

# Explorer des modèles par simulation

*Vincent Ginot*

INRA-Biométrie

Unité de statistiques spatiales d'Avignon

Vincent.ginot@avignon.inra.fr

<http://www.avignon.inra.fr/mobidyc>

## Pourquoi *explorer* un modèle ?

- Cohérence numérique (dans le temps et l'espace)
- **Analyses de sensibilité**
- Analyses de stabilité
- Analyses de robustesse
- Analyses des conditions initiales et de propagation des erreurs
- Identification des paramètres
- **Analyses d'incertitudes**
- Méta-modélisation : et si on faisait le modèle du modèle ?

Pourquoi par simulation ?

traditionnellement, étude analytique des systèmes  
d'équations différentiels EDO et EDP (+ matrice/Markov)

mais...

- Non-existence ou complexité des solutions analytiques (grande dimension des systèmes)
- Stochasticité des modèles
- Non-existence de formulation analytique (MIC-SMA)
- Propriétés émergentes : changement de variable d'état

Mais...

Quid *juste à côté* des points explorés ???

Pas de preuves formelles....

## Définition mathématique : sensibilité locale

Coefficient de sensibilité local :

$$s_{i,j,t} = \frac{\partial M_{j,t}}{\partial p_i}$$

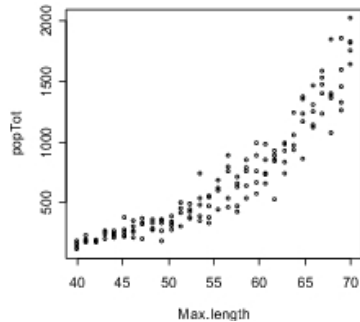
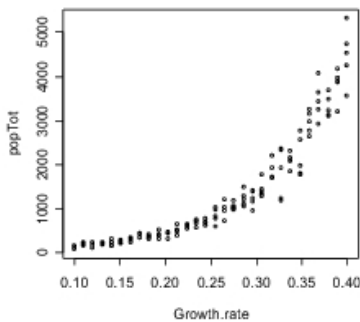
Coefficient normalisé :

$$ss_{i,j,t} = \frac{\partial M_{j,t}}{\partial p_i} \Delta p_i$$

Coefficient relatif  
ou « élasticité » :

$$rs_{i,j,t} = \frac{\partial M_{j,t}}{\partial p_i} \frac{p_i}{M_{j,t}} = \frac{\partial \ln(M_{j,t})}{\partial \ln(p_i)}$$

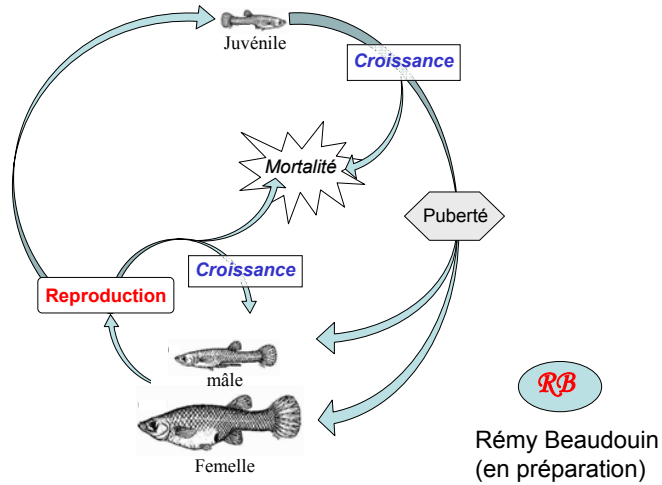
Sensibilité locale = pente de la réponse  
du modèle à une variation du paramètre



Conséquences :

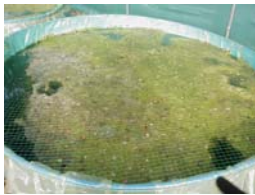
- La sensibilité dépend de la valeur du paramètre sauf si la courbe est une droite
- Elle dépend aussi de la valeur des autres paramètres, du temps, et de la sortie.
- La sensibilité « locale » ne s'applique qu'aux variables dérivables

