hippopus hippopus*, living in a UNESCO World Heritage Site, a clean area, on a reef 1.7 km from the shoreline, in the southern lagoon of New Caledonia, South Pacific, during a full growth season from September 2007 to August 2008. This one year of growth history was studied by combining, for the first time, two complementary techniques: High Frequency, Non Invasive (HFNI) valvometry (measuring shell gaping activity) and sclerochronology. The HFNI valvometry is a new generation remote valvometry set-up aimed at studying online behavior of mollusk bivalves, freely living in their natural habitat. It allows autonomous long term recordings (> 1 year) of valve movements at high frequency without interfering with normal behavior. Here, we applied it for the very first time to the online recording of growth rate. Basically, the technique involves a group of 16 animals each with one lightweight electrode glued on each valve, and then measuring, at 0.6 Hz/individual, 24 h a day, 7 days a week, the distance between the valves (Tran et al. 2003; Chambon et al. 2007). Behavior patterns with daily, weekly or monthly tendencies are then mathematically constructed and published online (http://www.domino.u-bordeaux.fr/molluscan_eye; Sow et al. 2011). To obtain the growth history on a daily scale, we took advantage of the fact that in mollusk bivalves, calcification takes place in the mantle cavity: if daily growth layers are produced, a consequence is that when valves close every day, the minimal distance between electrodes also increases every day. To obtain an online growth rate index, we isolated these daily values and plotted them as a function of time. Note that the same operation can be done with the maximum daily opening value which gives complementary behavioral information on the animal welfare (stress status), as full or partial closures are a major way to protect the soft body in a molluscan bivalve. Here we present both analyses in parallel to correlate growth rate and stress status with sclerochronology in the same clams.

As regards sclerochronology, the term was first applied to a radiographic study on coral skeletons by Buddemier and Maragos (1974) and Hudson et al. (1976) and its meaning has been recently reviewed by Gröcke and Gillikin (2008). The technique can be applied to mollusk shells (postmortem examination) and allows the historical reconstruction of shell layer thickness changes, i.e. growth rate, in the temporal context in which the shell layers formed. The main focus of sclerochronology is an attempt to characterize growth patterns as a function of time and constraints by environmental cues and/or zeitgeber and biological rhythms. It is routinely applied to deduce life history traits and largely used in paleoclimatology where well known examples are daily or annual growth in reef coral and mollusk shells.

In New Caledonia, water temperature closely follows air temperature changes, with a maximum in March (mean maximum, ~ 27-28°C) and a mean minimum in July-August (~ 21°C, Le Borgne et al. 2010). We demonstrate here that by the end of the summer period, when SST reaches 27-28°C and solar irradiance is still reaching maximum annual values, the daily shell growth rate of *Hippopus hippopus* is highly variable, its mean value is decreasing and associated to strong behavioral signs of stress symptoms.

METHODS

Samples.—Sixteen *Hippopus hippopus* giant clams were placed in late August 2007 on the Ioro reef (166°57'E-22°23'S), located in the Havannah channel, southern New Caledonia (Fig. 1A, B, C). One animal died on 6 February 2008, a few days after the hurricane GENE approached New Caledonia, and was excluded from all calculations. The Ioro reef is a permanently immersed reef, with a minimum depth of 0.5-1 m at low tide, and which *H. hippopus* naturally inhabits. The animals were collected by a local fisherman in early august 2007 from the Goro coral reef (~ 10 km from the experimental site) in front of the Goro tribe area. The animals were then placed on a concrete platform (1 x 1.1 m) supporting 17 Netlon cages (1 per animal + 1 for the submerged electronics), firmly fixed on a dead *Porites* coral. The depth of the full set-up varied from 3-4 m according to the tide amplitudes which vary from ~ 1-1.8 m. The full set-up was left in place from 24 August 2007 to 18 August 2008. During that time, in situ behavior and growth rate were studied by HFNI valvometry, both by online and post analysis. After collection, a sclerochronological analysis was performed on the shell of 5 specimens. At the beginning of the recording period, the *H. hippopus* sizes were homogenous (whole body mass, 575 ± 25 g; maximum length, 133 ± 8 mm) and the animals were around 2 years old (according to post-mortem analysis).

HFNI valvometry. HFNI valvometry allows the gathering of information on animal behavior in normal in situ environmental conditions, without formal distance limits between field and laboratory, and for as far as mobile phone network and internet connections are available. We used a homemade technology that was first described and critically assessed under laboratory conditions by Tran et al. (2003). Light-weight electrodes (60 mg) were attached to the two valves of the *H. hippopus* shell (Fig. 1D). The electrodes were linked by 1 m flexible wires (2.14 g.m^{-1} , $\varnothing = 0.9 \text{ mm}$) to a waterproof case containing the 1st level electronic control unit which measured valve activity. The 2nd electronic card (developed by

the Society EUKREA Electromatic, Pessac, France) was fixed outside the water on the Ioro reef lighthouse (distance from the animals, ~ 30 m). This 2^d card used the Linux operating system to drive the 1st card, which received the measurements of the distance between electrodes that arrived every 0.1 second from the pool of the equipped animals, i.e., the initial analysis, and the mobile phone emission protocols. With 16 bivalves, the initial number when the experiment starts, information was received from each animal (live or dead) every 1.6 sec and each day a data matrix of 864 000 lines was collected. Finally, for each animal each day, 54 000 pairs of values (1 distance between electrodes plus 1 sampling time) described every single animal behavior. The total electronic consumption of 2 watts was supplied by a solar panel positioned on the Ioro lighthouse. The measurements are based on the magnetic principle: Each laboratory-made electrode encloses a coil; one is a transmitter, the other a receiver. The strength of the electric field produced between the two coils decreases with distance, according to the law $1/D$, where D is the distance between the point of measurement and the centre of the transmitting coil. On each clam, the first distance measurements were performed following electrode fixation when the animals' valves were firmly shut in the air, and these were taken as 0 values.

Principle of the mathematical analysis. To model giant clam behavior, we used a nonparametric regression model based on the kernel estimator (Härdle 1990; Tran et al. 2003; Durrieu et al. 2009). The relationship between the two electrodes (Y) and the time of the measurement (T) is modeled using the following nonparametric regression model:

$$Y_i = m(T_i) + \varepsilon_i, \quad \text{for } i=1, \dots, n \quad (1)$$

where n , m and ε denote the sample size (total number of paired values), the unknown regression function to be estimated and the model error term respectively. The stochastic distribution f of ε is typically unknown and is unlikely to follow any familiar distribution,

such as the normal distribution, and is independent of T ; hence no distribution is assumed (nonparametric statistics). The distance between the two electrodes is measured at equally spaced measurements points. So, we follow the case of fixed equidistant design that implies that the T_i 's have a uniform known design density f_T . The variation ε_i is a random variable with a mean value equal to 0 and the stochastic distribution f of these random variations enables us to characterize the variation of the random variable Y around:

$$m(t) = E(Y/T = t) = \frac{\int y f_{Y,T}(y, t) dy}{f_T(t)}. \quad (2)$$

This regression function m (definition of the conditional expected value of Y given T) depends only on the unknown joint distribution function of (Y, T) , denoted by $f_{Y,T}$, because f_T is known. Using the kernel estimator of the joint distribution function, the estimator and the statistical properties of m describing the behavior of an animal are established (Härdle 1990; Briollais et al. 2007 and Sow et al. in press). The regression estimator of m is given by

$$\hat{m}_{h_n}(t) = \begin{cases} \frac{1}{n} \frac{\sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right) Y_i}{f_T(t)} & \text{if } f_T(t) \neq 0, \\ \frac{1}{n} \sum_{i=1}^n Y_i & \text{otherwise,} \end{cases} \quad (3)$$

where h_n is the bandwidth (the smoothing parameter) satisfying (C1) h_n tends to zero and nh_n tends to infinity as n tends to infinity. The kernel K is a symmetric function satisfying (C2) $\int K(t)dt = 1$, $\int tK(t)dt = 0$, $\int |t|K(t)dt = 0$, and the distribution function is twice differentiable (C3). Selecting a proper bandwidth parameter h_n is a critical step in estimating the regression function. Although in practice one can choose the bandwidth subjectively, this can lead to inaccurate estimation. The choice of h_n is much more important for the behavior of the estimators than the choice of the kernel function K . Small values of h_n make the estimate look

uneven and suggest features which don't really exist, whereas large values of h_n will lead to an estimate which is too smooth in the sense that it is heavily biased and may not reveal important structural features, like a bimodality for example. We used a cross-validation criterion to estimate the bandwidth parameter. This bandwidth is asymptotically optimal and yields the same speed of convergence as other techniques (Härdle 1990). Once the daily activity of each animal was modeled, different parameters specifying openings and closures (duration, number, degree of opening in mm) and also the number of partial or total closures could be studied for each day of recording by using the modeling approach. The computations were performed under the Linux operating system (version Fedora 12) on a bi-processor Xeon Workstation DELL using R (R Development Core Team 2008).

Hippopus hippopus growth model using valvometry. The model allows for estimating the growth vs. time using measurements collected by valvometry. The general form of the model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is the cumulated growth vector (in mm), $\mathbf{X}$ is a known design matrix of time condition, $\boldsymbol{\beta}$ is the vector of unknown regression parameters to be estimated and $\boldsymbol{\varepsilon}$ is the error term. If $\boldsymbol{\varepsilon}$ follows a normal $N(\mathbf{0}, \mathbf{I})$ distribution, then ordinary least squares provide the maximum likelihood estimator of $\boldsymbol{\beta}$. However, it is clear that the normality assumptions are often violated in an environmental context. To avoid the distributional assumptions and to protect against outlying measurements, we proposed a robust and nonparametric inferential method by using quantile regression. An overview of robust statistics literature can be found in Huber (2003), Jurečková and Sen (1996), and Durrieu and Briollais (2009). Our approach is based on regression quantile estimators of $\boldsymbol{\beta}$ (Dodge and Jurečková 1995). Instead of focusing on the changes in the mean cumulated growth, the quantile regression approach

allows one to test whether there is a change in the θ^{th} quantile of y for any given θ in $(0,1)$.

When the conditional distributions of Y are non-Gaussian, the mean might not be the best summary, and a change in distributions may not be detected. Inference for linear quantile regression models has become a subject of intense investigation in the past years. Any solution of the following minimization problem ($\mathbf{x}_i'$ denotes the i th row of the matrix $\mathbf{X}$)

$$\hat{\beta}_{\theta} = \arg \min_{\beta \in R^p} \sum_{i=1}^n \rho_{\theta}(Y_i - \mathbf{x}_i' \beta)$$

is called the θ -regression quantile with $\rho_{\theta}(x) = x(\theta - I(x < 0))$ where $I(P)$ takes the value 1 or 0 depending on whether the condition P is satisfied or not. Quantile regression reveals multiple rates of changes (slopes) from minimum to maximum response, providing a more complete picture of the relationships between variables missed by other regression methods. Most regression applications in the ecological sciences focus on estimating rates of change to the mean of the response variable distribution. For instance, $\theta = 0.5$ is chosen as an alternative to least squares estimators in the presence of heavy-tailed distribution (median regression). It can be proved that under regularity conditions, this estimator is asymptotically normally distributed (Jurečková 1984; Durrieu and Briollais 2009). The asymptotic variance of $\hat{\beta}_{\theta}$ is unknown because it depends on the distribution function f of the error term in the model, which is itself unknown. So, we estimate this asymptotic variance using the regression invariant and scale equivariant kernel-type estimator (Durrieu and Briollais 2009).

Sclerochronology. Fifteen animals were collected alive on 18 August 2008. Soft tissues were immediately separated from the shells which were then air dried. Five animals out of the 15 were selected for sclerochronology on the basis of their representativeness of the whole group's behavior following HFNI valvometry analysis (see below and Fig. 3). One valve of

each of the 5 selected specimens was cut, using a diamond precision saw (Buehler ®) in ~ 2 mm-thick slices along the maximum growth axis in the umbo and inner layer part of the shell (Fig. 1D, E). Indeed, it is in the inner layer of *H. hippopus* that growth patterns are the most easily seen compared to other portions of the shell (Aubert et al. 2009). The slices were embedded in epoxy and polished with diamond suspensions with decreasing grain size, the last polishing step being performed with a 0.25 µm grain-size suspension. In between each suspension, the sections were carefully washed in an ultrasonic water bath.

After the polishing steps, the growth increments were not clear enough for accurate measurements of their thicknesses. Consequently, an etching technique was used to reveal the shell growth pattern. The etching was accomplished following the procedure described by Aubert et al. (2009) which uses a slightly modified Mutvei's solution (Schöne et al. 2005) for 45 minutes. Here, an increment was defined as the amount of shell between each ridge revealed by the etching. The etched sections were photographed under an optical microscope under diffuse light at a magnification of 100X, choosing a part of the inner layer where the increments were the most easily seen and where the growth rate was maximal. Then, growth increment thicknesses were measured using a modified version of Analysis Visilog basic ® V. 6.841 (Noesis Co.).

Environmental data. To analyze *Hippopus hippopus* shell growth variations with respect to its environment, environmental data for the period of growth were collected. The sea surface temperature, irradiance, wind force and pluviometry were provided by Météo France Nouméa (<http://www.meteo.nc/>) for the southern island. Tide timetables and ephemerides were produced by the SHOM, French Service Hydrographique et Océanographique de la Marine (<http://www.shom.fr/>).

Results

Gaping behavior and growth as revealed by HFNI valvometry. The typical daily behavior of a permanently immersed *Hippopus hippopus* was mostly characterized, all year long, by a fairly reproducible basic pattern (Fig. 2). During daytime, the opening was large while at night it was reduced to about 20% of its maximum daily opening value. No relationship between tide (timing and amplitude) and shell closing and opening time was observed, showing that the primary zeitgeber was clearly the night and day cycle. Short term complete closures regularly occurred at night (Fig. 2) while they were exceptional during daytime. Note that they were also occasionally absent. During daytime, most valve movements, micro-closures and/or partial closures, were of small amplitude (in relative value), characterizing a settled and resting state.

To visualize the growth of the *Hippopus hippopus* specimens for the studied year, the daily shortest distance between valves, was plotted as a function of time (Fig. 3). From 27 August 2007 to 18 August 2008, at the electrode level, the distance between each specimen's valves increased from 4 to 18 mm (min, max) depending on the specimen. This increase was regular, demonstrating a continuous growth all year round. Most animals (13 out of 15, i.e. 87%; median group) exhibited a mean growth of 8-14 mm at the electrode position while two other congeners (13%) exhibited either a clearly higher or slower growth rate. The periodic peaks on each curve show that a complete closure every day (more precisely, night) is not a prerequisite in the species and does not occur daily.

For all individuals and for the whole year, no spawning activity, characterized by a stereotyped series of rapid valve contractions (Galsthoff 1938), was observed.

Cumulated growth and comparison between sclerochronology and HFNI valvometry.

Most animals exhibited a similar mean annual growth (Fig. 3). Consequently, we performed the sclerochronological analysis in 5 animals randomly sampled from the median group (n°a1, a3, a5, a6 and a15) to have a statistical insight into the most frequently observed growth patterns. Remarkably, the cumulated growth curves obtained from the sclerochronology and HFNI valvometry studies were similar. This is shown by the significant correlation coefficients (up to 0.998, $p < 2 \cdot 10^{-16}$) between the results obtained using both techniques (Table 1) although the two sets of measurements were not performed at the same shell position. The daily increment thickness of the 5 *H. hippopus* measured by sclerochronology in the inner layer varied from 5.5 to 50 μm with a mean value of $22.4 \pm 8.1 \mu\text{m}$. The results given by the HFNI valvometry were interestingly in the same range, from 12-27 $\mu\text{m/day}$. To our knowledge, this is the 1st empirical demonstration of daily growth band production in *Hippopus hippopus*. As every specimen deposits one increment per day and as the sampling date was known, each shell growth band measured by sclerochronology was dated. The last shell layer (cf. star on Fig. 1E) was deposited on 18 August 2008 (sampling date) and the former shell increments were then dated retroactively.

Sclerochronological profiles. The sclerochronological profiles of the 5 shell clams studied are presented in figure 4A and the mean value in figure 4B. As all animals were located on the same platform of $\sim 1 \text{ m}^2$ on the reef (uniqueness of time and location, i.e. simultaneity of exposition to environmental stressors), these data illustrate inter- and intra-individual change and variability as a function of time. First, the global pattern of daily thickness changes throughout the monitored year exhibits remarkable similarities for the 5 clams. This is furthermore illustrated by the mean curve for the 5 clams (Fig. 4B, daily mean values, SE not shown for clarity). Based on this full data set we suggest defining 3 periods (see vertical

dashed line on Fig. 4). Period 1, from September to early January, was characterized by global growth rate acceleration. Period 2, early January – early April, started with a deep general slowdown which reached a minimal value by the end of January. Then, shell growth started to increase again until mid-February. The maximum annual growth rate was observed during this period with a mean peak value at $42 \pm 5 \mu\text{m}.\text{day}^{-1}$ around mid-February. However, immediately after, and within 1.5 months, the growth rate decreased rapidly down to the minimal annual values. The analysis of the individual records for the same time period (Fig. 5A) clearly shows that the intra-individual variability was maximal during period 2, defining a very specific time period. Importantly, these maximum and minimum individual daily shell growth values were essentially asynchronous among animals as shown by the change of their mean value during that period (Fig 4B). This suggested the absence of a common drive, simultaneously accelerating or decelerating all growth rates, but rather an erratic individual behavior. Period 3, from April to 18 August 2008 (sampling date): the clams keep on growing but following a monotone pattern and rate, at minimal annual values and with a minimal variability. The mean daily growth line thickness during this period was quite close to that observed in late September 2007 (respectively $17.9 \pm 0.2 \mu\text{m}$ and 16.2 ± 0.4 , $p = 0.018 < 0.05$, Nonparametric Mann and Whitney U test).

To explain these global growth patterns, we then turned to the analysis of the corresponding changes in SST, solar irradiance and precipitation in the area (Fig. 4C to E).

Environmental changes. The New Caledonia climatic regime is characterized by two main periods, an austral rainy summer (Dec-March) and a dryer austral winter (May-Oct), each separated by transition phases. The experimental period, September 2007 to August 2008, thus started and ended in austral winter. The SST changes during that period can be divided in two parts (Fig. 4C). A warming period occurred from September until early April followed by a cooling period from mid-April to August. These two periods can be described by two linear

regressions that are respectively: $SST = 0.018 \text{ time} + 24.37$ ($R^2 = 0.82$, $p < 2.2 \cdot 10^{-16}$) and $SST = -0.04 \text{ time} + 24.37$ ($R^2 = 0.83$, $p < 2.2 \cdot 10^{-16}$). This annual temperature pattern is perfectly representative of the 10-year long survey reported by Le Borgne et al. (2010) in the same area. The annual change in solar irradiance can also be described as two main periods separated by an intermediary one (Fig. 4D). Maximum annual solar irradiance values were systematically observed from September to March (mean $1031 \pm 40 \mu\text{mol photons m}^{-2} \text{ sec}^{-1}$; maximum values, $1850 \mu\text{mol photons m}^{-2} \text{ sec}^{-1}$) which corresponds to the SST warming period. Then, from April to August, the solar irradiance was globally lower ($684 \pm 40 \mu\text{mol photons m}^{-2} \text{ sec}^{-1}$; significantly different from the previous period, $p < 0.001$), corresponding to the decrease in water temperature. In early February 2008 (vertical grey bar on Fig. 4), the tropical cyclone GENE passed to the east of New Caledonia. This climatic event generated, from January 31 to February 2, sea swell amplitude above 6-7 m in the Havannah channel where the Ioro reef and the giant clams were located (data Météo France).

Daily growth changes vs. environmental parameters. Regarding water temperature, a major drive in poikilotherms, we first looked for a global relationship between mean daily growth line thickness and SST during the whole period September 2007 - August 2008. A significant correlation was observed, characterized by the following linear regression equation, $\text{Growth increment thickness} = 2.33 \text{ SST} + 32.12$, $R^2 = 0.304$, $p = 2.2 \cdot 10^{-16}$ (Fig. 5A). We turned then to a closer step by step analysis, mainly based on the 3 periods defined following the sclerochronological study. Considering period 1 (24 September 2007 - 1 January 2008), the correlation between shell growth and SST was clearly better than for the whole year ($\text{Growth increment thickness} = 4.54 \text{ SST} - 89.9$, $R^2 = 0.55$, $p = 2.2 \cdot 10^{-16}$; Fig. 5B). Interestingly, not only did this acceleration period of growth rate clearly parallel the SST change but it also paralleled the slow down period recorded just after, at the beginning of period 2 (before GENE occurrence). Indeed, a virtually day by day correlation between growth rate

and SST was observed during this period ($r^2 = 0.45$; Figs. 5C1 to C3). Remarkably, the nature of this correlation during the 2nd part of period 2 (15 February – 6 April 2008), characterized by a highly variable shell growth and SST above 27°C, was completely different. Indeed, the positive relationship between increment thickness and SST vanished and the relationship even became inverse (Fig. 5D1). Specifically, when SST stayed above 27 °C, individual shell growth became erratic but its mean value exhibited a significant and rather monotone slow down (Fig. 4 and 5D2-3). During this second part of period 2, the relationship between daily growth increment thickness and SST is described by: Growth increment thickness = $-7.34 \text{ SST} + 228.8$ ($R^2 = 0.27$, $p < 0.003$). Finally, from April to August 2008 (i.e. period 3), there was a complete change of pattern. While SST progressively decreased (Fig. 4C), the clams kept on growing but at a low and constant rate. Daily shell growth was independent of SST (Growth increment thickness = $0.25 \text{ SST} + 11.76$, $R^2 = 0.016$, $p = 0.13$; Fig. 5E). For that last period, the shell growth pattern paralleled the solar irradiance regime that was at its lowest annual values.

Related behavioral changes. The above data set showed that shell growth variations were not a direct function of the water temperature and, more specifically, that growth was highly variable, even erratic, by the end of the hottest period. We next addressed the existence of behavioral signs associated with this altered growth pattern. Specifically, we examined whether the erratic growth period was associated with valve behavior, either revealing a protective behavior and/or a stress reaction. Consequently, we focused our attention on the changes of maximum opening status of the group of clams.

Figure 6B shows the change of maximal daily openings from September 2007 to August 2008 in the whole group. For clarity, data were combined with the mean growth line thickness from Fig. 4 (Fig. 6A). Besides the general growing pattern that confirms the continuous

growth as illustrated in Fig. 3, two striking events of major interest for our purposes were observed. The first one was the valve behavior associated with the presence of the hurricane GENE, i.e. when the swell amplitude reached 6-7 m in the southern lagoon that was obviously a stressful condition for clams living at - 3-4 m. Clam behavior was characterized by a sudden decrease of maximum opening state. This behavior abruptly started on 30 January 2008 and was back to reference values on 5-6 February 2008. The maximum valve opening status at the end of period 2 (large arrow) shows a similar change, although its time course was different. The maximum opening values decreased progressively by early March and suddenly recovered at the beginning of period 3. Interestingly, the recovery occurred right when the water temperature started to decrease (Fig. 4C). Figure 7 presents two typical records of daily valve activity during these two periods of reduced values of maximal daily openings. They are perfectly representative of the whole animal group and must be compared to the steady reference situation shown in Fig. 2. These behaviors are typically unsteady and disturbed. They are both characterized by numerous partial or nearly complete valve closures during daytime and nighttime. In late January - early February, the stress signature was clearly associated with the presence of the hurricane GENE (Fig 7, left). In late March - early April, at the end of the warmest period in the lagoon, the same disturbed pattern was observed, (Fig 7, right).

Discussion

In this one-year study of growth rate and gaping behavior in giant clams *Hippopus hippopus* in the southern lagoon of New Caledonia, HFNI Valvometry, a new remote sensing system that records molluscan bivalve gaping behavior, and classical sclerochronology, used to study shell growth variations, were used for the first time in combination. Our data show that the relationship between growth rate and temperature is not linear. A similar non-linearity was shown in a 4-year sclerochronological record of one *H. hippopus* specimen from the Southwest Nouméa lagoon (Aubert et al. 2009). Particular growth decreases in the course of the year were interpreted as reflecting either physiological or non-climatic external stress or upwelling-related environmental stress (for one of them). The growth-scheme observed here, based on one year of growth of 5 specimens located more northeastward in the lagoon, is quite different. We present evidence that at the end of the summer period increasing SST (summer values are above 27°C) together with elevated solar irradiance deeply impact, *H. hippopus* daily growth and shell gaping behavior. This strongly suggests that the *H. hippopus* giant clam species could, at 27-28°C, already live close to its upper thermal limits. Although the experimental conditions are completely different, this specific sensitivity is in agreement with conclusions reported by Blidberg et al. (2000) for giant clams. Indeed, they reported that *H. hippopus* has the lowest tolerance to increased temperature (ambient + 3°C for 24h) compared with *Tridacna gigas* and *Tridacna derasa*.

History of the technique. Oceanographers rely heavily upon in situ sensing observations to better understand oceans and how marine animals live in the field (Dickey et al. 2008). There is also a compelling need for the use of remote online sensors to instantly and widely distribute information on a daily basis, specifically in a marine environment (Kröger et al. 2009). Initially, HFNI valvometry was developed to solve fundamental questions in respiratory physiology in molluscan bivalves (Massabuau 2001; Chambon et al. 2007;

Massabuau and Abele 2011), ecotoxicology (Tran et al. 2003; Tran et al. 2010) and marine ethology (Sow et al. 2011; Tran et al. 2011). However, studying bivalve gaping behavior is a very old story, dating back to the pioneering work of Marceau (1909). Since then, various systems have been developed with variable success. The approaches for measuring valve behavior were quite diverse and many different types of recording apparatus were proposed. During the last 20-30 years, ‘valvometry’ techniques have been studied as an aquatic pollution biosensor and various devices for measuring valve behavior have been developed (e.g. Higgins 1980; Borcharding 1997; Curtis et al. 2000). More recently, systems using magnetic fields (Wilson et al. 2005), fiber optics (Frank et al. 2007), video recordings (Riisgard et al. 2006; Saurel et al. 2007) and magnetic switches (Garcia-March et al. 2008) have been presented. Some systems are commercially available, such as the Mussel Monitor (Kramer et al. 1989) and the Dreissena monitor (Borcharding 2006). We alone, however, have developed the ability to monitor gaping behavior in marine environments from 1000s of kilometers away, accessed the behavior *via* the web (exchange data and programs from the lab to the field and *vice versa*), and deployed the system up to 1-2 years in full autonomy, without in situ human intervention. Our method differs thus from previous efforts in a number of significant ways. First, it can work for long periods of time even in the high sea. Second, as we use light electrodes and thin cables, it allows bivalves to behave without constraint in their natural environment including burrowing into the sediment. Third, our technology allows monitoring bivalve behavior anyplace in the world using cellular, internet, or satellite networks to transmit information.

The present data demonstrate continuous growth in the giant clam *Hippopus hippopus*, although it is quite likely that an occasional absence of growth increment deposition of 1-2 days could perfectly well have been missed. Importantly, the comparison between measurements using sclerochronology and HFNI valvometry shows a main mean bias of 0.2

μm over one year. This is not to say that the accuracy of each technique independently is 0.2 μm but that the errors of both methods are identical.

Relationship between growth rate, temperature and irradiance. In the southern lagoon of New Caledonia, where SST ranges from 22 to 28 °C, we show that the relationship between growth rate and water temperature in *Hippopus hippopus* is not linear. Based on the 2007-2008 record, we suggest the existence of three different periods and we show that by the end of the summer period, *H. hippopus* could already be living close to its upper thermal limits. When the water temperature started to rise following austral winter, the growth rate continuously accelerated. When it surpassed 26-27 and until 28°C, the sclerochronology demonstrates that growth rates became erratic, which we suggest is a first sign of temperature-induced physiological disorder (Figs 4A). Then, the mean value of the daily growth increment started to slow down, which is a second sign of negative interaction between high temperature and growth rate (Fig. 4B). By the end of the 2nd period, just before the SST started to decrease and the maximum irradiance day values vanished, mollusc behavior was typically deeply disturbed and no longer tranquil (compare Fig 7B and 2). In fact, the behavior was as disturbed as during a very high sea record obtained during a hurricane alert with a swell amplitude estimated to be 7-8 m by Meteo France (Fig. 7A). This is a third sign of negative interaction. Importantly, we also recorded a progressive decrease in daily maximum valve-opening values, a sign of protective behavior in bivalves (Fig 6B). This behavior started to develop in the whole group in early March, parallel to the production of shell growth increments with erratic thicknesses (compare Fig 6B and 6A). This is the fourth sign of deep behavioral disturbance that we observed. As the animals we studied were (i) located on an isolated reef in a clean area, that is in an open lagoon of high natural value and diversity (a UNESCO World Heritage site, Andréfouët and Wantiez 2010), and as (ii) there was no possibility of contaminant surge at the time of the experiments and (iii) no other

particular climatic event (Fig 4E), we strongly suggest that the local changes of SST and local irradiance (Figs 4C, D) should be the most likely explanatory environmental parameters for the reported growth rate disturbances. Thermally-related growth reduction and/or cessation have been widely reported in bivalve shells, including summer SST-related stress on shell growth (Arnold et al., 1998; Lazareth et al., 2006; Rhoads and Lutz, 1980). Nevertheless, to our knowledge, this is the first time that perturbations of shell growth (erratic increment thicknesses), without growth cessation, is empirically connected to perturbed shell gaping behavior, and that both of those observations are linked with elevated SST and solar irradiance. The combination of all these observations can be at least partly explained by the specificities of the giant clams.

