Statistique simultanée, Statistique spatiale et maximum de processus

Rennes 24 mars 2009

Jean-Marc Azaïs

Laboratoire de Statistique et Probabilités, IMT,Toulouse

modèle signal + bruit

En statistique spatiale on est souvent amené à considérer le modèle "signal + bruit gaussien".
Des exemples de telles situations sont données par

- l'agriculture de précision
- les neurosciences
- les problèmes de modélisation des vagues

modèle signal + bruit

En statistique spatiale on est souvent amené à considérer le modèle "signal + bruit gaussien".
Des exemples de telles situations sont données par

- l'agriculture de précision
- les neurosciences
- les problèmes de modélisation des vagues

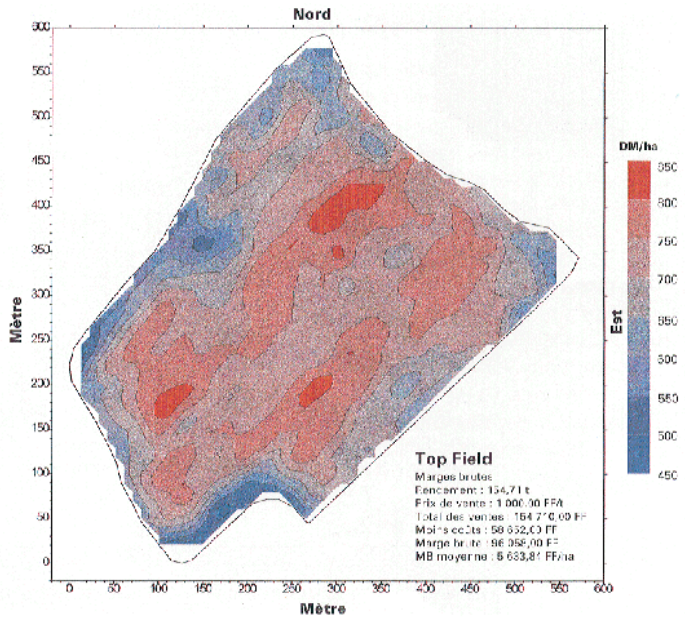
modèle signal + bruit

En statistique spatiale on est souvent amené à considérer le modèle "signal + bruit gaussien".
Des exemples de telles situations sont données par

- l'agriculture de précision
- les neurosciences
- les problèmes de modélisation des vagues

Agriculture de précision

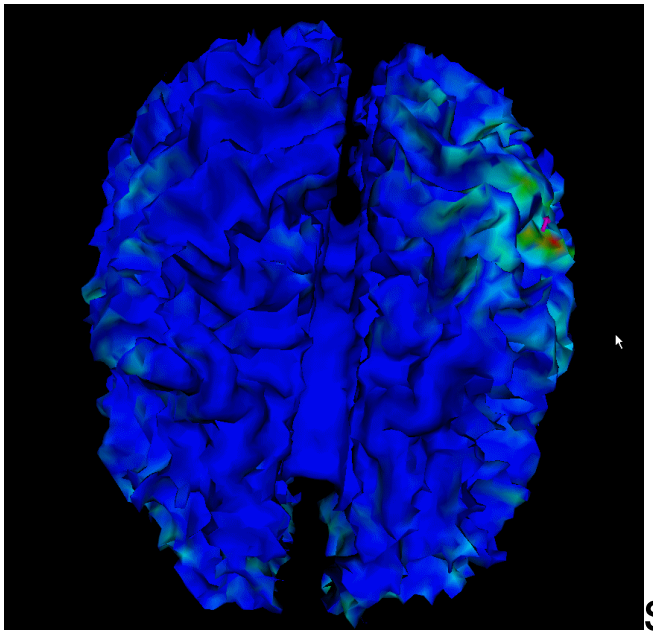
Mesure du rendement par moissonneuse GPS



On considère que l'on a mesuré la variable d'intérêt sur une grille si fine que l'on peut considérer qu'on l'observe sur $\mathbb{R}^2$.