## Cycle de vie de la gambusie



La gambusie est utilisée pour des expériences d'écotoxicologie

RB



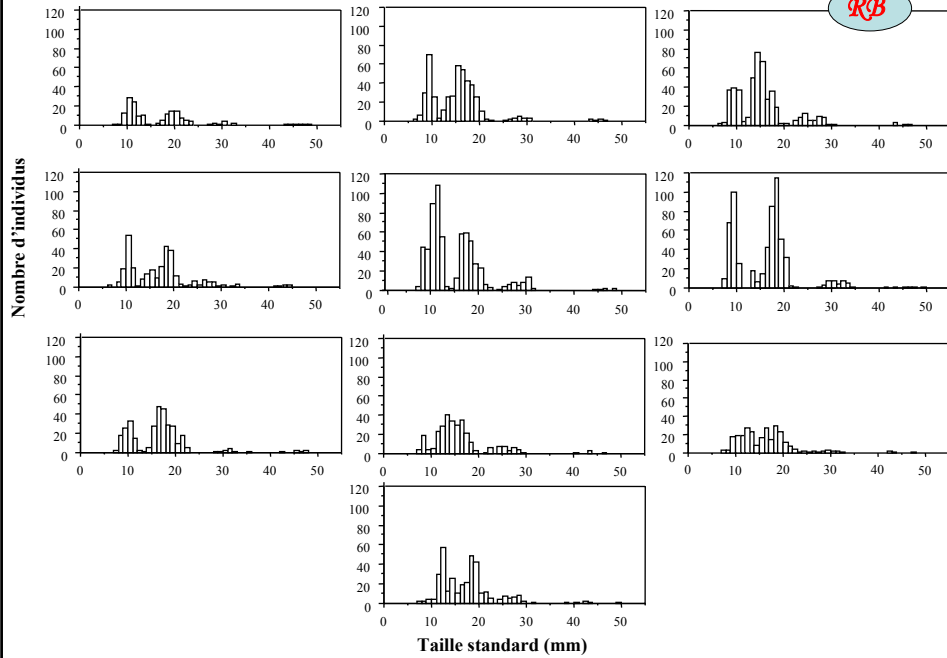
### Les mésocosmes :

minis eco-systèmes dans lesquels on développe des populations de gambusies soumises ou non à des pesticides



## Populations obtenues en mésocosmes en 2002

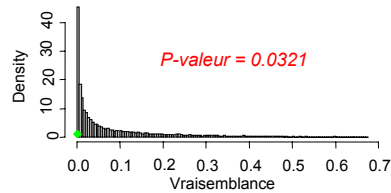
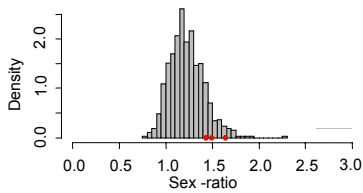
RB



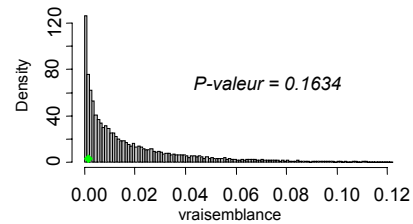
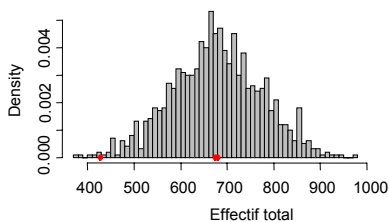
Le modèle (une fois validé !) sert de référentiel pour analyser les observations