Underlying physiological mechanisms. The disturbances we report were observed during the end of the warm period which strongly suggests a mechanism based on damage induced by excess light energy and photo-oxidative stress and/or imbalance between O₂ production by zooxanthellae, O₂-supply and O₂-consumption in the clam tissues. Indeed, the irradiance values we report were clearly in the upper range of the maximum required values in giant clams (Klumpp and Griffiths 1994). These authors showed that gross photosynthetic rates approach saturation at $723 \pm 75 \mu\text{mol photons m}^{-2} \text{sec}^{-1}$ in *Hippopus hippopus*, 488 ± 51 in *Tridacna gigas*, 644 ± 35 in *Tridacna crocea* and $535 \pm 76 \mu\text{mol photons m}^{-2} \text{sec}^{-1}$ in *Tridacna squamosa*, and here the mean summer value we reported was $1031 \pm 40 \mu\text{mol photons m}^{-2} \text{sec}^{-1}$. Interestingly, we also report physiological alterations at temperatures ranging from 27-28°C at the end of the summer period. Lesser (1996) showed that cultures of a symbiotic dinoflagellate, grown in an environment that attempts to simulate in situ light and temperatures at ecologically relevant levels, with and without UV exposure, exhibit a significant decline in maximum photosynthetic capacity and cellular growth rates also at 25 and 27°C (Fig 1 in Lesser, 1996). Symbiotic animals routinely experience elevated O₂ partial

pressures (hyperoxic PO_2) within their tissues as a result of photosynthetically produced oxygen (Dyken and Shick 1982). Physiological hyperoxia and UV exposure act with sublethal temperature perturbations to induce oxidative stress (Asada and Takahashi 1987; Lesser et al. 1990). It has been proposed as a mechanism that could lead to bleaching in symbiotic invertebrates (Lesser and Shick 1989; Lesser et al. 1990). The absorption of excitation energy in the presence of excess O_2 produces active oxygen, superoxide radicals and hydrogen peroxide compromising cell integrity (Asada and Takahashi 1987). Both elevated temperatures and photodynamic production of reactive oxygen species (ROS) induce cellular alteration that further disrupts primary photochemistry (Asada and Takahashi 1987; Kyle 1987; Ludlow 1987). An additional increase in the rates of ROS production, associated to temperature-induced metabolic enhancement, occurs primarily in the mitochondria of host and symbionts (Burdon et al. 1990). In extreme situations, the increase in the cellular production of O_2^- and H_2O_2 is not counterbalanced by the increased activities of antioxidant enzymes in zooxanthellae (Lesser and Shick 1989; Dyken and Shick 1992). In addition to the occurrence of hyperoxic blood PO_2 , it has been reported that the oxidative stress in the host tissues also occurs as a result of direct H_2O_2 diffusion from algal symbionts (Asada and Takahashj 1987). Reactive oxygen species affect cellular proteins and numerous physiological mechanisms such as oxidation of DNA, lipids and proteins (Morel and Barouski 1998). In giant clams this may lead to the loss of symbiotic dinoflagellates *via* expulsion or complete loss of gastrodermal cells as described by Gates et al. (1992). That ROS damage cells is a major hypothesis with respect to ageing (Finkel and Holbrook 2000; Kirkwood and Austad 2000). Superoxide dismutase, a specialized enzyme aimed at protecting cells against O_2 -toxicity, is thought to have appeared very early in evolution, but without doubt, preventing high PO_2 is also the simplest and most efficient tool for limiting ROS production. It is also fundamental to note that O_2 requirement is not only a matter of ATP (adenosine triphosphate)

production, even if 90 % of the O₂ supply is devoted to matching the mitochondrial demand. Organisms use O₂ and its derived forms as messengers, modulators, and even neuromodulator-like substances. Even small deviations from precise PO₂ set-points can have great effects, for example on the lobster neural networks, which highlight the necessity of well-tuned tissue PO₂ status (Clemens et al. 2001). As a whole, the *Hippopus hippopus* story appears in complete opposition to the low PO₂ strategy that is successfully performed in numerous water breathers. Indeed, many aquatic animals have a strategy which enables them to regulate PO₂ in their arterial blood, independent of many oxygenation conditions in the environment, to low levels of between 1 and 3 kPa (Massabuau 2001, 2003), levels which are similar to the PO₂ in most mammalian tissues (Vanderkooi et al. 1991). As a reminder, PO₂ in water equilibrated with air is 21 kPa (for reference 1 kPa = 10⁻² bar; in seawater when PO₂ = 1 kPa; it corresponds to an O₂ fraction of 1% and an O₂ concentration of 0.32 mg l⁻¹ at 25°C). This ability to regulate respiration has been demonstrated in crustaceans, fish and molluscs and the strategy appears to be a very ancient one, as it is already present in the primitive ostracods, which existed 420-450 million years ago (Corbari et al. 2004). It is unlikely however to apply to *H. hippopus* at the end of the summer period because, as we strongly suggested above, the evidence of SST- and solar irradiance-related stress could be related to a physiological oxidative stress triggered by the zooxanthellae symbionts. It is also unlikely that the present thermal tolerance threshold could be correlated to a respiratory impairment and the onset of anaerobic metabolism as proposed for marine species (Pörtner and Knust 2007). The precise nature of the mechanisms involved here remains consequently an open and challenging question for future research.

Recording online effects of climate change on invertebrates of coral reefs. Climate change is threatening tropical reefs across the world and most scientists are focusing largely on corals and fishes. The present work reports information on the effect of environmental conditions on

giant clams, focusing on a particular species that we believe to be more specifically amenable to an online analysis of behavioral activity than others. *Hippopus hippopus* is not sessile, not fixed in the reef, and it is often taken as a rustic clam, able to face emersion in the intertidal area. It is clear that tropical invertebrates show species-specific sublethal responses to environmental changes but *H. hippopus*, may be due to a specific thermal sensitivity (Blidberg et al. 2000), could be an interesting sentinel species. This report was also intended to show how a control framework could be developed to study and eventually predict the vulnerability of benthic invertebrates to climate change and/or their adaptive ability on a long term time scale.

References

- ADDESSI, L. 2001. Giant clam bleaching in the lagoon of Takapoto atoll (French Polynesia). *Coral Reefs* **19**: 220-220.
- ANDREFOUËT, S. AND L. WANTIEZ. 2010. Characterizing the diversity of coral reef habitats and fish communities found in a UNESCO World Heritage Site: The strategy developed for Lagoons of New Caledonia. *Mar. Poll. Bull.* **61**: 612-620.
- ARNOLD, W. S., T. M. BERT, I. R. QUITMYER, AND D. S. JONES. 1998. Contemporaneous deposition of annual growth bands in *Mercenaria mercenaria* (Linnaeus), *Mercenaria campechiensis* (Gmelin), and their natural hybrid forms. *J. Exp. Mar. Biol. Ecol.* **223**: 93-109.
- ASADA, K., AND M. TAKAHASHI. 1987. Production and scavenging of active oxygen in photosynthesis, p. 228-287. *In* D. J. Kyle, C. B. Osmond and C. J. Arntzen [eds.], *Photoinhibition*. Elsevier.
- AUBERT, A., C. E. LAZARETH, G. CABIOCH, H. BOUCHER, T. YAMADA, Y. IRYU, AND R. FARMAN.. The tropical giant clam *Hippopus hippopus* shell, a new archive of environmental conditions as revealed by sclerochronological and $\delta^{18}\text{O}$ profiles. *Coral Reefs* **28**: 989-998, doi :10.1007/s00338-009-0538-0
- BLIDBERG, E., T. ELFWING, P. PLANHNAN, AND M. TEDENGREN. 2000. Water temperature influences on physiological behaviour in three species of giant clams (Tridacnidae), *Proceedings 9 International Coral Reef Symposium*.
- BORCHERDING, J. 2006. Ten years of practical experience with the *Dreissena*-Monitor, a biological early warning system for continuous water quality monitoring. *Hydrobiologia* **556**: 417-426.

- BRIOLLAIS, L., G. DURRIEU, AND R. UPATHILAKE. 2007. Novel approach for genome scan meta-analysis of rheumatoid arthritis: a kernel-based estimation procedure. *BMC Genetics* **1**: 1-6.
- BROWN, B. E. 1997. Coral bleaching: causes and consequences. *Coral Reefs* **16**: S129-S138.
- BUCK, B. H., AND H. ROSENTHAL, AND U. SAINT-PAUL. 2002. Effect of increased irradiance and thermal stress on the symbiosis of *Symbiodinium microadriaticum* and *Tridacna gigas*. *Aquat. Living Resour.* **15**: 107-117.
- BUDDEMEIER, R. W., AND J. E. MARAGOS. 1975. Radiographic studies of reef coral exoskeletons; rates and patterns of coral growth. *Contributions of the Hawaii Institute of Geophysics* **1974**: 233-253.
- BURDON, R.H., V. GILL, AND C. EVANS. 1990. Active oxygen and heat shock protein induction, p 19-25. *In* M.J. Schlesinger et al. [eds.], *Stress proteins*, Springer.
- CHAMBON, C., A. LEGEAY, G. DURRIEU, P. GONZALEZ, P. CIRET, AND J.-C. MASSABUAU. 2007. Influence of the parasite worm *Polydora* sp on the behaviour of the oyster *Crassostrea gigas*: a study of the respiratory impact and associated oxidative stress. *Mar. Biol.* **152**: 329-338.
- CLEMENS, S., J.-C. MASSABUAU, P. MEYRAND, AND J. SIMMERS. 2001. A new neuromodulatory-like role for oxygen in shaping the activity of neuronal networks in lobsters. *Respir. Physiol.* **128**: 299-315.
- CORBARI, L., P. CARBONEL, AND J.-C. MASSABUAU. 2004. How a low tissue O₂ strategy could be conserved in early crustaceans: the example of the podocopid ostracods. *J. Exp. Biol.* **207**: 4415-4425.
- CURTIS, T. M., R. WILLIAMSON, AND M. H. DEPLEDGE. 2000. Simultaneous, long-term monitoring of valve and cardiac activity in the blue mussel *Mytilus edulis* exposed to copper. *Mar. Biol.* **136**: 837-846.

- DICKEY, T. D., E. C. ITSWEIRE, M. A. MOLINE, AND M. J. PERRY. 2008. Introduction to the Limnology and Oceanography Special Issue on Autonomous and Lagrangian Platforms and Sensors (ALPS). *Limnol. Oceanogr.* **53**: 2057-2061.
- DODGE, Y., AND J. JURECKOVA. 1995. Estimation of Quantile Density-Function Based on Regression Quantiles. *Stat. Probabil. Lett.* **23**: 73-78.
- DURRIEU, G., AND L. BRIOLLAIS. 2009. Sequential design for microarray experiments. *J. Am. Stat. Assoc.* **104**: 650-660.
- DYKENS, J. A., AND J. M. SHICK. 1982. Oxygen production by endosymbiotic algae controls superoxide-dismutase activity in their animal host. *Nature* **297**: 579-580.
- FRANK, D. M., J. F. HAMILTON, J. E. WARD, AND S. E. SHUMWAY. 2007. A fiber optic sensor for high resolution measurement and continuous monitoring of valve gape in bivalve molluscs. *J. Shell. Res.* **26**: 575-580.
- FINKEL, T., AND N. J. HOLBROOK. 2000. Oxidants, oxidative stress and the biology of ageing. *Nature* **408**: 239-245.
- GALTISOFF, P. S. 1938. Physiology of reproduction of *Ostrea virginica*. I. Spawning reactions of the female and male. *Biol. Bull.* **74**: 461-486.
- GARCIA-MARCH, J. R., M. A. S. SOLSONA, AND A. M. GARCIA-CARRASCOSA. 2008. Shell gaping behaviour of *Pinna nobilis* L., 1758: circadian and circalunar rhythms revealed by in situ monitoring. *Mar. Biol.* **153**: 689-698.
- GATES, R. D., G. BAGHDASARIAN, AND L. MUSCATINE. 1992. Temperature stress causes host-cell detachment in symbiotic cnidarians - implications for coral bleaching. *Biol. Bull.* **182**: 324-332.
- GOREAU, T. F., N. I. GOREAU, AND C. M. YONGE. 1971. Reef corals: autotrophs or heterotrophs? *Biol. Bull.* **141**: 247-260.

- GRÖCKE, D. R., AND D. P. GILLIKIN. 2008. Advances in mollusc sclerochronology and sclerochemistry: tools for understanding climate and environment. *Geo-Mar. Lett.* **1-4**: doi:10.1007/s00367-008-0108-4.
- GROTTOLI, A. G., L. J. RODRIGUES, AND J. E. PALARDY. 2006. Heterotrophic plasticity and resilience in bleached corals. *Nature* **440**: 1186-1189.
- HÄRDLE, W. 1990. Applied nonparametric regression. Cambridge University Press.
- HIGGINS, P. J. 1980. Effects of Food Availability on the Valve Movements and Feeding-Behavior of Juvenile *Crassostrea-Virginica* (Gmelin) .2. Feeding Rates and Behavior. *J. Exp. Mar. Biol. Ecol.* **46**: 17-27.
- HOEGH-GULDBERG, O. 2009. Climate change and coral reefs: Trojan horse or false prophecy? *Coral Reefs* **28**: 569-575.
- HUBER, P. J. 2004. Robust Statistics. John Wiley & Sons.
- HUDSON, J. H., E. A. SHINN, R. B. HALLEY, AND B. LIDZ. 1976. Sclerochronology - Tool for interpreting past environments. *Geology* **4**: 361-364.
- JOKIEL, P. L., AND S. L. COLES. 1990. Response of Hawaiian and other Indo-Pacific reef corals to elevated temperature. *Coral Reefs* **8**: 155-162.
- JUREČKOVČÁ, J. 1984. Regression quantiles and trimmed least squares estimator under a general design. *Kybernetika* **20**: 345-357.
- JUREČKOVČÁ, J., AND P. K. SEN. 1996. Asymptotics and interrelations. John Wiley & Sons.
- KIRKWOOD, T.B.B., AND S.N. AUSTAD. 2000. Why do we age. *Nature* **408**: 233-238.
- KLUMPP, D. W., AND C. L. GRIFFITHS. 1994. Contribution of phototrophic and heterotrophic nutrition to the metabolic and growth requirements of four species of giant clam (*Tridacnidae*). *Mar. Ecol. Prog. Ser.* **115**: 103-115.
- KRAMER, K. J. M., H. A., JENNER, AND D. ZWART. 1989. The valve movement response of mussels: a tool in biological monitoring. *Hydrobiologia* **188-189**: 433-443.

- KRÖGER, S., E. R. PARKER, J. D. METCALFE, N. GREENWOOD, R. M. FORSTER, D. B. SIVYER, AND D. J. PEARCE. 2009. Sensors for observing ecosystem status. *Ocean Science* **5**: 523-535, doi:10.5194/os-5-523-2009.
- KYLE, D. J. 1987. The biochemical basis for photoinhibition of photosystem II, p. 197-226. *In* D. J. Kyle, C. B. Osmond and C. J. Arntzen [eds.], *Photoinhibition*. Elsevier.
- LAZARETH, C. E., G. LASNE, AND L. ORTLIEB. 2006. Growth anomalies in *Protothaca thaca* (Mollusca, Veneridae) shells as markers of ENSO conditions. *Clim. Res.* **30**: 263-269.
- LE BORGNE, R., P. DOUILLET, R. FICHEZ, AND J.-P. TORRETON. 2010. Hydrography and plankton temporal variabilities at different time scales in the southwest lagoon of New Caledonia: A review. *Mar. Poll. Bull.* **61**: 297-308.
- LEGGAT, W., B. H. BUCK, A. GRICE, AND D. YELLOWLESS. 2003. The impact of bleaching on the metabolic contribution of dinoflagellate symbionts to their giant clam host. *Plant Cell Environ.* **26**: 1951-1961.
- LESSER, M. P. 1996. Elevated temperatures and ultraviolet radiation cause oxidative stress and inhibit photosynthesis in symbiotic dinoflagellates. *Limnol. Oceanogr.* **41**: 271-283.
- LESSER, M. P., AND J. M. SHICK. 1989. Effects of irradiance and ultraviolet-radiation on photoadaptation in the zooxanthellae of *Aiptasia pallida* - primary production, photoinhibition, and enzymic defenses against oxygen-toxicity. *Mar. Biol.* **102**: 243-255.
- LESSER, M. P., W. R. STOCHAJ, D. W. TAPLEY, AND J. M. SHICK. 1990. Bleaching in coral-reef anthozoans - effects of irradiance, ultraviolet-radiation, and temperature on the activities of protective enzymes against active oxygen. *Coral Reefs* **8**: 225-232.
- LUDLOW, M. M. 1987. Light stress at high temperature, p. 89-109. *In* D. J. Kyle, C. B. Osmond and C. J. Arntzen [eds.], *Photoinhibition*. Elsevier.

- MARCEAU, F. 1909. Recherches sur la morphologie, et l'histologie, et la physiologie compares des muscles adducteurs des mollusques acéphales. 2: 295-469. Archives de zoologie expérimentale et générale **2**: 295-469.
- MASSABUAU, J.-C. 2001. From low arterial- to low tissue-oxygenation strategy. An evolutionary theory. Resp. Physiol. **128**: 249-261.
- MASSABUAU, J.-C., AND D. ABELE. 2011 (In Press). Respiratory principles of oxygen exchange in water breathing animals. In T. Zenteno-Savin, J. P. Vázquez-Medina and D. Abele [eds.], Oxidative Stress in Aquatic Ecosystems. Blackwell Publishing.
- MASSABUAU, J.-C. 2003. Primitive, and protective, our cellular oxygenation status? Mech. Ageing Dev. **124**: 857-863.
- MAYNARD, J. A., A. H. BAIRD, AND M. S. PRATCHETT. 2008. Revisiting the Cassandra syndrome; the changing climate of coral reef research. Coral Reefs **27**: 745-749.
- MCCLANAHAN, T. R., M. ATEWEBERHAN, N. A. J. GRAHAM, S. K. WILSON, C. R. SEBASTIAN, M. M. M. GUILLAUME, AND J. H. BRUGGEMANN. 2007.. Western Indian Ocean coral communities: bleaching responses and susceptibility to extinction. Mar. Ecol. Prog. Ser. **337**: 1-13.
- MOREL, Y., AND R. BAROUKI. 1998. Influence du stress oxydant sur la régulation des gènes = Gene regulation by oxidative stress. Médecine sciences **14**: 713-721.
- MUSCATINE, L. 1990. The role of symbiotic algae in carbon and energy flux in reef corals, p. 75-87. In Z. Dubinsky [ed.], Coral Reefs. Ecosystems of the World. Elsevier.
- PORTNER, H. O., AND R. KNUST. 2007. Climate change affects marine fishes through the oxygen limitation of thermal tolerance. Science **315**: 95-97.
- PRZESLAWSKI, R., S. AHYONG, M. BYRNE, G. WÖRHEIDE, AND P. HUTCHINGS. 2008. Beyond corals and fish: the effects of climate change on noncoral benthic invertebrates of tropical reefs. Global Change Biol. **14**: 2773-2795.

- RHOADS, D. C., AND R. A. LUTZ. 1980. Skeletal growth of aquatic organisms: biological records of environmental change, p. 750. *In* F. G. Stehli [ed.], Topics in geobiology. Plenum Press.
- RIISGARD, H. U., J. LASSEN, AND C. KITTNER. 2006. Valve-gape response times in mussels: effects of laboratory preceding-feeding conditions and in situ tidally induced variation in phytoplankton biomass. *J. Shell. Res.* **25**: 901-911.
- SAUREL, C., J. C. GASCOIGNE, M. R. PALMER, AND M. J. KAISER. 2007. In situ mussel feeding behavior in relation to multiple environmental factors: Regulation through food concentration and tidal conditions. *Limnol. Oceanogr.* **52**: 1919-1929.
- SCHÖNE, B. R., E. DUNCA, J. FIEBIG, AND M. PFEIFFER. 2005. Mutvei's solution: An ideal agent for resolving microgrowth structures of biogenic carbonates. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **228**: 149-166.
- SOW, M., G. DURRIEU, L. BRIOLLAIS, P. CIRET, AND J.-C. MASSABUAU. 2011 (in press). Water quality assessment by means of valvometry and high-frequency data modeling. *Envir. Monit. Assess.*
- SUGGETT, D. J., AND D. J. SMITH. 2011. Interpreting the sign of coral bleaching as friend vs. foe. *Global Change Biol.* **17**: 45-55.
- TOSCANO, M. A., G. LIU, I. C. GUCH, K. S. CASEY, A. E. STRONG, AND J. E. MEYER. 2000. Improved prediction of coral bleaching using high-resolution HotSpot anomaly mapping, p. 1143-1147, Proceedings 9th International Coral Reef Symposium.
- TRAN, D., P. CIRET, A. CIUTAT, G. DURRIEU, AND J.-C. MASSABUAU. 2003. Estimation of potential and limits of bivalve closure response to detect contaminants: Application to cadmium. *Environ. Toxicol. Chem.* **22**: 914-920.

- TRAN, D., H. HABERKORN, P. SOUDANT, P. CIRET, AND J.-C. MASSABUAU. 2010. Behavioral responses of *Crassostrea gigas* exposed to the harmful algae *Alexandrium minutum*. *Aquaculture* **298**: 338-345.
- VANDERKOOI, J. M., M. ERECINSKA, and I. A. SILVER. 1991. Oxygen in mammalian tissue - Methods of measurement and affinities of various reactions. *Am. J. Physiol.* **260**: C1131-C1150.
- WILSON, R., P. REUTER, and M. WAHL. 2005. Muscling in on mussels: new insights into bivalve behaviour using vertebrate remote-sensing technology. *Mar. Biol.* **147**: 1165-1172.

Table 1. Comparison of cumulated growth obtained by sclerochronology and HFNI valvometry for the period September 2007 - August 2008 in 5 giant clams *H. hippopus*. Correlation data, r; Bias mean, mean difference between both techniques in μm ; SE, standard error of the bias mean in μm ; CI, confidence intervals, min and max values covering 95 and 99% of the distribution in μm .

Clam	Correlation (r)	Bias mean, μm	SE	CI 95%	CI 99 %
a1	0.994	0.178	0.03	[0.118; 0.237]	[0.099; 0.256]
a3	0.998	0.322	0.035	[0.253; 0.390]	[0.232; 0.411]
a5	0.998	0.024	0.012	[-0.0004; 0.049]	[-0.008; 0.056]
a6	0.974	0.479	0.047	[0.388; 0.571]	[0.360; 0.600]
a15	0.983	0.109	0.031	[0.048; 0.170]	[0.029; 0.189]

Figure captions

Fig 1. (A-C) Location of the Ioro reef in the southern lagoon of New Caledonia. (D) *Hippopus hippopus*, horse's hoof clam, equipped with 2 electrodes (open circles): animal n°a4 after 1 year on site. Parallel bars: sampling location of the shell slices used in sclerochronology. (E) Shell slice showing the growth-line deposition pattern and the umbo (star). Maps: copyright IRD – Cartography department.

Fig 2. Typical daily valve behavior in a resting giant clam *Hippopus hippopus* recorded *in-situ* in the southern lagoon of New Caledonia, on the Ioro reef (clam n°a4, 21 November 2007). The shell is largely opened during the day and partially closed at night. Complete shell closures occur briefly at night; they are rare during the day.

Fig 3. Growth index measured by HFNI valvometry from September 2007 to August 2008. The index corresponds to the minimum daily opening value measured between the electrodes shown in Fig. 1D in 15 giant clams *Hippopus hippopus*, living in the southern lagoon of New Caledonia. The growth is continuous.

Fig 4. Change of daily growth increment thickness at the umbo level in *Hippopus hippopus* as a function of time, from September 2007 to August 2008, sea surface temperature SST, solar radiation and precipitation. From top to bottom, (A) a1, 3, 5, 6 et15: giant clams1, 3, 5, 6, 15. (B) mean growth line thickness for the same clams $\pm$ 1SE. C, SST, (C). (D) Solar irradiance; (E) Pluviometry (B-E: data Météo France). Vertical grey line: the hurricane GENE was approaching, the swell amplitude reached 6-7 m in the southern lagoon. The growth

increment thickness became highly variable and its mean value started to decrease when the SST reached its maximal annual value.

Fig 5. (A-B) (C1-D1) (E) Relationships between mean daily growth increment thickness measured by sclerochronology, sea surface temperature, SST, and date. The sampling periods are mentioned in each panel. (A) Global relationship between growth rate and temperature; (B) (C1-D1) (E) Relationship for specific periods. (C2) and (C3) parallel changes of growth increment thickness (C2) and SST (C3) as a function of time from 1/1/-14/2/2008; (D2-3) Parallel changes of growth increment thickness (D2) and SST (D3) as a function of time from 15/2/-6/4/2008 (C2-3, D2-3, enlarged views of Fig. 4B and 4C). Vertical grey line: hurricane GENE, which generated sea swell amplitude of 6-7 m around the clams' location. See text for regression.

Fig. 6: Change of maximum valve opening values in the whole group of 15 clams during the 3 time periods defined in Fig 4B. (A) Mean growth line thickness in 5 giant clams (same data as in Fig. 4); (B) Maximum valve openings. Vertical grey line: hurricane GENE. Arrows, periods during which the maximum shell opening values decreased during period 2. See Text.

Fig. 7. Typical disturbed daily valve-behaviors in giant clams *H. hippopus* from the southern lagoon of New Caledonia (compare with Fig. 2). (A) Animal a6 facing the swell associated with hurricane GENE (31 January 2008). (B) Animal a5 after 3 weeks at temperatures ranging from 27-28°C (2 April 2008; TTS = 27.7°C). Compare with the resting pattern shown in Fig 2 and note the analogy between the numerous transient closures occurring all day under the stressful storm conditions and the numerous closures during the day at temperatures of 27-28°C.

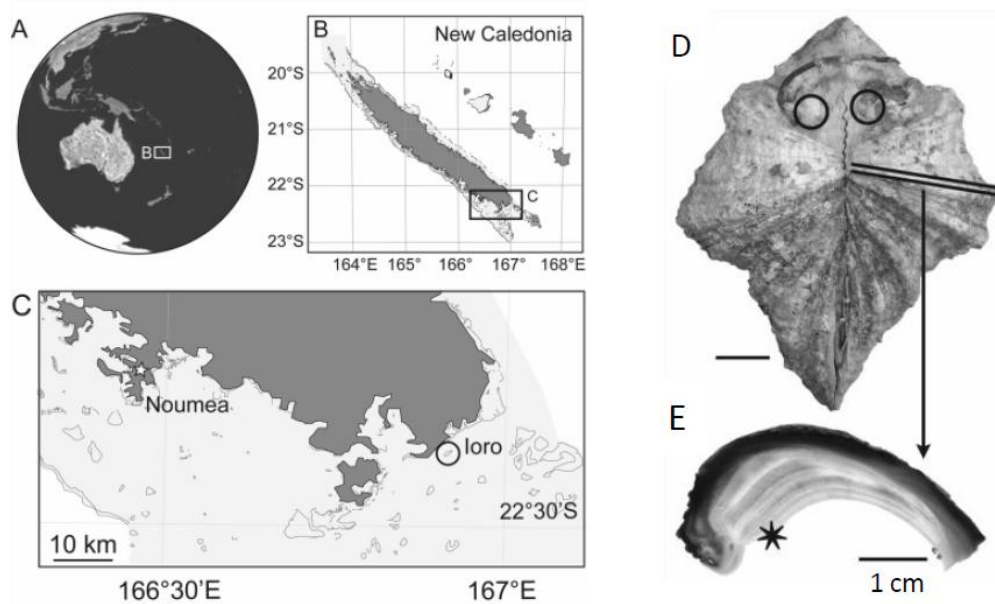


Figure 1

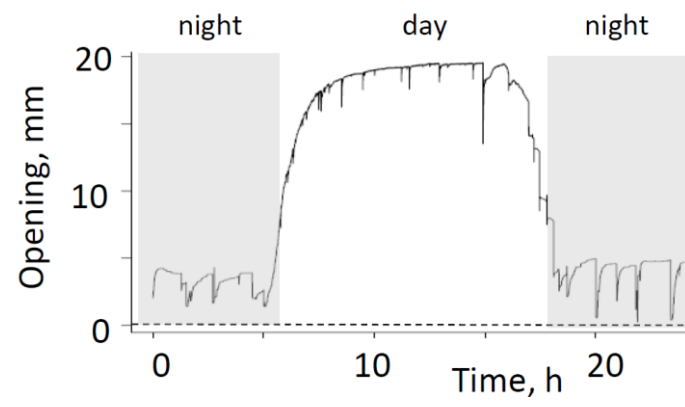


Figure 2

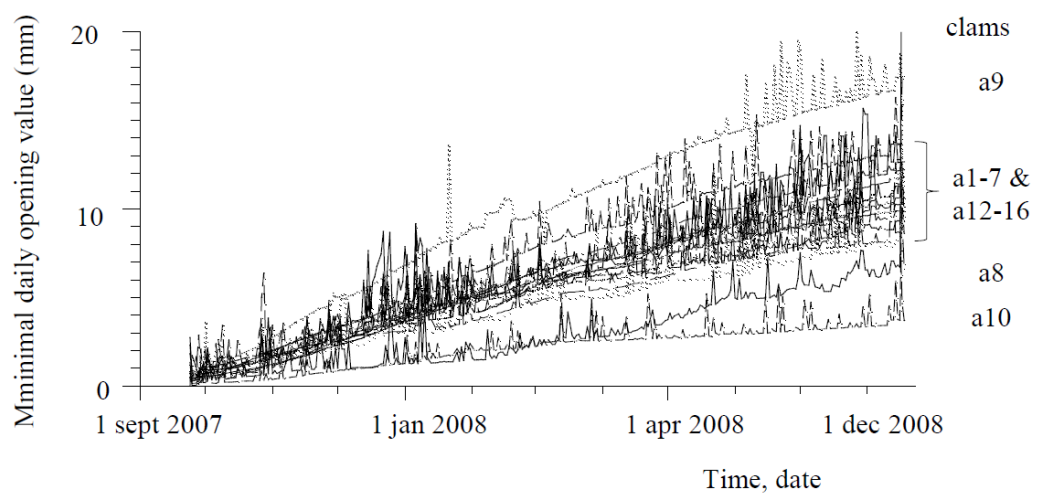


Figure 3

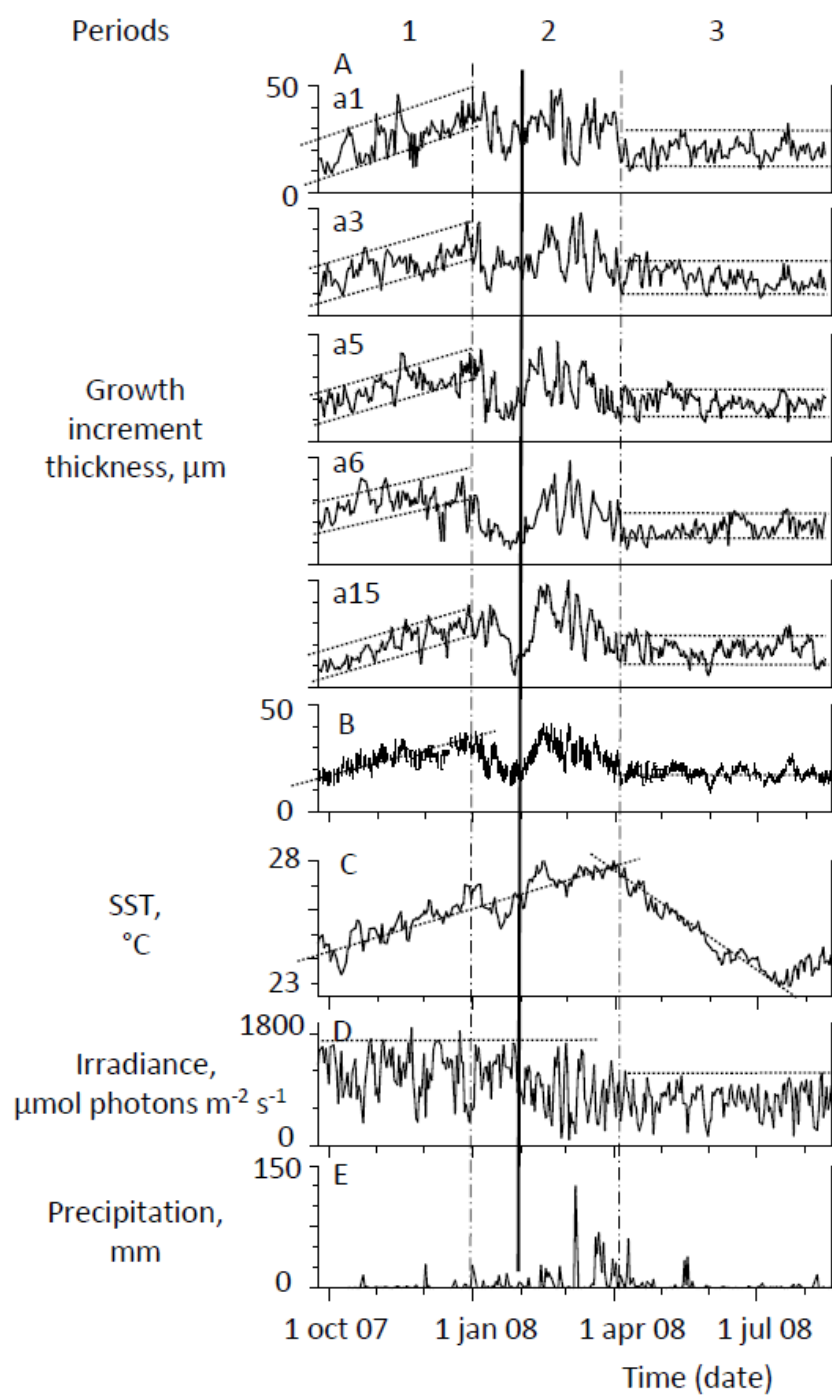


Figure 4

Chapitre 4

Estimation par la théorie des shot noise

4.1 Introduction

Dans le but de modéliser et caractériser la dynamique des micro-fermetures (ou fermetures partielles des valves), nous avons utilisé une approche basée sur des processus stochastique de type shot noise. Pour utiliser cette approche, nous avons besoin d'introduire la notion de chocs qui correspond à des fermetures spontanées chez l'huître. On parlera de choc lorsqu'à un temps donné, l'huître se ferme avec une amplitude supérieure à un seuil, ce seuil a été choisi de sorte que le bruit électronique n'intervienne pas dans nos résultats. A cette fermeture, appelée temps du déclenchement du choc, tant que l'huître continue à se fermer, on sera toujours concerné par le même choc. Nous pouvons extraire toutes les informations dont nous avons besoin telles que : le temps de départ du choc, sa durée, son amplitude et par déduction le temps entre deux chocs, appelé intertemps (voir figure 4.1). Notre travail se base sur les manifestations d'ouvertures et de fermetures des huîtres restreintes aux périodes non stationnaires. Mais finalement, ce qui est réellement intéressant afin de connaître comment elles se comportent, ce sont leur amplitude lors de fermeture et le temps de ce phénomène puis de les analyser, d'où l'intérêt et l'importance des chocs. Les shot noise (ou bruits de grenaille, en français) font l'objet de recherches depuis le début du 20ème siècle. Le premier papier traitant ce sujet fut écrit par Campbell (1909). Le bruit de grenaille est un bruit électrique dû au déplacement des porteurs de charges (électrons) dans les conducteurs. Il fait parti des concepts de modélisation appropriés pour des systèmes perturbés par des bruits discrets. Ce bruit se présente en fait, sous forme de coups externes (chocs) apparaissant de manière aléatoire. Le système de shot noise se particularise donc selon la façon dont la décroissance se fait et selon le processus d'apparition des chocs. A partir d'une généralisation du modèle des shot noise classiques, on peut mettre en évidence la notion du modèle des shot noise non linéaires (Eliazar et Klafter, 2005). Ce cadre théorique permet de considérer des chocs répartis de façon aléatoires, ayant une intensité également aléatoire, les deux phénomènes ayant une loi de probabilité quelconque. Le cadre non linéaire est donc le plus général et c'est cette approche, via les amplitudes et les intertemps qui est prise ici comme cadre théorique. De ceci, il sera défini l'amplitude du choc i ,

et le temps entre deux chocs i et $i + 1$, nommé intertemps du choc i . Cette approche focalisée sur un paramètre fin que sont les microfermetures valvaires permettra de caractériser un état de référence et des états atypiques caractéristiques du milieu. Ce travail a été réalisé en collaboration avec François Schmitt et Marie De Rosa de la station marine de Wimereux (Laboratoire d'Océanologie et de Géosciences, CNRS UMR 8187).

