# Estimation sous contraintes pour des données tronquées

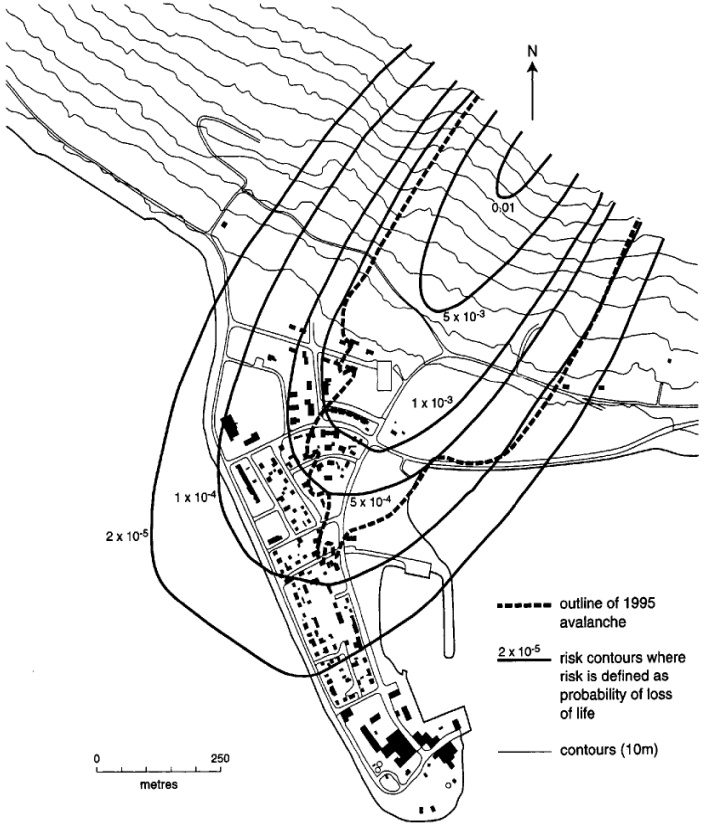
## Application au cas des altitudes d'arrêts en Haute-Savoie

Aurore Lavigne, Liliane Bel, Nicolas Eckert et Eric Parent

journée RESSTE, 11 mai 2015

# Cartes d'aléas/risques, des outils indispensables à la gestion du risque d'avalanches

Exemple de carte de risque : probabilité annuelle de décès (+/− proportionnelle à la probabilité d'atteinte)



# Objectif

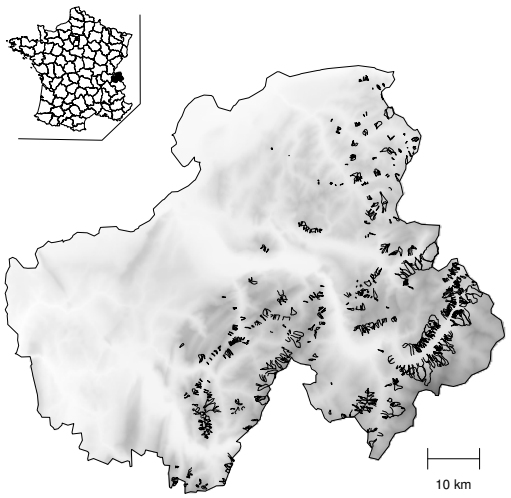
Proposer une nouvelle méthode de construction de cartes d'aléa à l'échelle régionale, s'appuyant sur la dépendance spatiale entre couloirs.
Objectifs :

- Démontrer la dépendance spatiale entre les altitudes d'arrêts moyenne des couloirs.
- Utiliser cette dépendance spatiale pour prédire la distribution des altitudes d'arrêts sur un nouveau couloir.

# Altitudes d'arrêts enregistrées par l'EPA

## Altitudes d'arrêts en Haute-Savoie

- 389 couloirs avec au moins une donnée
- 5331 avalanches
- Période d'étude : 1925-2012
- En 2002, un seuil d'observation $s_c$ est défini pour chaque couloir => **troncature à droite**