RB

⇒ La sex-ratio



⇒ Effectif total

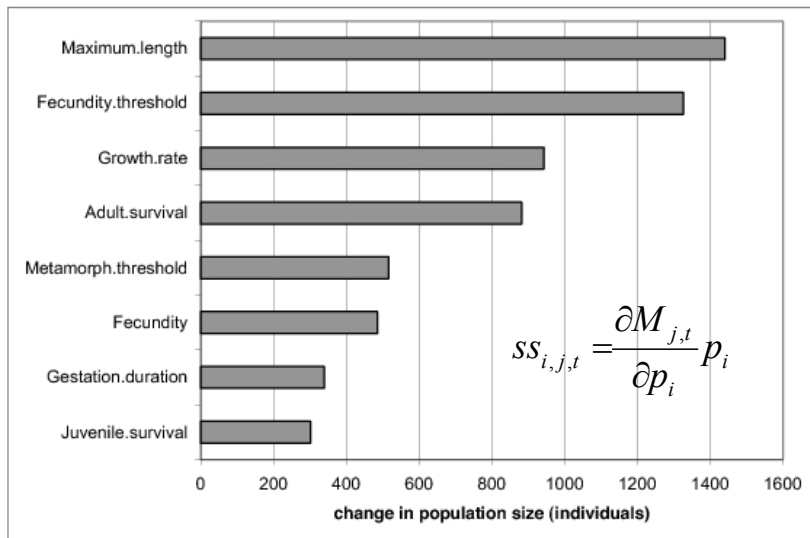


## Modèle gambusies : les paramètres

<i>Parameter</i>	<i>unit</i>	<i>Range</i>	<i>Réf. value</i>	<i>ANOVA levels</i>
Growth rate	month <sup>-1</sup>	0.10 – 0.4	0.25	0.15 ; 0.2 ; 0.25 ; 0.30
Maximal length	mm	40 – 70	60	50 ; 70
Metamorphosis threshold	mm	12 – 18	15	13 ; 17
Adult survival	month <sup>-1</sup>	50 – 100 %	90	60 ; 90
Juvenile survival	month <sup>-1</sup>	20 – 40	30	20 ; 40
Gestation duration	days	20 – 40	30	25 ; 35
Fecundity threshold	mm	20 – 26	24	22 ; 26
Fecundity	juveniles.mm <sup>-1</sup>	1 – 4	2	1.5 ; 3

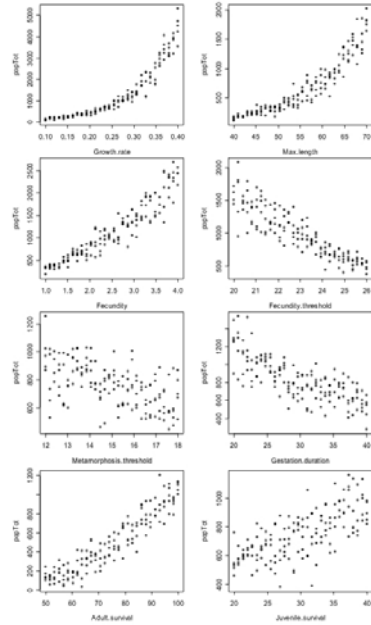
Question piège : donner à priori le tiercé des trois paramètres les plus influents sur la sortie étudiée, la population totale en fin de simulation

## Sensibilité locale : hiérarchie des paramètres (gambusies)



Remarque : n'est valable que dans les conditions de la simulation !!!

Modèle gambusies :  
 Signatures individuelles.  
 Sortie = population totale  
 au dernier pas de temps

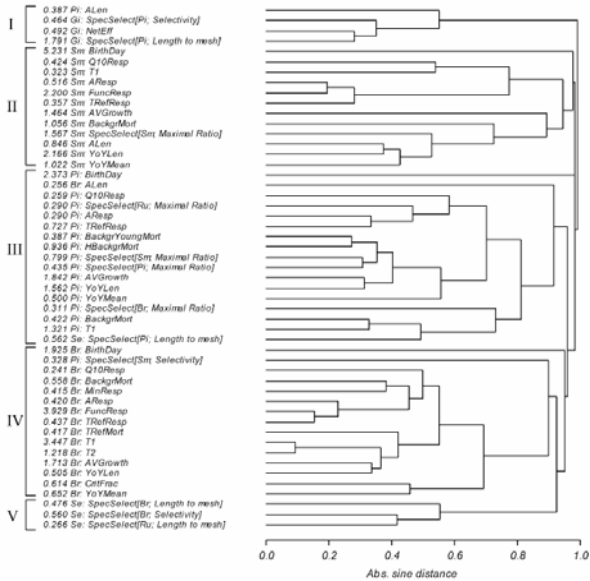


## Sensibilités locales : Explorer par Classification

Piscator  
 (van Nes et al. 2002)

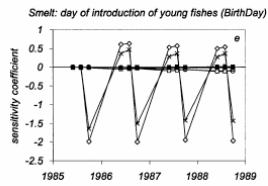
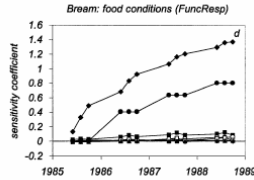
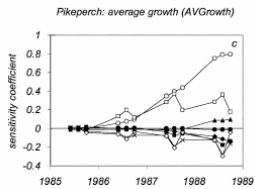
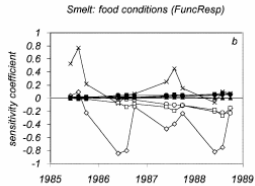
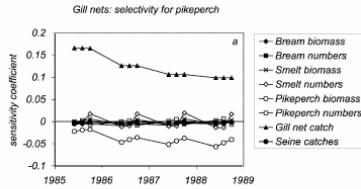
- IBM multi-spécifique
- ± 15 variables
- 380 paramètres dont  
 66 « efficaces »
- 12 pas de temps  
 mémorisés