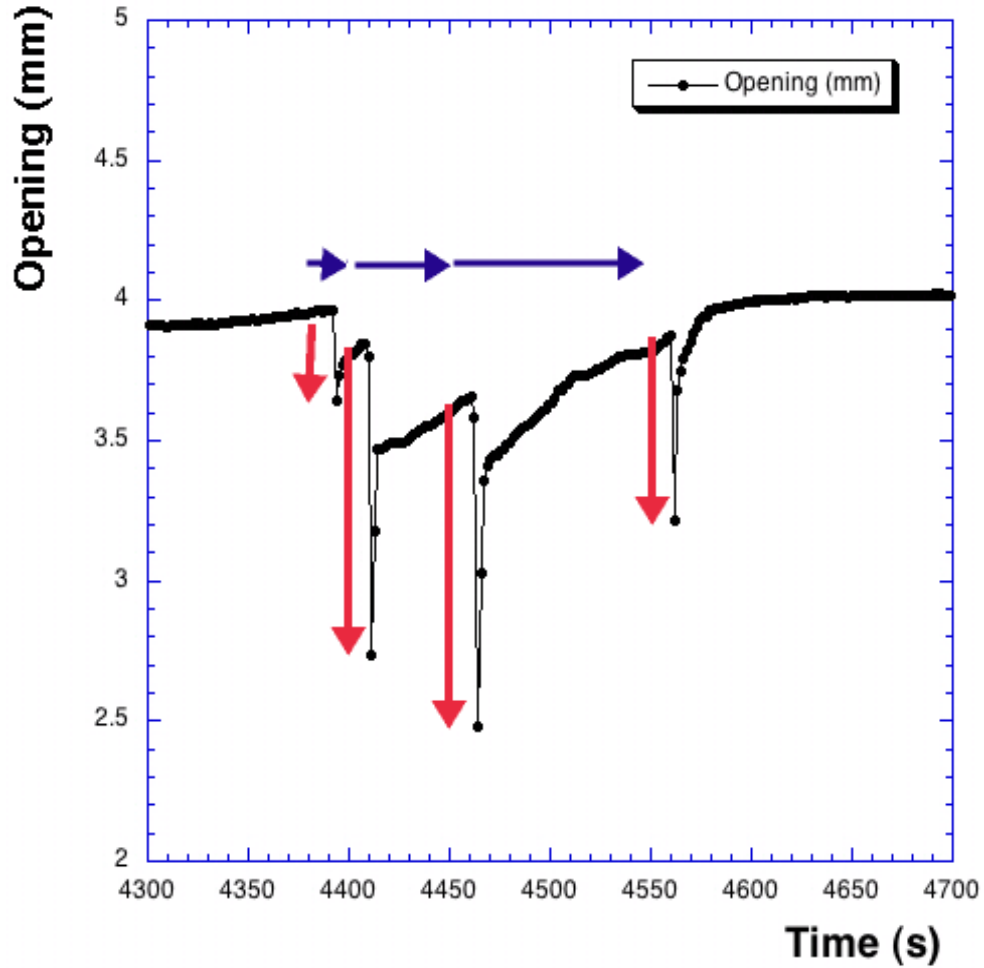


FIGURE 4.1 – Illustration des micro-fermetures avec 4 chocs et 3 inter-temps.

Données

Nous avons considéré ici le dispositif expérimental au niveau de la jetée d'Eyrac sur le bassin d'Arcachon. Nous rappelons que les écartements valvaires de 16 huîtres sont mesurés 24h/24. Les données sont accumulées et stockées pendant 24h (chaque fichier représente environ 18M octets pour une journée) puis transmises au laboratoire par protocole de transfert de données GPRS (General Packet Radio Service) en utilisant le réseau de la téléphonie mobile et Internet (FTP). Pour cette étude nous avons considéré les enregistrements de 16 huîtres sur 13 jours en Mai 2007. Les données présentent une forte variabilité comme le montre la figure 4.2a pour un bivalve sur 48 heures. Nous nous intéressons aux variations à haute fréquence pour une échelle comprise entre 1.6 seconde et plusieurs heures. La figure 4.2b montre un zoom sur une plus petite proportion de données (12 heures). Dans la suite nous considérons seulement des proportions de données comme celle ci où nous avons plusieurs micro-fermetures. Dans une série de 13 jours, on retrouve plusieurs proportions comme celle là.

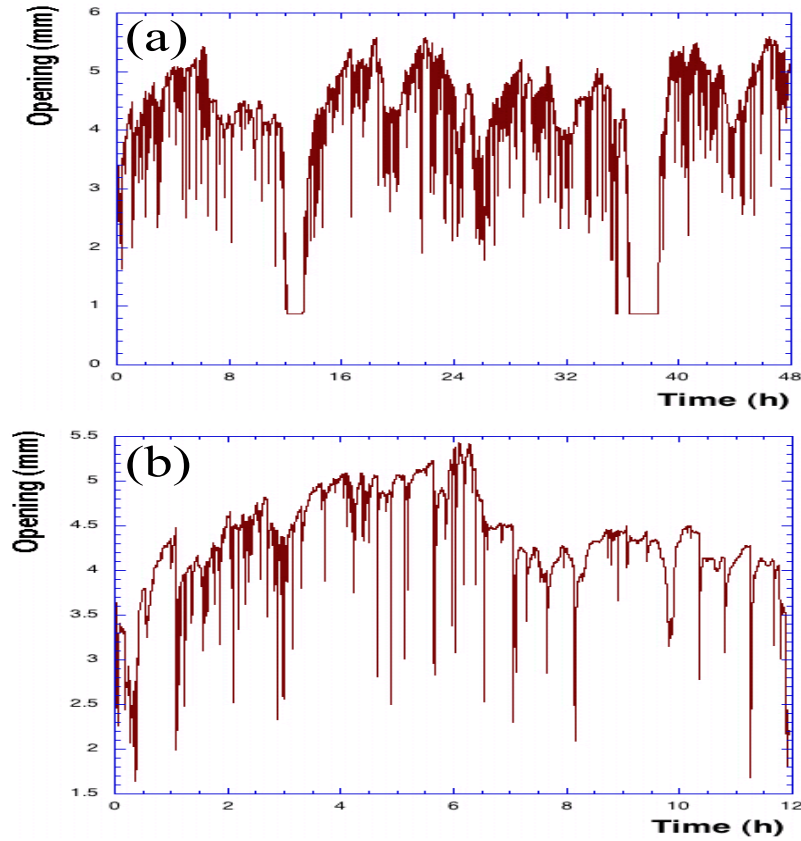


FIGURE 4.2 – Evolution de l'activité valvaire a haute fréquence. a) Représentation sur 48 heures; b) Zoom sur une plus petite proportions montrant plusieurs micro-fermetures.

4.2 Méthodes

Dans cette partie, au cours de l'analyse des données, nous faisons appel à certains résultats provenant du domaine de la turbulence. Nous rappelons ici certains de ces résultats sans vouloir entrer dans les détails de ce domaine, qui peut être complexe.

4.2.1 Densités de probabilité et extrêmes

Nous considérons ici l'estimation non paramétrique d'une densité par la méthode des noyaux (cf. Chapitre 1, 2). Les définitions données dans les chapitres 1 et 2 peuvent être généraliser dans le cas multivarié. Lorsque la variable explicative ou prédictive X est à valeurs dans $\mathbb{R}^d$, pour $d \in \mathbb{N}$ fixé, l'estimateur multivarié récursif de la densité de dimension d avec un noyau $\mathbf{K}$ et un paramètre de lissage h_i est défini par

$$\hat{f}_{h_n}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} \mathbf{K} \left(\frac{\mathbf{x} - \mathbf{X}_i}{h_i} \right), \quad \mathbf{x} \in \mathbb{R}^d. \quad (4.1)$$

La fonction noyau $\mathbf{K} : \mathbb{R}^d \rightarrow \mathbb{R}$ désigne une fonction multivariée satisfaisant

$$\int_{\mathbb{R}^d} \mathbf{K}(\mathbf{x}) d\mathbf{x} = 1.$$

les extrêmes

L'étude des extrêmes consiste à caractériser le comportement d'un processus stochastique pour les grandes valeurs. De façon générale, la probabilité de ces grandes valeurs décroît, ce qui fait que plus un extrême est grand, plus il est rare. Pour un processus gaussien, par exemple, les extrêmes décroissent très vite (en $\exp(-x^2)$) ce qui fait qu'ils entrent peu en compte dans les diverses applications de ce type de processus.

Par contre dans le domaine de l'environnement les processus sont rarement gaussien et présentent le plus souvent des événements extrêmes ayant une probabilité d'apparition bien plus élevée que pour un processus gaussien : Ce sont des lois à fortes queue de distribution ou "fat tail" en anglais.

Dans le domaine de l'environnement différentes loi de probabilité ont été proposées pour modéliser le comportement asymptotique de la distribution. On trouve par exemple le modèle en exponentielle étirée ("stretched exponential") du type Sor-nette (2000) :

$$f(x) = C_1 \exp(Ax^\beta).$$

Si $\beta = 2$, on retrouve les processus gaussiens. Pour $0 < \beta < 1$, on obtient des "fat tails". On trouve aussi les lois du type Fréchet, appelé aussi lois de Pareto ou lois hyperboliques. Pour ce type de loi, la densité de probabilité s'écrit :

$$f(x) = C_2 x^{-u}, \quad x \gg 1.$$

avec $u > 0$ et C_2 étant une constante. Ce type de distribution mène à des extrêmes très fréquents par la propriété de la fonction puissance. L'indice u caractérise la distribution des événements extrêmes ; plus u est grand, plus les extrêmes sont rares et lorsque u est petit, les valeurs extrêmes sont fréquentes, donnant à la série une apparence "violente" et extrêmement variable. L'estimation du paramètre u peut se faire en représentant la pdf (la fonction de densité de probabilité) en log-log ; une estimation directe par une droite décrivant la pente de la queue de probabilité donne u . Il existe également une méthode quantitative utilisant un estimateur des valeurs extrêmes (Hill, 1975 ; Beirlant et al., 1996).

4.2.2 Densité spectrale

La densité spectrale est un outil mathématique permettant de représenter les différentes composantes spectrales d'un signal et d'en effectuer l'analyse harmonique. Elle renseigne aussi sur la structure temporelle de nos données, telle que l'étude de l'invariance d'échelle. Elle est souvent utilisée dans l'environnement marin en

écologie marine (Legendre et Legendre, 1998, Schmitt et al., 2008). Pour une série $X(t)$, le spectre d'énergie $E_X(f)$ indique comment l'énergie ou la variance de la série est distribuée selon la fréquence f .

L'expression du spectre d'énergie utilise la transformée de Fourier de la fonction d'autocorrélation :

$$E(f) = 2 \int_0^\infty R(\tau) \cos(2\pi f\tau) d\tau$$

où E représente l'énergie, f la fréquence et $R(\tau) = \overline{x(t)x(t+\tau)}$ est la fonction d'autocorrélation. Le spectre d'énergie est représenté généralement en log-log ce qui permet une bonne visualisation de la dépendance en échelle. Le bruit blanc est caractérisé par l'absence de corrélation menant à un spectre plat ($E(f) = cte$). On considère ici l'invariance d'échelle, pour laquelle le spectre d'énergie peut être écrit sous la forme :

$$E(f) \approx f^{-\beta} \quad (4.2)$$

où β est l'exposant spectral caractérisant la zone invariante d'échelle. Nous rappelons que 12 heures d'enregistrement avec un pas d'échantillonnage de 1.6 seconde génère 27, 000 points de mesure, ce qui est assez pour faire une analyse spectrale. La figure 4.3 montre une estimation du spectre d'énergie pour 8 proportions de longueur comprise entre 24, 000 et 55, 000 points pour l'huître numéro 7. Le spectre d'énergie est représenté en log-log ce qui permet une bonne visualisation de la dépendance en échelle. Dans cet exemple, l'exposant spectral est $\beta = 1.54$ (Fig. 4.3). La même analyse a été faite sur 16 bivalves différents pour une période de 13 jours. Pour le paramètre caractérisant la zone invariante d'échelle, nous avons à peu près trouvé la même valeur pour chaque huître $\beta (= 1.60 \pm 0.05)$. Cet exposant spectral est proche de celui de la turbulence, $\beta = 5/3$ comme prédit pour la turbulence homogène (Kolmogorov, 1941). Ceci indique une possible dominance de la turbulence.

4.2.3 Shot noise

Nous rappelons ici les fonctions de structure, qui ont été introduites en turbulence par Kolmogorov (1962). Il s'agit de caractériser les fluctuations d'une série temporelle non stationnaire, mais dont les incréments sont stationnaire (leur moyenne est indépendante du temps).

Pour l'analyse à haute fréquence des micro-fermetures, nous avons considéré les shot noise non linéaire et envisageons deux lois de probabilité :

- Une loi $\phi_1(T)$ représentant les temps T entre 2 chocs successifs ;
- une loi $\rho_2(a)$ représentant les amplitudes à des chocs.

Le modèle que nous considérons ici consiste alors à supposer que les temps entre 2 chocs ($T_i = t_{i+1} - t_i$) et les amplitudes sont des variables aléatoires indépendantes. On cherchera alors à estimer les lois de probabilité de ces variables aléatoires.

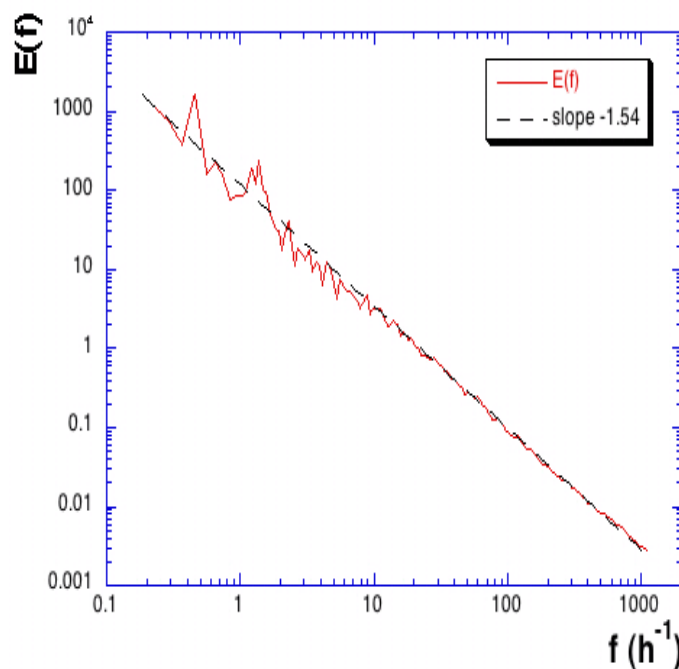


FIGURE 4.3 – Représentation du spectre d' énergie en log-log de l'activité valvaire au cours d'une journée et de la pente ($\beta = -1.54$) pour la comparaison.

Nous avons utilisé les estimateurs récursifs de la densité de type noyau en dimension d ($d = 1$ et $d = 2$) pour estimer les densités de probabilités des intertemps, des amplitudes et de la densité jointe.

Ces résultats montrent d'une part que les probabilités les plus élevées correspondent aux amplitudes les plus faibles et aux intertemps autour de 10 secondes, et une dépendance entre les valeurs logarithmiques de $T(s)$ et $A(mm)$ (corrélation = 0.14, $p < 0.05$). Et d'autre part qu'il y a une possible dominance de la turbulence dans les fluctuations.

- 4.3 Article 4 : Statistical study of bivalve high frequency microclosing behavior : scaling properties and shot noise analysis.** Schmitt, F. G., De Rosa, M., Durrieu, G., Sow, M., Ciret, P., Tran, D., Massabuau, J.C. (2011). *International Journal of Bifurcation and Chaos*. (sous presse).

STATISTICAL STUDY OF BIVALVE HIGH FREQUENCY MICROCLOSING BEHAVIOR: SCALING PROPERTIES AND SHOT NOISE ANALYSIS

François G. Schmitt,^{*} Marie De Rosa
*CNRS, UMR 8187, University of Lille 1, Station Marine, 28 av. Foch
F-62930 Wimereux, France
francois.schmitt@univ-lille1.fr*

Gilles Durrieu,[†] Mohamedou Sow, Pierre Ciret, Damien Tran, Jean-Charles Massabuau
*Université de Bordeaux I, CNRS, UMR 5805 EPOC
33120 Arcachon, France
gilles.durrieu@univ-ubs.fr*

Received (to be inserted by publisher)

We consider bivalve (oyster) behavior using high frequency valvometry data. These data are obtained in the field in the Arcachon bay, with light-weight electrodes of millimeters size, attached to the molluscs, and linked by flexible wires to an electronic control unit which measures valve activity. The latter are able to move freely and open their valve with an amplitude which is linked to environmental and physiological conditions, in order to filter surrounding waters. We consider here daily data recorded every 0.1 s for 16 different animals, corresponding to a total of 1,728,000 data points. Valvometry data show important intermittent activity at high frequency, with frequent and sudden “microclosing” events (meaning partial closures), at apparently random times and with random amplitudes. The overall frequency spectrum of valve opening time series show power-law behavior with a slope $\beta = 1.6 \pm 0.1$. The random succession of closing events with random amplitudes suggests to consider a shot-noise analysis, with two independent stochastic processes representing respectively the times of these discrete shot noise events and their amplitudes. This shows that methods and approaches devoted to extreme values, chaotic dynamics, recurrence time statistics in geosciences and climate studies can also be used to study living organisms and their interactions with complex environments such as their local ecosystem.

Keywords: Scaling; Valvometry; Animal behavior; Shot noise; Intermittency.

1. Introduction

A current estimate is that 23% of the world’s population lives in coastal regions, and that the population density in these regions is globally three times larger than the mean population density (112 people per square km and 44 respectively) [Small & Nicholls, 2003]. Most of the present marine pollution comes from land-based activities; surveillance and protection of the shores is thus becoming more and more

^{*}Corresponding author.

[†]Now at LMAM, University of South Brittany, 56017 Vannes, France

fundamental. New tools to inform people of water conditions are needed. The ability of oysters or other mollusk bivalves to taste their environment is one of the possible ways to monitor the quality of coastal waters. Monitoring their natural valve movements is another way to characterize water quality and read, throughout the year, the health of both the bivalves and their environment.

The interest in investigating the bivalve's activities by recording their valve movements (e.g. valvometry) is not recent. Marceau was certainly pioneer in this method and performed the first recordings of molluscan valve movements more than 100 years ago [Marceau, 1909]. This technique has now been explored in ecotoxicology for more than 20 years. The basic idea of valvometry is to use the bivalve's ability to close its shell when exposed to a contaminant as an alarm signal [Doherty *et al.*, 1987; Shumway *et al.*, 1987; Byrne *et al.*, 1990; Kramer & Botterweg, 1991; Borcharding, 1992; Sluyts *et al.*, 1996; Curtis *et al.*, 2000; Kadar *et al.*, 2001; Nagai *et al.*, 2006]. Nowadays, valvometers are available on the market and use the principle of electromagnetic induction [Sloff *et al.*, 1983; Jenner *et al.*, 1989] such as the Mossel Monitor [Kramer *et al.*, 1989] or the Dreissena Monitor [Borcharding, 1992]. There has been a clear research interest in the recent years to measuring the bivalve's behavior directly in real conditions [Riisgard *et al.*, 2006; Robson *et al.*, 2007; Garcia-March *et al.*, 2008].

In the same line, a novel technique and analytical tools to monitor and analyze quantitatively the bivalve's behavior has been developed recently by some of the present authors. One of our objectives is to model statistically the normal bivalve's activities under field conditions, and detect abnormal patterns due to a pollution such as the presence of toxic algae. For this purpose, since 2003, a new approach has been developed using a noninvasive valvometric technique¹, coupled with the integration of the bivalve respiratory physiology and finally an in-depth statistical analysis of the animal's responses. Two original features have been introduced. First, a new valvometric system was built using light-weight electrodes ($\approx 100mg$) linked by flexible wires to high performance electronic units. This system allows the bivalves to be studied in their natural positions with minimal experimental constraints. Second, a statistical approach based on kernel regression was developed and applied recently with success to evaluate the pros and cons in using a bivalve's closure response as a method of detecting trace amounts of metals [Tran *et al.*, 2003, 2004, 2007; Fournier *et al.*, 2004; Durrieu *et al.*, 2009a,b; Sow *et al.*, 2011]. We consider here the Bay of Arcachon (France), but other sampling sites were considered in the Havannah channel, New Caledonia. These field stations use the light electromagnetic and electronic principles as described in Chambon *et al.* [Chambon *et al.*, 2007].

This work is dedicated to the scientific achievements of Catherine Nicolis. She has studied extremes, recurrence times and complexity in chaotic dynamics, geosciences and climate studies [Balakrishnan *et al.*, 1995, 1997; Nicolis & Nicolis, 2007]. This work and also her work in the field of symbolic dynamics in meteorology [Nicolis *et al.*, 1997] have directly inspired other studies involving recurrences for rainfall fields [Schmitt *et al.*, 1998; Schmitt & Nicolis, 2002], and symbolic dynamics studies for considering swimming states of zooplankton [Schmitt *et al.*, 2006]. The present study is a continuation of these works, applying methods and approaches devoted to extreme values, chaotic dynamics, recurrence time statistics in geosciences, to ecology and animal behavior.

The paper is organized as follows. In Section 2, we briefly present the principle of valvometry and the data collected. In sections 3, scaling analysis of high frequency data is performed, using spectral analysis and structure functions. In section 4 we present a shot noise analysis of high frequency microclosing events. Section 5 is a conclusion and discusses perspectives.

2. Presentation of the data

2.1. *Experimental sites and animal species*

The monitoring site is located in the Bay of Arcachon, France, at the Eyrac pier (Latitude: 44°40 N, Longitude: 1°10 W). Sixteen Pacific oysters, *Crassostrea gigas* measuring from 8 to 10 cm in length were permanently installed on this site. These oysters are all from the same age group (1.5 years old) and come

¹HFNI valvometry, High Frequency, Non Invasive

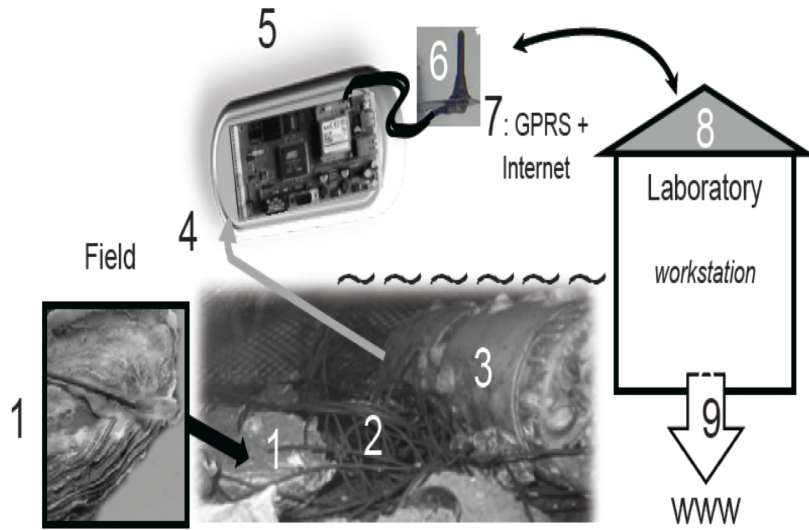


Fig. 1. Synoptic representation of the system, from field to laboratory. (1), oyster equipped with two electrodes; (2), connecting cables from oysters to 1st level electronic card; (3), 1st level electronic card in a waterproof case (immersed); (4), electrical connection between the 1st and the 2nd electronic card; (5), 2nd electronic card (emerged); (6), GPRS antennae; (7), GPRS and internet connection; (8), Marine Biological Station of Arcachon; (9), daily update on internet

from the same local supplier. They also all grew in the Bay of Arcachon. They were placed on the sea bottom and permanently immersed at 3 to 7 meters deep in the water, depending on the tide activity.

2.2. Measurements of valve activity

The electronic principle was first described by Tran et al. [Tran *et al.*, 2003] and then modified as described in Chambon et al. [Chambon *et al.*, 2007]. The modification was necessary for field application in the ocean and was purely mechanical. It involved the hardening of the monitoring system to allow its constant immersion in seawater, even in strong water current conditions, and eventually during storms². The main challenge was to ensure the complete autonomy of the system without in-situ human intervention for at least one full year. Each animal is equipped with two light coils (sensors), ≈ 53 mg each (unembedded), fixed on the edge of each valve. These coils measure $2.5 \times 2.5 \times 2$ mm and were coated with a resin-sealing before fixation on the valves. One of the coils emits a high-frequency, sinusoidal signal which is received by the other coil. In each group of animals, one measurement was received every 0.1 s (10 Hz) on one of the sixteen animals. This means that each animal's behavior was measured every 1.6 seconds. Every day, a data set with 864,000 pairs of values (1 distance value, 1 stamped time value) was generated representing a total of 1,728,000 triplet of values (1 distance, 1 stamped time value, 1 animal number). Each oyster's behavior was described by a total of 108,000 measurements per day. The strength of the electric field produced between the two coils decreases with the distance according to the transformation $1/D$, where D is the distance between the point of measurement and the center of the transmitting coil. The measured signal (D_{mv} in mV) is converted into distance in millimeters (D_{mm}) using the following calibration model:

$$D_{mv} = \frac{151}{D_{mm}^{0.35}} - 1.48. \quad (1)$$

The model parameters were estimated using least-squares estimators and the model was very significant ($R^2 = 0.99$, P-value < 0.001).

²Some information about these specific aspects can be found on <http://www.domino.u-bordeaux.fr/molluscan-eye>

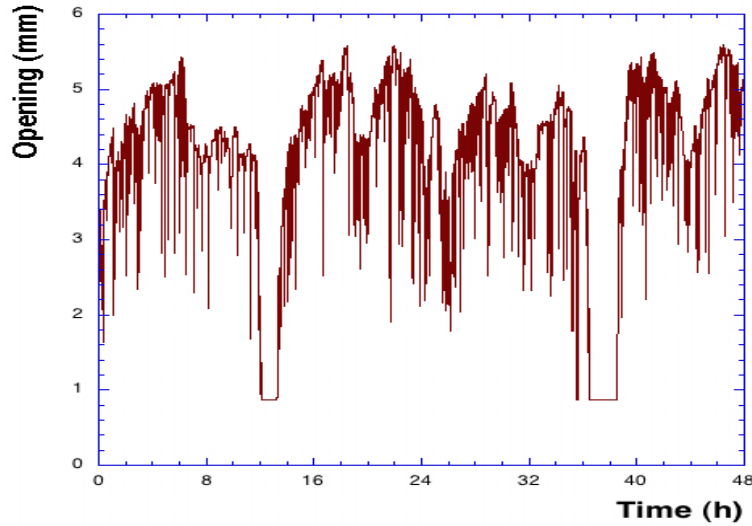


Fig. 2. A typical opening time series during two days, showing the high frequency activity. Long horizontal segments correspond to closing periods.

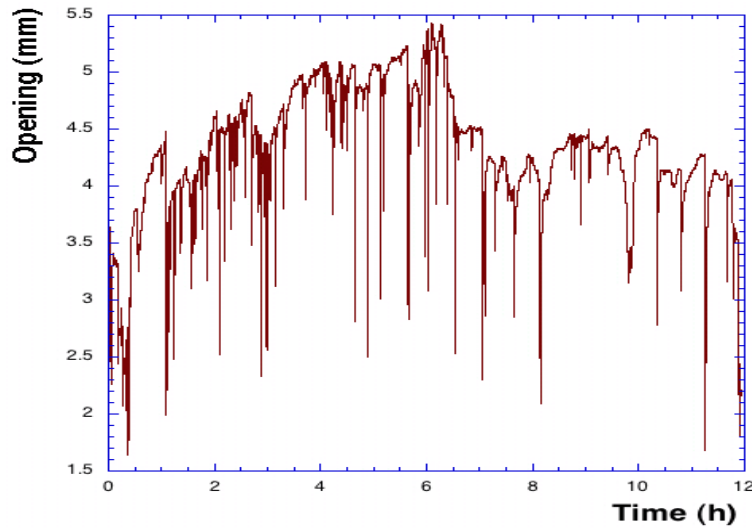


Fig. 3. Zoom on a short portion of the opening time series, showing several microclosing events.

2.3. Data collection and data transmission

On the shore (i.e. by a pier or lighthouse), a second electronic unit handles the data acquisition and the programmed emission. The unit is equipped with a GSM/GPRS modem and uses Linux operating system for driving the first control unit, managing the data storage with a time stamp, accessing the internet, and transferring the data. A self-developed software module runs on mobile phone technology. After finishing a working day (or any programmed experimental duration), the collected data are transmitted to a central workstation server located in the Marine Biological Station in Arcachon, France. The valve activity data are stored in a central database in a merged data set. Access to this database is provided via an Intranet or Internet connection (Fig. 1).