Neuroscience

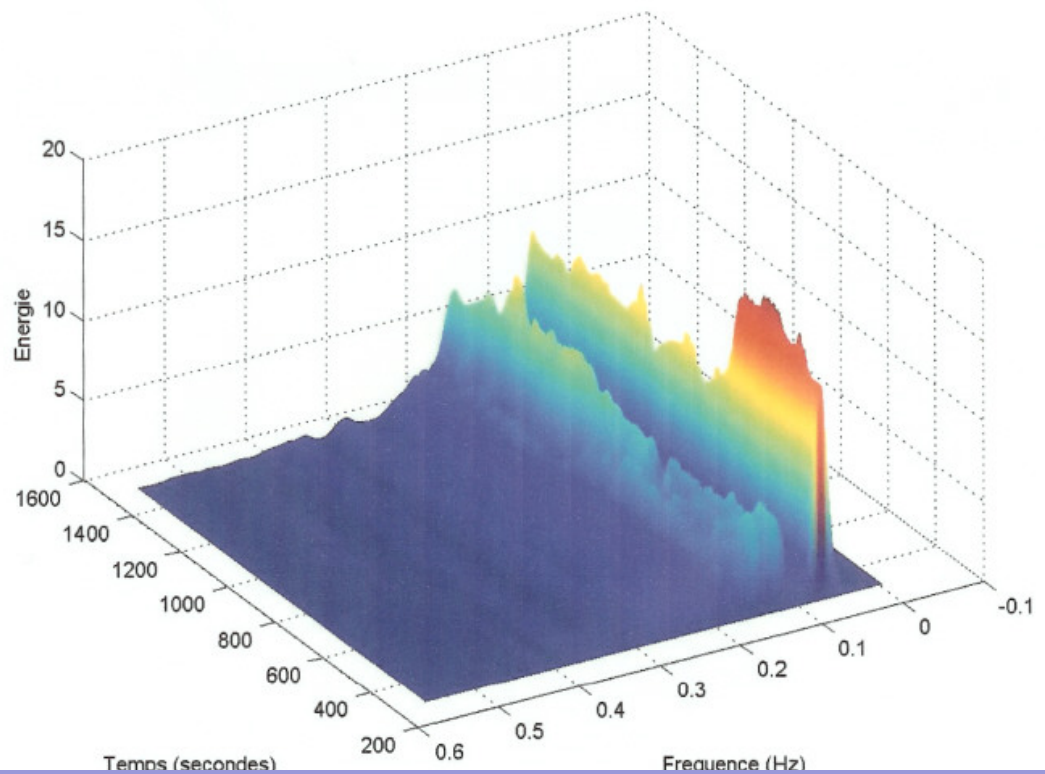
On utilise un modèle 2 dimensionnel ou 3 dimensionnel pour le cerveau et on désire savoir si il existe une zone particulièrement activée par une activité donnée.



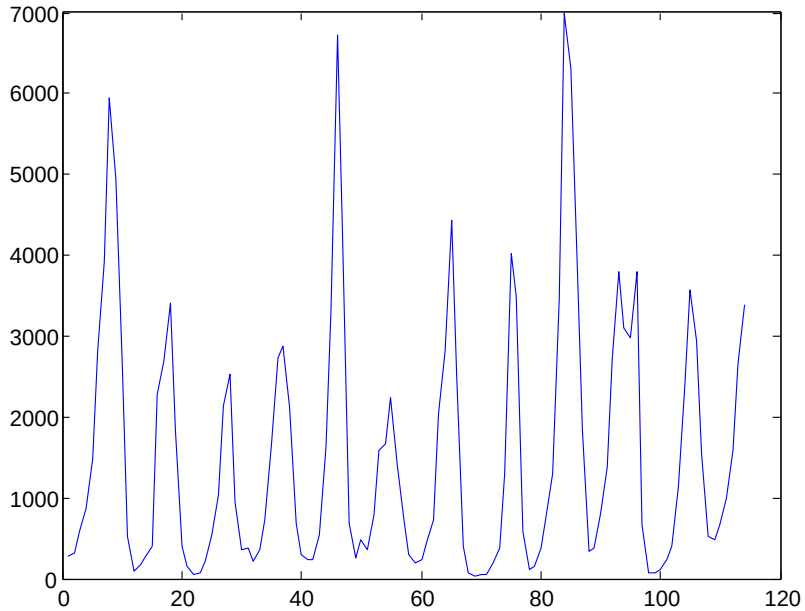
source : Maureen CLERC

Spectre de houle

On mesure localement, en temps, le spectre de vagues et on veut détecter des instants de changement : les transition entre les "états de mer".