$$rs_{i,j,t} = \frac{\partial M_{j,t}}{\partial p_i} \frac{p_i}{M_{j,t}}$$

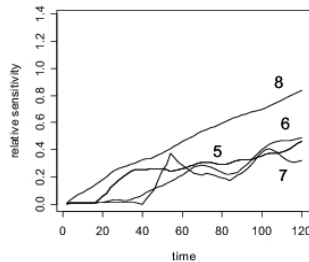
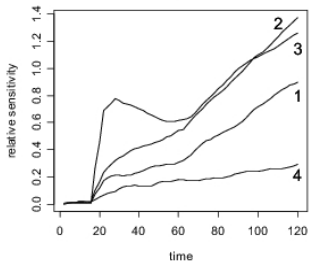
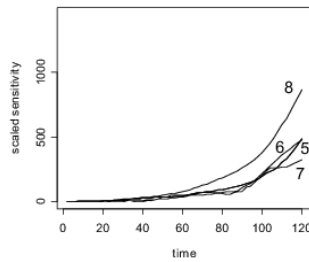
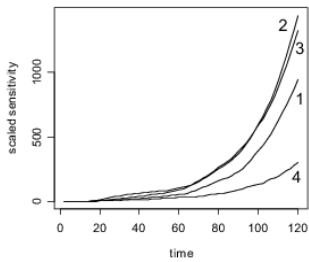


# Explorer par Classification

Piscator  
(van Nes et al. 2002)



## Modèle gambusies. Sensibilités locales dans l'espace des observations (ici le temps) : identifiabilité (1)



## Modèle gambusies. Identifiabilité (2). Corrélations entre fonctions de sensibilités

	Grth.rate	Max.lgth	Fec.thld	Juv.surv	Fecund	Met.thld	Gest.dur	Ad.surv
Growth.rate	1.000							
Max.length	0.995	1.000						
Fec.threshold	-0.906	-0.918	1.000					
Juv.survival	0.926	0.951	-0.930	1.000				
Fecundity	0.896	0.925	-0.914	0.968	1.000			
Metam.thld	-0.956	-0.954	0.797	-0.892	-0.864	1.000		
Gest.duration	-0.782	-0.792	0.636	-0.803	-0.731	0.867	1.000	
Adult.survival	0.975	0.986	-0.888	0.959	0.932	-0.954	-0.819	1.000

## Analyses locales et globales

	Local (A)	Global (B)
Faire varier Un par Un (I)	<ul style="list-style-type: none"> <li>- hiérarchie des paramètres</li> <li>- CAH, ACP si nombreux paramètres / sorties</li> <li>- fonctions de sensibilités, identifiabilité</li> </ul>	<ul style="list-style-type: none"> <li>- signatures graphiques individuelles</li> </ul>
Faire varier Tous ensemble (II)	<ul style="list-style-type: none"> <li>Idem I-A</li> <li>+ Analyses d'incertitudes</li> </ul>	<ul style="list-style-type: none"> <li>- signatures individuelles (tendances + interactions)</li> <li>- recherche des facteurs principaux</li> <li>- recherche des interactions</li> <li>- méta-modélisation ?</li> </ul>

**Case II-B :**  
Faire varier tous les paramètres simultanément  
dans une grande gamme de valeurs

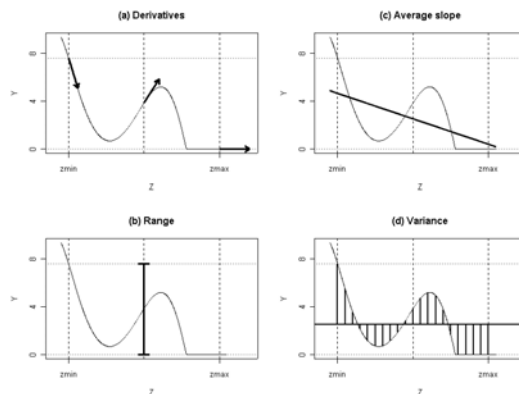
**Intérêts :**

- Non centré sur un point de fonctionnement particulier
- Hiérarchie des paramètres
- Décomposition en facteurs principaux et interactions

**Difficultés :**

- Comment « résumer » une fonction de sensibilité ?
- Comment explorer l'espace des paramètres ?

