Cartographie de l'abondance pour gros jeux de données et/ou représentation fine de l'espace

Vincent Garreta

INRA, Unité BIOGER-CPP et BioSP, basé à Avignon

Réunion ModStatSP, 27 et 28 Août 2012

Travail en partie réalisé avec l'aide de
John Haslett, Daniel Simpson et Brian Huntley

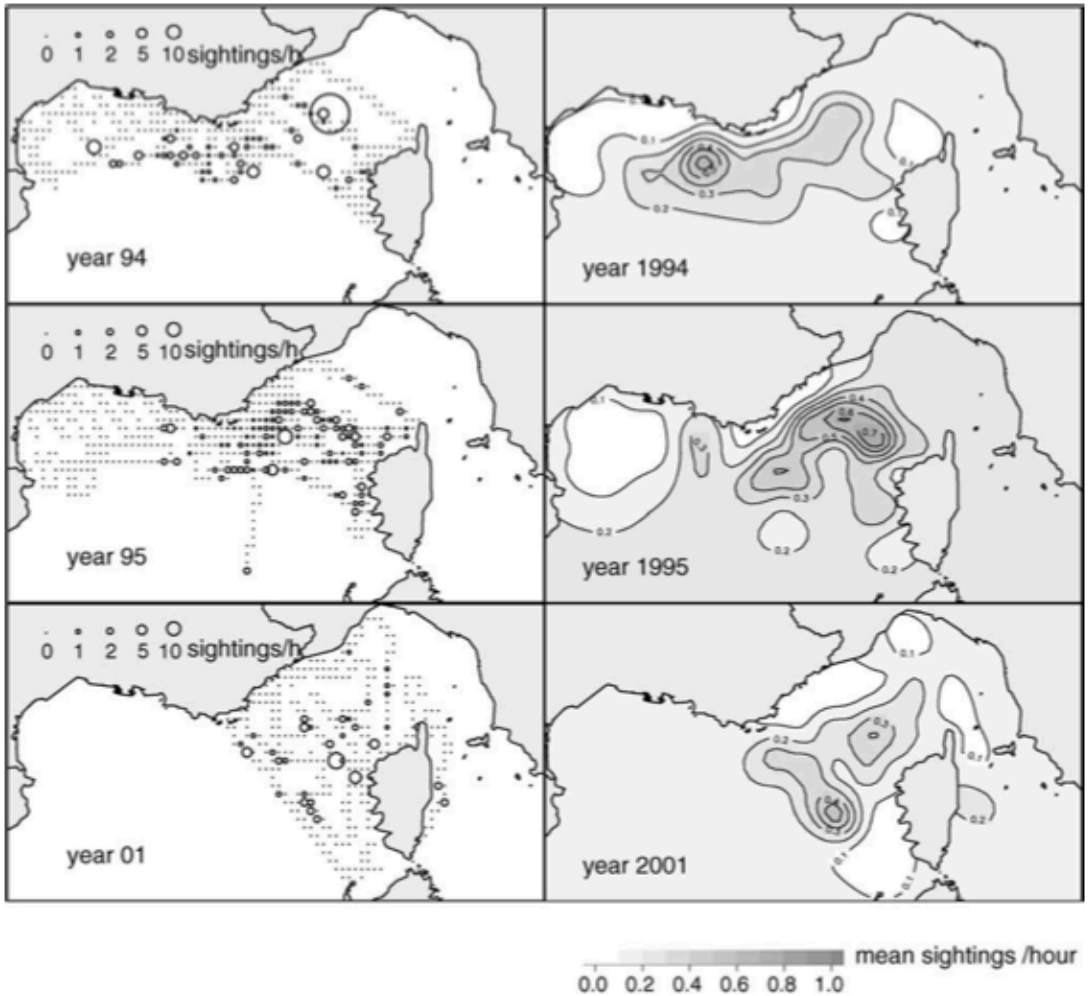
v.garreta@avignon.inra.fr

Postdoc financé par le projet européen PlantFoodSec

La "cartographie" en statistiques