3. Scaling properties

We consider here a data set consisting in valvometry data recorded for 16 oysters in the field, each one recorded during 13 days in May 2007. These data present a very high variability, as shown in Fig. 2 for a

typical bivalve, shown during 48h. Horizontal sections correspond to periods when the bivalve was closed: since the electrodes are fixed on the shelf, the minimum distance when bivalves are closed is not zero; this has no influence for the remaining of the study, since we consider below only portions of the time series with high microclosing activity; closed portions are removed (for a study of full closing duration in the framework of a pollution event, see [Sow *et al.*, 2011]). Here we are interested in high frequency fluctuations, for scales between 1.6s and several hours. Figure 3 shows a zoom of a shorter portion of the data, for a total duration of 12 hours: in the following we study only these portions. Let us note that in sites with weak tides, the 24h solar cycle is a major drive, whereas in sites with large or strong tides and strong currents, the tidal cycle is more important than the solar cycle. This is discussed in [Sow *et al.*, 2011] and [Tran *et al.*, 2011].

Several microclosing events are illustrated in this Figure. In a series of 13 days there are a large number of such events: we perform ensemble averages on all these events.

3.1. Scaling power spectral densities

We note that 12 hours of data at a sampling resolution of 1.6s generates 27,000 data points, which is more than enough to perform Fourier power spectral analysis. We estimate the Fourier power spectrum as the Fourier transform of the autocorrelation function of the opening time series $E(f) = \int_0^{+\infty} C(\tau) \cos(2\pi f\tau) d\tau$, where $C(\tau) = \langle (X(t) - \mu)(X(t - \tau) - \mu) \rangle$ is the autocorrelation function of the opening series X . This is shown in Figure 4 for one typical individual oyster, where the power spectrum is estimated for 8 sections of length from 24,000 to 55,000 data, representing a total of 333,732 data points. This spectrum is represented in log-log plot. In this figure, for display convenience, we performed band-average of high frequencies, in order to have a smoother curve and to better reveal scaling ranges. A quite remarkable result is obtained, with a scaling power spectra of the form:

$$E(f) \approx f^{-\beta} \quad (2)$$

where $E(f)$ is the energy density, f the frequency and β the power spectral exponent. In this example, the value of the spectral exponent is $\beta = 1.54$ ($R^2 = 0.99$). We have performed such analysis on 16 different animals, for a period of 13 days. We found a quite universal scaling behavior for all spectra, with a mean power law slope estimator of $\beta (= 1.60 \pm 0.05)$, where the error bar comes from statistics on power law slopes estimated on each oyster.

This very nice scaling property, found for scales between 1.6 s and about 1 hour, spans over 3 decades in frequency space. It seems possible that the driving force behind these series of microclosing events is external to the animal. There seems to be only one forcing mechanism presenting such scaling property: turbulence. This idea is enforced by the fact that the estimated scaling exponent which is found is relatively close to the scaling exponent of turbulence [Kolmogorov, 1941], $\beta = 5/3$. Let us note here that passive scalar turbulence is also associated to a -5/3 power spectrum [Obukhov, 1949; Corrsin, 1951]. There are therefore two different possible mechanisms that could explain these microclosing events: fluctuations in velocity, or fluctuations in some chemical quantity, being a passive or active scalar [Monin & Yaglom, 1975]. Here we cannot differentiate between velocity or scalar turbulence as a driving mechanism of microclosing events. However, since both velocity and passive scalar time series possess multifractal statistics [Schmitt *et al.*, 1992; Frisch, 1995; Schmitt, 2005, 2006], we perform below multifractal analysis of opening time series.

3.2. Structure functions analysis and the search for multifractal properties

Scaling properties in the spectral space may be also associated to scaling structure functions. Here we perform structure functions analysis, an analysis technique which is now classical in the turbulence literature [Monin & Yaglom, 1975; Frisch, 1995]. Structure function analysis is a way to characterize the intermittency of scaling time series: the structuration of huge fluctuations, at all scales. For a non-stationary time series $X(t)$ with stationary increments, we note the fluctuations amplitude $\Delta_\tau X = |X(t + \tau) - X(t)|$ and we consider the structure functions of order q as the q -moment of the increments:

$$\langle (\Delta_\tau X)^q \rangle = C_q \tau^{\zeta(q)} \quad (3)$$

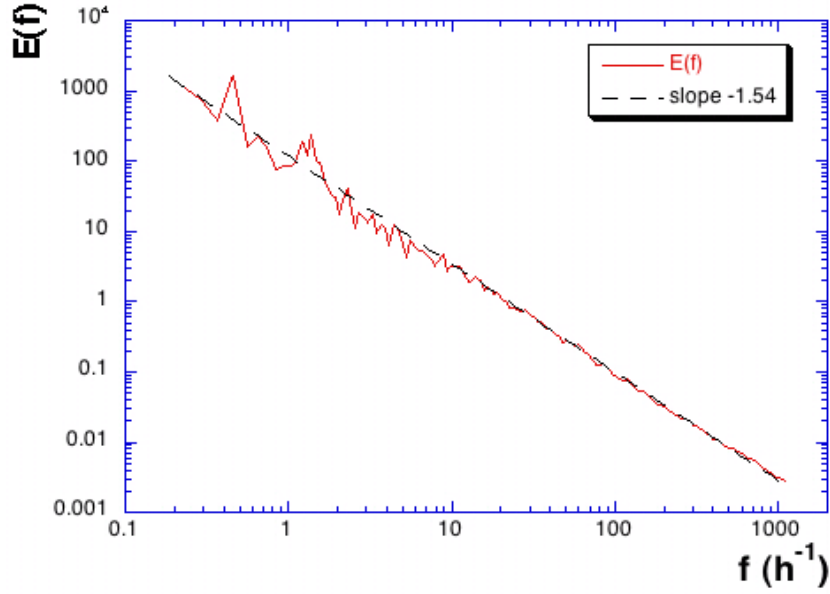


Fig. 4. Spectral energy density of a typical bivalve opening time series recorded during one day. The representation is in log-log and the very nice scaling which is obtained is found on more than 3 orders of magnitudes in the frequency range. A straight line of slope -1.54 is shown for comparison.

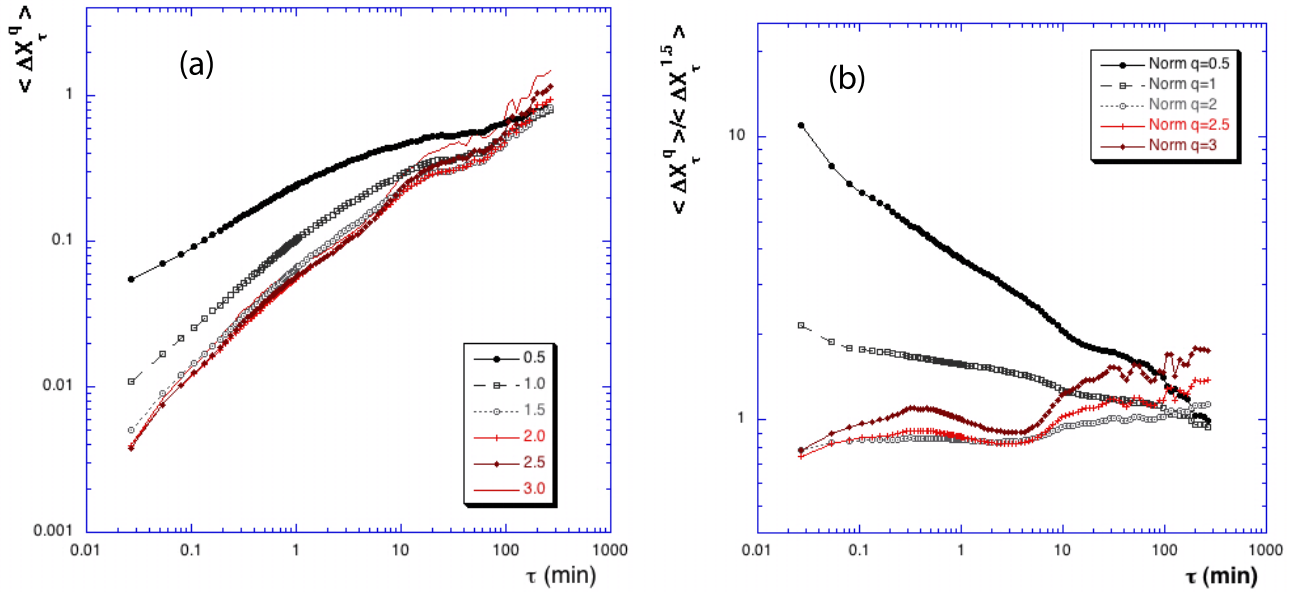


Fig. 5. (a) Log-log plot of the structure functions $\langle (\Delta_\tau X)^q \rangle$ versus τ for $q = 0.5$ to 3 . The straight line parts indicate the scaling region. (b) The same, normalized by the structure function of order 1.5 : this shows that for moments larger than 1.5 , there is a saturation and the slopes are almost constant.

where $\langle . \rangle$ means statistical average, C_q is a constant and $\zeta(q)$ a moment-scaling function which characterizes the scaling properties of the original time series. Due to the Wiener-Khinchine theorem, there is the relation $\beta = 1 + \zeta(2)$ showing that the power spectral exponent is a second order moment which characterizes “medium” fluctuations. Large order moments (i.e. larger than 3) correspond to extreme events.

The moment scaling function is useful to estimate which scaling process is compatible with the data: for Brownian motion, $\zeta(q) = q/2$ and for fractional Brownian motion of order H , $\zeta(q) = qH$. When $\zeta(q)$ is linear the process is said to be monofractal and for multifractal time series, it is nonlinear and concave.

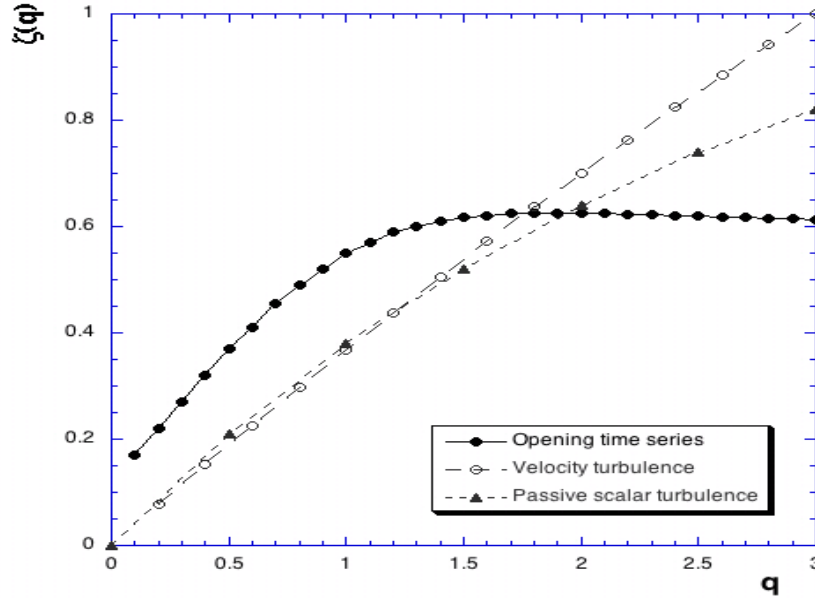


Fig. 6. Values of $\zeta(q)$ estimated for scales between 1.6 s and 10 min. The shape of this moment function is nonlinear, with a saturation for moments larger than 1.5. It is rather different from the moment function corresponding to velocity and passive scalar turbulence, which are shown as dotted lines for comparison (data from compilations done in [Schmitt, 2005, 2006])

For example for a scaling lognormal process, $\zeta(q)$ is quadratic. For Lévy process of order α ($0 \leq \alpha \leq 2$) the function is bilinear [Schmitt *et al.*, 1999; Nakao, 2000]: $\zeta(q) = q/\alpha$ for $q \leq \alpha$ and $\zeta(q) = 1$ for $q \geq \alpha$.

This analysis has been performed on opening time series. As expected, scaling is found for various order of moments. This is shown in Figs. 5a and 5b. Figure 5a shows the structure functions $\langle (\Delta_\tau X)^q \rangle$ versus q for $q = 0.5$ to 3 in log-log plot. Power-law is found from the smallest available scale (1.6 s) to about 10 minutes. For moments larger than 1.5 the curves seem to superpose rather well; we thus have shown in Fig. 5b the same curves normalized by the moment of order 1.5: the flat curves obtained illustrate the saturation of $\zeta(q)$ for moments larger than this value. This is also visible in Fig. 6, representing the scaling moment function $\zeta(q)$ and its saturation for moments larger than 1.5.

These results confirm the scaling properties of the opening time series, with some multifractal properties for low orders of moments. However, the comparison in Fig. 6 of $\zeta(q)$ for opening data with velocity and passive scalar turbulence (data from laboratory studies, compilations done in [Schmitt, 2005, 2006]), shows that the curves are very different. The values obtained are similar only for moments close to 2, a result in agreement with what was obtained from power spectral analysis. This different form of the function indicate that there may be some turbulence influence in the opening time series, but that the fluctuations have nevertheless quite different statistics. In fact, the peculiar patterns of the opening time series as can be seen in Figs. 2 and 3, suggest shot noise analysis, as we discuss in the next section.

4. Shot noise analysis

4.1. Nonlinear shot noises

Figure 3 shows that the opening time series is composed of many sudden microclosing events separated by apparent random times. This suggests the introduction of a shot noise analysis. A shot noise process is a continuous time process subject to discrete shocks having random amplitudes. In the classical formulation, impulses are derived from a Poisson distribution and thus have exponential distribution for interevent interval pdf [Lowen & Teich, 2005; Eliazar & Klafter, 2006]. A more general formulation has been called nonlinear shot noise, for which interevent times and shocks magnitudes have two (independent or not) different probability density functions (pdf) [Eliazar & Klafter, 2005, 2006].

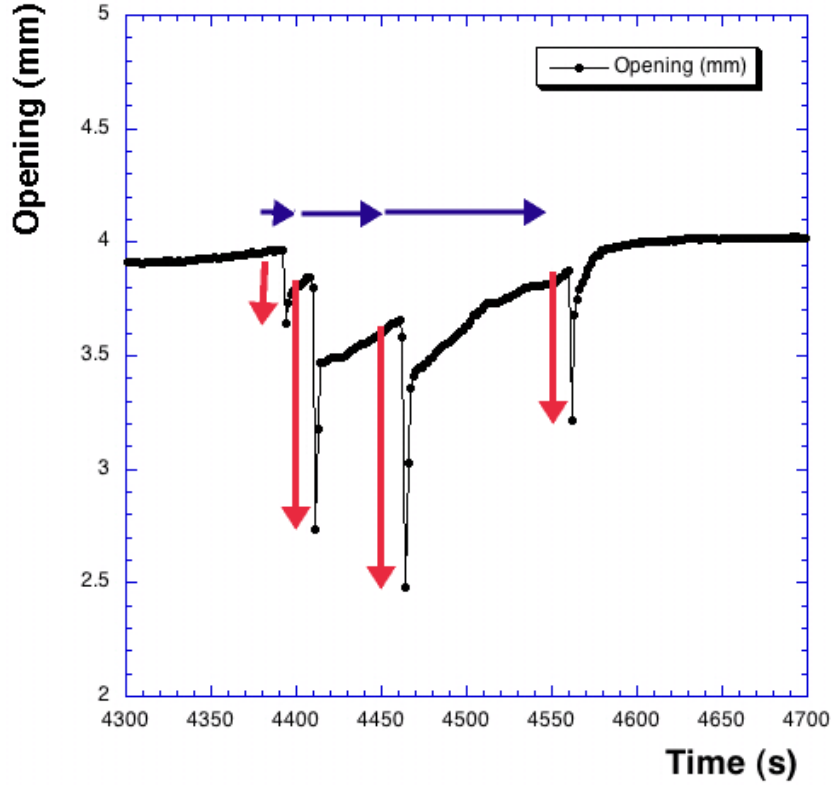


Fig. 7. Illustration of the shot noise modelling of the microclosing dynamics: 4 shocks and 3 interevent times are shown here. Red arrows show the amplitudes of the shocks and blue arrows the interevent intervals.

For this nonlinear shot noise modeling of high frequency microclosing dynamics, we thus propose to consider two experimental pdf:

- The interevent time pdf $\phi(T)$ estimated for $T_i = t_i - t_{i-1}$ where (t_i) are the discrete shock times. See Figure 7 for an illustration of a typical sequence.
- The amplitude pdf $\rho(A)$ corresponding to the total amplitude of the microclosing event: see Figure 7 for an illustration for 4 shocks.

The identification of micro-closing events was done using a simple algorithm: since these events can typically have a duration of a few time steps, we considered a threshold associated to noise (a decrease larger than 0.1 mm) and for a decrease larger than this threshold, the duration of the microclosing event is the time until the next increase in the series; the inter-time is the time between a negative peak and the beginning of the next event; the amplitude is the distance between the beginning and the end of a micro-closing event.

With such experimental pdf, it is then possible to realize a stochastic simulation using a relaxation function providing the evolution of the closing value after each shock: this relaxation function indicates how fast the series comes back to an equilibrium value after each shock. It is often assumed to be exponential. Here we do not estimate this relaxation function and focus on the two pdf mentioned above.

4.2. *Probability density function estimator*

To determine a general profile of valve activities during a given period of time, we estimate here the probability density of the time intervals between microclosing events and the amplitude of micro-closing events. In order to correctly estimate these probability density functions, we use a nonparametric kernel

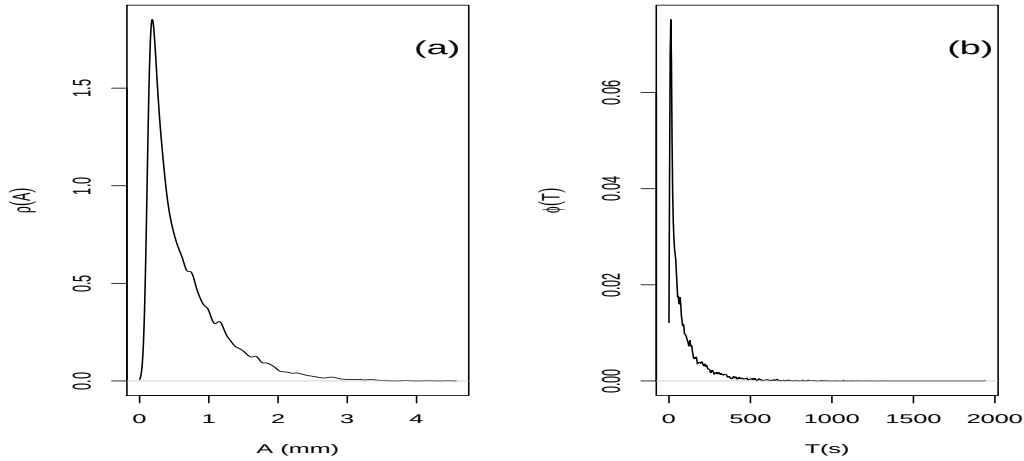


Fig. 8. Kernel density estimator of the valve opening activity of one individual oyster (a) with $\hat{h}_n = n^{-0.290}$ and of the inter-times events (b) with $\hat{h}_n = n^{-0.191}$ considering a 12 days period. The shape of the distribution is similar using Wolverton-Wagner or Deheuvels's estimates.

density estimator. A kernel estimator [Wand & Jones, 1995] of a function f is defined by :

$$\hat{f}_{h_n}(t) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{t - T_i}{h_n}\right), \quad (4)$$

where n is the sample size and h_n is the bandwidth (the smoothing scale or smoothing parameter) satisfying $(C1) h_n \rightarrow 0$ and $n h_n \rightarrow \infty$, as $n \rightarrow \infty$. The kernel K is a symmetric function satisfying $(C2) \int K(t) dt = 1$, $\int t K(t) dt = 0$, $\int |t| K(t) dt < +\infty$ and the distribution function f is twice differentiable $(C3)$. We introduce the notation $\int K^2(t) dt = \tau^2$. Under conditions $(C1) - (C3)$, the bias tends to zero and the kernel density estimator $\hat{f}_{h_n}$ is a consistent estimate of f asymptotically normal distributed as $n \rightarrow \infty$:

$$\sqrt{n h_n} \left(\hat{f}_{h_n}(t) - f(t) \right) \xrightarrow{D} N(0, \tau^2 f(t)). \quad (5)$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 . It is also possible and often preferred to use a recursive kernel density estimator given by :

$$\hat{g}_n(t) = (1 - \gamma_n) \hat{g}_{n-1}(t) + \frac{\gamma_n}{h_n} K\left(\frac{t - T_n}{h_n}\right), \quad (6)$$

where (γ_n) and (h_n) are two positive sequence of number converging to 0. For instance, it is possible to use the Wolverton-Wagner estimate or the Deheuvels's estimate ([Barty *et al.*, 2007] and [Mokkadem *et al.*, 2009]). The bandwidth estimator has been determined by cross-validation (CV). The objective is to find h_n minimizing the CV function given by $CV(h_n) = \int \hat{f}_{h_n}^2(x) dx - \frac{2}{n} \sum_{j=1}^n \hat{f}_{h_n,j}(T_i)$, where $\hat{f}_{h_n,j}(x)$ is the kernel density estimate given by (4) with the j th observation deleted from the sample. The statistical properties of $\hat{h}_n$ have been studied and the bandwidth estimator is consistent and asymptotically normal [Hall *et al.*, 1987]. We represent in Figs 8(a-b) the kernel density estimators for the valve opening activity and the inter-times events of only one individual oyster. It is worth noting that similar density plots were also obtained with the other oysters. The estimated standard error of density estimates considering the 16 oysters on this period 13 days are respectively 0.0034 for the amplitude and 0.0026 for the inter-times microclosing events.

The previous definition of univariate kernel estimator can be easily generalized to the multivariate case. For instance, the d -multivariate recursive kernel density estimator with kernel $\mathbf{K}$ and window width

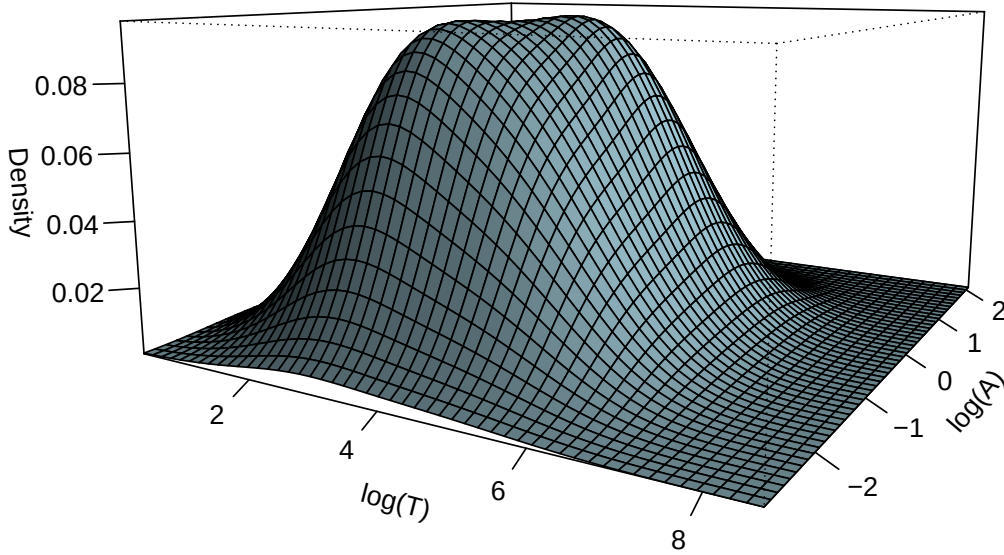


Fig. 9. One individual recursive two-dimensional ($d = 2$) density estimate for the logarithmic transform of valve opening activity and inter-times events considering a 12 days period with $\hat{h}_n = n^{-0.055}$.

h_i is defined by

$$\hat{f}_{h_n}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} \mathbf{K} \left(\frac{\mathbf{x} - \mathbf{X}_i}{h_i} \right). \quad (7)$$

The kernel function $\mathbf{K}$ is now a function, defined for d -dimensional $\mathbf{x}$ (d is the dimension parameter), satisfying $\int_{\mathbb{R}^d} \mathbf{K}(\mathbf{x}) d\mathbf{x} = 1$.

4.3. Experimental estimates of the two probability density functions

The method presented above has been used to estimate the probability density functions $\phi(T)$ and $\rho(a)$, shown in Figs. 8 (a-b). They show clearly that the probability to obtain the lowest amplitude is the highest and small inter-times events (around 10 seconds) occurs with the maximum probability. This can be also estimated in $2d$, as shown in Fig. 9. Fig. 9 reveals also a small but significant dependence between the logarithmic transform of $T(s)$ and $A(mm)$ (correlation = 0.14, $p < 0.05$). Nonlinear shot noise processes can generate scaling statistics when the intertime or amplitude pdf has a power-law tail; in this case the process has been called fractal shot noise [Lowen & Teich, 2005]. In order to consider this, we plot in Figs 10(a-b) the average probability density functions $\phi(T)$ and $\rho(A)$ in log-log plot. Power-law tails are shown as very significant linear fit $R^2 = 0.95$ for the extremes oysters amplitudes and $R^2 = 0.92$ for the extremes inter-times microclosing ($p < 2.2e-16$). This shows that power-law tails for these probability density functions are found, with exponents $\mu = 2.32$ and $\mu = 2.95$ for the interval and amplitudes respectively. For the amplitudes the power-law tail begins at an amplitude value of about 1 – 2 mm; this characteristic amplitude value could be associated to a change of regime.

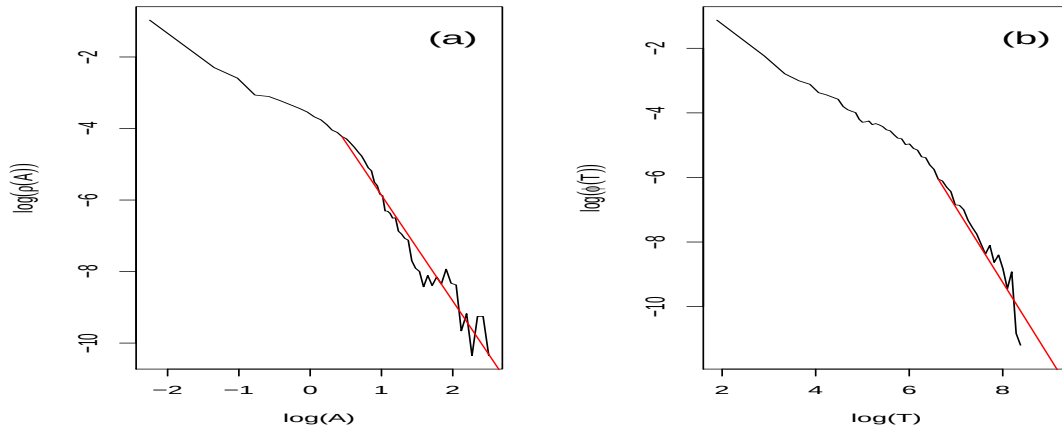


Fig. 10. Log-log plots of the average density oysters amplitudes (a) and of the average probability density functions of the oysters inter-times microclosing (b). Power-law fits are shown for the extremes of these probability density functions with exponents $\mu = 2.32$ and $\mu = 2.95$ for the interval and amplitudes respectively.

5. Conclusion

Our article presents a general statistical framework in the context of the survey of aquatic systems using high frequency and non invasive valvometry. We show that these data present a high variability with frequent and sudden microclosing events. The overall frequency spectrum of valve opening time series show power-law behavior with a spectral slope very close to turbulence, but with multifractal scaling moment function quite different from passive scalar or velocity turbulence. The random succession of closing events (with random amplitudes) allows us to consider a shot-noise and fractal shot noise modeling, with the generation of random variables with a pdf give by the experimental results, in order to obtain two stochastic processes representing respectively the times of these discrete shot noise events and their amplitudes. We plan to study by simulation the behavior of shot noise model to assess specifically the sensitivity of the model to two factors: (1) the independence assumption between the time and amplitude stochastic processes and the effect of different levels of dependence (different values of correlation) on the results and (2) the effect of specific choice of relaxation function. Here, we do not estimate the relaxation function but we consider non parametric kernel estimators of the two probability density functions. We think that this statistical approach provides a very useful tool for quantifying and characterizing the behaviors of oysters directly exposed to their natural environment and exposed to change in the water quality of their natural environment and so this approach could have an important contribution to the survey of aquatic systems.

Let us note here that we have compared multifractal properties of the opening time series with fully developed velocity or passive scalar turbulence, with data coming from laboratory studies. We plan in the future to install *in situ*, close to the oysters, high frequency velocity and temperature measuring devices in order to compare their scaling properties and detect some possible correlations between micro closing events and turbulent bursts. In the near future, we plan also to use these techniques to detect harmful algae contamination considering the fact that toxic algae, when absorbed by mollusks rejects amnesic toxins (domoic acid, ASP) in their flesh that could affect the locomotor activity of oysters. In this case we expect that our modeling here using shot noises could help to detect some changes in the pdf of amplitudes or inter-times micro-closings. Thus such modeling could contribute to producing new indicators in an ecotoxicological framework.

Acknowledgements

We thank Ifremer Arcachon, and more specifically Isabelle Auby and Danièle Maurer, for providing informations about the presence of harmful algae in the Bay of Arcachon. M. Sow was supported by a grant

from the Aquitaine Region and the Samuel Lunenfeld Research Institute of the University of Toronto. We also thank Arnaud Hequette for providing us a reference and the anonymous referees for their useful suggestions.

