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

[a]Mohamedou Sow

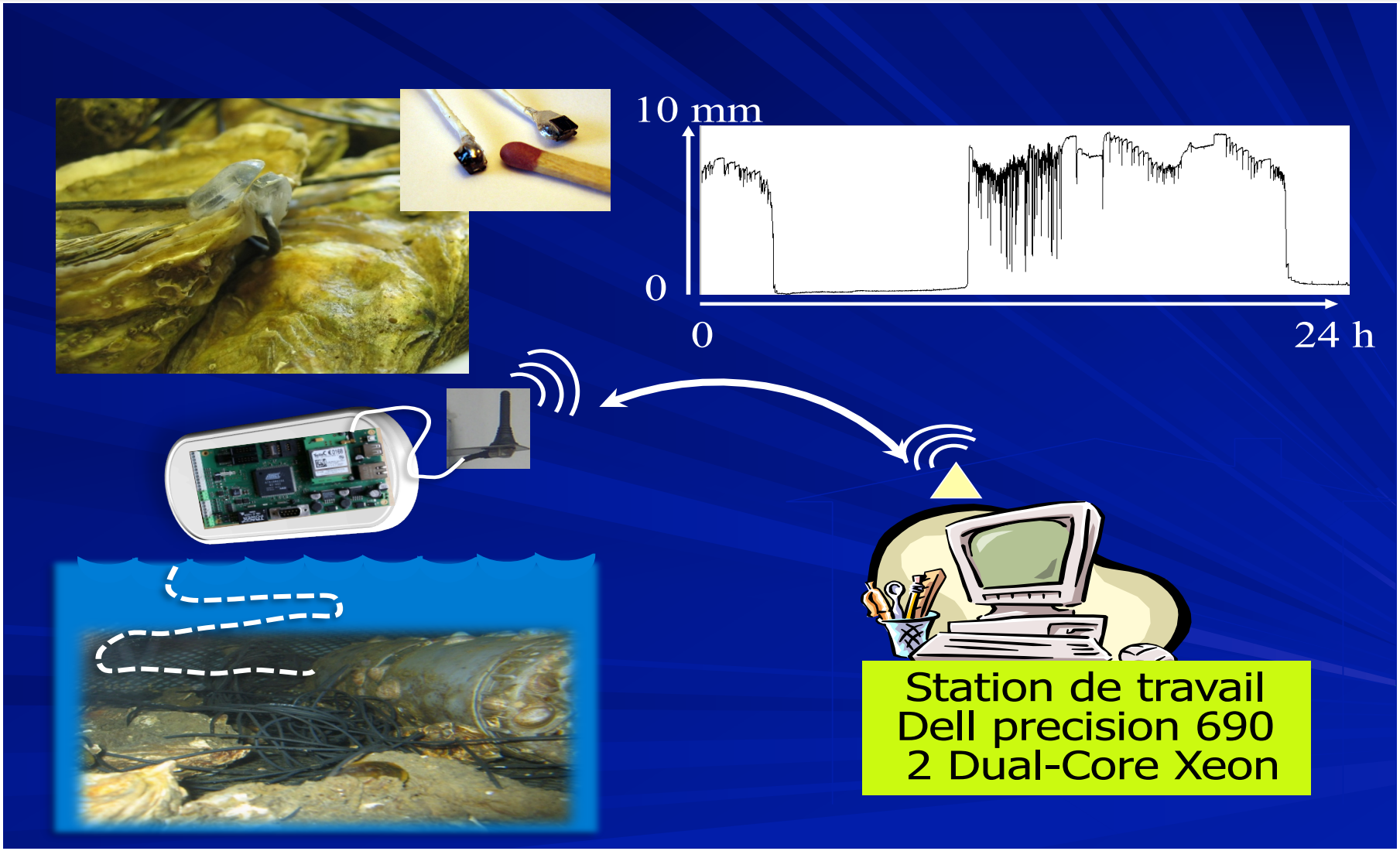
[b]Gilles Durrieu, [a]Damien Tran, [a]Pierre Ciret et [a]Jean-Charles Massabuau

[a]EPOC, UMR CNRS 5805, Université de Bordeaux
[b]LMBA, UMR CNRS 6205, Université de Bretagne sud