Lissage et d'interpolation de données réparties dans l'espace. Exploitation de

- La liaison spatiale "intrinsèque" entre les données

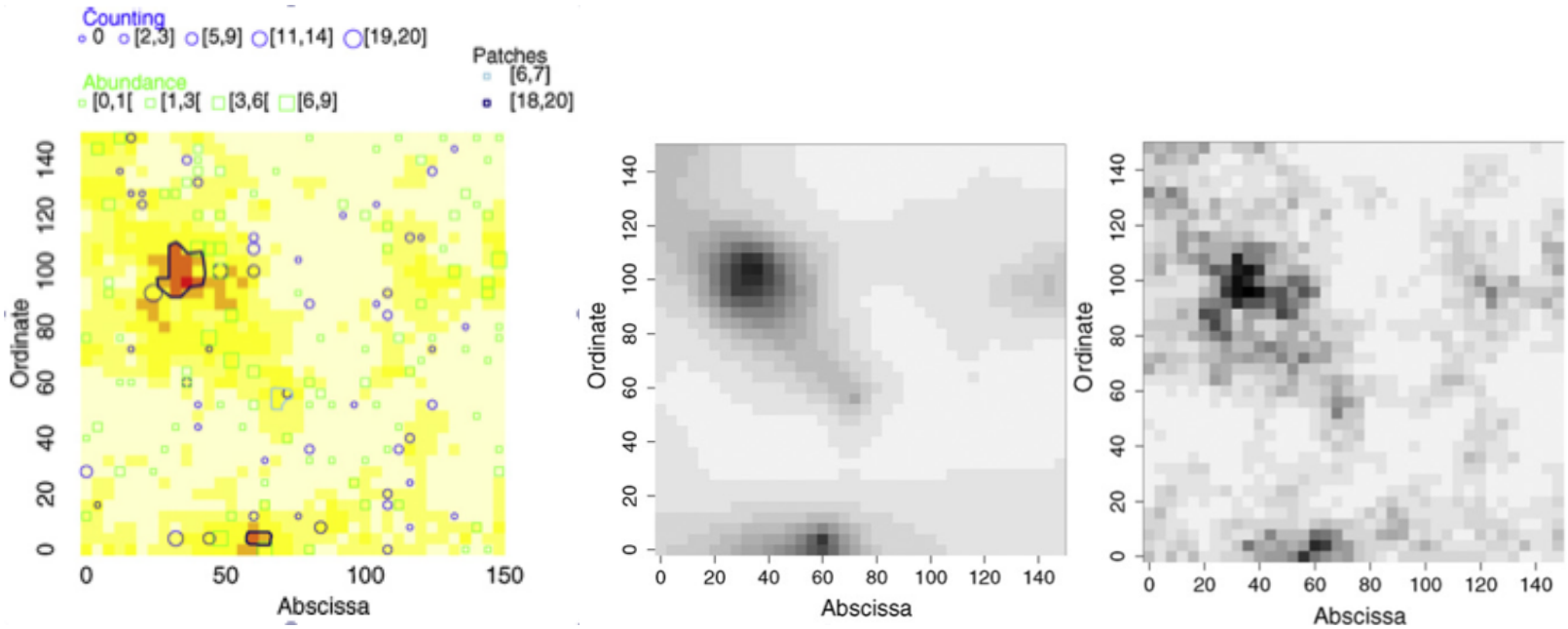


Monestiez et al, (2005)

La “cartographie” en statistiques

Lissage et d’interpolation de données réparties dans l’espace. Exploitation de

- La liaison spatiale “intrinsèque” entre les données
- Le lien avec des covariables spatialisées