References

- Balakrishnan, V., Nicolis, C. & Nicolis, G. [1995] "Extreme value distributions in chaotic dynamics," *J. Stat. Phys.* **80**, 307–336.
- Balakrishnan, V., Nicolis, G. & Nicolis, C. [1997] "Recurrence time statistics in chaotic dynamics I. Discrete time maps," *J. Stat. Phys.* **86**, 191–212.
- Barty, K., Roy, J.S. & Strugarek, C. [2007] "Un usage de l'approximation stochastique pour l'estimation recursive," *C. R. Acad. Sci. Paris* **1344**, 199–204.
- Borcherding, J. [1992] "Another early warning system for the detection of toxic discharges in the aquatic environment based on valve movements of the freshwater mussel *Dreissena polymorpha*," *The zebra mussel dreissena polymorpha: Ecology, Biological Monitoring and First Applications in the Water Quality Management*, eds. Neumann, D. & Jenner, H.A., (Gustav Fischer Verlag, New York), pp. 127–146.
- Byrne, R., Gnaiger, E., McMahon, R. & Dietz, T. [1990] "Behavioral and metabolic responses to emersion and subsequent reimmersion in the freshwater bivalve *Corbicula fluminea*," *Biol. Bull.* **178**, 251–259.
- Chambon, C., Legeay, A., Durrieu, G., Gonzalez, P., Ciret, P. & Massabuau, J.C. [2007] "Influence of the parasite worm *Polydora* sp. on the behaviour of the oyster *Crassostrea gigas*: a study of the respiratory impact and associated oxidative stress," *Mar. Biol.* **152**, 329–338.
- Corrsin, S. [1951] "On the spectrum of isotropic temperature fluctuations in an isotropic turbulence," *J. Appl. Phys.* **22**, 469–1951.
- Curtis, T.M., Williamson, R. & Depledge, M.H. [2000] "Simultaneous, long-term monitoring of valve and cardiac activity in the blue mussel *Mytilus edulis* exposed to copper," *Mar. Biol.* **136**, 837–846.
- Doherty, F.G., Cherry, D.S. & Cairns, J. [1987] "Valve closure responses of the Asiatic clam *C. Fluminea* exposed to cadmium and zinc," *Hydrobiologia* **153**, 159–167.
- Durrieu, G., Nguyen, T.M.N. & Sow, M. [2009] "Comparaison d'estimateurs de regression non paramétriques : application en valvometrie," <http://hal.inria.fr/docs/00/38/67/16/PDF/p147.pdf>.
- Durrieu, G. & Briollais, L. [2009] "Sequential design for microarray experiments," *J. Am. Stat. Assoc.* **104(86)**, 650–660.
- Eliazar, I. & Klafter, J. [2005] "On the nonlinear modeling of shot noise," *Proc. Natl. Acad. Sci. U.S.A.* **102(39)**, 13779–13782.
- Eliazar, I. & Klafter, J. [2006] "Non-linear Shot Noise: Lévy, Noah, & Joseph," *Physica A* **360**, 227–260.
- Fournier, E., Tran, D., Denison, F., Massabuau, J.C. & Garnier-Laplace, J. [2004] "Valve closure response to uranium exposure for a freshwater bivalve *Corbicula fluminea*: quantification of the influence of pH," *Envir. Tox. Chem.* **23**, 1108–111.
- Frisch, U. [1995] *Turbulence; the legacy of A. N. Kolmogorov*, (Cambridge University Press, Cambridge).
- Garcia-March, J.R., Sanchis Solsona, M.A. & Garcia-Carrascosa, A.M. [2008] "Shell gaping behavior of *Pinna nobilis* L., 1758: circadian and circalunar rhythms revealed by in situ monitoring," *Mar. Biol.* **153**, 689–698.
- Hall, P. & Marron, J.S. [1987] "Extent to which least-squares cross-validation minimises integrated square error in non parametric density estimation," *Probab. Thory Rel. Fields* **74**, 567–581.
- Jenner, H.A., Noppert, F. & Sikking, T. [1989] "A new system for the detection of valve movement response of bivalves," *Kema Sci. Techn. Rep.* **7(2)**, 91–98.
- Kadar, E., Salanki, J., Jugdaohsingh, R., Powell, J.J., McCrohan, C.R. & White, K.N. [2001] "Avoidance responses to aluminium in the freshwater bivalve *Anodonta cygnea*," *Aquat. Toxicol.* **55**, 137–148.
- Kolmogorov, A. N. [1941] "The local structure of turbulence in incompressible viscous fluid for very large Reynolds' numbers," *C. R. Acad. Sci. URSS* **30** 301–305.
- Kramer, K.J.M., Jenner, H.A. & De Zwart, D. [1989] "The valve movement response of mussels: a tool in biological monitoring," *Hydrobiologia* **188/189**, 433–443.

- Kramer, K.J.M. & Botterweg, J. [1991] "Aquatic biological early warning systems: An overview," *Bioindicators and environmental management*, eds. Jeffrey, D.W. & Madden, B., (Academic Press, London, pp. 95–126.
- Lowen, S.B. & Teich, M.C. [2005] *Fractal-Based Point Processes*, (Wiley, reprint 2007).
- Marceau, F. [1909] "Contraction of molluscan muscle," *Arch. Zool. Exp. Gen.* **2**, 295–469.
- Mokkadem, A., Pelletier, M. & Slaoui, Y. [2009] "The stochastic approximation method for the estimation of a multivariate probability density," *J. Statist. Plann. Inference* **139**, 2459–2478.
- Monin, A.S. & Yaglom, A.M. [1975] *Statistical Fluid Mechanics; Mechanics of Turbulence, vol. II*, (Dover, reprint 2007).
- Nagai, K., Honjo, T., Go, J., Yamashita, H. & Oh, S.J. [2006] "Detecting the shellfish killer *Heterocapsa circularisquama* (Dinophyceae) by measuring the bivalve valve activity with Hall element sensor," *Aquaculture* **255**, 395–401.
- Nakao, H. [2000] "Multi-scaling properties of truncated Lévy flights," *Phys. Lett. A* **266**(7), 282–289.
- Nicolis, G., & Nicolis, C. [2007] *Foundations of Complex Systems; Nonlinear Dynamics, Statistical Physics, Information and Prediction*, (World Scientific, 2007).
- Nicolis, C., Ebeling, W. & Baraldi, C. [1997] "Markov processes, dynamic entropies and the statistical prediction of mesoscale weather regimes," *Tellus* **49A**, 108–118.
- Obukhov, A. M. & Akad. Izv. [1949] "Structure of the temperature field in turbulent flows," *Nauk. SSSR Geogr. Geofiz.* **13**, 58.
- Riisgard, H.U., Lassen, J. & Kittner, C. [2006] "Valve-gape response times in mussels (*Mytilus edulis*): effects of laboratory preceding feeding conditions and in situ tidally induced variation in phytoplankton biomass," *J. Shellfish Res.* **25**, 901–911.
- Robson, A., Wilson, R. & Garcia de Leaniz, C. [2007] "Mussels flexing their muscles: a new method for quantifying bivalve behavior," *Mar. Biol.* **151**, 1195–1204.
- Schmitt, F. G. [2005] "Relating Lagrangian passive scalar scaling exponents to Eulerian scaling exponents in turbulence," *Eur. Phys. J. B* **48**, 129–137.
- Schmitt, F. G. [2006] "Linking Eulerian and Lagrangian structure functions scaling exponents in turbulence," *Physica A*, **368**, 377–386.
- Schmitt, F. G., D. Lavallée, S. Lovejoy, D. Schertzer, and C. Hooge [1992] "Estimations directes des indices de multifractals universels dans le champ de vent et de température," *C. R. Acad. Sci. ser. II* **314**, 749–754.
- Schmitt, F. G. & Nicolis, C. [2002] "Scaling of return times for a high resolution rainfall time series," *Fractals* **10**, 285–290.
- Schmitt, F., Schertzer, D. and Lovejoy, S. [1999] "Multifractal analysis of foreign exchange data," *Appl. Stoch. Mod. Data Anal.* **15**, 29–53.
- Schmitt, F. G., Seuront, L., Hwang, J.-S., Souissi, S. & Tseng, L.C. [2006] "Scaling of swimming sequences in copepod behaviour: data analysis and simulation," *Physica A*, **364**, 287–296.
- Schmitt, F. G., Vannitsem, S. & Barbosa, A. [1998] "Modeling of rainfall time series using two-state renewal processes and multifractals," *J. Geophys. Res.* **103**(D18), 23181–23193.
- Shumway, S.E. & Cucci, T.L. [1987] "The effects of the toxic dinoflagellate *protogonyaulax tamarensis* on the feeding and behavior of bivalve mollusks," *Aquat. Toxicol.* **10**, 9–27.
- Sloff, W., De Zwart, D. & Marquenie, J.M. [1983] "Detection limits of a biological monitoring system for chemical water pollution based on mussel activity," *Environ. Contam. Tox.* **30**(4), 400–405.
- Sluyts, H., Van Hoof, F., Cornet, A. & Paulussen, J. [1996] "A dynamic new alarm system for use in biological early warning systems," *Environ. Toxicol. Chem.* **15**, 1317–1323.
- Small, C. and Nicholls, R.J. [2003] "A global analysis of human settlement in coastal zones," *J. of Coastal Res.* **19**, 584–599.
- Sow, M., Durrieu, G., Briollais, L., Ciret, P. & Massabuau, J.C. [2011] "Water quality assessment by means of HFNI valvometry and high-frequency data modeling," *Environmental Monitoring and Assessment*, in press.
- Tran, D., Ciret, P., Ciutat, A., Durrieu, G. & Massabuau, J.C. [2003] "Estimation of potential and limits of bivalve closure response to detect contaminants: application to cadmium," *Environ. Toxicol. Chem.*

22, 116–122.

- Tran, D., Fournier, E., Durrieu, G. & Massabuau, J.C. [2004] “Copper detection in the asiatic clam *Corbicula fluminea*: optimum valve closure response,” *Aquat. Toxicol.* **65**(3), 317–327.
- Tran, D., Fournier, E., Durrieu, G. & Massabuau, J.C. [2007] “*Corbicula fluminea* valve closure response as sensitivity threshold to inorganic mercury contamination,” *Environ. Toxicol. Chem.* **26**(7), 1545–1551.
- Tran, D., Nadau, A., Durrieu, G., Ciret, P., Parisot, J.P. & Massabuau, J.C. [2011] “Field chronobiology in molluscan bivalves: how moon and sun cycles interactions deeply drive oyster activity rhythms,” *Chronobiol. Int.* , in revision.
- Wand, M.P. & Jones, M.C. [1995] *Kernel Smoothing*, (London, Chapman and Hall).

PARTIE 2. Développement de méthodes statistiques pour l'analyse de liaison génétique

Dans cette partie, nous nous intéressons au développement et à la comparaison d'estimateurs robustes dans le domaine de la génétique humaine, plus particulièrement à l'analyse de liaisons génétiques. Cette partie se compose de deux chapitres. Dans le chapitre 5, nous présentons les différentes méthodes étudiées dans la littérature, puis nous nous intéressons aux estimateurs robustes développés dans cette thèse et comparons les différents estimateurs sur des données réelles de QTL et eQTL. Enfin, le chapitre 6 s'intéresse au quantile de régression en utilisant des copules spécifiant les dépendances entre les variables. Une application sur des données eQTL montre l'intérêt pratique de cette approche. Ce chapitre fait l'objet d'un article publié dans le proceeding de l'American Statistical Association JSM (joint statistical meeting).

Chapitre 5

Approches robustes pour l'analyse de paires de germains : application à l'analyse de QTL et eQTL

5.1 Introduction

L'analyse de liaison génétique ("linkage analysis") vise à identifier et localiser sur le génome des régions chromosomiques contenant un ou plusieurs gènes d'intérêt, soit en étudiant un nombre restreint de gènes candidats ou de régions candidates, soit en criblant le génome entier. Le but du statisticien est de définir la stratégie permettant de modéliser l'ensemble des informations disponibles de manière optimale dans le cadre d'un modèle choisi. Jusqu'à tout récemment, deux types d'approches ont été successivement considérées pour l'analyse de liaison génétique (paires de germains). La régression Haseman-Elston (HE), basée sur un modèle de régression linéaire simple (Haseman et Elston, 1972), teste si les germains (frères/soeurs) qui se ressemblent pour le trait quantitatif se ressemblent également pour le génotype au niveau du marqueur. Le principe de la méthode HE consiste à régresser pour une paire de germains donnée la différence phénotypique quadratique sur la proportion d'allèles partagés par la paire de germains IBD (Identical By Descent). Sous l'hypothèse nulle d'absence de liaison entre le locus testé et le phénotype étudié, la pente de la régression est nulle. Sous l'hypothèse alternative de liaison entre le locus et le phénotype, la pente de régression est négative puisque des valeurs phénotypiques proches (et par conséquent des petites différences quadratiques) sont associées à une plus grande proportion d'allèles partagés IBD. Le test de liaison consiste donc à tester si la pente de régression est égale à zéro par un test unilatéral de Student.

La méthode HE a été développée sous certaines hypothèses (indépendance des paires de germains, distribution normale du terme d'erreur et homogénéité des variances). Par conséquent, il est important d'envisager des méthodes robustes pour l'analyse de QTL et eQTL (expression Quantitative Trait Loci), applicable à un contexte plus large d'études génétiques et génomiques moderne, qui soit robuste aux hypothèses de normalité et à la présence de valeurs aberrantes, avec une puis-

sance suffisante pour détecter les QTLs et eQTLs.

Une approche robuste consiste à s'intéresser à des méthodes peu sensibles à la présence de valeurs aberrantes. Nous considérons ici des modèles de régression non paramétriques basés sur différents estimateurs robustes (quantiles de régression, M-estimateur, MM-estimateur et S-estimateur).

Nous avons comparé les différentes approches sur des données réelles de type QTL et eQTL afin d'analyser : (1) le facteur V pouvant exhiber une mutation dite de Leiden, intervenant dans la coagulation sanguine et qui conduit les porteurs de cette mutation à un risque de thrombose veineuse, (2) l'expression génétique de lignées cellulaires lymphoblastiques résultant d'une infection par le virus d'Epstein-Barr à partir des données du CEPH (Centre d'Etude du Polymorphisme Humain). Nos principaux résultats montrent l'intérêt des estimateurs robustes, pour l'analyse de ces données.

5.2 Article 5 : Robust Alternatives to Sib-Pair Methods for QTL and eQTL analyses. Sow, M., Durrieu, G., Gagnon, F., Briollais, L. (en préparation).

ROBUST ALTERNATIVES TO SIB-PAIR METHODS FOR QTL AND EQTL ANALYSES

Mohamedou SOW^{1,3}, Gilles DURRIEU¹, France Gagnon² & Laurent Briollais³

¹ *Université Bordeaux 1 et UMR CNRS 5805 & Samuel Lunenfeld Research Institute, Mount Sinai Hospital.*

² *Dalla Lana School of Public Health, University of Toronto.*

³ *Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto.*

Abstract

Abstract: Detection of linkage to genes for quantitative traits loci (QTLs) remains a challenging statistical issue. Classical approaches such as Haseman and Elston's sib-pair method and the variance component approach rely strongly on distribution assumptions and can be affected by outlier observations. Non-normal distributions generally exhibit skewness and kurtosis with very large values in studies such as expression quantitative trait loci (eQTL) experiments. Our goal is to propose alternative approaches that offer robustness against distribution assumptions, efficiency and flexibility for general genetic and genomic applications. We studied the properties of various non-parametric robust regression approaches. While most methods have comparable performance under the null hypothesis of no linkage, their respective power is quite sensitive to the presence of kurtosis in the trait distribution. Our real applications to a QTL analysis in the French-Canadian factor V Leiden family study and to an eQTL analysis of the CEPH families confirm these results and also demonstrate the good robustness properties of the robust approach.

Keywords: QTL analysis; eQTL experiment; sib-pair method; Haseman-Elston; kurtosis; skewness; robust estimator

1 Introduction

Many important human disease-related phenotypes are quantitative by nature (e.g. blood pressure, body mass index, etc...). Study of quantitative trait loci (QTLs) associated with complex diseases are critical to dissect

their underlying genetic mechanisms. Until recently, there were two major statistical approaches for QTL analysis using sibships: Haseman-Elston regression (HE) based on least squares estimation (Haseman and Elston, 1972) and variance component (VC) analysis using maximum likelihood estimation (see for example, Amos and Elston, 1989; Fulker and Cardon, 1994; Almasy and Blangero, 1998). The HE method was developed under the assumption of a population sample of normally distributed trait values, but it has been shown to be quite robust to the normality assumption (see Feingold, 2002). Moreover this approach is simple to use but does not optimally utilize information from data (Xu and Atchley, 1995). The VC approach on the other hand has higher power than the HE method and can fully exploit all the information in nuclear and extended families (Goldgar, 1990; Amos, 1994; Amos et al., 1996; Blangero and Almasy, 1997), but is not robust to selected sampling and deviations from distributional assumptions (Feingold, 2001). A major reason for this assumption is certainly the unrivalled mathematical tractability of the normal distribution (Azzalini and Capitanio, 1999). Violation of this assumption can lead to inflated type I error, reduced linkage power and biases in parameter estimates (Amos et al., 1996; Allison et al., 1999; Epstein et al., 2003). Extensions of HE and VC approaches have been proposed to circumvent these limitations. Some attempts have been made to robustify the VC approach, using for instances score statistics, but this has led to mixed successes (T.Cuenca et al., 2003). On the other hand, the HE approach has also been extended to provide power similar to the VC method while retaining its robustness (see Feingold, 2001; 2002). These recent extensions also yielded contrasted results, especially when the trait distribution is highly perturbed or when the sample of sibs is not random but ascertained through affected individuals.

The emergence of new technologies such as eQTL experiments, which generate thousands of gene expression related quantitative traits in a single experiment raises even more challenges because many gene distributions are likely to show extreme departure from normality and be affected by outlier observations (Durrieu and Briollais, 2009), exceeding what is usually observed in classical QTL analyses. Many of the recent HE extensions that rely on an optimal weighting of two statistics (squared trait difference and mean-corrected squared trait sum) for a sib-pair and commonly used trait transformations, such as Box-Cox, would need to be optimized for each gene (or each probe), which is not practical in real experiments.

Therefore, there is a compelling need to develop a more general statistical

framework for QTL and eQTL analyses, applicable to the wider context of modern genomics studies, with appropriate robustness against distribution assumptions and presence of outliers and with sufficient power to detect the QTLs underlying the traits of interest. Here we investigate a non-parametric robust regressions based on regression quantile estimators, M-, MM-, and S-estimators. The performance of these different estimators are assessed through two real applications: one on QTL analysis and the other one on eQTL analysis.

2 Methods

2.1 HE approach and the Linear Model Framework

Linear models belong to a classical statistical framework for the analysis of QTLs (Feingold, 2002), gene expression data (Smyth, 2004) and eQTLs (Schadt et al., 2005). For QTL analysis, it is generally assumed that the trait vector $\mathbf{Y}_n$ follows the general model

$$\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n. \quad (1)$$

In HE method $\mathbf{Y}_n = \mathbf{Y}_n^D = (\mathbf{Y}_1 - \mathbf{Y}_2)_n^2$ is the squared sib-pair trait difference, $\mathbf{X}_n$ is a known $n \times p$ design matrix of shared risk factors within a sib including the estimated number of alleles $\hat{\pi}$ shared identically by descent (*IBD*) at a particular locus, $\boldsymbol{\beta}$ is the vector of parameters of interest to be estimated, $\boldsymbol{\varepsilon}_n$ is the error term whose distribution is unknown and n is the number of observations (i.e. sib-pairs, here). HE approach is based on least-squares estimation (LSE) of the $\boldsymbol{\beta}$ parameters. The principle consists of testing that the regression slope parameter on the *IBD* is significantly lower than 0 to detect linkage at a particular locus. The reason is that similar trait values (and thus small squared trait differences) should be associated with high *IBD* sharing. Linkage can be tested with a one-sided t statistic of the regression slope estimate. This is similar to test the null hypothesis that the genetic variance at the major locus is 0. If $\boldsymbol{\varepsilon}$ follows a normal distribution, LSE provides the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$. However, it is clear that the normality assumptions are often violated in the context of QTL analysis and the presence of outliers might also affect the results of the test (Durrieu et al., 1999). In the following, we will denote $\hat{\boldsymbol{\beta}}_{LSE}^1$ the regression slope estimate from HE approach. Haseman and Elston (1972)

showed that the regression slope parameter is equal to $-2\sigma_a^2$, where σ_a^2 is the additive genetic variance at the major locus.

Various alternatives to HE method have been proposed to increase the power of the test. Wright (1997) and Drigalenko (1998) suggested to use the trait sum and trait difference in a regression framework. The new trait vector $\mathbf{Y}_n^S$ is defined as

$$\mathbf{Y}_n^S = ((\mathbf{Y}_1 - \mu) + (\mathbf{Y}_2 - \mu))_n^2$$

where μ is the trait sample mean.

Drigalenko (1998) further noted that this was equivalent to fitting a linear model for the mean-corrected trait product (or covariance)

$$\mathbf{Y}_n^P = ((\mathbf{Y}_1 - \mu)(\mathbf{Y}_2 - \mu))_n.$$

Finally other variants of this method use a weighting of the two regression slopes (of $\mathbf{Y}_n^D$ and $\mathbf{Y}_n^S$) by their variance estimates (Xu et al., 2000; Forrest, 2001; Sham and Purcell, 2001). The properties of these different estimators have been studied extensively by Feingold (2001; 2002), especially their robustness against distributional assumptions. The robustness against outliers has been less investigated. However, it is well known that the behavior of the least-squares estimator is not good from the point of robustness. The effect of even a single outlier can have a large effect on the result. Compared to other estimators such as classical least squares estimators, S -estimators are of great interest with respect with the breakdown point properties. The breakdown point is the smallest fraction of contaminated data that can cause the estimator to take one value arbitrary far from the true value. Therefore, the higher the breakdown point, the more robust the estimator is. For the least-squares estimator, the breakdown point is $1/n$ which tends to 0 when n tends to infinity. Rousseeuw and Yohai (1984) showed that the breakdown point of S -estimator can achieve $1/2$ which is the best that one can expect.

In the following we consider non parametric inferential methods. They have interesting robustness properties and can also be advantageous from the power perspective. An overview of robust statistics literature can be found in Hampel et al. (1986), Huber (2004), Jurečková and Sen (1996) and Durrieu and Briollais (2009).

2.2 Quantile Regression

Our approach is based on regression quantile estimators of β (Koenker and D'Orey, 1987; Dodge and Jurečková, 1995). Instead of focusing on the changes in the mean of $\mathbf{Y}_n$, the quantile regression approach allows one to test whether there is a change in the θ th quantile of $\mathbf{Y}_n$ for any given $\theta \in (0,1)$. When the conditional distributions of $\mathbf{Y}_n$ are non-Gaussian, the mean might not be the best summary, and any change in distributions may not be detected. Inference for linear quantile regression models has become a subject of intense investigation in the past years. Any solution of the following minimization problem ($\mathbf{x}_i^T$ denotes the i^{th} row of the matrix $\mathbf{X}$)

$$\hat{\beta}_{QR}(\theta) = \arg \min_{\beta \in R^p} \sum_{i=1}^n \rho_{\theta}(Y_i - \mathbf{x}_i^T \beta) \quad (2)$$

is called a θ -regression quantile with $\rho_{\theta}(x) = x(\theta - I(x < 0))$ where $I(\mathcal{P})$ takes the value 1 or 0 depending on whether the condition $\mathcal{P}$ is satisfied or not. Quantile regression can estimate the conditional median or other quantiles of the response variable, rather than ordinary least squares regression which estimates the conditional mean. Thus quantile regression is generally more robust to outliers. For instance, the L^1 -norm estimator, also known as *LAD* (*least absolute deviation*) estimator, obtained by taking $\theta = 1/2$, chosen as an alternative to least squares estimators and perform better in the presence of heavy-tailed distributions.

The asymptotic normality for the quantile estimate $\hat{\beta}_{QR}(\theta)$ was established by Koenker and Bassett (1978) under the assumption of independent and identically distributed (iid) error models, that is, ε_i are iid variables in the model (1). The asymptotic representation was given in He and Shao (1996) for independently, but not necessarily, identically distributed errors. If ε are independent with the probability density function f_i , then the asymptotic variance-covariance matrix of $\hat{\beta}_{QR}(\theta)$ is given by

$$\Sigma_{\theta} = (\theta(1 - \theta)) (\mathbf{X}^T \mathbf{F} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{F} \mathbf{X})^{-1} \quad (3)$$

where $\mathbf{F} = \text{diag}\{f_1(Q(\theta)), \dots, f_n(Q(\theta))\}$ with $Q(\theta)$ denoting the quantile function. In an iid error model the expression (3) reduces to

$$\Sigma_{\theta} = (\theta(1 - \theta)/f^2(Q(\theta))) (\mathbf{X}^T \mathbf{X})^{-1}, \quad (4)$$

where $1/f(Q(\theta))$ is the quantile density function denoted by $q(\theta)$. The asymptotic variance of $\hat{\beta}^n(\theta)$ is thus unknown because it depends on the distribution function f of the error term in the model, which itself is unknown. So, we estimate this asymptotic variance using the regression invariant and scale equivariant kernel-type estimator of $q(\theta)$ in model (1) based on the regression quantiles. This estimator of $q(\theta)$, for $0 < \theta < 1$, is defined by

$$\hat{Z}_n(\theta) = \frac{1}{\nu_n^2} \int_0^1 \hat{\beta}_1^n(w) k\left(\frac{\theta - w}{\nu_n}\right) dw,$$

where ν_n is a sequence of properly chosen bandwidths and k denotes the kernel function. It was proven that under regular conditions (Dodge and Jurečková, 1995), as $n \rightarrow \infty$,

$$\sqrt{n\nu_n} \left(\hat{Z}_n(\theta) - q(\theta) \right) = O_p(1) \text{ and } \sqrt{n\nu_n} \left(\hat{Z}_n(\theta) - q(\theta) \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, q^2(\theta) \overline{K} \right), \quad (5)$$

where $\overline{K} = \int K^2(x) dx$ with $K(x) = \int_{-\infty}^x k(y) dy$.

To test the hypothesis of no linkage, i.e., to test the linear hypothesis $H_0 : \mathbf{R}\beta = \mathbf{r}$, where $\mathbf{R}$ is a pre-specified matrix with full rank, and $\mathbf{r}$ is a vector of hypothetical values (of length 2 here). We consider the Wald test. From the asymptotic normality of regression parameter $\hat{\beta}_{QR}(\theta)$ and (4), we have under the null hypothesis, as $n \rightarrow \infty$,

$$\sqrt{n}\mathbf{R} \left(\hat{\beta}_{QR}(\theta) - \beta(\theta) \right) = \sqrt{n} \left(\mathbf{R}\hat{\beta}_{QR}(\theta) - \mathbf{r} \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(\mathbf{0}, \mathbf{R}\Sigma_\theta\mathbf{R}^T \right). \quad (6)$$

Using (5), we have as $n \rightarrow \infty$,

$$\hat{Z}_n(\theta) \xrightarrow{P} \frac{1}{f(Q(\theta))}. \quad (7)$$

It follows that as $n \rightarrow \infty$,

$$\hat{\Gamma}_\theta^{-1/2} \sqrt{n} \left(\mathbf{R}\hat{\beta}_{QR}(\theta) - \mathbf{r} \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(\mathbf{0}, \mathbf{I}_2 \right),$$

where $\hat{\Gamma}_\theta = \mathbf{R}\hat{\Sigma}_\theta\mathbf{R}^T$ with

$$\hat{\Sigma}_\theta = \theta(1 - \theta) \hat{Z}_n^2(\theta) (\mathbf{X}^T \mathbf{X})^{-1}, \quad (8)$$

for iid errors and $\mathbf{I}_2$ is the identity matrix of dimension 2. We obtain asymptotically, under the null hypothesis, the Wald statistic given by

$$n \left(\mathbf{R} \hat{\boldsymbol{\beta}}_{QR}(\theta) - \mathbf{r} \right)^T \hat{\boldsymbol{\Gamma}}_\theta^{-1} \left(\mathbf{R} \hat{\boldsymbol{\beta}}_{QR}(\theta) - \mathbf{r} \right) \xrightarrow{\mathcal{D}} \chi^2(2). \quad (9)$$

Under the null hypothesis we test the regression slope parameter, $\hat{\beta}_{QR}^1(\theta)$ being equal to zero by the Z statistic

$$Z = \sqrt{n} \times \hat{\beta}_{QR}^1(\theta) / \sqrt{\hat{\Sigma}_\theta^{22}} \xrightarrow{\mathcal{D}} N(0, 1),$$

where $\hat{\Sigma}_\theta^{22}$ is the second diagonal element of $\hat{\Sigma}_\theta$. The estimator can be deduced for non iid errors using (3).

2.3 M-estimator

M-estimation in the context of regression was first introduced by Huber (1973). The M-estimator of location parameter extends to the linear regression model (Huber, 1964; 2004; Hampel et al., 1986). Given an absolutely continuous function ρ with derivative ψ , we define a M-estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_M$, as a solution to the minimization problem

$$\hat{\boldsymbol{\beta}}_M = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \quad (10)$$

Here, we used the Huber function given by

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \leq k \\ k|x| - \frac{1}{2}k^2 & \text{for } |x| > k \end{cases} \quad (11)$$

and its derivative $\psi(x)$ given by

$$\psi(x) = \rho'(x) = \begin{cases} x & \text{for } |x| \leq k \\ k \operatorname{sgn}(x) & \text{for } |x| > k, \end{cases} \quad (12)$$

is continuous everywhere but is not differentiable at $\pm k$. Considering the regression model (1) and under regular conditions on the design matrix and on ψ (Stephanski and Boos, 2002), we have as $n \rightarrow \infty$,

$$\hat{\boldsymbol{\beta}}_M \xrightarrow{\mathcal{P}} \boldsymbol{\beta} \quad \text{and} \quad \sqrt{n} \left(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta} \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma}_M),$$

where $\mathbf{Q}$ is a positive definite matrix satisfying $\lim_{n \rightarrow \infty} 1/n \mathbf{X}^T \mathbf{X} = \mathbf{Q}$ and

$$\Sigma_M = \frac{E(\psi^2(\varepsilon_i))}{(E(\psi'(\varepsilon_i)))^2} \mathbf{Q}^{-1}.$$

If the errors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are identically distributed. If the errors are not identically distributed

$$\Sigma_M = \frac{1}{n} \frac{\sum_{i=1}^n E(\psi(\varepsilon_i)^2) \mathbf{x}_i^T \mathbf{x}_i}{\left(\sum_{i=1}^n E(\psi'(\varepsilon_i)) \mathbf{x}_i^T \mathbf{x}_i \right)^2}. \quad (13)$$

To test for a linkage, we proceed exactly in the same way as for quantile regression method. Under the null hypothesis, we obtain as $n \rightarrow \infty$,

$$\sqrt{n} \hat{\beta}_M^1 / \sqrt{\hat{\Sigma}_M^{22}} \xrightarrow{\mathcal{D}} N(0, 1),$$

where $\hat{\Sigma}_M^{22}$ is the second diagonal element of $\hat{\Sigma}_M$ given by

$$\hat{\Sigma}_M = \frac{\int \psi^2(x) \hat{f}_h(x) dx}{\left(\int \psi'(x) \hat{f}_h(x) dx \right)^2} \mathbf{Q}^{-1} \quad (14)$$

for iid error terms. We denotes by $\hat{f}_h$ the kernel density estimator of f (see Silverman (1986) among others) and the bandwidth parameter h is estimated using the plug-in procedure developed by Sheather and Jones (1991). The estimator can be derived in the same way for non iid errors.

2.4 S-estimator

Considering the model (1), the S-estimator (Rousseeuw and Yohai, 1984; Hossjer, 1992) is defined by minimizing a robust measure s of scale of the residuals $r_i = r_i(\boldsymbol{\beta}) = Y_i - \mathbf{x}_i^T \boldsymbol{\beta}$. So, the estimator of $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_S = \arg \min_{\boldsymbol{\beta}} s(r_1(\boldsymbol{\beta}), \dots, r_n(\boldsymbol{\beta})).$$

Usually, s is an M-estimator of scale defined by a function ρ symmetric, non-decreasing on $(0, \infty)$, absolutely continuous with bounded derivative ψ such

that $\rho(0) = 0$ and there exists a $c > 0$ such that ρ is constant on $[c, \infty)$. The scale estimator is defined by means of

$$\frac{1}{n} \sum_{i=1}^n \rho(r_i/s) = E_F(\rho).$$

Under regularity conditions, Hossjer (1992) and Davies (1990) proved that as $n \rightarrow \infty$,

$$\sqrt{n} \left(\hat{\beta}_S - \beta \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma_S) \quad (15)$$

where

$$\Sigma_S = \frac{\int \psi^2(x) f(x) dx}{\left(\int \psi(x) \Lambda(x) f(x) dx \right)^2} \mathbf{Q}^{-1} \quad (16)$$

and $\Lambda(x) = -f'(x)/f(x)$ is the likelihood score function. We then replace $f(x)$ by the kernel density estimator (Silverman, 1986) to estimate Σ_S using the plug-in procedure (Sheather and Jones, 1991) to estimate the bandwidth parameter. So, the estimator of (16) is given by

$$\Sigma_S = \frac{\int \psi^2(x) \hat{f}_h(x) dx}{\left(\int \psi(x) \hat{\Lambda}(x) \hat{f}_h(x) dx \right)^2} \mathbf{Q}^{-1}. \quad (17)$$

To test for linkage, we proceed as before.

2.5 MM-estimator

The MM-estimator is a combination of high breakdown value estimation and efficient estimation. This estimator proposed by Yohai et al. (1991) is an M-estimator starting at the coefficients given by the S-estimator. Since the MM estimate is computed as a M estimate at the last step, the asymptotic covariance of the M estimate can be used for the asymptotic covariance of the MM estimator.

