

Application de modèle non paramétrique sous R : analyse et suivi de la qualité de l'eau

^aMohamedou Sow

^bGilles Durrieu, ^aDamien Tran, ^aPierre Ciret et ^aJean-Charles Massabuau

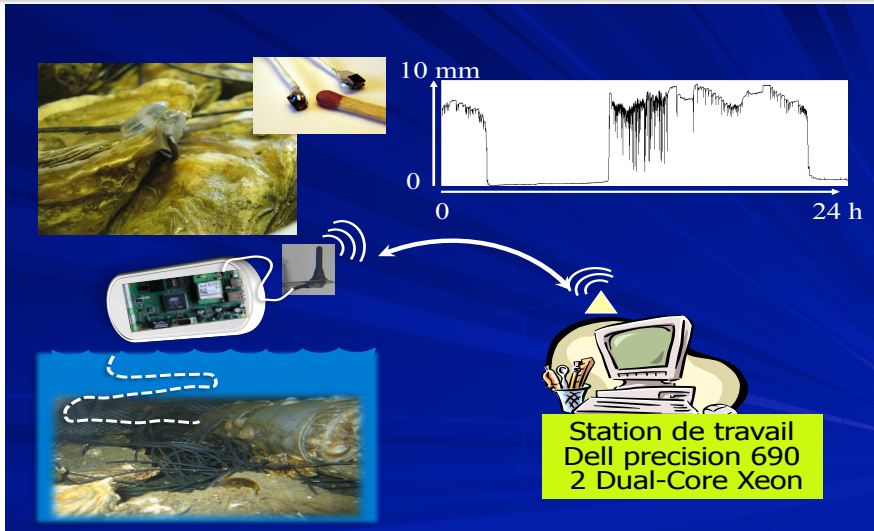
^aEPOC, UMR CNRS 5805, Université de Bordeaux

^bLMBA, UMR CNRS 6205, Université de Bretagne sud

Objectifs

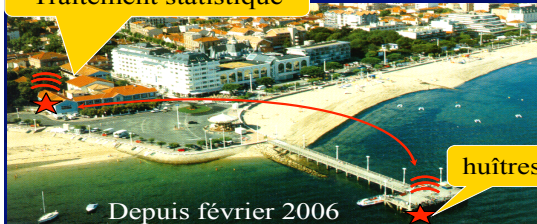
- Caractériser et quantifier le comportement de bivalves dans leur milieu naturel afin d'extraire des **invariants**, des **tendances** et des **rythmes biologiques**.
- Étudier les **perturbations** du comportement et des rythmes que peut induire une modification du milieu (pollution, changement climatique, phytoplancton, ...).
- Mettre en place et utiliser des outils statistiques permettant de modéliser et traiter de grands volumes de données en environnement.

Enregistrement et acquisition des données



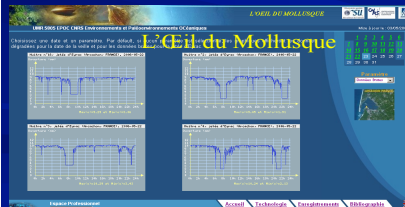
Site expérimental : jetée d'Eyrac

Traitement statistique



huîtres

Depuis février 2006



Diffusion/web

huître	amplitude	temps
1	9,86	100
2	9,52	100
3	6,8	100
4	12,7	100
5	8,68	100
6	9,48	100
7	8,65	100
8	4,78	100
9	6,51	100
10	7,58	100
11	7,96	100
12	11,32	100
13	10,87	100
14	5,64	100
15	10,91	100
16	8,65	100

Extrait de l'enregistrement
des données brutes

Données

Biologiques :

- Fréquence d'échantillonnage : une mesure toutes les 0.1s, chaque bivalve est questionné chaque 1.6 s ($N = 16$);
- Tous les jours, nous acquérons 54 000 mesures (couples de valeurs $(T_i; Y_i)$) pour chaque bivalve;
- $n=864\ 000$ mesures par jour pour 16 bivalves;
- Plus de 300 000 000 mesures effectuées par an / site.