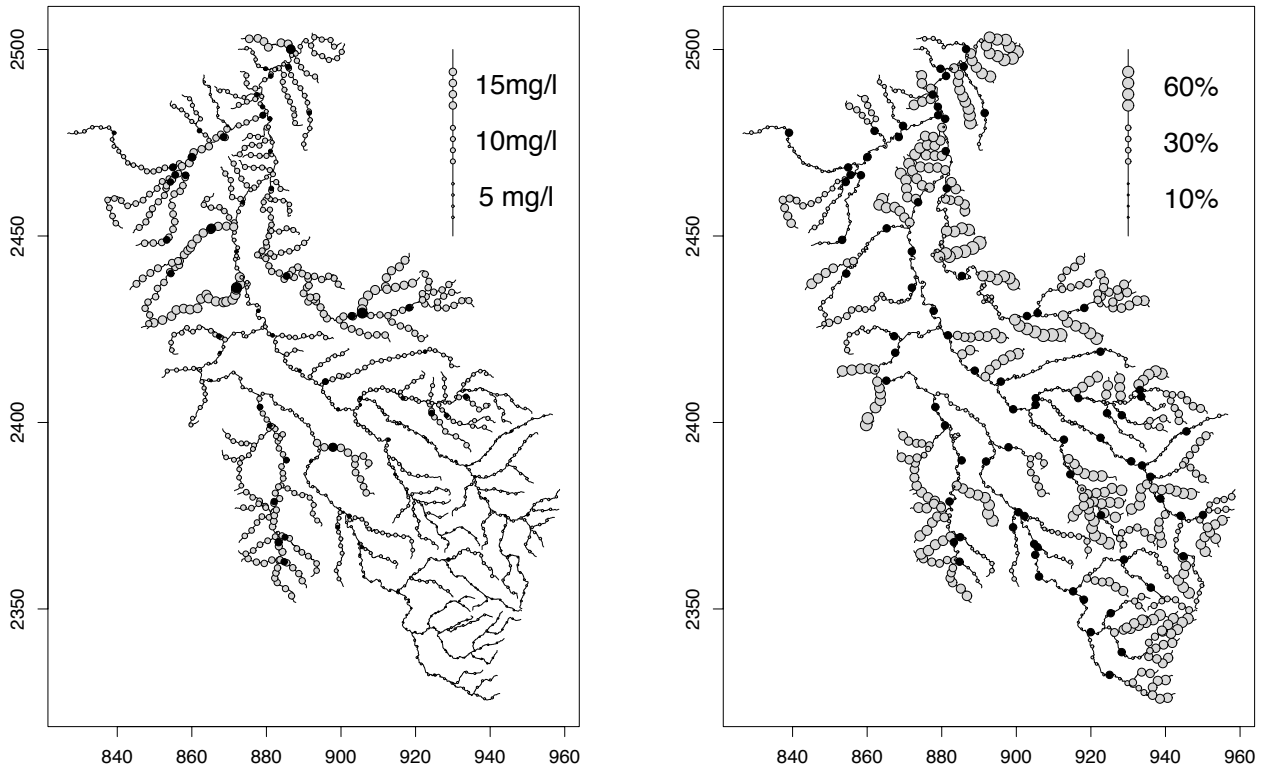


Bourgeois et al, (2012)

La "cartographie" en statistiques

Lissage et d'interpolation de données réparties dans l'espace. Exploitation de

- La liaison spatiale "intrinsèque" entre les données
- Le lien avec des covariables spatialisées



- Accompagné d'une "erreur d'estimation"

Garreta, Monestiez et Ver Hoef, (2010)

Cartographie de l'abondance

L'abondance est prise dans un sens vague. Elle sera plus ou moins bien définie selon le type de données disponibles

- Donnée = nombre de parasite par m^2
  $\Rightarrow$ quantification du nombre absolu de parasite dans l'espace
- Donnée = quantité de la souche 1 par rapport à la souche 2
  $\Rightarrow$ quantification du ratio entre souche 1 et 2
- Donnée = détection de la présence du parasite
  $\Rightarrow$ quantification de la probabilité de présence
- etc

Les modèles dont on dispose varient peu, c'est l'interprétation que l'on en fait qui varie

Gros jeux de données et représentation de l'espace

Les gros jeux de données apparaissent lorsque

- on fait beaucoup de mesures (radar, image satellite)
- on s'intéresse à plusieurs variables à la fois
- on récupère des données dans l'espace et le temps

Si on n'agrège pas ces données, i.e. si on s'intéresse à des variations à toutes les échelles, alors il faut représenter les interactions pour chaque couple de données.

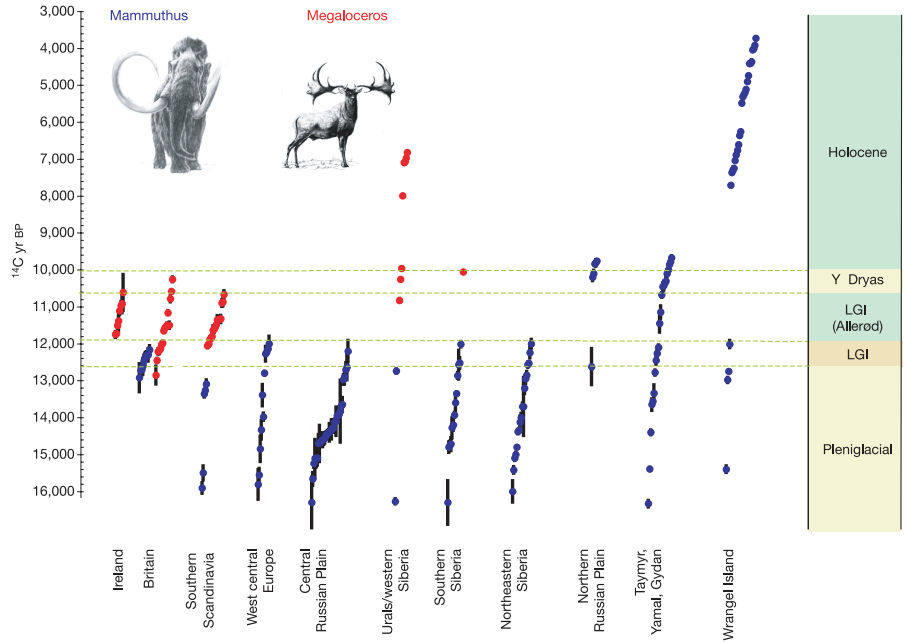
- Pour n données on a $n * (n-1)/2$ interactions possibles
- Cela tue les algorithmes pour l'inférence (inversion matricielle efficace jusqu'à $\approx 10^4$ données)

plan

1. Modèle générique pour la cartographie : modèle Gaussien latent
   - Example du mammouth laineux
   - Processus ponctuel
   - Géostatistique
2. Modélisation des liens spatiaux
   - Les différentes possibilités
   - Discussion : Processus predictif ou GMRF ?
3. Applications au mammouth et discussion