3 Application 1: QTL analysis in the French-Canadian Factor V Leiden Family Study

Study: The general objective of the French-Canadian Factor V Leiden (F5L) family study is to identify quantitative trait loci (QTLs) underlying plasma

levels of hemostatic proteins and to investigate their relationships to each other and to venous thromboembolism (VTE, i.e. blood clot in a vein). VTE is a common complex disorder for which a well-characterized major gene has been identified i.e., F5L. The F5L thrombophilia phenotype refers to an increased tendency of developing blood clots in carriers of F5L mutations, which occurs via a mechanism characterized as Activated Protein C resistance (APCR). The F5L variant is also known to interact with common environmental factors, such as oral contraceptive and hormonal replacement therapy, and increases the risk of VTE. Despite the success in identifying this gene, its predictive clinical value remains limited. Experimental evidence suggest that additional genes are implicated in VTE. As several genetically determined plasma levels of hemostatic proteins have been associated with VTE, our main strategy is to identify QTLs underlying hemostatic quantitative traits as a way to identify genetic variants implicated in VTE. For this study, many quantitative traits from the hemostatic pathways, including APCR, FVIII and vWF, have been collected on 253 individuals from 5 extended French-Canadian families. The families were ascertained through single probands with VTE and carriers of F5L. Details of this study have been described elsewhere (Antoni et al., 2010 under review). For the purpose of our analysis, the extended families were decomposed into 396 independent sib-pairs with the program SIBPAIR (Duffy, 1997) and the IBD sharing probability was estimated with this program. In our analyses, we focused on the von Willebrand Factor (vWF), which shows high heritability (from 0.30 to 0.75; 0.40 in the French-Canadian F5L families) and good predictive value for VTE occurrence, suggesting that vWF is a good intermediate phenotype in the search of new VTE susceptibility genes. In our data, this trait exhibits skewness and kurtosis of 1.20 and 1.84 (average of both sibs) and the HE's transformed trait $(Y_1 - Y_2)^2$ of 2.97 and 10.09, respectively.

Candidate genetic markers: We selected a few markers previously linked to vWF or other phenotypes related to hemostasis. D1S196 and D1S452 are microsatellite markers under the linkage peak on chromosome 1 identified in the genome-wide linkage scan (GWLS) of APCR. This linkage peak includes the well-characterized major gene variant F5L underlying variation in plasma APCR levels. ABO blood group genotypes are well known to be associated with plasma levels of vWF, as well as to other phenotypes related to hemostasis. In this context, SNPs (ABO261, ABO297, ABO681, ABO703, ABO802, AB803, all these SNPs serve to determine the ABO genotypes from their

combinations) in the ABO locus were tested, as well as the two microsatellites D9S1793 and D9S1830 located under the GWLS peak signal for vWF on chromosome 9, also corresponding to the ABO locus. A common variant of the factor XII (F12) gene (rs1801020) known to be associated with factor XII levels was also tested. Factor XII is a trait of the hemostatic pathway, as APCR, FVIII and vWF. Finally, two SNPs located in the BAI3 gene (BAI3_SNP1 and BAI3_SNP2), which is a locus potentially contributing to early onset idiopathic venous thromboembolism through a mechanism modulating Factor VIII and von Willebrand Factor (Antoni et al., 2010 under review), were also genotyped.

Linkage results: For this application we selected the Haseman-Elston’s method (LSE), the quantile regression (QR) ($\theta = 0.5$), the M-estimator and the MM-estimator. The results are summarized in Table 3. First, our results confirm some previous findings for these traits and markers. The linkage between vWF and the Chr9 microsatellites and blood group are well known and is found with all the estimators. Other SNPs that serve to determine the ABO genotypes are also found systematically with the various estimators, for example the SNP ABO261 and ABO group.

4 Application 2: eQTL analysis of the CEPH families

Study: The data comprises 14 three-generation Centre d’Etude du Polymorphisme Humain (CEPH) Utah families (approximately 8 offspring per sibship and 14 individuals per family). Gene expression data were collected from 30 ”HapMap” trios: these are ”grandparentparent” trios that are partly included among those in the 14 families, plus approximately 12 additional grandparent-parent trios of CEPH Utah individuals. The expression level of genes expressed in lymphoblastoid cells (EBV-transformed B-cells) were obtained for the above subjects, using the Affymetrix Human Focus Arrays that contain probes for 8500 transcripts. Data were analyzed for the 3554 genes for which we found greater variation among individuals than between replicates for the same individual (Morley et al., 2004).

Candidate genetic markers and traits: Genotypes of 2882 autosomal

Table 1: QTL analysis of the trait vWF in the Factor V Leiden Family Study

Marker	LSE			QR ($\theta = 0.5$)			M-estimator		
	p -value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$	p -value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$	p -value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$
D1S196	0.497	-0.001	0.179	0.781	0.064	0.083	0.759	0.052	0.074
D1S452	0.841	0.181	0.181	0.670	0.039	0.089	0.799	0.063	0.075
D9S1793	0.229	-0.195	0.257	0.110	-0.148	0.117	0.123	-0.121	0.104
D9S1830	0.490	0.001	0.220	0.045*	-0.176	0.104	0.207	-0.075	0.091
F12	0.955	0.407	0.239	0.991	0.288	0.121	0.995	0.246	0.096
BAI3_SNP1	0.092	-0.292	0.219	0.717	0.061	0.107	0.310	-0.045	0.091
BAI3_SNP2	0.061	-0.399	0.257	0.347	-0.048	0.123	0.165	-0.104	0.106
ABO261	0.006**	-0.599	0.236	3.4E-05***	-0.319	0.079	0.002**	-0.270	0.094
ABO297	0.723	0.166	0.280	0.324	-0.063	0.137	0.668	0.051	0.117
ABO681	0.620	0.077	0.251	0.114	-0.158	0.131	0.129	-0.117	0.103
ABO703	0.673	0.676	1.506	0.776	0.550	0.723	0.742	0.407	0.625
ABO802	0.742	0.194	0.297	0.454	-0.017	0.145	0.665	0.052	0.123
ABO803	0.641	0.462	1.277	0.730	0.282	0.459	0.854	0.559	0.530
ABO group	0.002**	-0.693	0.240	0.007**	-0.305	0.095	0.003**	-0.272	0.347

NOTE: SE denotes the Standard Error of the regression estimator $\hat{\beta}$

We denote * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

MM-estimator				S-estimator		
Marker	p-value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$	p-value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$
D1S196	0.815	0.033	0.036	0.788	0.033	0.041
D1S452	0.52	0.002	0.037	0.058	0.0092	0.042
D9S1793	0.0599	-0.083	0.054	0.085	-0.08	0.058
D9S1830	0.107	-0.062	0.051	0.062	-0.082	0.053
F12	0.888	0.083	0.0685	0.77	0.045	0.06
BAI3_SNP1	0.7862	0.040	0.050	0.84	0.055	0.054
BAI3_SNP2	0.73	0.036	0.057	0.72	0.035	0.06
ABO261	4.3E-03***	-0.188	0.056	0.0018**	-0.17	0.058
ABO297	0.17	-0.058	0.061	0.15	-0.068	0.066
ABO681	0.052	-0.099	0.06	0.059	-0.09	0.059
ABO703	0.88	0.26	0.22	0.825	0.29	0.31
ABO802	0.299	-0.04	0.076	0.24	-0.052	0.074
ABO803	0.706	0.11	0.209	0.55	0.042	0.31
ABO group	0.013*	-0.124	0.056	0.028*	-0.11	0.057

NOTE: SE denotes the Standard Error of the regression estimator $\hat{\beta}$

We denote * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

Table 2: eQTL analysis of the CEPH families: Top gene expressions under each estimator

LSE			QR ($\theta = 0.5$)			M-estimator		
Marker	p-value	Marker	p-value	Marker	p-value	Marker	p-value	Marker
X212377	0.012	X201909	$< 10^{-16}$	X211367	$2.8 \cdot 10^{-3}$			
X2000049	0.016	X214218	$< 10^{-16}$	X209496	0.032			
-	-	X210632	1.310^{-4}	X203216	0.033			
-	-	X204254	1.410^{-4}	X210094	0.033			
-	-	X218200	6.210^{-3}	-	-			
-	-	X204011	7.910^{-3}	-	-			
-	-	X205848	0.015	-	-			
-	-	X209829	0.024	-	-			
-	-	X210094	0.029	-	-			
-	-	X202133	0.033	-	-			

Table 3: eQTL analysis of the CEPH families: Gene expressions with the largest skewness and kurtosis

Marker	Skew	Kurt	LSE			QR ($\theta = 0.5$)			M-estimator		
			$\hat{\beta}^1$	p-value	$SE(\hat{\beta}^1)$	p-value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$	p-value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$
X216268	7.1	61.2	-0.085	0.463	0.913	1.3.10 ^{-4***}	-0.685	0.186	0.001***	-0.605	0.200
X219767	6.1	48.8	-0.124	0.007**	0.050	0.138	-0.025	0.023	0.031*	-0.037	0.020
X217225	5.8	48.6	-0.109	0.013*	0.049	0.045*	-0.041	0.024	0.062	-0.031	0.020
X202530	6.0	48.0	-0.078	0.056	0.049	0.957	0.030	0.017	0.889	0.025	0.021
X204544	5.3	45.1	-0.077	0.097	0.059	0.294	-0.014	0.026	0.794	0.020	0.024
X219412	5.7	43.5	-0.431	0.224	0.568	0.309	-0.087	0.174	0.050	-0.330	0.201
X206141	5.0	35.4	-0.264	0.095	0.201	0.076	-0.114	0.079	0.076	-0.112	0.078

We denote * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

Skew: Skewness; Kurt: Kurtosis.

Marker	MM-estimator					S-estimator		
	Skew	Kurt	p-value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$	p-value	$\hat{\beta}^1$	$SE(\hat{\beta}^1)$
X216268	7.1	61.2	0.012*	-0.451	0.2	0.026*	-0.317	0.16
X219767	6.1	48.8	0.15	-0.015	0.015	0.198	-0.011	0.0135
X217225	5.8	48.6	0.059	-0.024	0.015	0.078	-0.021	0.014
X202530	6.0	48.0	0.97	0.029	0.145	0.98	0.030	0.014
X204544	5.3	45.1	0.131	-0.016	0.014	0.095	-0.020	0.015
X219412	5.7	43.5	0.142	-0.162	0.15	0.35	-0.05	0.13
X206141	5.0	35.4	0.33	-0.023	0.055	0.31	-0.025	0.051

We denote * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

Skew: Skewness; Kurt: Kurtosis.

and X-linked SNPs for members of the 14 CEPH Utah families described above were provided. These marker genotypes were used for the mapping of gene expression phenotypes described in (Morley et al., 2004). In our analysis, we selected only one SNP in the chromosomal region 14q32 previously found associated with many gene expressions (Morley et al., 2004). Across the 3554 gene expressions, the skewness varies between 0.4 and 7.4 with mean 3.0 and the kurtosis between -1.5 and 63.8 with mean 12.5.

Linkage results: In table 2, we present the top 10 gene expressions associated with the SNP, when the P -value was smaller than 0.05 after Bonferroni correction for multiple testing. Our results show that the rank and the significance of the best hits vary extremely across the different estimators. The quantile regression (QR) and M-estimator seem to be the most powerful. In Table 4, we studied more particularly the difference between the estimators for the eQTLs with the largest skewness and kurtosis. For example, the gene expression X216268 which has largest skewness and kurtosis is found with all estimator except for HE approach.

The gene expression X219767 is only found significant with the HE approach (p -value is equal to 0.007) and the M-estimator (p -value is equal to 0.031) and the gene expression X217225 is only found by the HE approach (p -value is equal to 0.013) and the quantile regression (p -value is equal to 0.045).

The scatter plots of the distribution of the trait $(Y_1 - Y_2)^2$ as function of the IBD in Figure 1 indicate that these estimators can be affected by some outlier or extreme observations.

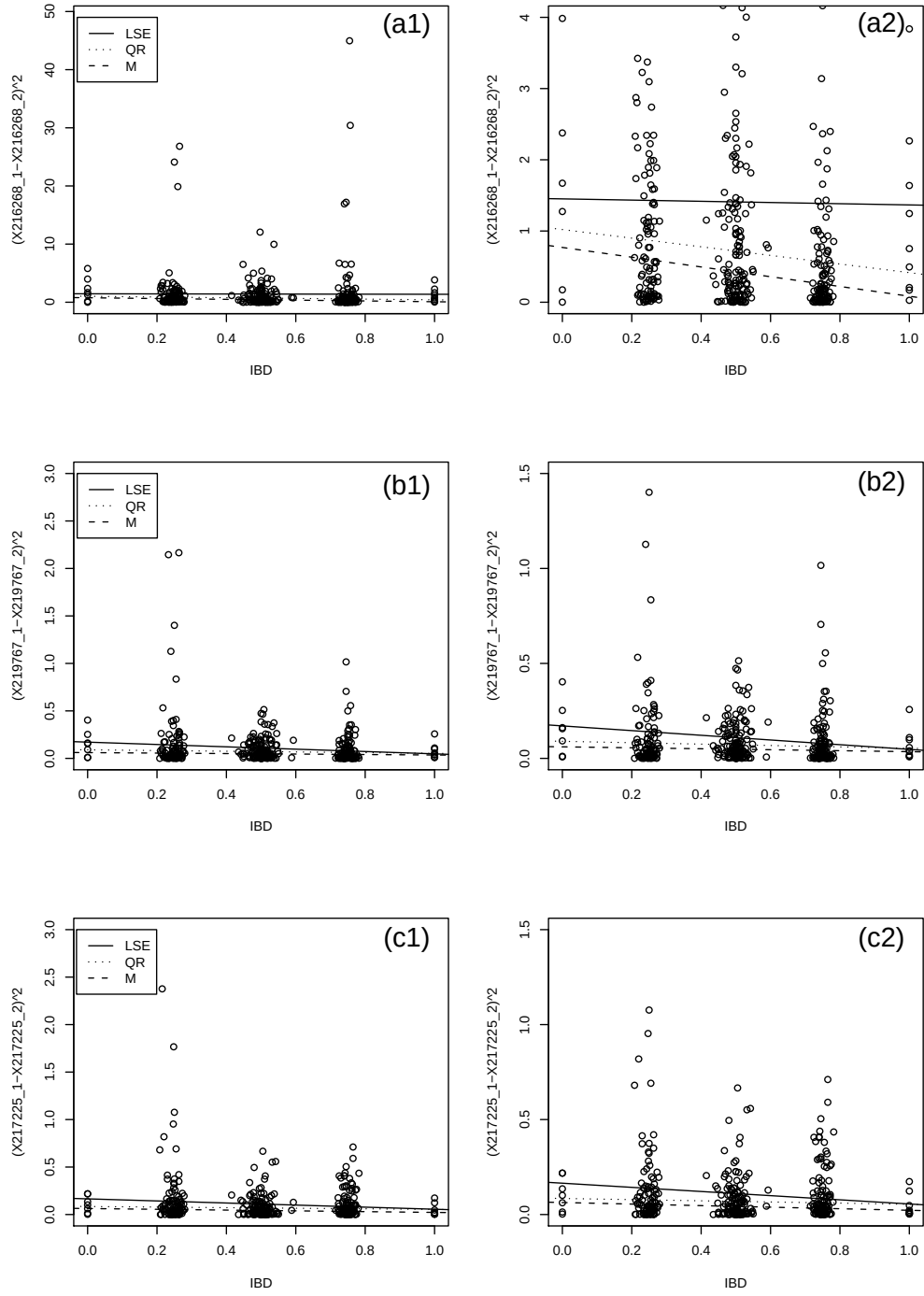


Figure 1: Regression of the squared sib-pair trait differences upon the estimated proportion of alleles shared identical-by-descent by the sib-pair for the gene expressions: (a1) X216268, (b1) X219767, (c1) X217225. The Plots (a2), (b2), (c2) are identical to (a1), (b2), (c2) but with a different scale for the y-axis.

5 Conclusion

We proposed alternative approaches to classical sib-pair analyses in the context of both QTL analysis of candidate genetic markers and genome-wide analysis of eQTL data. Our goal was to investigate statistical approaches that could have a broad application in the field of genetic and genomic high-throughput experiments. In particular we were interested in their good robustness properties against distribution assumptions and the presence of outlying observations and also in their efficiency to detect the locus of interest. Among the different application studies, the robust estimators were found to be the most promising. For instance, the Haseman-Elstons method can be affected by large kurtosis and skewness values. Its application to studies such as eQTL experiments can therefore be problematic. In conclusion, robust estimators are very flexible approaches that can be applied to a wide context of high-throughput genetic and genomic experiments, allowing many trait distributions to be fitted within the same inferential approach.

5.1 Acknowledgements

This work was supported by a PhD fellowship from the Aquitaine Region, France and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. The French-Canadian F5L family study is supported by grants from the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). France Gagnon holds a Canada Research Chair. We also thank David A. Tregouet for his useful comments on the manuscript.

References

- Allison, D. B., M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos, and J. Blangero (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *The American Journal of Human Genetics* 65(2), 531-544.
- Almasy, L. and J. Blangero (1998). Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees. *The American Journal of Human Genetics* 62, 1198-1211.
- Amos, C. I. and R. C. Elston (1989). Robust methods for the detection

of genetic linkage for quantitative data from pedigrees. *Genetic Epidemiology* 6(2), 349-360.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *The American Journal of Human Genetics* 54, 535-543.

Amos, C. I., D. K. Zhu, and E. Boerwinkle (1996). Assessing genetic linkage and association with robust components of variance approaches. *The American Journal of Human Genetics* 60, 143-160.

Antoni, G., P. E. Morange, Y. Luo, N. Saut, G. Burgos, S. Heath, M. Germain, C. Biron-Andreani, J. F. Schved, G. Pernod, P. Galan, D. Zelenika, M. C. Alessi, L. Drouet, S. Visvikis-Siest, P. S. Wells, M. Lathrop, J. Emmerich, D. A. Tregouet, and F. Gagnon (2010). A multi-stage multi-design strategy provides strong evidence that the XXX locus is associated with early-onset venous thromboembolism. Submitted.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society: series B* 61, 579-602.

Blangero, J. and L. Almasy (1997). Multipoint oligogenic linkage analysis of quantitative traits. *Genetic Epidemiology* 14, 959-964.

Blangero, J., J. T. Williams, and L. Almasy (2001). Variance component methods for detecting complex trait loci. *Advances in Genetics* 42, 151-181.

Davies, L. (1990). The Asymptotics of S-Estimators in the Linear Regression Model. *The Annals of Statistics* 18(4), 1651-1675.

Dodge, Y. and J. Jurečková (1995). Estimation of quantile density function based on regression quantiles. *Statistics and Probability Letters* 23, 73-78.

Drigalenko, E. (1998). Letters to the Editor: How Sib Pairs Reveal Linkage. *The American Journal of Human Genetics* 63(4), 1243-1245.

Duffy, D. (1997). Sib-pair: a program for non-parametric linkage/association analysis. *American Journal of Human Genetics Supp.* 61, A197.

Durrieu, G. and L. Briollais (2009). Sequential determination of sample size for robust linear regression: application to microarray experimental designs. *Journal of the American Statistical Association* 104(486), 650-660.

Durrieu, G., F. Meunier, J. O'Connel, M. Martinez, and F. De-

- menais (1999). Detection of quantitative trait loci associated with alcohol-dependence: use of model-free sib-pair method and combined segregation-linkage analysis based on regressive models. *Genetic Epidemiology* 17, 145-150.
- Epstein, M. P. and G. A. Satten (2003). Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data. *The American Journal of Human Genetics* 73(6), 1316-1329.
- Feingold, E. (2001). Methods for linkage analysis of quantitative trait loci in humans. *Theoretical Population Biology* 60, 167-180.
- Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *The American Journal of Human Genetics* 71, 217-222.
- Forrest, W. (2001). Weighting improves the "new Haseman-Elston" method. *Human Heredity* 52, 47-54.
- Fulker, D. W. and L. R. Cardon (1994). A sib-pair approach to interval mapping of quantitative trait loci. *The American Journal of Human Genetics* 54(6), 1092-1103.
- Fulker, D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt (1999). Combined linkage and association analysis for quantitative traits. *The American Journal of Human Genetics* 64, 259-267.
- Goldgar, D. E. 1990. Multipoint analysis of human quantitative genetic variation. *The American Journal of Human Genetics* 47(6), 957-967.
- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley: New York.
- Haseman, J. K. and R. C. Elston (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavioral Genetics* 2, 3-19.
- He, X. and Q. Shao (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Annals of Statistics* 24, 2608-2630.
- Hopper, J. L. and J. D. Mathews (1982). Extensions to multivariate normal models for pedigree analysis. *Annals of Human Genetics* 46, 373-83.
- Hossjer, O. (1992). On the optimality of S-estimators. *Statistics and Probability Letters* 14, 413-419.
- Huber P. J. (2004). *Robust Statistics*. John Wiley and sons: New Jersey.
- Huber P. J. (1973). Robust regression: Asymptotics, conjectures, and

- Monte Carlo. *Annals of Statistics* 1, 799-821.
- Huber P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35, 73-101.
- Jurečková, J. and P. K. Sen (1996). *Robust Statistical Procedures: Asymptotics and Interrelations* (1st edn). John Wiley and sons: New York.
- Lange, K. L., J. A. Little, and M. G. Taylor (1989). Robust statistical modelling using the t-distribution. *Journal of the American Statistical Association* 84, 881-896.
- Koenker, R. W. and G. Bassett (1978). Regression Quantiles. *Econometrica* 46, 33-50.
- Koenker, R. W. and V. D'Orey (1987). Computing Regression Quantiles. *Applied Statistics* 36, 383-393.
- Morley, M., C. M. Molony, T. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743-747.
- Rousseeuw, P. J. and V. Yohai (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics* 26, 256-272. Springer: New York.
- Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* 37, 710-717.
- Sheather, S. J. and M. C. Jones. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: series B* 53, 683-690.
- Sham, P. C. and S. Purcell (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs, *The American Journal of Human Genetics* 68, 1527-1532.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall: London.
- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3(1), Article 3.
- Stephanski, L. A. and D. D. Boos (2002). The calculus of M-estimator. *The American Statistician* 56(1), 29-38.

- T.Cuenca K. T., J. P. Szatkiewicz, and E. Feingold (2003). Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. *The American Journal of Human Genetics* 73, 863-873.
- Wright, F.A. (1997). The phenotypic difference discards sib-pair QTL linkage information, *The American Journal of Human Genetics* 60, 740-742.
- Xu, S. and W. Atchley (1995). A Random Model Approach to Interval Mapping of Quantitative Trait Loci. *Genetics* 141, 1189-1197.
- Xu, X., S. Weiss, Xu. X, L. J. Wei (2000). A unified Haseman-Elston method for testing linkage with quantitative traits. *The American Journal of Human Genetics* 67, 1025-1028.
- Yohai, V., W. A. Stahel, R. H. Zamar (1991). A Procedure for Robust Estimation and Inference in Linear Regression. *In Directions in Robust Statistics and Diagnostics, Part II*, eds Stahel WA and Weisberg SW. Springer-Verlag: New York.

Chapitre 6

Copules et régression quantile : application à l'analyse de eQTL

6.1 Introduction

Dans ce chapitre, nous abordons la dépendance entre variables aléatoires en utilisant la théorie des copules afin d'analyser des données de eQTL (QTL d'expression). Nous rappelons d'abord que la notion de eQTL se rapporte à la détection de régions contrôlant la variabilité de l'expression d'un gène.

En statistique, la théorie des copules permet une décomposition de la loi multidimensionnelle en ses marginales univariées et en une fonction de dépendance.

L'outil fondamental de la théorie des copules est le théorème de Sklar (Sklar, 1959). Il établit le lien entre la définition formelle des copules et les variables aléatoires, permettant ainsi l'application des copules dans la modélisation statistique. En 1959 Abel Sklar établit que si $F(x, y)$ est une fonction de répartition bivariée, avec $F_X(x) = P(X \leq x)$ et $F_Y(y) = P(Y \leq y)$ ses fonctions de répartitions marginales, alors il existe une copule $C(., .)$ telle que

$$C(u, v) = F(F_X^{-1}(u), F_Y^{-1}(v)), \quad u, v \in (0, 1).$$

et connecte $F(x, y)$ à $F_X(x)$ et $F_Y(y)$ via

$$F(x, y) = C(F_X(x), F_Y(y)), \quad x, y \in (-\infty, \infty).$$

Les copules sont utilisées dans plusieurs domaines, en finance (Bouyé et Salmon, 2002), sciences environnementales (Joe, 1994) et en génétique (Kim et al., 2008). Les copules constituent un outil statistique qui présente de nombreux avantages, tant pour les statisticiens que pour les généticiens, écologistes, et financiers. Outre une grande souplesse dans la mise en oeuvre de l'analyse multivariée, les copules autorisent une sélection plus étendue des distributions conjointes des séries de données. Elles permettent en outre la construction de distributions multidimensionnelles assez générales et ce, indépendamment des lois marginales qui peuvent avoir des lois

différentes et quelconques. Ainsi, elles permettent de s'affranchir de certaines hypothèses peu réalistes faites dans les études empiriques.

6.2 Modèles et estimateurs

Nous rappelons que le trait $\mathbf{Y}_n$ suit le modèle de régression suivant

$$\mathbf{Y}_n = \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n. \quad (6.1)$$

Dans le cas de la méthode de HE, $\mathbf{Y}_n = \mathbf{Y}_n^D = (\mathbf{Y}_1 - \mathbf{Y}_2)_n^2$ est la différence des traits au carré, $\mathbf{X}_n$ est une matrice connue $n \times p$ représentant ici l'estimée de la proportion d'allèles du marqueur partagée pour cette paire $\hat{\boldsymbol{\pi}}$ (*IBD*), $\boldsymbol{\beta}$ est le vecteur inconnu des paramètres de régression à estimer, $\boldsymbol{\varepsilon}_n$ est le terme d'erreur de loi inconnue, et n est le nombre d'observations (i.e. nombre de paires de germains). Dans le cas de HE, $\boldsymbol{\beta}$ est estimé par les moindres carrés (LSE).

Copules et dépendence

Nous explorons la régression quantile (cf. chapitre 3 et 5) en utilisant les copules pour spécifier les dépendances entre les variables d'intérêt.

Dans cette étude, nous nous sommes limités aux copules bivariées et avons considéré la copule gaussienne bivariée définie de la façon suivante

$$C(x, y; \rho) = \Phi_2 \left(\Phi^{[-1]}(x), \Phi^{[-1]}(y); \rho \right)$$

où ρ est le coefficient de corrélation, Φ_2 une distribution gaussienne bivariée, Φ une distribution univariée et $\Phi^{[-1]}$ le pseudo-inverse de Φ . Par le théorème de Sklar la distribution conjointes s'écrit

$$F_{X,Y}(x, y) = \Phi_2 \left(\Phi^{[-1]}(F_X(x)), \Phi^{[-1]}(F_Y(y)); \rho \right)$$

où F_X et F_Y désignent les distributions marginales.

Le lien entre les fonctions copules et les quantiles conditionnels a été introduit par une importante observation de Bouyé et Salmon, 2002. Cette observation permet d'estimer les quantiles conditionnels à partir de familles de copules.

Soit $\frac{\partial C(u,v)}{\partial u}$ la dérivée de $C(u, v)$ par rapport u , la fonction de répartition conditionnelle $F(y|x)$ est donnée par

$$F(y|x) = \frac{\partial C(F_X(x), F_Y(y))}{\partial F_X(x)},$$

Pour la copule paramétrique $\mathbf{C}(\cdot, \cdot; \rho)$, le θ -copule quantile de y conditionnelle à x est définie par

$$\theta = C^*(F_X(x), F_Y(y); \rho) \quad (6.2)$$

où $C^*(x, y; \rho) = \frac{\partial}{\partial x} \mathbf{C}(x, y; \rho)$. La relation entre X et Y peut être décrite en utilisant (6.2) par

$$y = q(x, \theta; \rho) \quad (6.3)$$

où $q(x, \theta; \rho) = F_Y^{[-1]}(D(F_X(x), \theta; \rho))$ avec D l'inverse de la dérivée partielle de C^* et $F_Y^{[-1]}$ le pseudo-inverse de F_Y . La relation (6.3) peut être exprimée avec les marginales

$$v = r(u, \theta; \rho) \quad (6.4)$$

avec $u = F_X(x)$ et $v = F_Y(y)$. A partir de (6.2) et (6.4), le θ -copule quantile est définie par

$$\theta = \Phi \left(\frac{\Phi^{[-1]}(v) - \rho \Phi^{[-1]}(u)}{\sqrt{1 - \rho^2}} \right),$$

et la relation entre y et x est

$$y = F_Y^{[-1]} \left(\Phi \left(\rho \Phi^{[-1]}(F_X(x)) + \sqrt{1 - \rho^2} \Phi^{[-1]}(\theta) \right) \right). \quad (6.5)$$

Le θ -copule quantile régression $q(x_i, \theta, \rho)$ est la solution du problème de minimisation suivant

$$\min_{\rho} \left(\sum_{i \in \mathcal{F}_{\theta}} \theta |y_i - q(x_i, \theta; \rho)| + \sum_{i \in \mathcal{F}_{1-\theta}} (1 - \theta) |y_i - q(x_i, \theta; \rho)| \right)$$

avec $\mathcal{F}_{\theta} = \{i : y_i \geq q(x_i, \theta; \rho)\}$ et $\mathcal{F}_{1-\theta}$ son complément.

6.3 Application

Les données que nous avons analysées portent sur 3 marqueurs de données CEPH, le choix de ces marqueurs est dû au fait qu'ils présentent les coefficients d'asymétrie et d'aplatissement les plus élevés (avec des points aberrants).

Nous avons considéré l'approche HE basée sur les moindres carrés (LSE), la régression quantile classique (QR) et la copule régression quantile (C-QR).

- 6.4 Article 6 : A Robust Statistical Framework for eQTL Analysis.** Sow, M., Durrieu, G., Briollais, L. (2010). Joint Statistical Meeting proceedings.

A ROBUST STATISTICAL FRAMEWORK FOR eQTL ANALYSIS

Mohamedou SOW^{1,3}, Gilles DURRIEU¹, & Laurent BRIOLLAIS³

¹ *Université Bordeaux 1 et UMR CNRS 5805 & Samuel Lunenfeld Research Institute, Mount Sinai Hospital.*

³ *Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto.*

Abstract

Abstract: Detection of linkage to genes for quantitative traits loci (QTLs) remains a challenging statistical issue. Classical approaches such as Haseman and Elston's sib-pair method and the variance component approach rely strongly on distribution assumptions and can be affected by outlier observations. Non-normal distributions generally exhibit skewness and kurtosis with very large values in studies such as expression quantitative trait loci (eQTL) experiments. Our goal here is to propose an alternative approach that offers robustness against distribution assumptions, efficiency and flexibility for general genetic and genomic applications. We describe a novel approach based on quantile regression that exploits the form of a copula function to measure the dependence between phenotype similarity and the number of alleles shared identical by descent (IBD) at a given marker. We show the interest of the method through an application to eQTL data from the CEPH study.