Environnementales :

- hauteur d'eau, jour/nuit, température, courant,

Codes de calculs (Script bash, C, R)

```
#!/bin/bash
```

```
login=`sed -n -e '1,$p' /valvo/SANTANDER/identkrea`
pass=`sed -n -e '1,$p' /valvo/SANTANDER/passwdkrea`
```

```
jj=`date -d $y +_ %d% m%y`
j=`date -d $y +%A`
m=`date -d $y +%B`
numenm=`date -d $y +%m`
jm=`date -d $y +%d`
a=`date -d $y +%Y`
b=`date -d $y +%y`
cd /valvo/SANTANDER
wget -t inf
ftp://\$login:\$pass@stamar010.epoc.u-bordeaux1.fr:21/santander/\$b\$numenm\$j\*
gzip $b$numenm$j*
mkdir $a"-"$numenm"-"$jm
cd /valvo/SANTANDER/$a"-"$numenm"-"$jm
cp
/valvo/SANTANDER/extraitvoies.c
gcc -o extraitvoies extraitvoies.c -lm
...
cp /valvo/SANTANDER/calcul1.s
cp /valvo/SANTANDER/cumul.s
cp
/valvo/SANTANDER/p_fermeture.s
cp
/valvo/SANTANDER/p_ouverture.s
...
R --no-save < calcul1.s >/dev/null
R --no-save < cumul.s >/dev/null
R --no-save < Tab1_Sites_Web.s
>/dev/null
```

```
File Edit Options Buffers Tools Asm Help
d = read.table("Tableau_2_Rythmes_Tour_Nuit_Maree.txt", h=T)
jours = as.numeric(levels(as.factor(d$nbjours)))
nbj = length(jours)
numhuitre = as.numeric(levels(as.factor(d$huitre)))
nbhuitre = length(numhuitre)
v=jours
nbj=length(v)
pdf("Fig_Rythmes.pdf", paper="a4")
plot(0,0, type="n", xlim=c(0, 24), yaxt="n", ylim=c(0, (nbhuitre+1)*nbj),
     xlab="temps sur 24H", ylab="")
decal <- nbj*nbhuitre+(nbj-1)
i=decal
for(l in v)
{
  for(k in numhuitre)
  {
    # on recupere data de l'huitre k et du jour d
    j=dim(d[d$huitre==k & d$nbjours == 1,])[1]
    da2 <- d[d$huitre==k & d$nbjours == 1,]
    # on trace les segments pour l'huitre k et le jour d
    while(j > 0){
      segments(da2[j,5], i, da2[j,6], i)
      j = j - 1
    }
    # On change d'huitre
    i=i-1
  }
  decal <- decal - (nbhuitre+1)
  i = decal-1
  abline(h=decal,col="red")
}
dev.off()
```

Automatisation des scripts pour les différents sites

```
#Crontab
MAILTO=valvo
HOME1=/Data/valvo/EYRAC
HOME4=/Data/valvo/SANTANDER
HOME5=/Data/valvo/LOCMARIAQUER
HOME6=/Data/valvo/TROMSO
HOME7=/Data/valvo/NYALESUND
```

```
# minute heure jour_mois mois jour_semaine commande_Shell
00 02 * * * $HOME1/Relance_script_Calcul_EYRAC_Jour.sh
00 03 * * * $HOME4/Relance_script_Calcul_SANTANDER_Jour.sh
00 04 * * * $HOME5/Relance_script_Calcul_LOCMARIAQUER_Jour.sh
00 04 * * * $HOME6/Relance_script_Calcul_TROMSO_Jour.sh
00 05 * * * $HOME7/Relance_script_Calcul_NYALESUND_Jour.sh
```

- * Nombres d'ouvertures et de fermetures journalières
- * Durées d'ouverture et de fermeture journalières
- * Périodes d'ouvertures et de fermetures journalières
- * Moments d'ouvertures et de fermetures journalières
- * Amplitude d'ouverture minimum et maximum
- * Vitesse d'ouverture et de fermeture
- * Nombres de micro-fermetures
- * Relation avec les paramètres du milieu
- * Evolution de la croissance au cours du temps
- * Distribution des périodes d'ouverture au cours du temps
- * Distribution des périodes de fermeture au cours du temps

```
2012-01-03 2012-02-11 2012-03-21 2012-04-29 2012-06-07
2012-01-04 2012-02-12 2012-03-22 2012-04-30 2012-06-08
2012-01-05 2012-02-13 2012-03-23 2012-05-01 2012-06-09
2012-01-06 2012-02-14 2012-03-24 2012-05-02 2012-06-10
2012-01-07 2012-02-15 2012-03-25 2012-05-03 concat_Tab.sh
2012-01-08 2012-02-16 2012-03-26 2012-05-04 concat_Tab.sh
2012-01-09 2012-02-17 2012-03-27 2012-05-05 datatemperature.txt
2012-01-10 2012-02-18 2012-03-28 2012-05-06 Figure_Classe Horaire
2012-01-11 2012-02-19 2012-03-29 2012-05-07 Figure.sh
2012-01-12 2012-02-20 2012-03-30 2012-05-08 identkrea
2012-01-13 2012-02-21 2012-03-31 2012-05-09 mise_a_jour_maree.sh
2012-01-14 2012-02-22 2012-04-01 2012-05-10 passw@krea
2012-01-15 2012-02-23 2012-04-02 2012-05-11 Relance_script_Calcul_SANTANDER_Jour.sh
2012-01-16 2012-02-24 2012-04-03 2012-05-12 Relance_script_Calcul_SANTANDER_Jour.sh
2012-01-17 2012-02-25 2012-04-04 2012-05-13 replace_home_data.sh
2012-01-18 2012-02-26 2012-04-05 2012-05-14 Resu_Min_Max_Jour.txt
2012-01-19 2012-02-27 2012-04-06 2012-05-15 taille_no_ek
2012-01-20 2012-02-28 2012-04-07 2012-05-16 taille_ok
2012-01-21 2012-02-29 2012-04-08 2012-05-17 valvo_no_ek
2012-01-22 2012-03-01 2012-04-09 2012-05-18 valvo_ok
```