Répartition du mammouth au quaternaire

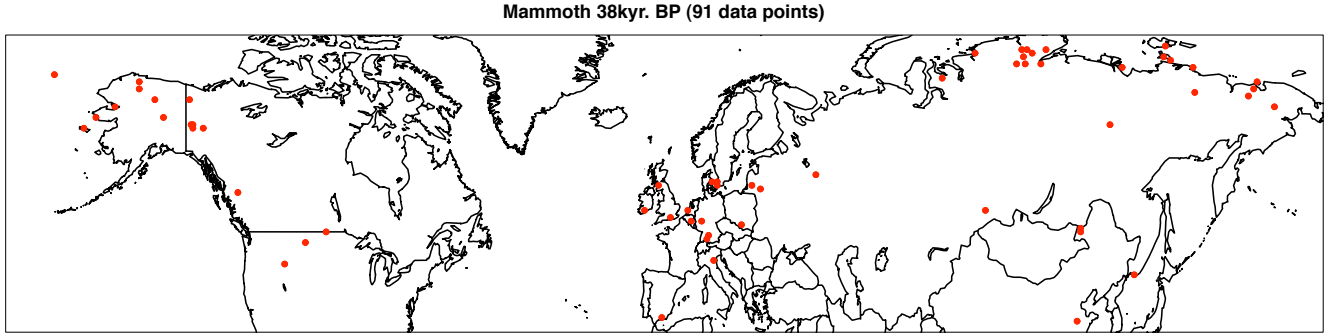
- Chaque fossile est daté au ^{14}C
- Arrivées / extinctions dans l'espace et le temps ?



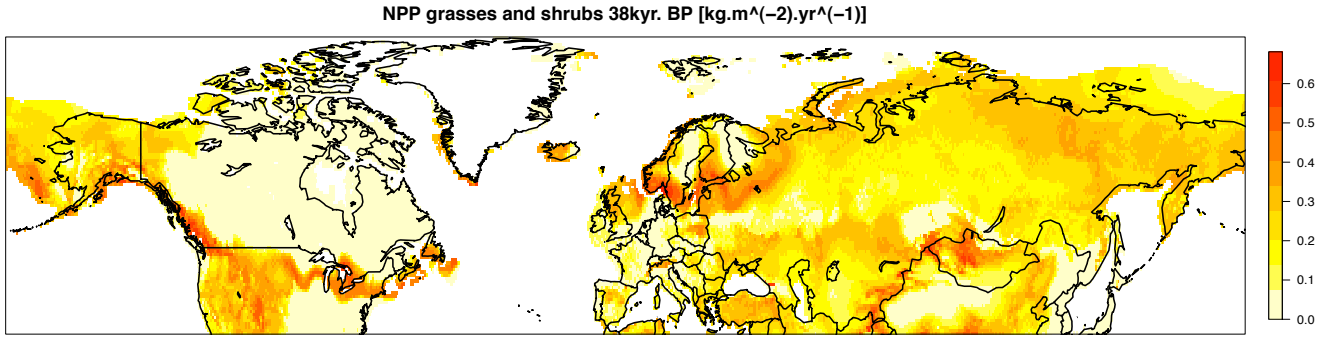
Stuart et al. (2004)

Données

- Données mammouth sont des points de présence par tranche de temps



- Simulation du climat avec GCM
- Simulation de la vegetation avec DGVM

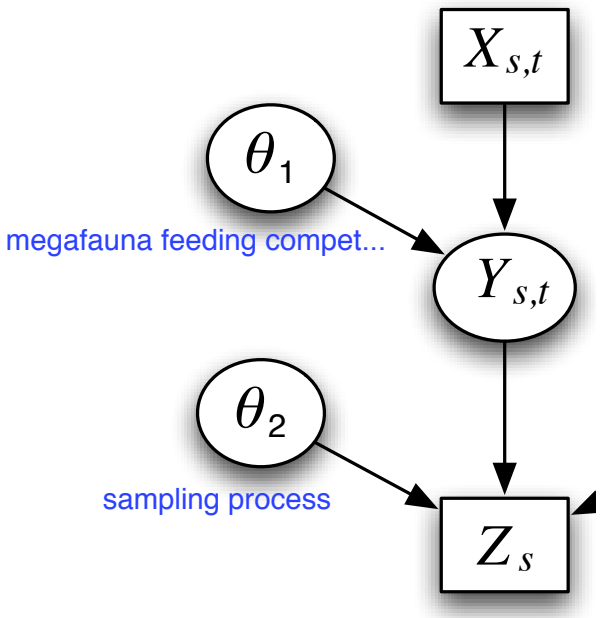


Singarayer et Valdes (2010); Allen et al. (2010)

Processus ponctuel, modélisation hiérarchique

Dans le cas du mammouth

- on a des données de présence, pas d'absence réellement mesurée
- on aimerait séparer le bruit dut à l'échantillonnage du processus de présence



- Proba. de présence mammouth $p(\mathbf{Y}|\mathbf{X}, \theta_1) =$ proc. Gaussien
- Erreur de mesure $p(Z_{s,t}|Y_{s,t}, \theta_2) = Poisson(\theta_2 e^{Y_{s,t}})$