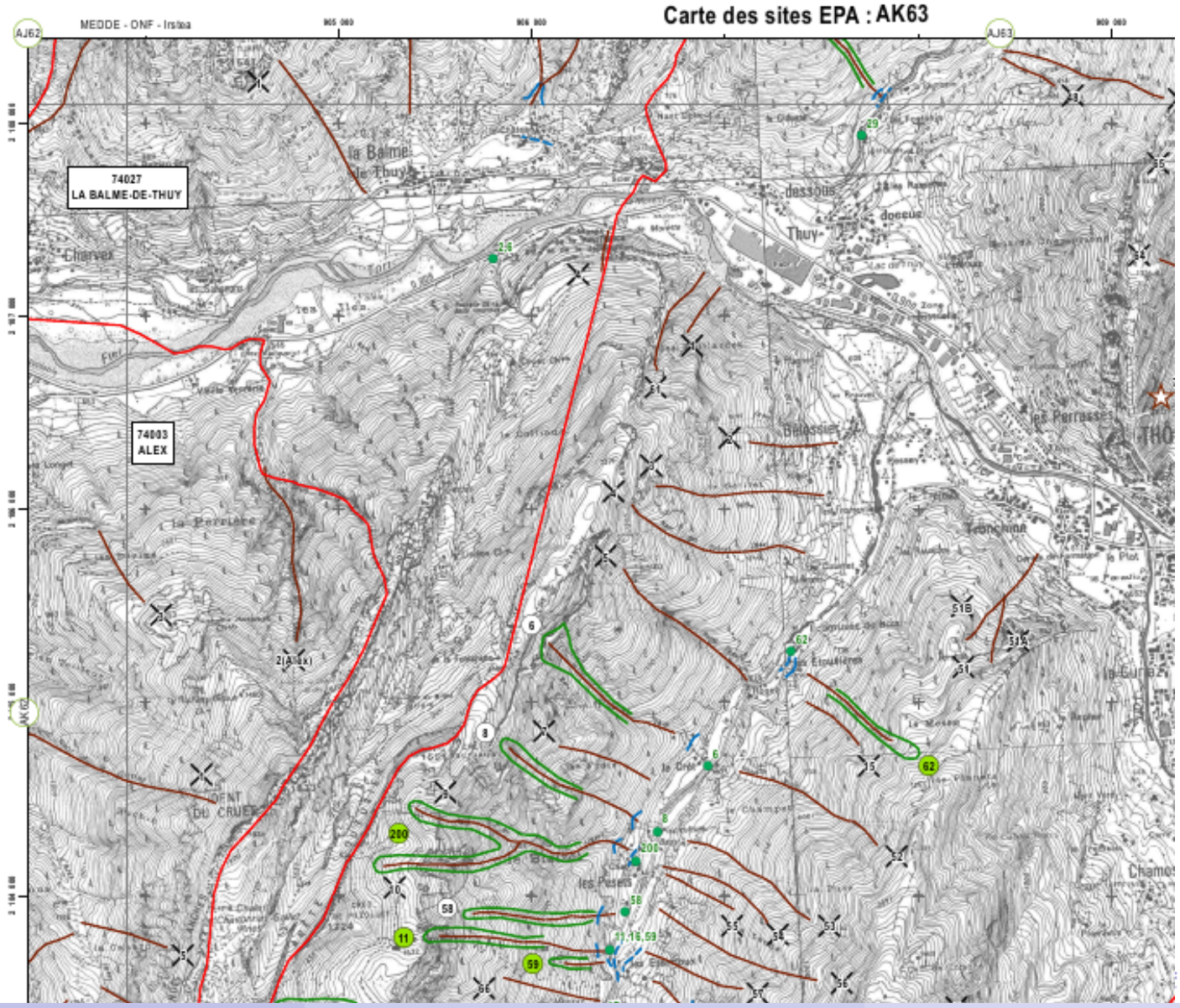



Couloirs d'avalanches en Haute-Savoie

## Données topologiques

- Altitude de la vallée au pied du couloir $h_c$.
- Exposition de la zone de départ.

# Exemples de tracés EPA



# Des données hétérogènes

## Activité des couloirs

- 162 couloirs avec moins de 5 avalanches.
- 51 couloirs avec plus de 30 événements.
- 177 avalanches enregistrées pour le couloir le plus actif.

## Variabilités de la variance inter-couloir

- Dépends de la topographie et des caractéristiques de la zone d'arrêt.
- Les écart-types s'étendent de 0 ! à 250 m.
- L'erreur d'observation est d'environ 10 m - 50 m.

## Différence de dénivelé entre le seuil d'observation et la vallée

- autour de 300 m,
- s'étend de 0 à 1000 m.
- Les modèles de dépassement de seuil ne peuvent pas être utilisés.