Keywords: quantile regression, copula, sib-pair method, eQTL
quantile regression, copula, sib-pair method, eQTL

1 Introduction

Understanding the pattern of gene expression variation in humans is important, as it is likely to underlie a significant proportion of the risk for many common disorders. Gene expression levels can be considered as quantitative trait (eQTL) and thus be dissected using standard genetic approaches such

as quantitative linkage analysis. In human genetics, linkage analysis of quantitative trait loci (QTL) is an approach that aims to correlate genetic similarity at a given marker, commonly measured by identity by descent (IBD) status, with similarity of phenotypes. Until recently, there were two major statistical approaches for QTL analysis using sibships: Haseman-Elston regression (HE) based on least squares estimation (Haseman and Elston, 1972) and variance component (VC) analysis using maximum likelihood estimation (see for example, Amos and Elston, 1989; Fulker and Cardon, 1994; Almasy and Blangero, 1998). Haseman and Elston (1972) proposed to regress the squared sib-pair trait difference $(Y_1 - Y_2)^2$ on the estimated proportion of marker alleles the sibs share (IBD) and to declare linkage whenever one finds evidence for negative slope of the regression. The HE approach was developed under the assumption of a population sample of normally distributed trait values, but it has been shown to be quite robust to the normality assumption (see Feingold, 2002). Moreover this approach is simple to use but does not optimally utilize information from data (Xu and Atchley, 1995). The VC approach on the other hand has higher power than the HE method and can fully exploit all the information in nuclear and extended families (Goldgar, 1990; Amos, 1994; Amos et al., 1996; Blangero and Almasy, 1997), but is not robust to selected sampling and deviations from distributional assumptions (Feingold, 2001). A major reason for this assumption is certainly the unrivaled mathematical tractability of the normal distribution (Azzalini and Capitanio, 1999). Violation of this assumption can lead to inflated type I error, reduced linkage power and biases in parameter estimates (Amos et al., 1996; Allison et al., 1999; Epstein et al., 2003).

The emergence of new technologies such as eQTL experiments, which generate thousands of gene expression related quantitative traits in a single experiment raises even more challenges because many gene distributions are likely to show extreme departure from normality and be affected by outlier observations (Durrieu and Briollais, 2009), exceeding what is usually observed in classical QTL analyses. Many of the recent HE extensions that rely on an optimal weighting of two statistics (squared trait difference and mean-corrected squared trait sum) for a sib-pair and commonly used trait transformations, such as Box-Cox, would need to be optimized for each gene (or each probe), which is not practical in real experiments.

Therefore, there is a compelling need to develop a more general statisti-

cal framework for eQTL analyses, applicable to the wider context of modern genomics studies, with appropriate robustness against distribution assumptions and presence of outliers. In statistical literature, a flexible way to describe dependence between correlated random variables is to use copulas (Sklar, 1959). Copulas are multivariate distribution functions whose one-dimensional margins are uniform on the $[0, 1]$ interval. Using a copula to build multivariate distributions is a flexible and powerful technique, because it separates choice of dependence from choice of marginals, on which no restrictions are placed (Clemen and Reilly, 1999). Copulas functions have been used in many fields, in finance to handle the interaction between the markets and risk factors in a flexible way (Cherubini et al., 2004) and to assess the conditional dependency between foreign exchange rates. In biology, a gaussian copula has been applied in quantitative trait linkage (Mingyao et al., 2006).

To overcome the limitations of classical methods for QTL and eQTL analyses, we propose a general approach based on quantile regression that exploits the form of a copula function to measure the dependence between the phenotype similarity and the IBD information. The link between copula functions and conditional quantile functions is provided by Bouyé and Salmon (2002). In this paper, we compare the performance of the new copula quantile regression approach (C-QR) with the classical HE method and quantile regression through an application to the human CEPH data.

2 Methods

2.1 HE approach and the Linear Model Framework

Linear models belong to a classical statistical framework for the analysis of QTLs (Feingold, 2002), gene expression data (Smyth, 2004) and eQTLs (Schadt et al., 2005). For QTL analysis, it is generally assumed that the trait vector $\mathbf{Y}_n$ follows the general model

$$\mathbf{Y}_n = \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n. \quad (1)$$

In HE method $\mathbf{Y}_n = \mathbf{Y}_n^D = (\mathbf{Y}_1 - \mathbf{Y}_2)_n^2$ is the squared sib-pair trait difference, $\mathbf{X}_n$ is a known $n \times p$ design matrix of shared risk factors within a

sib including the estimated number of alleles $\hat{\pi}$ shared identically by descent (*IBD*) at a particular locus, β is the vector of parameters of interest to be estimated, ε_n is the error term whose distribution is unknown and n is the number of observations (i.e. sib-pairs, here). HE approach is based on least-squares estimation (LSE) of the β parameters. The principle consists of testing that the regression parameter β on the *IBD* is significantly lower than 0 to detect linkage at a particular locus. The reason is that similar trait values (and thus small squared trait differences) should be associated with high *IBD* sharing. Linkage can be tested with a one-sided t statistic of the regression slope estimate. This is similar to test the null hypothesis that the genetic variance at the major locus is 0. If ε follows a normal distribution, LSE provides the maximum likelihood estimate (MLE) of β . However, it is clear that the normality assumptions are often violated in the context of QTL analysis and the presence of outliers might also affect the results of the test (Durrieu et al., 1999). In the following, we will denote $\hat{\beta}_{LSE}$ the regression slope estimate from HE approach. Haseman and Elston (1972) showed that the regression slope parameter is equal to $-2\sigma_a^2$, where σ_a^2 is the additive genetic variance at the major locus.

2.2 Quantile Regression

Our approach is based on regression quantile estimators of β (Koenker and D'Orey, 1987). Instead of focusing on the changes in the mean of $\mathbf{Y}_n$, the quantile regression approach allows one to test whether there is a change in the θ th quantile of $\mathbf{Y}_n$ for any given $\theta \in (0,1)$. When the conditional distributions of $\mathbf{Y}_n$ are non-Gaussian, the mean might not be the best summary, and any change in distributions may not be detected. Inference for linear quantile regression models has become a subject of intense investigation in the past years. Any solution of the following minimization problem ($\mathbf{x}'_i$ denotes the i^{th} row of the matrix $\mathbf{X}$)

$$\hat{\beta}_{QR}(\theta) = \arg \min_{\beta \in R^p} \sum_{i=1}^n \rho_{\theta}(Y_i - \mathbf{x}'_i \beta) \quad (2)$$

is called a θ -regression quantile with $\rho_{\theta}(x) = x(\theta - I(x < 0))$ where $I(\mathcal{P})$ takes the value 1 or 0 depending on whether the condition $\mathcal{P}$ is satisfied or not. Quantile regression can estimate the conditional median or other quantiles of the response variable, rather than ordinary least squares regression

which estimates the conditional mean. Thus quantile regression is generally more robust to outliers. For instance, the L^1 -norm estimator, also known as *LAD* (*least absolute deviation*) estimator, obtained by taking $\theta = 1/2$, chosen as an alternative to least squares estimators and perform better in the presence of heavy-tailed distributions. The asymptotic normality for the quantile estimate $\hat{\beta}_{QR}(\theta)$ was established by Koenker and Bassett (1978) under the assumption of independent and identically distributed (iid) error models, that is, ε_i are iid variables in the model (1). The asymptotic representation was given in He and Shao (1996) for independently, but not necessarily, identically distributed errors. If ε are independent with the probability density function f_i , then the asymptotic variance-covariance matrix of $\hat{\beta}_{QR}(\theta)$ is given by

$$\Sigma_{\theta} = (\theta(1-\theta)) (\mathbf{X}'\mathbf{F}\mathbf{X})^{-1} (\mathbf{X}'\mathbf{X}) (\mathbf{X}'\mathbf{F}\mathbf{X})^{-1} \quad (3)$$

where $\mathbf{F} = \text{diag}\{f_1(Q(\theta)), \dots, f_n(Q(\theta))\}$ with $Q(\theta)$ denoting the quantile function. In an iid error model the expression (3) reduces to

$$\Sigma_{\theta} = (\theta(1-\theta)/f^2(Q(\theta))) (\mathbf{X}'\mathbf{X})^{-1}, \quad (4)$$

where $1/f(Q(\theta))$ is the quantile density function denoted by $q(\theta)$. The asymptotic variance of $\hat{\beta}^n(\theta)$ is thus unknown because it depends on the distribution function f of the error term in the model, which itself is unknown. So, we estimate this asymptotic variance using the regression invariant and scale equivariant kernel-type estimator of $q(\theta)$ in model (1) based on the regression quantiles. This estimator of $q(\theta)$, for $0 < \theta < 1$, is defined by

$$\hat{Z}_n(\theta) = \frac{1}{\nu_n^2} \int_0^1 \hat{\beta}_1^n(w) k\left(\frac{\theta - w}{\nu_n}\right) dw,$$

where ν_n is a sequence of properly chosen bandwidths and k denotes the kernel function. It was proven that under regular conditions (Dodge and Jurečková (1995), as $n \rightarrow \infty$,

$$\sqrt{n\nu_n} \left(\hat{Z}_n(\theta) - q(\theta) \right) = O_p(1) \text{ and } \sqrt{n\nu_n} \left(\hat{Z}_n(\theta) - q(\theta) \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, q^2(\theta) \overline{K} \right), \quad (5)$$

where $\overline{K} = \int K^2(x) dx$ with $K(x) = \int_{-\infty}^x k(y) dy$.

To test the hypothesis of no linkage, i.e., to test the linear hypothesis $H_0 : \mathbf{R}\beta = \mathbf{r}$, where $\mathbf{R}$ is a pre-specified matrix with full rank, and $\mathbf{r}$ is

a vector of hypothetical values (of length 2 here). From the asymptotic normality of regression parameter $\widehat{\boldsymbol{\beta}}_{QR}(\theta)$ and (4), we have under the null hypothesis, as $n \rightarrow \infty$,

$$\sqrt{n}\mathbf{R} \left(\widehat{\boldsymbol{\beta}}_{QR}(\theta) - \boldsymbol{\beta}(\theta) \right) = \sqrt{n} \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{QR}(\theta) - \mathbf{r} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{R}\boldsymbol{\Sigma}_{\theta}\mathbf{R}'). \quad (6)$$

Using (5), we have as $n \rightarrow \infty$,

$$\widehat{Z}_n(\theta) \xrightarrow{P} \frac{1}{f(Q(\theta))}. \quad (7)$$

It follows that as $n \rightarrow \infty$,

$$\widehat{\boldsymbol{\Gamma}}_{\theta}^{-1/2} \sqrt{n} \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{QR}(\theta) - \mathbf{r} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{I}_2),$$

where $\widehat{\boldsymbol{\Gamma}}_{\theta} = \mathbf{R}\widehat{\boldsymbol{\Sigma}}_{\theta}\mathbf{R}'$ with

$$\widehat{\boldsymbol{\Sigma}}_{\theta} = \theta(1 - \theta) \widehat{Z}_n^2(\theta) (\mathbf{X}'\mathbf{X})^{-1}, \quad (8)$$

for iid errors and $\mathbf{I}_2$ is the identity matrix of dimension 2. We obtain asymptotically, under the null hypothesis, the Wald statistic given by

$$n \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{QR}(\theta) - \mathbf{r} \right)' \widehat{\boldsymbol{\Gamma}}_{\theta}^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{QR}(\theta) - \mathbf{r} \right) \xrightarrow{\mathcal{D}} \chi^2(2). \quad (9)$$

Under the null hypothesis we test the slope regression parameter, $\widehat{\boldsymbol{\beta}}_{QR}(\theta)$ being equal to zero by the Z statistic

$$Z = \sqrt{n} \times \widehat{\boldsymbol{\beta}}_{QR}(\theta) / \sqrt{\widehat{\Sigma}_{\theta}^{22}} \xrightarrow{\mathcal{D}} N(0, 1), \quad (10)$$

where $\widehat{\Sigma}_{\theta}^{22}$ is the second diagonal element of $\widehat{\boldsymbol{\Sigma}}_{\theta}$. The estimator can be deduced for non iid errors using (3).

2.3 Copula and dependence

For the analysis of sib-pair designs and the treatment of QTL and eQTL data, we use a quantile regression modeling approach based on a copula function that handles the dependency structure between Y the sib-pair trait difference and X the estimated proportion of alleles shared. The idea is that the joint

distribution $F_{X,Y}$ can be decomposed into the marginal distributions of X and Y denoted respectively by F_X and F_Y and the dependence function is specified by the copula (Bouyé and Salmon, 2002). The form of the linear quantile relationship implied by the copula is deduced. According to the Sklar's theorem (1959), there exists a unique bivariate copula $\mathbf{C} : [0, 1]^2 \rightarrow [0, 1]$ satisfying

$$F_{X,Y}(x, y) = \mathbf{C}(F_X(x), F_Y(y))$$

if F_X and F_Y are continuous and represent the marginal distribution functions of X and Y , respectively. Different families of copula correspond to different types of dependence structure (Nelsen, 1998, for general introduction to copulas). In the following, we decide to present and use the Gaussian copula where the dependence pattern is measured by correlation. The bivariate gaussian copula is

$$C(x, y; \rho) = \Phi_2(\Phi^{[-1]}(x), \Phi^{[-1]}(y); \rho)$$

where Φ_2 corresponds to the bivariate gaussian distribution, Φ to the univariate distribution and $\Phi^{[-1]}$ to the pseudo-inverse of Φ . So, by Sklar's theorem, for marginal distribution functions F_X and F_Y , the joint distribution

$$F_{X,Y}(x, y) = \Phi_2(\Phi^{[-1]}(F_X(x)), \Phi^{[-1]}(F_Y(y)); \rho)$$

is a bivariate distribution function with marginals F_X and F_Y and the copula that connects $F_{X,Y}(x, y)$ to $F_X(x)$ and $F_Y(y)$ is the Gaussian copula. For a parametric copula $\mathbf{C}(\cdot, \cdot; \rho)$, the θ -th copula quantile curve of y conditional on x is defined by

$$\theta = C^*(F_X(x), F_Y(y); \rho) \quad (11)$$

where $C^*(x, y; \rho) = \frac{\partial}{\partial x} \mathbf{C}(x, y; \rho)$. The relationship between X and Y can be expressed using (11) by

$$y = q(x, \theta; \rho) \quad (12)$$

where $q(x, \theta; \rho) = F_Y^{[-1]}(D(F_X(x), \theta; \rho))$ with D the partial inverse in the second argument of C^* and $F_Y^{[-1]}$ the pseudo-inverse of F_Y . The relationship (12) can be also expressed using uniform margins as

$$v = r(u, \theta; \rho) \quad (13)$$

with $u = F_X(x)$ and $v = F_Y(y)$. By (11) and (13), the θ -th copula quantile is defined by

$$\theta = \Phi\left(\frac{\Phi^{[-1]}(v) - \rho\phi^{[-1]}(u)}{\sqrt{1 - \rho^2}}\right),$$

and the relationship between y and x is

$$y = F_Y^{[-1]} \left(\Phi \left(\rho \Phi^{[-1]}(F_X(x)) + \sqrt{1 - \rho^2} \Phi^{[-1]}(\theta) \right) \right). \quad (14)$$

The θ -th copula quantile regression $q(x_i, \theta, \rho)$ is defined as any solution to the minimization problem

$$\min_{\rho} \left(\sum_{i \in \mathcal{F}_{\theta}} \theta |y_i - q(x_i, \theta; \rho)| + \sum_{i \in \mathcal{F}_{1-\theta}} (1 - \theta) |y_i - q(x_i, \theta; \rho)| \right)$$

with $\mathcal{F}_{\theta} = \{i : y_i \geq q(x_i, \theta; \rho)\}$ and $\mathcal{F}_{1-\theta}$ its complement. The asymptotic behavior of θ -th copula quantile regression is closely analogous to the non linear quantile regression case given in Koenker 2005 (section 4.4 pages 123 and 124).

3 Application

The data consist of 14 three-generation Centre d'Etude du Polymorphisme Humain (CEPH) Utah families (approximately 8 offspring per sibship and 14 individuals per family). Data were analyzed for the 3554 genes for which we found greater variation among individuals than between replicates for the same individual (Morley 2004). In our analysis, we selected only one SNP in the chromosomal region 14q32 previously found associated with many gene expressions (Morley et al., 2004). Across the 3554 gene expressions, the skewness varies between 0.4 and 7.4 with mean 3.0 and the kurtosis between -1.5 and 63.8 with mean 12.5. From this set of gene expressions, we selected three that exhibited possible outlier observations and had extreme skewness and kurtosis values. In table 1, we present the p -value and parameter estimates corresponding to the HE approach based on least-squares estimation (LSE), the classical quantile regression approach (QR), and the copula quantile regression method C-QR.

First, our results confirmed that the SNP studied might be associated with several gene expressions. Second, we found subsequent discrepancy between the various estimators. The gene expression X216268 is highly significant with the QR and C-QR estimators but not with the LSE while the

gene X219767 is only significant with LSE and not with the two quantile regression estimators.

The gene expression X217225 is not significant at 0.01 level for the 3 estimators. Figure 1 represents $q(IBD, \theta, \rho)$ with respect to the estimated proportion of alleles shared identical-by-descent by the sib-pair for different values of θ . The results show that only the gene expression X216268 is significant ($p < 0.05$).

Table 1: eQTL analysis of the CEPH families: p -values and parameter estimates for the largest skewness and kurtosis.

Marker	Skew	Kurt	LSE			QR ($\theta = 0.5$)			C-QR ($\theta = 0.5$)		
			p -value	$\hat{\beta}$	$SE(\hat{\beta})$	p -value	$\hat{\beta}$	$SE(\hat{\beta})$	p -value	$\hat{\rho}_{CQR}$	$SE(\hat{\rho}_{CQR})$
X216268	7.1	61.2	0.463	-0.085	0.913	$1.3 \cdot 10^{-4***}$	-0.685	0.186	0.04*	-0.2	0.115
X219767	6.1	48.8	0.007**	-0.124	0.050	0.138	-0.025	0.023	0.126	-0.097	0.0853
X217225	5.8	48.6	0.013*	-0.109	0.049	0.045*	-0.041	0.024	0.317	-0.033	0.072

We denote * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

Skew: Skewness; Kurt: Kurtosis.

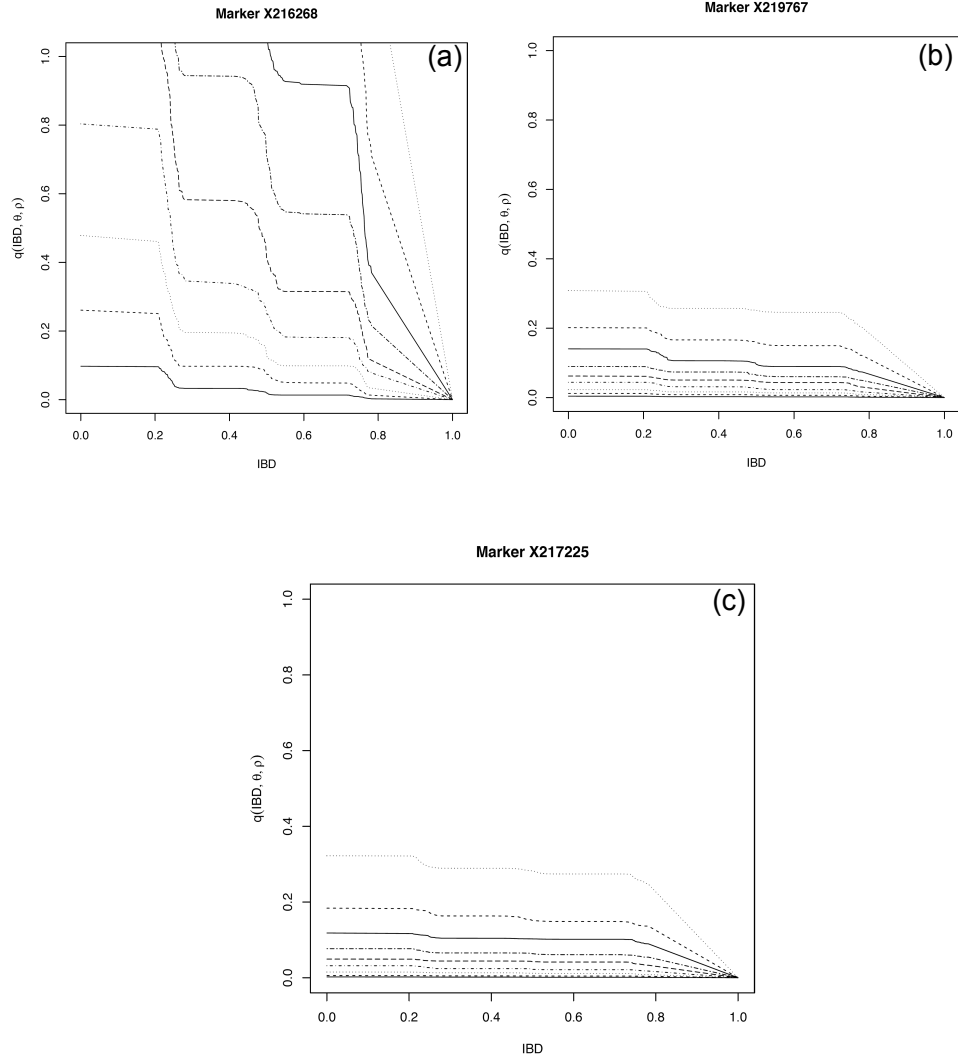


Figure 1: Representation of $q(IBD, \theta, \rho)$ with respect to the estimated proportion of alleles shared identical-by-descent by the sib-pair for different values of θ (0.1, 0.2, ..., 0.9) for the gene expressions: (a) X216268, (b) X219767, (c) X217225.

4 Conclusion

In this paper, we proposed a semiparametric approach for eQTL analyses based on a copula quantile regression. This approach is semiparametric in the sense that a parametric function is specified for the copula function but the marginal distributions are estimated empirically. This approach may be able to overcome the limitation of completely parametric approaches, in particular when the traits have marginal distributions that are very far from normal. Copula functions are useful for studying general forms of stochastic dependence. When outliers are suspected or the distribution of the trait is kurtotic, the copula quantile regression is a better choice for eQTL analysis than the traditional Haseman-Elston's method. Our future plan is to investigate the performance of the copula quantile regression through simulation studies and compare different copula functions.

References

- Allison, D. B., Neale, M. C., R. Zannolli, N. J. Schork, C. I. Amos, and Blangero, J. (1999), "Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure," *The American Journal of Human Genetics*, 65(2), 531-544.
- Almasy, L., and Blangero, J. (1998), "Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees," *The American Journal of Human Genetics*, 62, 1198-1211.
- Amos, C. I. (1994), "Robust variance-components approach for assessing genetic linkage in pedigrees," *The American Journal of Human Genetics*, 54, 535-543.
- Amos, C. I., and Elston, R. C. (1989), "Robust methods for the detection of genetic linkage for quantitative data from pedigrees," *Genetic Epidemiology*, 6(2), 349-360.
- Amos, C. I., Zhu, D. K., and Boerwinkle, E. (1996), "Assessing genetic linkage and association with robust components of variance approaches," *The American Journal of Human Genetics*, 60, 143-160.
- Azzalini, A., and Capitanio, A. (1999), "Statistical applications of the multivariate skew-normal distribution," *Journal of the Royal Statistical Society*, series B 61, 579-602.
- Blangero, J., and Almasy, L. (1997), "Multipoint oligogenic linkage analysis of quantitative traits," *Genetic Epidemiology*, 14, 959-964.
- Bouyé, E., and Salmon, M. (2002), "Dynamic Copula Quantile Regressions and Tail Area Dynamic Dependence in Forex Markets," *The European Journal of Finance*, 15(7), 721-750.
- Clemen R. T., and Reilly, T. 1999, "Correlations and Copulas for Decision and Risk Analysis," *Management Science*, 45, 208-224.
- Cherubini, U., Luciano, E. Vecchiato, W. (2004), *Copula Methods in Finance*, New York, John Wiley & Sons.
- Dodge, Y., Jurečková J. (1995), "Estimation of quantile density function based on regression quantiles", *Statistics and Probability Letters*, **23**, 73-78.
- Durrieu, G., and Briollais, L. (2009), "Sequential determination of sample size for robust linear regression: application to microarray experimental designs," *Journal of the American Statistical Association*, 104(486), 650-660.

- Durrieu, G., Meunier, F., O'Connel, J., Martinez, M., and Deme-nais, F. (1999), "Detection of quantitative trait loci associated with alcohol-dependence: use of model-free sib-pair method and combined segregation-linkage analysis based on regressive models," *Genetic Epidemiology*, 17, 145-150.
- Epstein, M. P., and Satten, G. A. (2003), "Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data," *The American Journal of Human Genetics*, 73(6), 1316-1329.
- Feingold, E. (2001), "Methods for linkage analysis of quantitative trait loci in humans," *Theoretical Population Biology*, 60, 167-180.
- Feingold, E. (2002), "Regression-based quantitative-trait-locus mapping in the 21st century," *The American Journal of Human Genetics*, 71, 217-222.
- Fulker, D. W., and Cardon, L. R. (1994), "A sib-pair approach to interval mapping of quantitative trait loci," *The American Journal of Human Genetics*, 54(6), 1092-1103.
- Goldgar, D. E. (1990), "Multipoint analysis of human quantitative genetic variation," *The American Journal of Human Genetics*, 47(6), 957-967.
- Haseman, J. K., and Elston, R. C. (1972), "The investigation of linkage between a quantitative trait and a marker locus," *Behavioral Genetics*, 2, 3-19.
- He, X., and Shao, Q. (1996), "A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs," *Annals of Statistics*, 24, 2608-2630.
- Koenker, R. W., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33-50.
- Koenker, R. W., and D'Orey, V. (1987), "Computing Regression Quantiles," *Applied Statistics*, 36, 383-393.
- Mingyao, L., Michael, B., Gonalo, R. A. and Peter, X.-K. S. (2006), "Quantitative Trait Linkage Analysis Using Gaussian Copulas," *the Genetics Society of America*, 173, 2317-2327.
- Morley, M., Molony, C. M., Weber, T., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004), "Genetic analysis of genome-wide variation in human gene expression," *Nature*, 430, 743-747.
- Nelsen, R. B. (1998), *An introduction to copulas*, New york, Springer-Verglas.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusis, A. J. (2005), “An integrative genomics approach to infer causal associations between gene expression and disease,” *Nature Genetics*, 37, 710-717.

Sklar, A. (1959). “Fonctions de repartition à n dimensions et leurs marges,” *Publications de l’institut de Statistique de l’Université de Paris*, 8, 229-231.

Smyth, G. K. (2004), “Linear models and empirical Bayes methods for assessing differential expression in microarray experiments,” *Statistical Applications in Genetics and Molecular Biology*, 3(1), Article 3.

Xu, S., and Atchley, W. (1995), “A Random Model Approach to Interval Mapping of Quantitative Trait Loci,” *Genetics*, 141, 1189-1197.

Chapitre 7

Discussion générale

Notre objectif principal dans ce travail de thèse était de modéliser des données complexes issues d'expériences de biologie modernes utilisant des méthodes de haute technologie. Dans ce cadre, nous avons envisagé des méthodes statistiques qui permettent de mieux gérer, interpréter et comprendre les données générées. En première application, ces résultats ont permis aux chercheurs de mieux comprendre les mécanismes et comportements mis en jeu. Une deuxième visée était d'aboutir à une présentation aisée des données afin que des acteurs/décideurs puissent prendre des décisions appropriées.

Nous avons exploré des domaines de la statistique qui ne sont pas toujours bien développés pour les données abordées. Nous avons montré comment certaines méthodes pouvaient s'adapter avantageusement à ces expériences et souvent proposer des méthodes statistiques permettant de tester l'effet de variables d'intérêt. Nous avons ainsi jeté les prémices d'un cadre statistique là où souvent des analyses descriptives, forcément trop réductrices, s'appliquent généralement.

Nous avons travaillé sur deux types d'études (environnementales et génétiques) qui peuvent paraître de nature assez différente, *a priori*, mais qui en fait possèdent certaines similarités. Nous dressons ci-après un bilan sur ces ressemblances mais aussi sur les particularités inhérentes aux deux domaines. Comme tout travail de thèse, nous n'avons pu échapper à certaines limitations, et une suite logique est de stimuler des recherches futures, tant sur le plan de la théorie statistique que sur son application à des problèmes variés de biologie. Cette dernière est en constante évolution, proposant des défis de plus en plus complexes aux statisticiens.

Conclusion générale

En environnement, le "bio-monitoring" ou suivi en ligne d'espèces animales est devenu un outil essentiel d'analyse dans les milieux naturels. Nous avons travaillé dans cette thèse sur l'analyse du comportement de bivalves par l'utilisation de la valvométrie HFNI (Haute Fréquence, Non Invasive) qui utilise une échelle de temps très fine (une mesure toutes les 0.1 s). Les enjeux de la modélisation de telles données dépassent le simple cadre de la connaissance du comportement des bivalves au

laboratoire (cadre expérimental) ou sur le terrain (éthologie) mais sont aussi économiques et sociaux face aux enjeux de la surveillance de la qualité de l'environnement aquatique marin.

Dans ce contexte, nous avons fait plus que proposer un outil de modélisation statistique. Nous avons d'abord préconisé un plan expérimental d'échantillonnage où plusieurs huîtres immergées sont équipées d'appareils de mesure de leur activité. Nous avons ensuite construit un système de mesure en ligne de ces activités en temps quasi réel. Ces données peuvent être consultées en ligne sur notre site web (http://www.domino.u-bordeaux.fr/molluscan_eye).

Notre travail a démontré qu'un tel système de "bio-monitoring" et d'analyse statistique pouvaient être combinés et pouvait être appliqués à l'étude du comportement de bivalves pour détecter des changements de comportements qui peuvent être associés à la perturbation du milieu. Notre première démarche était de transformer la masse énorme de données générées par l'expérience de valvométrie en un profil d'activité de l'animal afin de discerner des changements d'état qui décrivent différents moments de la vie d'une huître. Ces profils ont été corrélés à des données externes qui décrivent les caractéristiques environnementales. Nous avons ainsi pu mettre en évidence différentes relations entre l'activité valvaire et les caractéristiques du milieu. Nous avons en particulier décrit l'impact fort imposé par l'onde de marée chez des animaux immergés en permanence. Nous avons caractérisé pour la première fois le rythme tidal de l'huître *Crassostrea gigas* (chap. 1 ; Sow et al., 2011) et le rythme journalier d'ouverture fermeture du bénitier *Hippopus hippopus* (Chap. 3 ; Schwartzmann et al., en révision favorable). Nous avons caractérisé chez l'huître le phénomène de ponte, enregistré *in situ* dans les bassins d'Arcachon et de Marennes Oléron, par le calcul des vitesses (chap. 1 ; Bernard et al., en préparation). Nous avons également proposé un estimateur de type quantile de régression pour estimer la croissance des bivalves (Chap. 3 ; Schwartzmann et al., en révision favorable). Nous avons décrit les micro-fermetures ou fermetures partielles par une approche basée sur la théorie des shot noise (Chap. 4 ; Schmitt et al., sous presse).

L'ensemble de ces données pose les bases conceptuelles et analytiques qui permettent maintenant de décrire et donc mieux comprendre le comportement de bivalves en période « normale » ou période de référence. C'est à partir de ces observations que nous avons pu aborder la description d'évènements « anormaux ». Chez l'huître nous avons caractérisé la présence des fermetures anormales à Eyrac, c'est-à-dire à des périodes non associées à l'étal de marée basse, lors de tests souris positifs (et donc d'interdictions de ventes d'huîtres) dans le Bassin d'Arcachon (Chap. 1 ; Sow et al. 2011). Chez le bénitier, nous avons décrit des modifications de croissance et des altérations d'état d'ouverture maximale associées à des températures supérieures à 27°C pendant plusieurs semaines en fin de période estivale ainsi que des modifications de comportement par grosse mer lors d'une alerte cyclonique (Schwartzmann et al., en révision favorable).