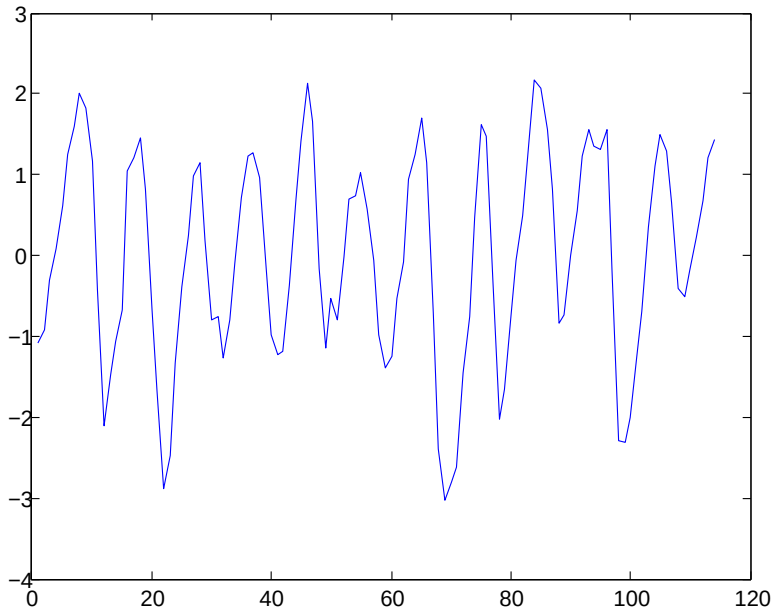


Lynx



Prises annuelles de lynx dans la rivière Mackenzie, Nord ouest du Canada durant la période 1821 - 1934, (Elton and Nicholson, 1942)

On passe en log et on centre.



Test unidimensionnel

On fait les hypothèses discutables suivantes

- les observations sont gaussiennes
- La série des erreurs est **stationnaire et mélangeante**, La pseudo-périodicité est aléatoire, due à un modèle proie prédateur.
- La taille de la série 114 est suffisante pour estimer la variance avec une erreur négligeable par

$$\widehat{\sigma^2} := \frac{1}{n}\sum X_i^2$$

qui donne $1.6387 = 1.28^2$.

Test unidimensionnel

On fait les hypothèses discutables suivantes

- les observations sont gaussiennes
- La série des erreurs est **stationnaire et mélangeante**, La pseudo-périodicité est aléatoire, due à un modèle proie prédateur.
- La taille de la série 114 est suffisante pour estimer la variance avec une erreur négligeable par

$$\widehat{\sigma^2} := \frac{1}{n}\sum X_i^2$$

qui donne $1.6387 = 1.28^2$.

Test unidimensionnel

On fait les hypothèses discutables suivantes

- les observations sont gaussiennes
- La série des erreurs est **stationnaire et mélangeante**, La pseudo-périodicité est aléatoire, due à un modèle proie prédateur.
- La taille de la série 114 est suffisante pour estimer la variance avec une erreur négligeable par

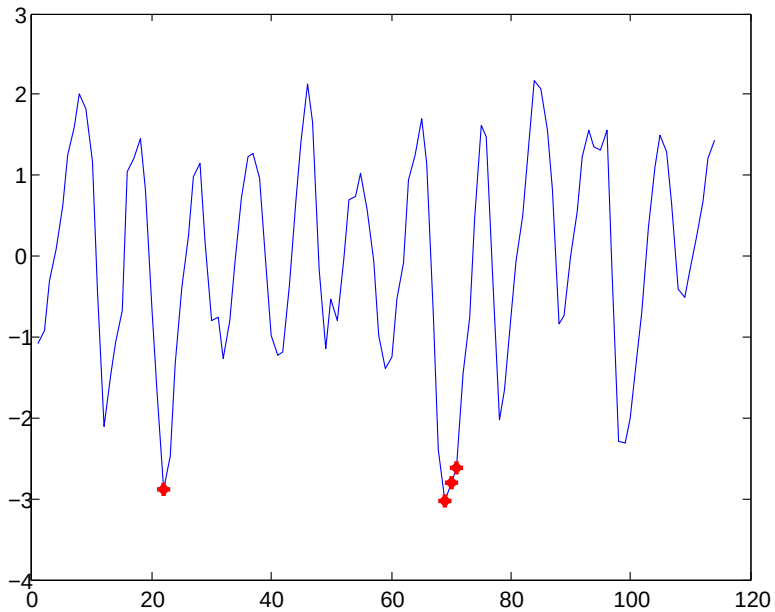
$$\widehat{\sigma^2} := \frac{1}{n}\sum X_i^2$$

qui donne $1.6387 = 1.28^2$.