**Sensibilité globale : comment résumer la variabilité ?**



(Hervé Monod, 2005, à paraître)

## Sensibilité globale : comment échantillonner ?

- 1- Les plans factoriels  
complets  
fractionnés
- 2- L'échantillonnage aléatoire (Monte Carlo)  
aléatoire pur  
hyper-cubes latins  
marches gironnées

**Mais l'espace est rapidement vide !!!!!**

$$4^{10} = 10^6 \text{ simulations}$$

$$4^{20} = 10^{12} \text{ simulations}$$

## Trois grandes classes de méthodes

- 1- La régression multiple (surtout en incertitude)
- 2- L'analyse de variance
- 3- Les décompositions de variance « sans modèles »

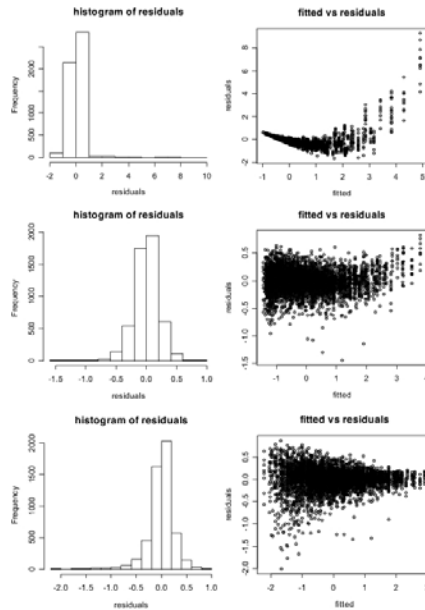
## Gambusies : Analyse de variance

Plan factoriel complet,  
7 paramètres à deux niveaux,  
1 paramètre à 4 niveaux, 10  
réplications.  
 $2^7 \cdot 4 \cdot 10 = 5120$  simulations.

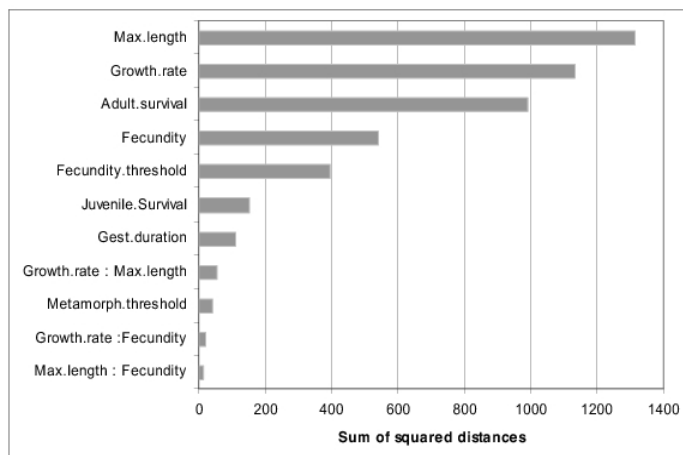
Sortie : population totale au  
dernier pas de temps.

Avec deux niveaux seulement,  
il faut «linéariser» la relation.  
Haut, données brutes, milieu  
transformation Box-Cox, bas,  
transfo log. On retient le log.

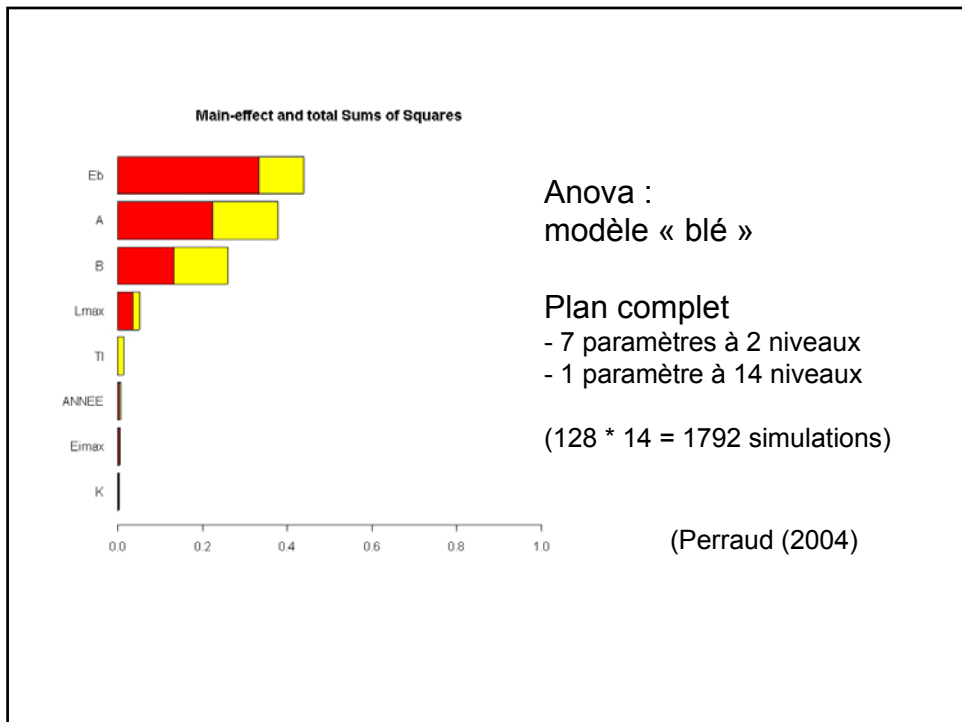
(Ginot et al., 2005 à paraître)