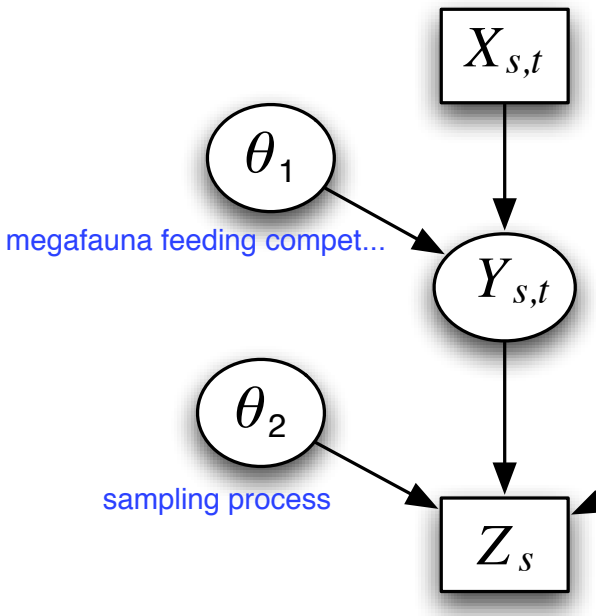
  (pour tout points de l'espace, en incluant les 0)

$\Rightarrow$ processus de Cox log Gaussien

Geostatistique, GLMM, modélisation hiérarchique

De la même façon, lorsque l'on travaille avec des mesures P/A ou quantitatives

- on aimerait séparer le bruit dut à l'échantillonnage du processus d'intérêt



- Proc. d'abondance
  $p(\mathbf{Y}|\mathbf{X}, \theta_1) =$ proc. Gaussien
- Erreur de mesure
  $p(Z_{s,t}|Y_{s,t}, \theta_2) = f(I(Y_{s,t}), \theta_2)$

  (seulement aux points de mesure)

$\Rightarrow$ Generalised Linear Mixed Model

Modèle Gaussien latent

Dans les deux cas nous utilisons exactement la même structure

$$p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}, \theta_1, \theta_2) = \left(\prod_{i=1..n} p(Z_i|Y_i, \theta_2) \right) \mathcal{N}(\mathbf{Y}|\beta\mathbf{X}, \Sigma(\theta_1))$$

L'inférence n'est pas triviale à cause des variables latentes $\mathbf{Y}$. L'obtention des paramètres θ_1 et θ_2 par maximisation de la vraisemblance nécessite une intégration

$$p(\mathbf{Z}|\mathbf{X}, \theta_1, \theta_2) = \int p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}, \theta_1, \theta_2)\, d\mathbf{Y}$$

Quelle que soit la méthode, elle nécessite l'inversion et le calcul du déterminant de $\Sigma(\theta)$, de taille $n * n$. Les algorithmes disponibles sont en $\mathcal{O}(n^3)$ et, en pratique, efficaces jusqu'à 10^3-10^4 données.

Comment induire la structure spatiale dans le GP ?

La structure spatiale du proc. Gaussien est portée par la covariance

$$p(Y|X,\theta_1) = \mathcal{N}(\beta X, \Sigma(\theta_1)) \qquad \text{et} \qquad \Sigma_{i,j}(\theta_1) = f(s_i, s_j, \theta_1)$$

Comment remplacer cette représentation ?

Comment induire la structure spatiale dans le GP ?

La structure spatiale du proc. Gaussien est portée par la covariance

$$p(Y|X,\theta_1) = \mathcal{N}(\beta X, \Sigma(\theta_1)) \qquad \text{et} \qquad \Sigma_{i,j}(\theta_1) = f(s_i, s_j, \theta_1)$$

Comment remplacer cette représentation ?

1. modéliser uniquement liens entre points proches (modèle CAR, SAR, etc)

$$Y_i|Y_{-i} \sim \mathcal{N}(\textstyle\sum_{j:j\neq i} \alpha_{ij} Y_j, \kappa_i) \qquad \Rightarrow \qquad p(Y|X,\theta_1) = \mathcal{N}(\beta X, Q^{-1}(\theta_1))$$

Comment induire la structure spatiale dans le GP ?

La structure spatiale du proc. Gaussien est portée par la covariance

$$p(Y|X,\theta_1) = \mathcal{N}(\beta X, \Sigma(\theta_1)) \qquad \text{et} \qquad \Sigma_{i,j}(\theta_1) = f(s_i, s_j, \theta_1)$$

Comment remplacer cette représentation ?

1. modéliser uniquement liens entre points proches (modèle CAR, SAR, etc)

   $Y_i|Y_{-i} \sim \mathcal{N}(\sum_{j:j\neq i} \alpha_{ij} Y_j, \kappa_i) \qquad \Rightarrow \qquad p(Y|X,\theta_1) = \mathcal{N}(\beta X, Q^{-1}(\theta_1))$

2. utiliser une représentation spectrale du processus

   (pb : Données sur grille régulière + processus stationnaire)

Comment induire la structure spatiale dans le GP ?