Sous l'hypothese nulle d'absence de signal $\frac{Y_i}{\widehat{\sigma}}$ suit approximativement un loi normale standard (centrée réduite) D'ou la règle de test

$$\text{si } |\frac{Y_i}{\widehat{\sigma}}| > 1.96$$

on déclare qu'il y a un signal au point i considéré.



Risque simultané

En statistique, pour être schématique on teste généralement à 5%, → probabilité de fausse alerte de 5 %.
Un seul test **OK ! !**
Ici 114 tests avec une proba de fausse alerte de 5 % pour chacun d'entre eux.
La proba de fausse alerte totale mais certainement **très importante ! ! !**
La **statistique simultanée** a pour but de contrôler la probabilité globale de fausse alerte.
La méthode le plus rudimentaire (mais pas toujours la plus mauvaise) est la **méthode de Bonferroni** qui consiste à faire chaque test élémentaire au niveau $\alpha' = \alpha/114$ dans notre cas

$$qnorm(1 - 0.025/114) = 3.5157$$

et après multiplication par l'écart type 1.28 donne **4.5**. Ce qui ne détecte rien **Peut on faire mieux ? ?**

Le maximum de la valeur une série gaussienne

De manière générale la **distribution du maximum** est inconnue même dans les cas les plus simples :
marche aléatoires, processus auto-régressif d'ordre 1. On peut faire des simulations mais c'est souvent long et peu précis .
Un méthode est d'écrire la densité du vecteur gaussien

$$(2\pi)^{-n/2}\frac{1}{\det(\Sigma)}\exp-\left(\frac{\mathbf{x}'\Sigma^{-1}\mathbf{x}}{2}\right).$$

et de l'intégrer sur un hyper-rectangle $[-u,u]^n$. Cela se fait numériquement par des méthodes fort complexes que je ne vais pas décrire **pour des tailles jusqu'à 1000**

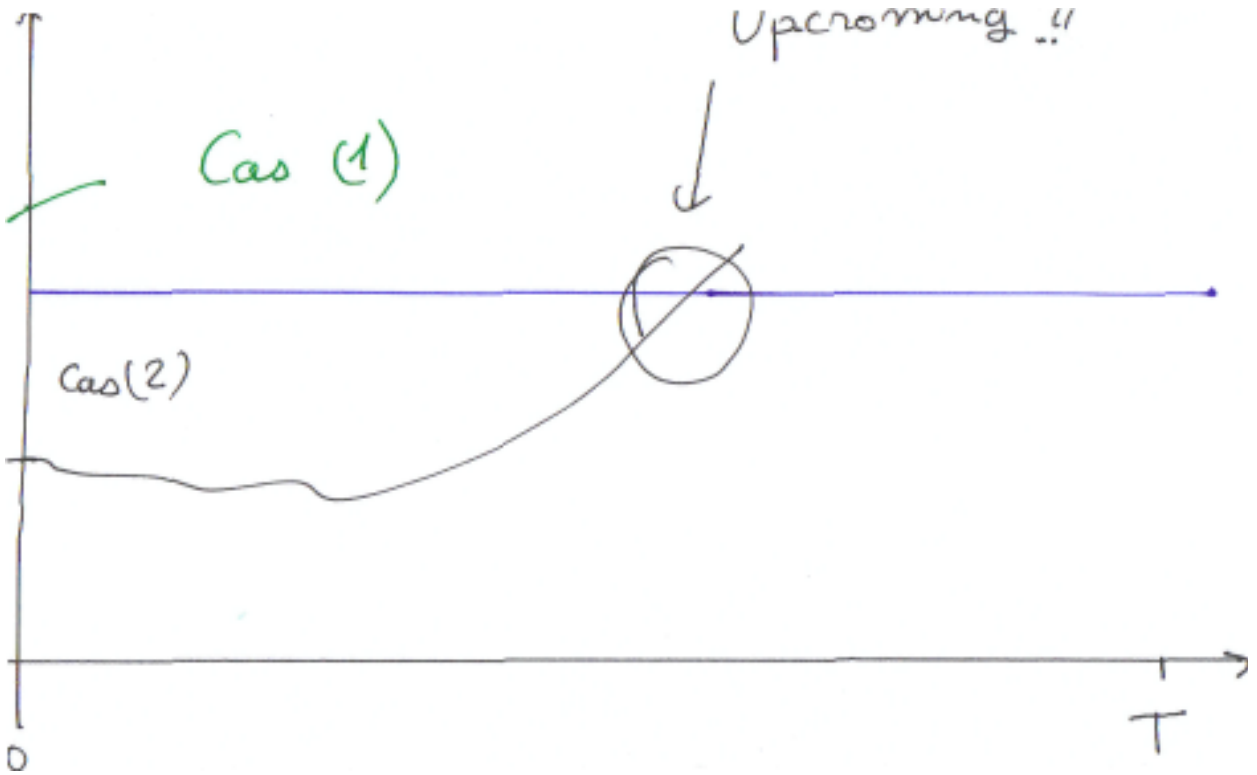
On trouve en utilisant la matrice de variance estimée, un niveau de signification de **0.4978** ce qui est clairement non significatif.