# Plan

# Plan

# Modélisation des observations par une distribution tronquée

- $Y_{cti}$ altitude d'arrêt de l'avalanche $i$, sur le couloir $c$, l'année $t$
- $s_c$ seuil d'observation du couloir $c$
- $\mu_{ct}$ espérance de la distribution complète (non tronquée)
- $\sigma_c^2$ variance du couloir $c$ pour la distribution complète

$$f(y_{cti}|\mu_{ct},\sigma_c^2) = \frac{1}{\Phi\left(\frac{s_c-\mu_{ct}}{\sigma_c}\right)} \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-1}{2\sigma_c^2}(y_{cti}-\mu_{ct})^2\right) \mathbf{1}_{y_{cti}<s_c}$$

# Modélisation des observations par une distribution tronquée

- $Y_{cti}$ altitude d'arrêt de l'avalanche $i$, sur le couloir $c$, l'année $t$
- $s_c$ seuil d'observation du couloir $c$
- $\mu_{ct}$ espérance de la distribution complète (non tronquée)
- $\sigma_c^2$ variance du couloir $c$ pour la distribution complète

$$f(y_{cti}|\mu_{ct},\sigma_c^2) = \frac{1}{\Phi\left(\frac{s_c-\mu_{ct}}{\sigma_c}\right)} \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-1}{2\sigma_c^2}(y_{cti}-\mu_{ct})^2\right) \mathbf{1}_{y_{cti}<s_c}$$

On tient compte de la troncature et de l'hétéroscédasticité.

# Un modèle spatio-temporel additif

$$\mu_{ct} = \alpha + B_t + C_c. \tag{1}$$

- $\alpha$ altitude d'arrêt moyenne
- $B_t$ terme temporel, modélisé par une marche aléatoire à l'ordre deux
- $C_c$ terme spatial

On ajoute les contraintes d'identifiabilité : $\sum_t B_t = \sum_c C_c = 0$

# Modèle spatial

Le couloir est réduit à un point $c$ dont les coordonnées appartiennent au domaine spatial $\mathcal{D}$. Les termes spatiaux $C_c$ sont modélisés par un processus gaussien, avec une covariance exponentielle et une pépite.

## Dépendance spatiale

$$\mathrm{cov}(C_c, C_{c'}) = \begin{cases} \tau^2 \exp(-\frac{h_{cc'}}{\phi}) \text{ if } c \neq c' \\ \rho^2 + \tau^2 \text{ otherwise,} \end{cases}$$

avec

- $h_{cc'}$ distance entre $c$ et $c'$,
- $\phi$ paramètre de portée, $\tau^2$ pallier, et $\rho^2$ pépite.

## Régression

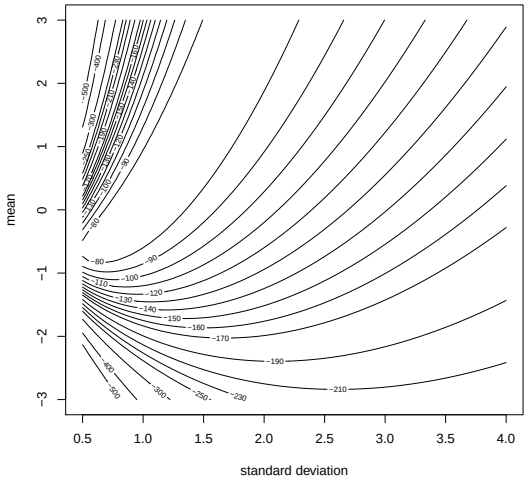
$$E(C_c) = a + bh_c + c_{o_c} \begin{cases} h_c \text{ altitude de la vallée} \\ o_c \in \{N, S\} \text{ exposition du couloir} \end{cases}$$

# Plan

# Problèmes d'estimation engendrés par la troncature

## Cas d'une variable normale tronquée à droite

- Espérance et variance d'une distribution gaussienne tronquée sont corrélées.
- La vraisemblance est distordue et présente une large zone plate.




Log-vraisemblance d'un échantillon Gaussien tronqué à droite à 0.1.

## Conséquences sur l'inférence fréquentiste

- La variance de l'estimateur du maximum de vraisemblance est grande
- Ajout d'un biais pour diminuer la variance : vraisemblance restreinte (Cope, 2011), vraisemblance pénalisée (A'Hearn, 2004)

# Inférence bayésienne

## Priors vagues

- $[\alpha] \propto \mathrm{cte}$, $a$, $b$, $c \sim N(0, 10^7)$
- $\phi \sim U[0, 30000]$
- $\tau^2 \sim \mathrm{IG}(0.001, 0.001)$ et $\rho^2 \sim \mathrm{IG}(0.001, 0.001)$

## Prior informatif, élicité

- $\sigma_c^2 \sim$ IG avec pour espérance et écart-type $25^2$ tronquée avant $30^2$ et après $316^2$.