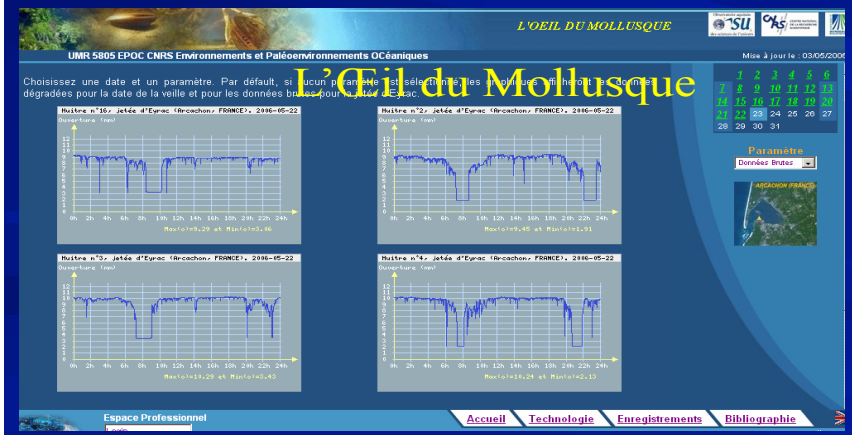
Objectifs

- 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



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 (Ti ;Yi)) pour chaque bivalve ;
- n=864 000 mesures par jour pour 16 bivalves ;
- Plus de 300 000 000 mesures effectuées par an / site.

Environnementales :

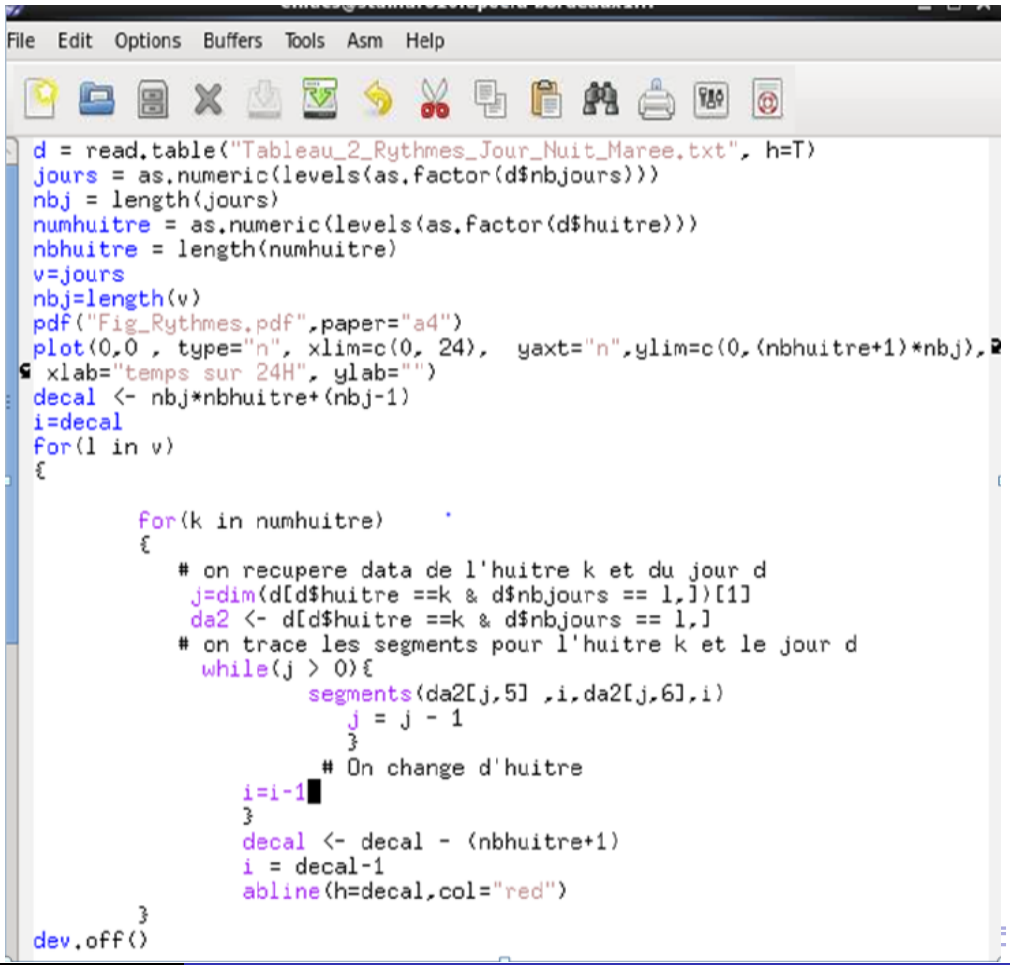
- hauteur d'eau, jour/nuit, température, courant, ...