On suppose

- Que l'on a tant de point que l'on peut considérer que l'on observe entièrement la fonction de la variable réelle $X(t)$.
- Que le phénomène aléatoire considéré est régulier (dérivable)
- on considère le maximum M (sans valeur absolue pour simplifier) sur un intervalle borné par exemple $[0, T]$.

on utilise la méthode de Rice qui est basée sur les inégalités basiques suivantes

$$\mathbb{P}\{M > u\} \leq \mathbb{P}\{X(0) > u\} + \mathbb{P}\{U_u > 0\} \leq \mathbb{P}\{X(0) > u\} + \mathrm{E}(U_u)$$


Upcrossing !!
Cas (1)
Cas(2)
0
T

Sur les franchissements d'un processus la seule chose que l'on sache compter : **moments**

$$\mathrm{E}(U_u) = \int_0^T \mathrm{E}\big((X')^+(t)\big|X(t) = u\big)p_{X(t)}(u)dt$$

qui a une version très simple dans le cas stationnaire, centré, variance 1.

$$\mathrm{E}(U_u) = T\sqrt{\frac{\lambda_2}{2\pi}}\phi(u)$$

ou λ_2 est le second moment spectral : la variance de la dérivée.

Une précision super-exponentielle

Sous certaines hypothèses

$$\mathbb{P}\{M_T > u\} = 1 - \Phi(u) + T\sqrt{\frac{\lambda_2}{2\pi}}\phi(u) + O\big(\phi(u(1+\delta))\big)$$

Montre la qualité de la borne, montre également que la forme exacte de la covariance importe peu.

Retour aux problèmes de l'introduction

On considère une fonction aléatoire sur $\mathbb{R}^2$ (pour simplifier) : un champ aléatoire. Le nombre de franchissement $\Rightarrow$ courbe (de niveau) : ne permet pas de construire de bornes.
Il faut remplacer par un autre caractéristique géométrique, par exemple le **nombre de maxima locaux au dessus d'un certain niveau**.
En négligeant les effets bords

$$\mathbb{P}\{M > u\} \simeq \mathbb{P}\{\exists \text{ maximum local au dessus de } u\}$$

$$\simeq \mathrm{E}(\ \#(\text{maxima locaux au dessus de } u))$$

qui peut être calculé par un formule de Rice comme précédemment

Theorème

Considérons le carré $[O,T]^2$, alors si le champ est centré et de variance 1, sous certaines conditions de régularité :

$$\mathbb{P}\{M > u\} = \frac{u\phi(u)}{2\pi} T^2 |\Lambda|^{1/2} + \frac{\phi(u)}{\sqrt{2\pi}} T\left(\sqrt{\Lambda_{11}} + \sqrt{\Lambda_{22}}\right)$$

$$+1 - \Phi(u) + O\left(\phi(u(1+\delta))\right)$$

où Λ est la matrice de variance-covariance du gradient.

Résultats dû a divers auteurs sous diverses conditions. Piterbarg (1981), Taylor Takemura Adler (2005) , Azaïs Wschebor (2008). On peut même "facturer" le δ

Conclusion

En supposant que dans les exemples présentés début le bruit véifie nos hypothèses de régularité, nous sommes capables de calculer une valeur qu'il ne devrait pas dépasser c'est à dire **la valeur critique du test simultané**

MERCI