## Anova : modèle « gambusies » Effets principaux et interactions



(Ginot et al., à paraître)



Anova :  
modèle « blé »

Plan complet

- 7 paramètres à 2 niveaux
- 1 paramètre à 14 niveaux

(128 \* 14 = 1792 simulations)

## ANOVA

### Intérêts :

- Echantillonnage « contrôlé » : plans complets ou non avec choix du niveau d'interaction
- Données discrètes et continues
- Modèles stochastiques ou déterministes
- Accès à la forme de la réponse (polygones orthogonaux)
- Résultats fiables, dans le cadre des hypothèses (on peut tester si la méthode s'applique !)

### Inconvénients :

- Il faut définir les niveaux pour chaque paramètres
- Réponses « pathologiques » mal prises en compte

Sobol

- Principe : paire de N échantillons avec
- une valeur commune sur un paramètre
  - deux valeurs communes (interaction / 1)
  - toutes communes sauf 1 (total)

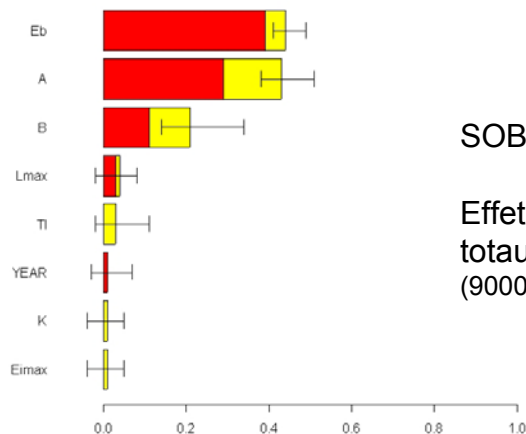
Nécessite  $2 \cdot N \cdot p$  ou  
 $N \cdot (p+1)$  simulations

$$\hat{f} = \frac{1}{2N} \sum_{k=1}^N [f(\theta_{1,k}) + f(\theta_{2,k})]$$

$$\hat{D} = \frac{1}{2N} \sum_{k=1}^N [f(\theta_{1,k})^2 + f(\theta_{2,k})^2] - \hat{f}^2$$

$$\hat{D}_i = \frac{1}{N} \sum_{k=1}^N [f(\theta_{1,k}) * f(\theta_{2,k})] - \hat{f}^2$$

First-order and global sensitivity indices



SOBOL :

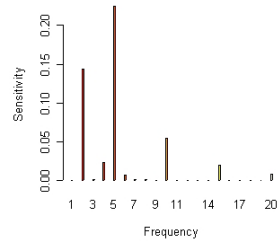
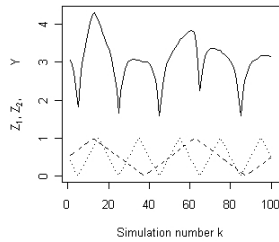
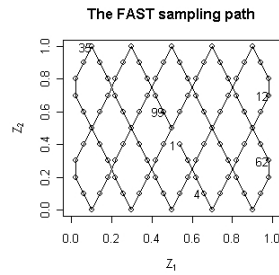
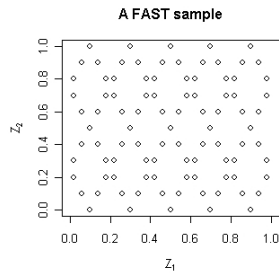
Effets principaux et  
totaux  
(9000 \* 20 simulations)