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$numenm$jm*
gzip $b$numenm$jm*
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
```



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

Modèle de régression non paramétrique

Pour $\mathrm{i} = 1, \ldots, \mathrm{n}$,

$$\mathbf{Y_i} = \mathbf{m}(\mathbf{T_i}) + \varepsilon_\mathbf{i}$$

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)

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

Théorème 1. (Convergence en loi)

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

- $\forall \alpha \in]1/3, 1[$ 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

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

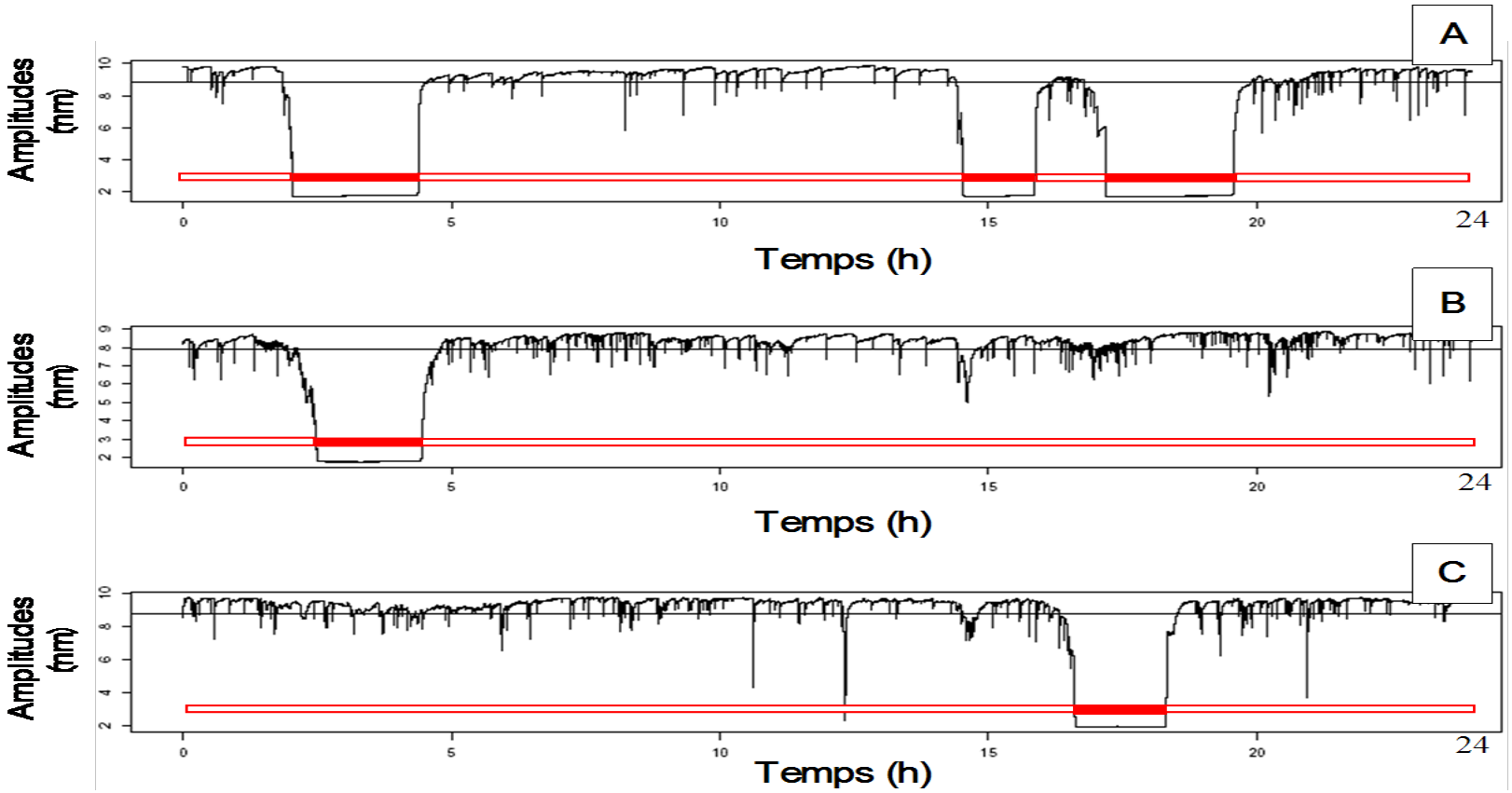
$$\tilde{\sigma}^2(t) = \frac{1}{\tilde{f}_n(t)} \sum_{i=1}^{n} K\left(\frac{t - T_i}{h_i}\right)(Y_i - \tilde{m}_n(t))^2.$$

Recherche de rythme biologique

A

B

C

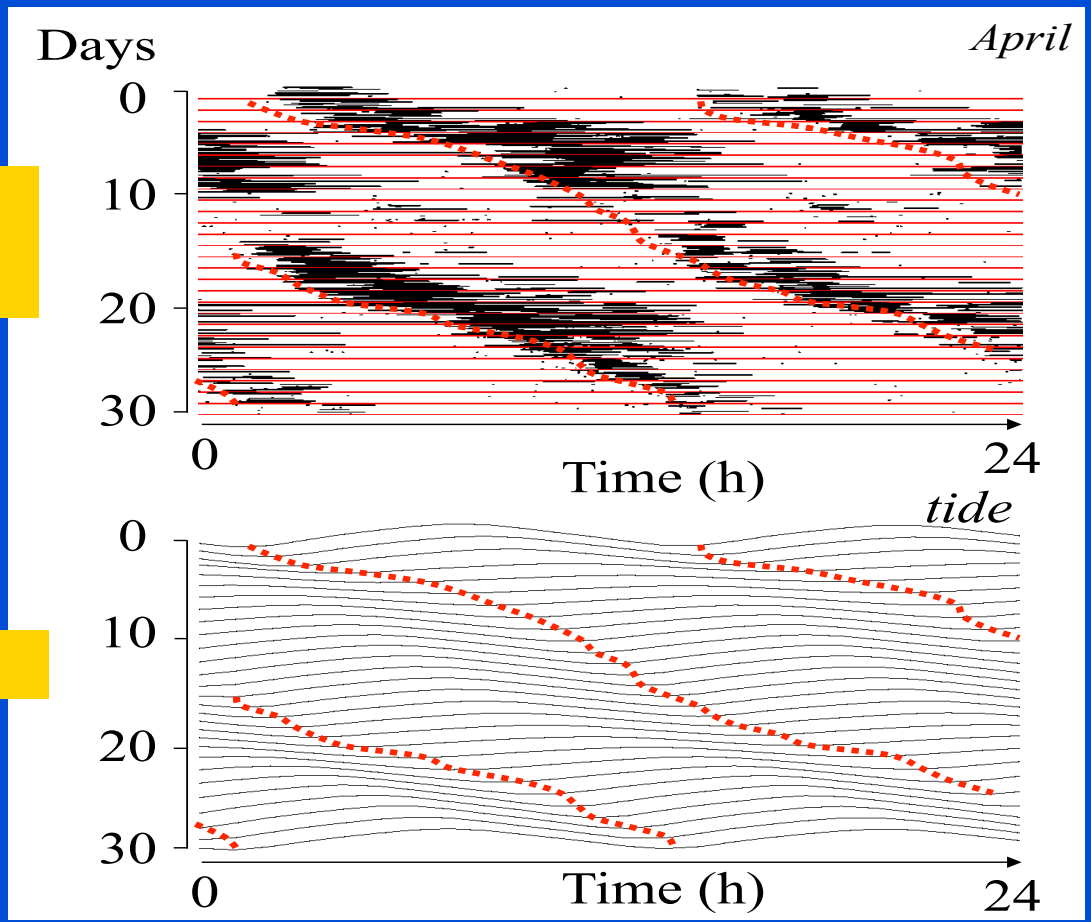


Rythme biologique

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) des données et résultats (tableaux et graphiques)