La structure spatiale du proc. Gaussien est portée par la covariance

$$p(Y|X,\theta_1) = \mathcal{N}(\beta X, \Sigma(\theta_1)) \qquad \text{et} \qquad \Sigma_{i,j}(\theta_1) = f(s_i, s_j, \theta_1)$$

Comment remplacer cette représentation ?

1. modéliser uniquement liens entre points proches (modèle CAR, SAR, etc)

   $Y_i|Y_{-i} \sim \mathcal{N}(\sum_{j:j\neq i} \alpha_{ij} Y_j, \kappa_i) \qquad \Rightarrow \qquad p(Y|X,\theta_1) = \mathcal{N}(\beta X, Q^{-1}(\theta_1))$

2. utiliser une représentation spectrale du processus

   (pb : Données sur grille régulière + processus stationnaire)

3. lisser une base de fonctions aléatoires (représentation classique de la dispersion)

   (pb : précision de la représentation)

Comment induire la structure spatiale dans le GP ?

La structure spatiale du proc. Gaussien est portée par la covariance

$$p(Y|X,\theta_1) = \mathcal{N}(\beta X, \Sigma(\theta_1)) \qquad \text{et} \qquad \Sigma_{i,j}(\theta_1) = f(s_i, s_j, \theta_1)$$

Comment remplacer cette représentation ?

1. modéliser uniquement liens entre points proches (modèle CAR, SAR, etc)

   $Y_i|Y_{-i} \sim \mathcal{N}(\sum_{j:j\neq i} \alpha_{ij} Y_j, \kappa_i) \qquad \Rightarrow \qquad p(Y|X,\theta_1) = \mathcal{N}(\beta X, Q^{-1}(\theta_1))$

2. utiliser une représentation spectrale du processus

   (pb : Données sur grille régulière + processus stationnaire)

3. lisser une base de fonctions aléatoires (représentation classique de la dispersion)

   (pb : précision de la représentation)

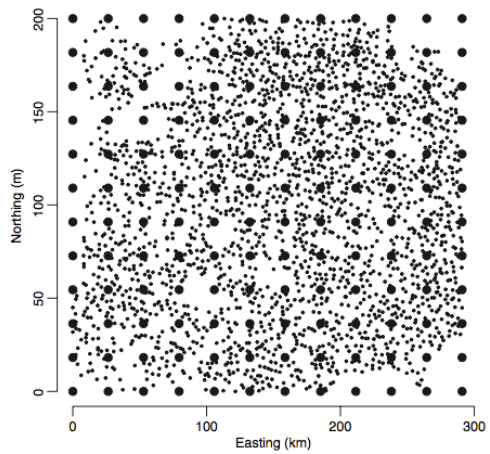
4. interpoler un proc. Gaussien défini sur peu de points

Processus sur grille grossière (predictive process)

L'idée : ne plus définir le processus dans tout l'espace mais sur une grille "réduite"

$$Y = \Sigma_{s_0,s}(\theta)\Sigma^{-1}_{s_0,s_0}(\theta)\, Y_{s_0}$$

$$p(\mathbf{Z},\mathbf{Y}|\mathbf{X},\theta_1,\theta_2) = \left(\prod_{i=1..n} p(Z_i|\Sigma_{s_0,s}(\theta_1)\Sigma^{-1}_{s_0,s_0}(\theta_1)\, Y_{s_0},\theta_2)\right) \mathcal{N}(\mathbf{Y_{s_0}}|\beta\mathbf{X},\Sigma_{s_0,s_0}(\theta_1))$$



- l'interpolation est un krigeage
- les calculs se font sur matrice $n_0 * n_0$ (+ multiplication)

- perte d'information sur liaisons à une échelle moindre que la grille des s_0 (mais modifications existent)

Banerjee et al. (2008)

Champs aléatoires de Markov

L'idée, très naturelle en temporel (AR), est moins utilisée en spatial : on spécifie

$$Y_i | Y_{-i} \sim \mathcal{N}(\sum_{j:j\neq i} \alpha_{ij} Y_j, \kappa_i)$$

où les α_{ij} sont non-nuls uniquement pour le voisinage de i.

Dans le cas Gaussien, la loi jointe des $\mathbf{Y}$ est Gaussienne et les voisinages se codent dans $Q = \Sigma^1$.

La puissance vient du fait que Q est creuse (sparse). En dimension 2 les algorithmes effectuent l'inversion en $\mathcal{O}(n^{3/2})$.