# Inférence bayésienne

## Priors vagues

- $[\alpha] \propto \mathrm{cte}$, $a$, $b$, $c \sim N(0, 10^7)$
- $\phi \sim U[0, 30000]$
- $\tau^2 \sim \mathrm{IG}(0.001, 0.001)$ et $\rho^2 \sim \mathrm{IG}(0.001, 0.001)$

## Prior informatif, élicité

- $\sigma_c^2 \sim$ IG avec pour espérance et écart-type $25^2$ tronquée avant $30^2$ et après $316^2$.

On tient compte de l'erreur de mesure.

# Échantillonneur de Gibbs

Objectif : tirer dans les distributions *a posteriori* conditionnelles complètes des paramètres et des variables latentes.

Deux principales difficultés :

- Covariance exponentielle.
- Troncature
  Utilisation d'une variable auxiliaire $Z_{cti}$, correspondant à l'observation sous la loi complète, telle que $\mathbf{P}(Y_{cti} < y) = \mathbf{P}(Z_{cti} < z)$ (Griffiths, 2004).

$$Z_{cti} = \mu_{ct} + \sigma \Phi^{-1} \left( \frac{\Phi(\frac{Y_{cti} - \mu_{ct}}{\sigma_c})}{\Phi(\frac{s_c - \mu_{ct}}{\sigma_c})} \right)$$

# Problème rencontré

Dès la première itération de l'algorithme :

```
Chaine 1 iteration 1
Z[4969] is not a number
mu = 2014.806636
seuil = 1000
y[i] = 995
sig = 118.9297427
```

Sur un couloir bien documenté :

| Altitude d'arrêt | 1240 | 1250 | 1260 | 1270 | 1280 | 1290 | 1300 | 1305 |
|---|---|---|---|---|---|---|---|---|
| Nb d'obs | 2 | 13 | 9 | 7 | 12 | 4 | 24 | 3 |

# Solution : contraindre $\alpha + C_c$ à un domaine raisonnable

Nous savons que :

- au moins 50% des avalanches sont observées sur chaque couloir,
- l'altitude de la vallée $h_c$ est une borne inférieure pour toutes les avalanches.

# Solution : contraindre $\alpha + C_c$ à un domaine raisonnable

Nous savons que :

- au moins 50% des avalanches sont observées sur chaque couloir,
- l'altitude de la vallée $h_c$ est une borne inférieure pour toutes les avalanches.

Ainsi, $h_c \leq \alpha + C_c \leq s_c$.

# Tirage de $C_c$, partie 1 : contraintes d'identifiabilités

On cherche à tirer $C$ dans $[C|\theta, Y]$, connaissant la contraintes $\sum_{c=1}^{n_c} C_c = 0$. ($\theta$ : autres paramètres et variables)

Idée (Gelman, 2005) : Tirer $C_1, \cdots, C_{n_c-1}$ dans $C|\theta, Y, \mathbf{1}_c' C = 0$ et prendre $C_{n_c} = -\sum_{c=1}^{n_c-1} C_c$.

Si $C|\theta, Y \sim \mathcal{N}(m, \Sigma)$ alors la loi jointe de $C|\theta, Y$ et de $\mathbf{1}_c C = 0$ s'écrit

$$\begin{bmatrix} C|\theta, Y \\ \mathbf{1}_c' C \end{bmatrix} \sim N\left( \begin{bmatrix} m \\ \mathbf{1}_c' m \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma \mathbf{1}_c' \\ \mathbf{1}_c' \Sigma & \mathbf{1}_c' \Sigma \mathbf{1}_c' \end{bmatrix} \right).$$

La distribution conditionnelle est alors

$$C|\theta, Y, \mathbf{1}_c' C = 0 \sim \mathcal{N}(m_0, \Sigma_0) \text{ with } \begin{cases} m_0 = m - \frac{\Sigma \mathbf{1}_c \mathbf{1}_c' m}{\mathbf{1}_c' \Sigma \mathbf{1}_c} \\ \Sigma_0 = \Sigma - \frac{\Sigma \mathbf{1}_c \mathbf{1}_c' \Sigma}{\mathbf{1}_c' \Sigma \mathbf{1}_c} \end{cases}.$$