Modèle de régression non paramétrique

Pour $i = 1, \dots, n$,

$$Y_i = m(T_i) + \varepsilon_i$$

où :

- Y : amplitude d'ouverture en millimètre,
- m : activité inconnue du bivalve à estimer,

$$m(t) = \mathbb{E}[Y \mid T = t]$$

- T : temps en heure,
- ε : terme aléatoire d'erreur de loi inconnue, indépendant de T .

- Estimateur de Nadaraya-Watson récursif (Duflo, 1997)

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

Théorème 1. (Convergence en loi)

Sous les hypothèses de régularités, nous avons quand $n \rightarrow \infty$:

- $\forall \alpha \in]1/3, 1[$ et $f(t) > 0, \forall t \in \mathbb{R}$,

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

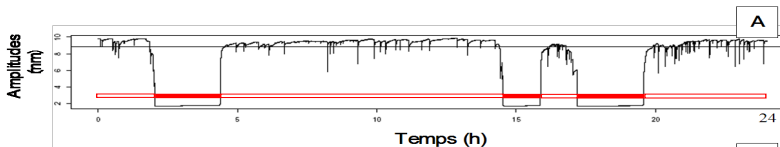
Avec

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

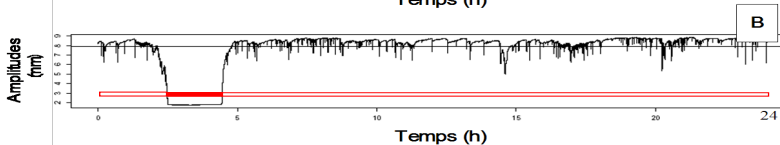
$$\tilde{\sigma}^2(t) = \frac{1}{\tilde{f}_n(t)} \sum_{i=1}^n K\left(\frac{t - T_i}{h_i}\right) (Y_i - \tilde{m}_n(t))^2.$$

Recherche de rythme biologique

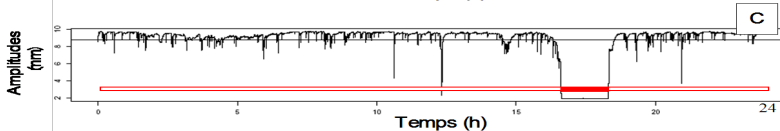
A



B



C



Day 1

Day 2

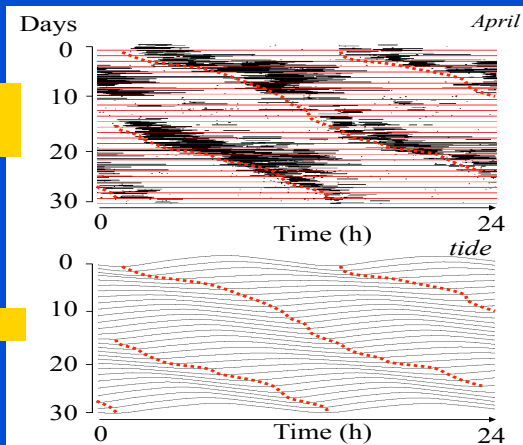


Rythme biologique

un rythme ouverture/fermeture lié à la marée

Activité
ouverture-fermeture

Ondes de marée



Conclusion

- Développement d'algorithmes et des codes de calculs (R, C, Script Bash, PHP...) pour l'analyse de gros volume de données
- Développement de modèle et estimateur non paramétrique sous R
- Transfert automatique en ligne ([http ://www.domino.u-bordeaux.fr/molluscan_eye](http://www.domino.u-bordeaux.fr/molluscan_eye)) des données et résultats (tableaux et graphiques)