En génétique, l'avancée des connaissances s'est faite elle aussi au gré d'avancées technologiques majeures, telles que les expériences utilisant les micro-puces plus connues sous le nom de microarray en anglais. Des milliers et même millions de données peuvent être générées maintenant en une seule expérience. Le but est alors dans ces données à haute densité génétique, d'identifier certains gènes ou certains variants génétiques susceptibles d'expliquer des maladies ou des traits quantitatifs associés à des maladies connues. Le travail du statisticien consiste à résumer ces informations à partir de méthodes telles que l'analyse de paire de germain puis de faire de l'inférence statistique afin de mettre en évidence des liaisons génétiques avec les traits ou maladies étudiées. L'apport de méthodes nouvelles est déterminant dans ces études où les données ont des variabilités importantes non expliquées par les facteurs génétiques causaux. Dans ce cadre nous avons abordé divers aspects de la robustesse. Ainsi nous avons développé des estimateurs peu sensibles aux données aberrantes, utilisé une approche semi-paramétrique, où nous avons abordé la dépendance entre variables aléatoires via le concept de copules.

Nos principaux résultats ont montré l'intérêt pratique de ces estimateurs sur des données réelles de QTL et eQTL. Parmi les différentes approches considérées, M-estimateur et la régression quantile semblent plus prometteurs. L'approche classique de Haseman-Elston (HE) peut être robuste à la non normalité des traits mais reste très sensible à l'aplatissement et aux points aberrants.

Similarité/particularisme des approches statistiques appliquées en génétique et études environnementales

Les études environnementales basées sur l'utilisation de la valvométrie tout comme les expériences modernes de génétique quantitative utilisant les techniques de micro-puces génèrent de gros volumes de données à analyser (nombre de variables qui peut être très élevé). Faire de l'inférence statistique sur ces données brutes est très difficile. Dans les deux domaines, il paraît avantageux de résumer l'information collectée en des profils généraux puis de faire de l'inférence statistique à partir de ces profils. Dans l'exemple des profils temporels générés par la valvométrie, notre stratégie consistait dans un premier temps à comprendre d'abord comment les facteurs du milieu (marée, alternance jour/nuît) règlent le rythme biologique pour discriminer ensuite un comportement qui n'est pas régi par ces facteurs, mais résulte d'une perturbation du milieu ambiant. L'inférence statistique peut être réalisée ensuite en corrélant ces profils généraux à des variables environnementales externes caractérisant des variations du milieu.

Dans le cadre génétique, la problématique est assez similaire. Les approches courantes visent souvent dans une première étape à extraire un signal intéressant à partir d'une masse de données très importante. Ce premier signal peut par exemple orienter le chercheur vers des régions chromosomiques particulières ou des réseaux de gènes qui seront explorés plus finement dans une deuxième étape (on parle souvent de "fine mapping" en anglais). Dans cette deuxième étape, la localisation du gène ou des variants causaux du gène est précisée. Une troisième étape qui intervient aussi fréquemment dans les deux domaines est de faire une prédiction sur les

comportements futurs du bivalve ou sur la survenue d'une maladie, par exemple.

Pour ces différentes raisons, on peut faire un parallèle assez clair entre l'application statistique aux domaines environnemental et génétique puisqu'il s'agit toujours de manipuler de grandes masses de données. Il y a cependant aussi des différences notables entre les deux domaines, liées aux niveaux de développement différents des deux techniques (la valvométrie HFNI est très jeune) et cela débouche sur des méthodes statistiques qui peuvent être différentes. Dans le cas des études génétiques, on recherche les facteurs causaux associés à une maladie ou un trait quantitatif associés à une maladie. On s'appuie alors sur l'importante base de données déjà acquise en biologie humaine et médecine. Dans les applications environnementales de la valvométrie HFNI, on cherche actuellement à mettre en évidence des corrélations écologiques pour décrire des comportements de base et de référence. L'aspect causal est encore difficile à mettre en évidence en l'absence d'analyses chimiques appropriées de l'eau et d'études expérimentales comme le cas d'algues toxiques spécifiques (Tran et al. 2010), de plusieurs métaux traces (Tran et al, 2003, 2007) et de présence de parasites spécifiques (Chambon et al. 2007). Cela implique souvent en génétique de nombreuses études de validation et réplifications avant qu'un gène particulier ou variant génétique puisse être déclaré comme causal ou ayant une forte évidence de l'être. Dans les études de type environnemental, on poursuit aussi l'étude des facteurs d'environnement impliqués même si la recherche des causes, si elles sont passées, est et sera souvent complexe sur le terrain.

Finalement, dans les deux domaines des expériences de biologie ou génétique, des analyses multiples sont conduites pour vérifier les résultats des analyses statistiques et cela reste la seule validation réelle de ces résultats.

Originalité de notre travail

Le caractère novateur de notre travail a été d'envisager des méthodes statistiques très générales, robustes et non paramétriques, à des domaines où elles s'appliquent en général assez peu. Dans le domaine de la valvométrie HFNI, totalement nouveau, les estimateurs de type noyau, la régression quantile et les approches type shot noise n'avaient jamais été appliquées. Dans le cadre des applications génétiques, les estimateurs robustes et méthodes des copules ont été aussi très peu utilisés. Il y a pourtant de réels avantages à l'utilisation de ces méthodes sur les deux technologies. Les données générées dans les deux domaines sont souvent très perturbées, avec des distributions complexes, parfois multimodales et avec des points aberrants. Les propriétés des méthodes classiques supposant la normalité des données ou des estimateurs ne sont plus vérifiées dans ces cas. Par contre les méthodes de régression robuste ou estimateurs robustes ont toujours des propriétés satisfaites dans cette situation et permettent une inférence statistique correcte. Lorsqu'on connaît le coût de résultats faux positifs dans le domaine génétique, on perçoit rapidement l'avantage qu'il y a à utiliser des méthodes robustes. Dans le cadre de notre étude environnementale, les enjeux économique, sociétal et humain sont tels, que là encore générer des résultats faux positifs serait très dommageable.

D'une façon général, nos développements statistiques ont été grandement mo-

tivés par ces contraintes importantes et nos résultats prometteurs montrent aussi que cette priorité pour des méthodes robustes ne s'accompagne pas nécessairement d'une perte de puissance pour des tests statistiques.

Un autre aspect important de notre travail a été de mettre en place en plus des développements statistiques, des algorithmes efficaces et des programmes informatiques (R, Script Bash et C) pour la gestion et le traitement des données.

Limites de nos méthodes statistiques et développements futurs

Dans le chapitre 2, nous avons proposé et comparé trois estimateurs non paramétriques de la fonction de lien, il serait avantageux de considérer aussi d'autres estimateurs de type quantile de régression de la fonction de lien et de comparer le comportement de l'ensemble de ces estimateurs au sens du critère de la variance asymptotique minimum.

Nous avons proposé un estimateur de la vitesse obtenu par dérivation par rapport au temps des estimateurs de type Nadaraya-Watson et Nadaraya-Watson récursif de la fonction de lien. Les premières simulations montrent un bon comportement de ces estimateurs de vitesse. Il sera aussi intéressant d'étudier ses principales propriétés asymptotique.

Dans le chapitre 4, pour l'exposant caractérisant la zone d'invariance d'échelle (β), nous avons trouvé un exposant qui est proche de celui de la turbulence homogène. Nous pensons aussi appliquer la théorie des shot noise sur d'autres cas de figure et d'autres sites pour confirmer les résultats trouvés.

Dans le contexte de la génétique, il serait utile de comparer les estimateurs développés avec les extensions de HE et de VC. Pour l'extension de HE, il existe par exemple la méthode de Xu et al. (2000) qui est considérée comme la meilleur extension de HE en terme de puissance (T.Cuenco et al., 2003). Pour l'extension de VC, nous pouvons par exemple considérer la méthode de Chen et al (2004, 2005) qui spécifie la matrice de variance covariance en incluant l'asymétrie et l'aplatissement de la distribution.

Il serait intéressant de comparer la performance (erreur de type 1 et la puissance) de l'ensemble des estimateurs par simulation. Ce travail est actuellement en cours. Dans le chapitre 6, nous avons considéré la copule gaussienne pour la dépendance. Il serait intéressant de considérer d'autres fonctions copules comme par exemple la copule de Gumbel qui permet de mesurer la dépendance entre valeurs extrêmes (Bouyé, 2002).

Les défis statistiques pour des données de plus en plus complexes

Des nouvelles méthodes voient le jour mais leur adaptation aux problèmes biologiques actuels prend souvent du temps. Par exemple, les méthodes de correction pour tests multiples "False Discovery Rate" (Benjamini et al, 1995) et ses variantes ont connu une évolution incroyable au cours des dernières années, elle-même grandement motivée par la révolution génomique. Pourtant, il semble que beaucoup de

progrès sont encore à faire pour que ces méthodes puissent prendre en compte les corrélations complexes existantes entre les variables étudiées. Sans vouloir donner une liste complète de domaines statistiques qui pourraient se développer dans les années à venir pour l'analyse de données à grande dimension, nous donnons quelques pistes de recherche intéressantes en particulier pour une application plus large aux domaines génétique et environnemental. 1) Les méthodes de régression de type lasso (Tibshirani, 1996) ou régression ridge (Arthur et al, 1970) sont très prometteuses pour l'analyse d'un grand nombre de variables. La méthode du lasso par exemple combine une étape de sélection de variables avec une étape de réduction des coefficients du modèle de régression ("shrinkage" en anglais) qui sont faites simultanément. Les variables non sélectionnées dans le modèle ont des coefficients de régression fixés à 0. La méthode du lasso introduit une pénalité de type L1. La méthode du lasso a été appliquée avec succès aux études génomiques (Shuangge et al, 2007) et plus récemment aux études génétiques sur l'ensemble du génome ("GWAS" en anglais, Tong et al, 2009). L'une des difficultés de cette méthode est de prendre en compte des structures de corrélation complexes des données. 2) L'existence d'informations *a priori* sur le comportement des bivalves ou sur des données génétiques/génomiques est très fréquente. Celles-ci pourraient être prises en compte dans l'inférence statistique dans un cadre bayésien. Les méthodes de ré-échantillonnage basées sur les chaînes de Markov et Gibbs sampling permettent de balayer efficacement l'espace des modèles et de sélectionner ceux qui ont une probabilité *a posteriori* élevée. L'association des techniques bayésiennes et des méthodes de régression robuste, telle que les quantiles de régression, a récemment vu le jour (Brian et al, 2010) et nous paraît une direction de recherche en statistique très intéressante pour l'analyse de données génétiques et environnementales à grande dimension. 3) Enfin, ce travail de thèse a montré l'intérêt des méthodes de régression non-paramétriques et estimateurs robustes. L'un des challenges futurs sera de développer ces méthodes de manière à prendre en compte des structures de corrélations données. Dans ce cadre, l'utilisation des copules est une voie de recherche prometteuse car il s'agit d'une méthodologie très générale. Dans ce travail de thèse, nous avons par exemple montré comment les quantiles de régression conditionnels pouvaient être estimés à partir de fonctions copules. Des méthodes de type GEE ou GLM qui ont des bonnes propriétés de robustesse pourraient être aussi employées. Dans le cadre des données de valvométrie, l'analyse des profils de l'animal pourrait se faire et tout comportement "anormal" pourrait être détecté en estimant un paramètre d'"overdispersion".

Perspectives

La majorité des études réalisées aujourd'hui dans le domaine de l'environnement ou de la génétique continuent soit de faire abstraction, soit plus généralement de ne pas prendre assez en compte, la présence potentielle d'interactions entre les facteurs génétiques et les facteurs environnementaux étudiés. Le problème majeur posé est en fait la difficulté de comparer des séries de données échantillonnées à des fréquences différentes, voir très différentes.

Le manque d'outils statistiques efficaces pour prendre en compte ces interactions peut être l'une des explications. En effet, les quelques méthodes proposées sont soit peu puissantes nécessitant donc des tailles d'échantillon irréalistes, soit difficiles à mettre en oeuvre car s'appuyant sur des schémas d'étude différents de ceux généralement utilisés dans ces contextes.

Le développement de nouvelles méthodes pour prendre en compte les facteurs environnementaux dans les études génétiques apparaît donc comme une étape importante vers une meilleure compréhension des phénomènes. Donc il serait bien d'envisager d'étendre nos méthodes dans le cas multivarié et de tenir en compte d'interactions gène-environnement dans l'étude du comportement de bivalves ou dans les maladies complexes :

- dans le contexte environnemental, il faudra inclure à terme d'autres paramètres du milieu (température, oxygène, pH, salinité, fluorescence, pression hydrostatique, vitesse du courant, chlorophylle a). On sait que les fréquences de mesure seront totalement différentes et que la technologie nécessaire à l'acquisition de ces données est difficile à mettre en place.
- dans le contexte génétique, il faudra généraliser nos résultats au cas de plusieurs paires de germains non indépendantes et de grandes familles.

De telles approches nécessitent de combiner les connaissances pluridisciplinaires issues de la génétique moléculaire, de l'épidémiologie, de la bioinformatique, mais aussi de la physiologie et d'autres domaines (Rebbeck et al. 2004). C'est tout un pan de la recherche qui n'en est qu'à ses débuts et qui vise à intégrer les informations issues de l'étude de SNPs (Single Nucleotide Polymorphisms), de l'ontologie des gènes, des voies biologiques, mais également des facteurs environnementaux en relation avec la maladie, des paramètres métaboliques et des informations toxicologiques. L'intégration des métadonnées est devenue une étape essentielle à la compréhension des phénomènes. Plus important encore, l'intégration des métadonnées, avec publication en ligne de grandes quantités de données sous formes graphiques lisibles rapidement, participe à l'émergence d'une nouvelle biologie intégrative encore plus puissante, dont l'objectif est une meilleure description et compréhension du fonctionnement des systèmes biologiques.

Table des figures

1	Description du protocole expérimental de 2 sites : Eyrac (Bassin d'Arcachon, France) et récif Ioro (lagon sud de Nouvelle Calédonie, Pacifique sud). Le principe est le même pour les autres sites.	22
1.1	Description de la distribution des états de fermetures/ouvertures au cours de la journée, avec en abscisse l'amplitude d'ouverture et en ordonnée la densité. Avec les bornes de l'intervalle de confiance à 95%.	30
1.2	Principe du rythme biologique lié à la marée : a) activité de fermetures/ouvertures pour 3 différentes huîtres (trait noir fermeture et trait blanc ouverture) ; b) superposition de fermetures/ouvertures de 3 huîtres différentes (16 et 17 Mai) ; c) représentation de l'activité de fermeture/ouverture sur 31 jours (pour chaque jour, il y a 16 lignes représentant les périodes de fermetures/ouvertures pour 16 huîtres) ; d) Représentation de l'évolution de la hauteur d'eau par heure. (Mai 2006)	31
1.3	Relation entre l'heure de la première fermeture et l'heure de la marée basse.	32
1.4	Ponte chez l'huître <i>Ostrea virginica</i> . a) premier enregistrement publié (Galtsoff 1938). b1) ponte d'une huître femelle. b2, b3) Enregistreur papier.	52
1.5	Extrait de l'enregistrement des données brutes correspondant à la ponte de 8 huîtres femelles, le 15 juillet 2007, en abscisse le temps (heures décimales) et en ordonné l'amplitude d'ouverture (mm). Jetée d'Eyrac, Arcachon, France.	53
1.6	Index de la vitesse correspondant aux enregistrements des données brutes, le 15 juillet 2007, en abscisse le temps (heures décimales) et en ordonné l'index de vitesse (u.a, unité arbitraire). L'index visuel est obtenu en calculant la valeur absolue à la puissance 5 de la vitesse réelle.	54
2.1	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).	60
2.2	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).	61

3.1	Evolution du minimum d'ouverture par jour : b�nitier num�ro 7 du 24 septembre au 31 d�cembre 2007.	64
3.2	(a) croissance (en mm) des b�nitiers mesur�e par la mesure des stries ; (b) croissance estim�e par un estimateur de type quantile de r�gres- sion et (c) corr�lation entre les deux croissances.	69
4.1	Illustration des micro-fermetures avec 4 chocs et 3 inter-temps.	112
4.2	Evolution de l'activit� valvaire a haute fr�quence. a) Repr�sentation sur 48 heures ; b) Zoom sur une plus petite proportions montrant plusieurs micro-fermetures.	114
4.3	Repr�sentation du spectre d' �nergie en log-log de l'activit� valvaire au cours d'une journ�e et de la pente ($\beta = -1.54$) pour la comparaison.	117

Bibliographie

- [1] Arthur E. Hoerl, Robert, W. Kennard (1970). Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics*, Vol. 12, No. 1, pp. 55-67.
- [2] Auby I., et Maurer D. (2004). Etude de la reproduction de l'huître creuse dans le bassin d'Arcachon. Rapport final, Edition Ifremer.
- [3] Beirlant, J., Vynckier, P., Teugels, J. L. (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association* 91, 1659-1667.
- [4] Benjamini, Yoav ; Hochberg, Yosef. (1995). "Controlling the false discovery rate : a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society, Series B (Methodological)* 57 (1) : 289-300.
- [5] Borcharding, J. (1992). *Another early warning system for the detection of toxic discharges in the aquatic environment based on valve movements of the freshwater mussel Dreissena polymorpha*. In D. Neumann and H.A. Jenner (Eds.), *The zebra mussel Dreissena polymorpha : Ecology, Biological Monitoring and First Applications in the Water Quality Management* (pp. 127-146). New York : Gustav Fischer Verlag.
- [6] Borcharding, J. (2006). Ten Years of Practical Experience with the Dreissena-Monitor, a Biological Early Warning System for Continuous Water Quality Monitoring. *Hydrobiologia* 556(1) :417-426.
- [7] Bouyé, E. (2002). Multivariate extremes at work for portfolio risk measurement, *Finance*, vol. 23, n°2, pp. 125-144.
- [8] Bouyé, E., and Salmon, M. (2002), "Dynamic Copula Quantile Regressions and Tail Area Dynamic Dependence in Forex Markets," *The European Journal of Finance*, 15(7), 721-750.
- [9] Brian J. Reich, Howard D. Bondell, Huixia J. Wang (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, 11, 2, pp. 337-352.
- [10] Briollais, L., Durrieu, G., Upathilake, R. (2007). Novel approach for genome scan meta-analysis of rheumatoid arthritis : a kernel-based estimation procedure. *BMC Genetics*, 1, 1-6.
- [11] Brown, R.L and Peet, R.K. (2003). Diversity and invasibility of southern Appalachian plant communities. *Ecology* 84 : 32-39.
- [12] Buddemier, R.W., Maragos, J.E., & Knutson, D.W., (1974). Radiographic studies of reef coral exoskeletons : rates and patterns of coral growth. *J Exp Mar Biol Ecol* 14 :179-200.

- [13] Cade, B.S., Terrel, J.W., Schoeder, R.L. (1999). Estimating effects of limiting factors with regression quantiles. *Ecology* 80 : 311-323.
- [14] Cao, J. et Ramsay, J.O. (2007). Parameter cascades and profiling in functional data analysis *Comp. Stat.* 22 (3) 335-351.
- [15] Campbell, N. (1909). The study of discontinuous phenomena, *Proceedings Cambridge Philosophical Society*, 15, 117-136.
- [16] Cardot, H., Crambes, C. and Sarda, P. (2006). Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. *Statistical methods for biostatistics and related fields*, Härdle, Mori and Vieu (Eds.), Springer.
- [17] Chambon, C., Legeay, A., Durrieu, G., Gonzalez, P., Ciret, P., Massabuau, J.C. (2007). Influence of the parasite worm *Polydora sp.* on the behaviour of the oyster *Crassostrea gigas* : a study of the respiratory impact and associated oxidative stress. *Marine Biology*, 152, 329-338.
- [18] Chen, W.M., Broman, K.W., Liang, K.Y. (2004). Quantitative trait linkage analysis by generalized estimating equations : unification of variance components and Haseman-Elston regression. *Genet Epidemiol* 26 :265-272.
- [19] Chen, W.M., Broman, K.W., Liang, K.Y. (2005). Power and Robustness of Linkage Tests for Quantitative Traits in General Pedigrees. *Genet Epidemiol* 28 :11-23.
- [20] Clot, D. (2002). Using functional PCA for cardiac motion exploration *Proceedings of the IEEE International Conference on Data Mining* 91-98.
- [21] Damon, J. et Guillas, S. (2002) The inclusion of exogenous variables in functional autoregressive ozone forecasting *Environmetrics* 13 759-774.
- [22] Dodge, Y. and J. Jurečková (1995). Estimation of quantile density function based on regression quantiles. *Statistics and Probability Letters* 23, 73-78.
- [23] Davies, L. (1990). The Asymptotics of S-Estimators in the Linear Regression Model. *The Annals of Statistics* 18(4), 1651-1675.
- [24] Durrieu G., Meunier F., O'Connell J., Martinez M. and Demenais F. (1999). Detection of quantitative trait loci associated with alcohol-dependence : Use of model-free sib-pair method and combined segregation-linkage analysis based on regressive models, *Genetic Epidemiology*, 17, 145-150.
- [25] Durrieu, G., & Briollais, L. (2009). Sequential design for microarray experiments. *Journal of the American Statistical Association*, 486, 650-660.
- [26] Eliazar I. and Klafter J. (2005). On the nonlinear modelling of shot noise, *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13779-13782.
- [27] Embrechts, P., McNeil, A., Straumann, D. (1999). Correlation : Pitfalls and alternatives A short, non-technical article, *RISK Magazine*, May, 69-71.
- [28] Epstein, M. P. and G. A. Satten (2003). Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data. *The American Journal of Human Genetics* 73(6), 1316-1329.

- [29] Falk, M. (1986). On the estimation of the quantile density function. *Statist. Probab. Lett.* 4, pp. 69-73.
- [30] Febrero, M., Galeano, P. and Gonzalez-Manteiga, W. (2007). A functional analysis of NOx levels : location and scale estimation and outlier detection *Comp. Statist.* 22 (3) 411-428.
- [31] Galtsoff, P. S. (1938). Physiology of reproduction of *Ostrea virginica*. I. Spawning reactions of the female and male. *Biological Bulletin* 74 : 461-486.
- [32] Gröcke, D. R., and D. P. Gillikin. (2008). Advances in mollusc sclerochronology and sclerochemistry : tools for understanding climate and environment. *Geo-Marine Letters* 1-4 :DOI 10.1007/s00367-00008-00108-00364.
- [33] Haseman, J. K. and R. C. Elston (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavioral Genetics* 2, 3-19.
- [34] Härdle, W., & Marron, J.S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika*, 72, 481-484.
- [35] Härdle, W., & Marron, JS (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann Stat* 13 :1465-1481.
- [36] Härdle, W. (1990). *Applied nonparametric regression*. Cambridge : Cambridge University Press.
- [37] Henderson, B. (2006). Exploring between site differences in water quality trends : a functional data analysis approach. *Environmetrics* 17 : 65-80.
- [38] Hill, B. M., 1975. A simple general approach to inference about the tail of distribution. *Annals of Statistics* 3. 1163-1174.
- [39] Hudson, J.H., Shinn, E.A., Halley, R.B., Lidz, B. (1976). Sclerochronology : a tool for interpreting past environments. *Geology* 4 :361-364.
- [40] Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *La revue canadienne de statistique*. 22, 47-64.
- [41] Jurečková, J. and P. K. Sen (1996). *Robust Statistical Procedures : Asymptotics and Interrelations* (1st edn). John Wiley and sons : New York.
- [42] Kim, J.-M., Jung, Y., Sungur, E. A., Han, K., Park, C., and Sohn, I. (2008). A copula Method for Modeling Directional Dependence of Genes. *BMC Bioinformatics*, Vol. 9 :225.
- [43] Koenker R. (1994). Confidence intervals for regression quantiles. In : Mandl P and Hušková M (Eds). *Asymptotic statistics : proceedings of the 5th Prague Symposium*. Physica-Verlag : Heidleberg. p349-359.
- [44] Koenker, R. W., and Bassett, G. (1978). Regression Quantiles. *Econometrica* 46, 33-50.
- [45] Koenker, R. W. and V. D'Orey (1987). Computing Regression Quantiles. *Applied Statistics* 36, 383-393.
- [46] Koenker, R. (2005). *Quantile Regression* , Econometric Society Monograph Series, Cambridge University Press.

- [47] Kolmogorov, A. N. (1941). The local structure of turbulence in incompressible viscous fluid for very large Reynolds number. *Compte rendus de l'Academie Science de l'URSS* 30, 299pp.
- [48] Kolmogorov, A. N. (1962). A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. *Journal of Fluid Mechanics* 13, 82-85.
- [49] Kramer, K.J.M., Jenner, H.A., De Zwart, D. (1989). The valve movement response of mussels : a tool in biological monitoring. *Hydrobiologia*, 188(189), 433-443.
- [50] Kramer, K.J.M., Foekema, E.M. (2001). The "Musselmonitor" as biological early warning system : the first decade. In Butterworth, FM, Gunatilaka, MA, Gonsebatt ME (eds). *Biomonitoring and Biomarkers as Indicators of Environmental Change*, Vol. 2. Kluwer Academic Publishers, New York : 59-87.
- [51] Legendre, P., Legendre, L. (1998). *Numerical Ecology*. Elsevier, Amsterdam.
- [52] Leng X. and Müller, H.G. (2006). Classification using functional data analysis for temporal gene expression data.
- [53] Lockhart, D. J., Dong, H., Byrne, M. C. et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat.Biotechnol.*, 14 : 1675-1680.
- [54] Mantripragada, K. K., Buckley, P. G., de Stahl, T. D., et Dumanski, J. P. (2004). Genomic microarrays in the spotlight. *Trends Genet.*, 20 : 87-94.
- [55] Marceau, F. (1909). Recherches sur la morphologie, et l'histologie, et la physiologie compares des muscles adducteurs des mollusques acphales. *Archives de Zoologie expérimentale et générale*, 2, 295-469.
- [56] Nadaraya, E.A. (1964). On estimating regression. *theor. Prob. Appl.* 10 186-196.
- [57] Narula, S.C and Wellington, J. F. (1977). Prediction, linear regression and the minimum sum of relative errors. *Technometrics*, 19, 2, 185-190.
- [58] Nerini, D., Ghattas, B. (2007). Classifying densities using functional regression trees : Applications in oceanology *Comp. Stat. & Data Anal.* 51 (10) 4984-4993.
- [59] Opgen-Rhein, R. et Strimmer, K. (2006). Inferring gene dependency networks from genomic longitudinal data : a functional data approach. *REVSTAT- Statistical Journal*. 4 (1) 53-65.
- [60] Parzen, E. (1962). On estimation of probability density function and mode. *Ann. Math. Statist.*, 33, 1065-1076.
- [61] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis (Second Edition)* Springer-Verlag, New York
- [62] Ratclie, S.J., Heller, G.Z. et Leader, L.R. (2002a). Functional data analysis with application to periodically stimulated foetal heart rate data. I : Functional regression. *Stat Med.* 21 (8) 1103-1114.
- [63] Ratclie, S.J., Heller, G.Z. et Leader, L.R. (2002b) Functional data analysis with application to periodically stimulated fetal heart rate data : II : functional logistic regression. *Stat Med.* 21 (8) 1115-1127.

- [64] Rebbeck, T.R., Spitz, M., Wu, X. (2004). Assessing the function of genetic variants in candidate gene association studies. *Nat Rev Genet* 5 : 589-97.
- [65] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.* 42 43-47.
- [66] Schena, M., Shalon, D., Davis, R. W., et Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 : 467-470.
- [67] Schmitt, F. G., Dur, G., Souissi, S., Zongo, S. B. (2008). Statistical properties of turbidity oxygen and pH fluctuations in the Seine river estuary (France). *Physica A* 387, 6613-6623.
- [68] Schmitt, F. G., De Rosa, M., Durrieu, G., SOW, M., Ciret, P., Tran, D., Massabuau, J.C. (2011). Statistical study of bivalve high frequency microclosing behavior : scaling properties and shot noise analysis. *International Journal of Bifurcation and Chaos*, in press.
- [69] Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman and Hall, London.
- [70] Sheather, S. J. and M. C. Jones. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society : series B* 53, 683-690.
- [71] Shuangge Ma, Xiao Song, Jian Huang. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, Vol. 8, No. 1.
- [72] Sklar, A. (1959). "Fonctions de repartition 'a n dimensions et leurs marges," Publications de l'institut de Statistique de l'Université de Paris, 8, 229-231.
- [73] Song, J.J., Lee, H.J., Morris, J.S. et Kang, S. (2007). Clustering of time-course gene expression data using functional data analysis *Computational Biology and Chemistry* 31 (4), 265-274.
- [74] Sornette, D. (2000). Critical Phenomena in Natural Sciences. Springer, New-york.
- [75] Sow, M., Durrieu, G., Briollais, L. (2010). A Robust Statistical Framework for eQTL Analysis. *JSM proceedings*.
- [76] Sow, M., Durrieu, G., Briollais, L., Ciret, P., Massabuau, J.C. (2011). Water quality assessment by means of HFNI valvometry and high-frequency data modeling. *Environ Monit Assess*. DOI 10.1007/s10661-010-1866-9.
- [77] Stephanski, L. A. and D. D. Boos (2002). The calculus of M-estimator. *The American Statistician* 56(1), 29-38.
- [78] Terrell, J.W., Cade, B.S., Carpenter, J. and Thompson, J.M. (1996). Modeling stream fish habitat limitations from wedged-shaped patterns of variation in standing stock. *Trans Am Fish Soc* 125 :104-117.
- [79] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288.

- [80] Tong Tong Wu., Yi Fang Chen., Trevor Hastie., Eric Sobel and Kenneth Lange (2009). Genomewide Association Analysis by Lasso Penalized Logistic Regression We develop efficient computational procedures for screening large-scale genome-wide association studies. *Bioinformatics* 25(6) : 714-721.
- [81] Tran, D., Boudou, A., Massabuau, J.C. (2000). Mechanism of oxygen consumption maintenance under varying levels of oxygenation in the freshwater clam *Corbicula fluminea*. *Canadian Journal of Zoology*, 78, 2027-2036.
- [82] Tran, D., Boudou, A., Massabuau, J.C. (2001). How water oxygenation level influences cadmium accumulation pattern in the Asiatic clam *Corbicula fluminea* : a laboratory and field study. *Environmental Toxicology and Chemistry*, 20(9), 2073-2080.
- [83] Tran, D. (2001). "Etude des mécanismes de régulation de la ventilation chez le bivalve filtreur *Corbicula fluminea*. Influence de l'oxygène et de l'apport trophique. Application aux problèmes de bioaccumulation et de biodétection du cadmium" thèse Université Bordeaux1.
- [84] Tran, D., Ciret, P., Ciutat, A., Durrieu, G., Massabuau, J.C. (2003). Estimation of potential and limits of bivalve closure response to detect contaminants : application to cadmium. *Environmental Toxicology and Chemistry*, 22, 116-122.
- [85] Tran, D., Fournier, E., Durrieu, G., Massabuau, J.C. (2004). Copper detection in the asiatic clam *Corbicula fluminea* : optimum valve closure response. *Aquatic Toxicology*, 65(3), 317-327.
- [86] Tran, D., Fournier, E., Durrieu, G., Massabuau, J.C. (2007). *Corbicula fluminea* valve closure response as sensitivity threshold to inorganic mercury contamination. *Environmental Toxicology and Chemistry*, 26(7), 1545-1551.
- [87] Tran, D., Haberkorn, H., Soudant, P., Ciret P., Massabuau, J.C. (2010). Behavioral responses of *Crassostrea gigas* exposed to the harmful algae *Alexandrium minutum*. *Aquaculture*, 298, 338-345.
- [88] Velculescu, V. E., Zhang, L., Vogelstein, B., et Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270 :484-487.
- [89] Watson, G.S. (1964). Smooth regression analysis. *Sankhya Series A*, 26, 359-372.
- [90] Wand, M.P., & Jones, M.C. (1995). *Kernel Smoothing*, London : Chapman and Hall.
- [91] Zeltermann, D. (1990). Smooth nonparametric estimation of the quantile function. *J. Statist. Plann. Inference* 26, pp. 339-352.