Ces modèles nommés CAR et SAR sont peu utilisés car

- ils collent à une représentation fixée de l'espace
- il ne paramétrent pas directement la covariance ni la variance

Rue et Held (2005)

Champs aléatoires de Markov : approche mécaniste

Lindgren et al. (2011) proposent de dériver ces champs de Markov comme la solution par élément finis d'une équation aux dérivées partielles stochastique

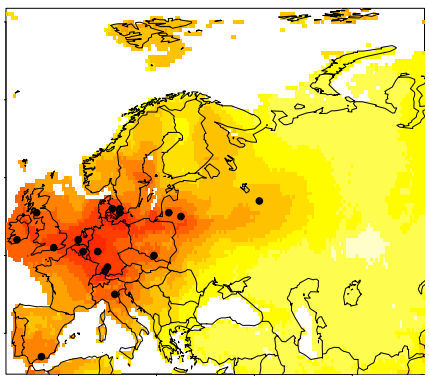
$$(\kappa^2 - \Delta)^{\alpha/2} x(s) = \mathcal{W}(s)$$

- le champs à une interprétation mécaniste et en espace continu
- cette équation différentielle définit un champ dont on connait la covariance
- la grille peut être aussi détaillée que nécessaire
- des condition eaux limites, instationarités etc peuvent être incluses...

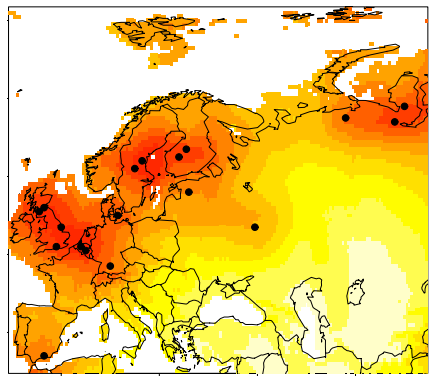
Mammouth : processus sur grille grossière

- Jeu de données réduit (Europe, 10,000 pts), s_0 250 pts
- Modèle de covariance : exponentiel ($\theta_1 = 1$ paramètre)
- Inférence en utilisant INLA (Rue et al, 2009)

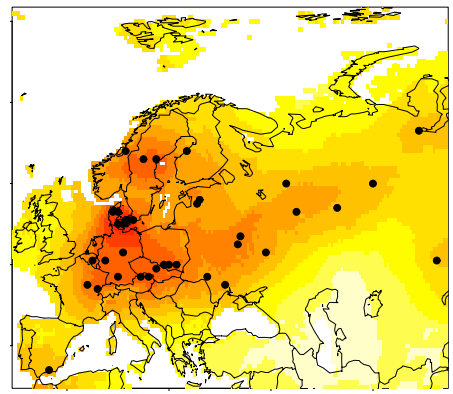
Mammoth 38kyr. BP



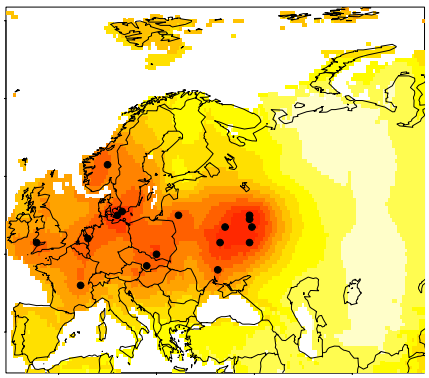
Mammoth 34kyr. BP



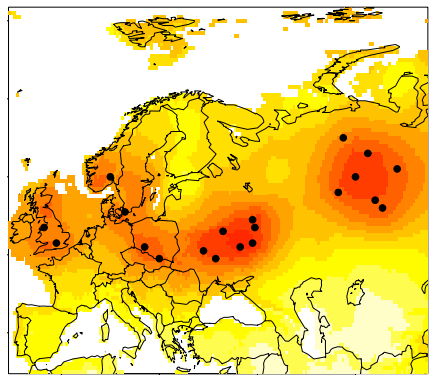
Mammoth 30kyr. BP



Mammoth 26kyr. BP



Mammoth 22kyr. BP

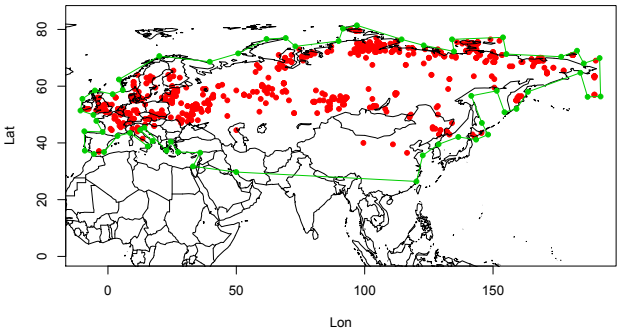


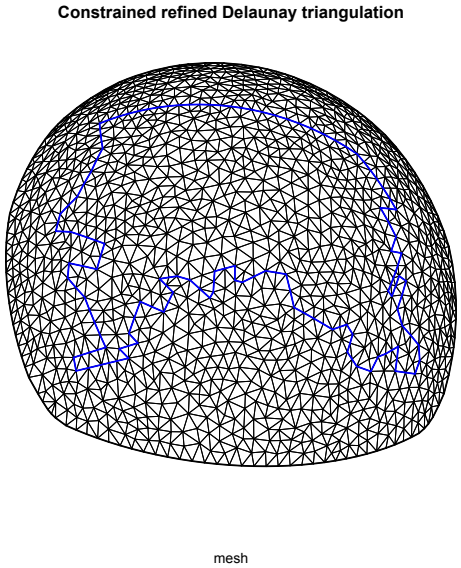
- C'est faisable (2 semaines à partir de rien)
- Instable, rapide
- pb de frontière