(Perraud (2004))

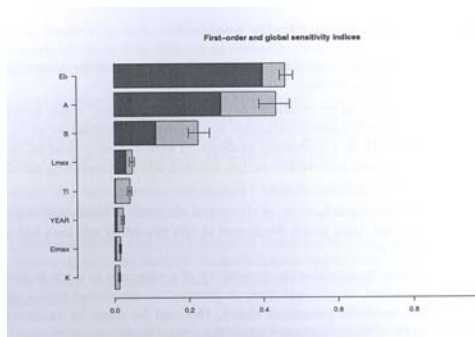
# FAST

Principe d'exploration de l'espace

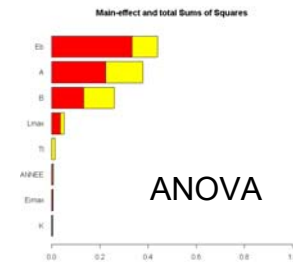
(Hervé Monod, 2005, à paraître)



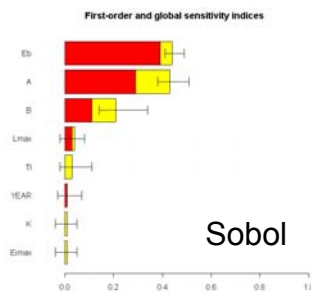
Modèle « blé » (Perraud (2004))  
FAST étendue : facteurs principaux et interactions



FAST  
(8000 \* 20 simulations)



ANOVA



Sobol

## Méthodes « libres » de modèles

### Avantages :

- Fonctionne quelle que soit le lien paramètre / réponse
- Facile et rapide à mettre en œuvre

### Inconvénients :

- Incertitude difficile à chiffrer (mais cela devrait s'arranger ?)
- Gourmand en simulations
- Applicabilité au stochastique et au discret ?  
(en principe oui, mais en pratique...)

## Résumé

### **Local :**

- Hiérarchies au voisinage d'un point singulier
- Etudes d'identifiabilité
- Facile à mettre en œuvre
- Pas trop gourmand en simulation  
(=> **dizaines voire centaines de paramètres**)

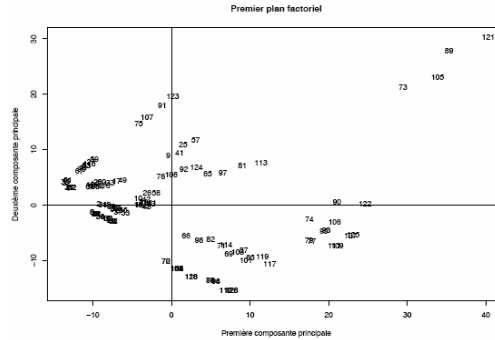
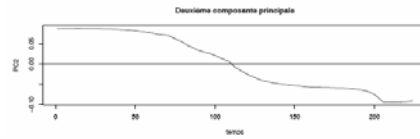
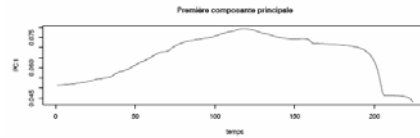
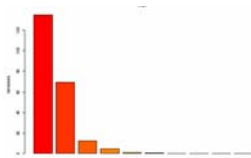
### **Global :**

- Non restreint à un point
- Accès aux interactions
- Difficile d'explorer tout l'espace  
(=> différents choix stratégiques...)
- Gourmand en simulation  
(=> **une dizaine de paramètres**)

## Sensibilité globale : Explorer directement l'espace des sorties par ACP

(Perraud 2004)

- EDO journalier déterministe
- 2 variables état (MS et ISF)
- 7 paramètres, 2 niveaux
- => 128 simulations
- ACP centrée réduite



=> Combiner les méthodes

*Analyses locales* + analyses factorielles

Ou

Algo de sélection

=> sélection des plus importants

Puis

*Analyses globales* sur le sous groupe

## Conclusion

Les méthodes sont souvent simples, ne pas hésiter à en utiliser plusieurs !

=> Mais il faut les rendre accessibles !!!

Encore trop peu de logiciels spécifiques à la simulation...

- R, SASS, SPSS (BAO statistique générique)
- SimLab (Saltelli, ISPRA) : méthodes Sobol et Fast
- ASSA (Jansen, Wageningen)
- MCAT (Imperial College, Londres)
- ...