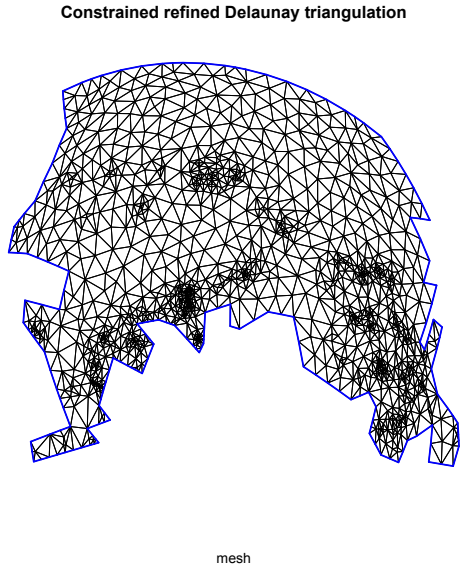
Eidsvik et al. (2012)

Mammouth : champs de Markov

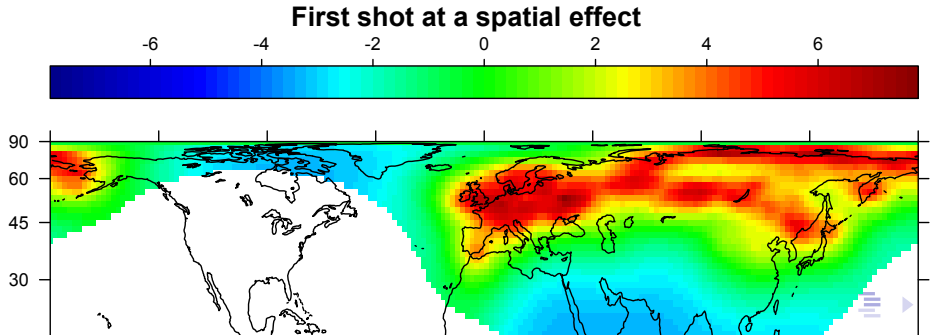
- Jeu de données Eurasie ($\approx$ 100,000 pts)
- Inférence en utilisant le package INLA (en fait Daniel Simpson, postdoc chez H. Rue, NTNU Norvège)







- Donner un contour
- Equivalent à la covariance
- En respectant le contour (non codé)



Lindgren, Rue, Lindström (2011)

Conclusion

Le problème du spatial de grande dimension est à la mode, beaucoup de choses nouvelles sortent. 2 options me semblent intéressantes

- Processus sur grille :
  - reste dans le monde de la covariance
  - on peut tout maitriser avec un (petit) bagage en statistique spatiale

- Processus de Markov basé sur équations différentielles :
  - très prometteur, plus riche que la grande dimension (frontières, non-stationnarité, spatio-temporel, etc)
  - demande des connaissances pluridisciplinaires (stat spatiale, processus aléatoire, équa. diff. stochastique, éléments finis)
  - il existe un package qui fait tout ça (et se développe à tout vitesse)

Je vous ai caché les questions liées à la méthode d'inférence :)

Merci de votre attention !

Inférence pour les modèles Gaussien latent

Le problème d'inférence avec les modèles Gaussien latent est la nécessaire intégration sur la couche latente pour obtenir la vraisemblance (dans le sens fréquentiste)

$$p(\mathbf{Z}|\mathbf{X}, \theta_1, \theta_2) = \int p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}, \theta_1, \theta_2)\, d\mathbf{Y}$$

Dans le monde fréquentiste, des algorithmes de type MCEM (Expectation-Maximisation par Monte-Carlo) sont utilisés (voir Zhang, 2002). Ils requièrent l'inversion et la décomposition de Choleski de la matrice de covariance. Peu utilisé en pratique.

Les Bayésiens semblent s'être emparé du problème (les variables latentes leur appartiennent :)

- Solution classique : MCMC (Marche aléatoire dans l'espace de $\mathbf{Y}$ et θ pour générer des données suivant la loi a posteriori)
- Approximation fonctionnelle : INLA (Integrated Nested Laplace Approximation)

Approches Bayésiennes

Méthode basées sur MC

- Flexible, facile à construire
- Facilité pour obtenir une fonction non linéaire de l'a posteriori
- Très lent (très grand nombre d'inversion de matrice)
- Problèmes de convergence
- Des packages existent pour le modèle "classique"

Approximation fonctionnelle : INLA

- Seulement modèle Gaussien latent
- Rapide quand dim(θ) petit (< 8)
- Très dur d'obtenir autre chose que des a posteriori Gaussiens et leurs combinaisons linéaires
- Disponible sous forme de paquet R, (presque) aussi facile d'utilisation que la fonction `lm()` ! Tant que l'on est dans un cas classique

Diggle et al. (1998); Rue et al. (2009)