### Et des initiatives en France !

- projet de réseau INRA
- projet SimExplorer du Cemagref
- projet collectif TEAMSS (Toulouse - Amblard)
- projet VLE (Lille - Ramat)
- (Association Experimentique (IUT Orléans))
- ...

### Les développements à surveiller...

- La géostatistique dans l'espace des paramètres
- sensibilités d'ordre 2 (dérivées secondes) ?
- La méta-modélisation (faire le modèle du modèle...)
- Intelligence Artificielle et Analyses de Sensibilités (projet Dream/Sacadeau, INRIA/INRA)
- Les données non scalaires...
- ...

## Bibliographie (1 - Analyses de sensibilité)

- *Les grands classiques (mais il n'y a pas foule d'ouvrages...)*
- Kleijnen, J.P.C., 1987. Statistical tools for simulation practitioners. Marcel Dekker, inc., NY, 429 pp. (excellent mais épuisé)
- Banks, J., ed., 1998. Handbook of Simulation. Principles, Methodology, Advances, Applications, and Practice. Wiley & Sons, NY, 849 pp. (un peu moins bon que le premier)
- Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. *Statistical Science* 4 (4) 409-435.
- *La référence la plus actuelle dans le domaine*
- Saltelli, Chan et Scott (Eds) (2000). Sensitivity analysis. Wiley
- *plus livre d'applications avec SimLab, logiciel libre développé dans leur équipe*
- Saltelli, Tarantola, Campolongo, Ratto, 2004. Sensitivity Analysis in practice. Wiley.
- *Analyses en appliquant les principes de la géo-statistique dans l'espace des paramètres*
- Santner T.J., Williams B.J., Notz W.I. (2003). The design and analysis of computer experiments. Springer.
- *Un exemple d'analyse de sensibilité et de classification*
- Van Nes, E.H., Lammens, E.H.R., Scheffer, M., 2002. Piscator, an individual-based model to analyze the dynamics of lake fish communities. *Ecol. Model.* 152 (2002) 261-278.
- *Sensibilité et ANOVA*
- Ginot V., Gaba S., Baudouin R., Ariès F., Monod H. (2005, sous presse). Combined use of local and ANOVA-based global sensitivity analyses for the investigation of a stochastic dynamic model: application to the case study of an Individual Based Model of a fish population. *Ecological Modelling* (sous presse)
- *Mise en oeuvre de la méthode globale de Sobol*
- Salvador, R., Pinol, J., Tarantola, S., Pla, E., 2001. Global sensitivity analysis and scale effects of a fire propagation model used over Mediterranean shrublands. *Ecol. Model.* 136 (2001) 175-189.
- Fieberg, J., Jenkins, K.J., 2005. Assessing uncertainty in ecological systems using global sensitivity analyses: a case example of simulated wolf reintroduction effects on elk. *Ecol. Model.* (2005) in press.

## Bibliographie (2)

- **Plans d'expériences (sur conseils d'Hervé Monod, INRA-MIA Jouy)**
- *Une bonne introduction :*
- Box, Hunter et Hunter (1978). Statistics for experimenters. An introduction to design, data analysis, and model building. Wiley
- *Le standard :*
- Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Designs. Oxford University Press.
- *En français, avec prise en compte de la décomposition en polygones orthogonaux :*
- Kobilinsky, A., 1997. Les plans factoriels. In: Plans d'Expériences, Applications à l'Entreprise. J.-J. Drosbeke, J. Fine and G. Saporta (Eds.), pp. 69-209. Editions Technip, Paris.
- Bons articles (vu par un statisticien):
- Sacks, Welch, Mitchell & Wynn (1989). Design and analysis of computer experiments. *Statistical Science*, 4, 409-435.
- Welch, Buck, Sacks, Wynn, Mitchell & Morris (1992). Screening, predicting, and computer experiments. *Technometrics*, 34, 15-47.
- **Un peu de pub pour de bons bouquins en français !**
- Coquillard P. & Hill D. R. C., 1997. Modélisation et simulation d'écosystèmes. Des Modèles déterministes aux simulations à événements discrets. Masson, Paris. 274 p.
- Ferber J., 1995. *Les systèmes multi-agents. Vers une intelligence collective*, InterEditions, Paris.
- (version anglaise, 1997, Addison-Wesley, 509 p)
- Walter Eric & Pronzato Luc, 1994. Identification de modèles paramétriques à partir de données expérimentales. Masson, Paris (version anglaise, Springer 97)