# Tirage de $C_c$, partie 2 : $h_c - \alpha < C_c < s_c - \alpha$

Idées (Rodriguez-Yam et al, 2004) :

- Ecrire les contraintes sous forme vectoriel $\mathbf{V}C < v$, $\mathbf{V}$ de dimension $2n_c \times n_c - 1$.
- Enlever la dépendance entre les composantes de $C$ en considérant $w = LC$ avec $L$ telle que $L\Sigma_0 L' = I_{n_c-1}$
- Tirer successivement et conditionnellement aux autres, chaque composant de $w$ en respectant ses contraintes $\mathbf{V}L^{-1}w < v$.

Pour ne pas recalculer $\mathbf{V}L^{-1}w$ à chaque composante on remarque que pour la contrainte $i$ sur la composante $j_0$ :

$$(\mathbf{V}L^{-1}w)_i = \underbrace{\sum_{j \neq j_0}^{n_c-1} \left( \sum_{l=1}^{n_c-1} \mathbf{V}_{il} L^{-1}_{lj} \right) w_j}_{u_i^{-j_0}} + \underbrace{\left( \sum_{l=1}^{n_c-1} \mathbf{V}_{il} L^{-1}_{lj_0} \right)}_{u_i^{j_0}} w_{j_0} < v_i.$$

La contrainte $i$ pour la composante $j_0$ de $\mathbf{w}$ s'écrit $w_{j_0} < \frac{v_i - u_i^{-j_0}}{u_i^{j_0}}$.

# Algorithme pour tirer $C_1, \cdots, C_{n_c-1}$ dans sa distribution tronquée

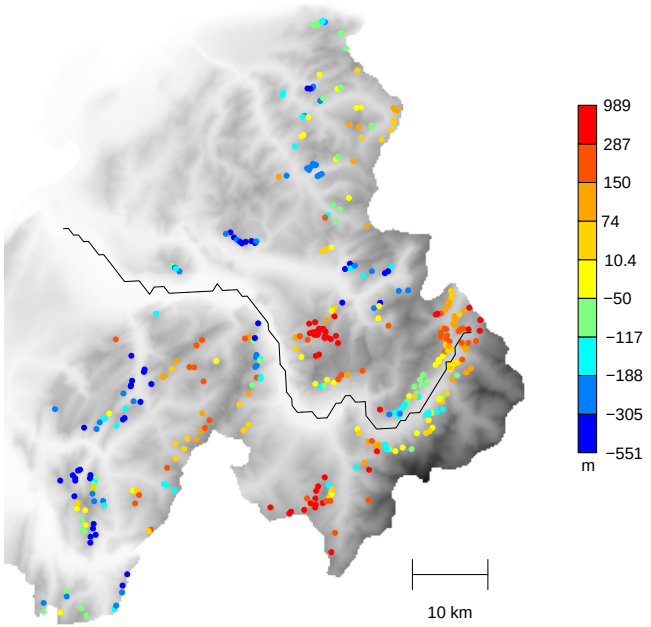
- Calculer le vecteur $\mathbf{u} = \mathbf{VC}$ ;
- Initialiser la matrice $\mathbf{L}$, son inverse $\mathbf{L}^{-1}$, et le vecteur $\mathbf{w} = \mathbf{LC}$ ;
- Pour chaque $j$ dans $\{1, \cdots n_c - 1\}$
  - Calculer le vecteur de dimension $2n_c$, $\mathbf{u}^j = (u_1^j, \cdots u_k^j)$, et $\mathbf{u}^{-j} = \mathbf{u} - \mathbf{u}^j w_j$ ;
  - Trouver l'intervalle $[a_j, b_j]$ dans lequel tirer $w_j$ pour satisfaire les $2n_c$ contraintes ;
  - Tirer $w_j$ dans la distribution normale tronquée $\mathcal{N}((\mathbf{Lm}_0)_j, 1)$ de support $[a_j, b_j]$.
  - Mettre à jour $\mathbf{u}$ comme $\mathbf{u} = \mathbf{u}^{-j} + \mathbf{u}^j w_j$ ;
- Obtenir $\mathbf{C}$ avec $\mathbf{C} = \mathbf{L}^{-1}\mathbf{w}$.

# Plan

# Résultats-Spatial

- Terme spatial représente 98.5% de la variabilité totale
- Dans le terme spatial, le terme de régression représente 86% de la variabilité.
- Portée effective estimée à 10 km.

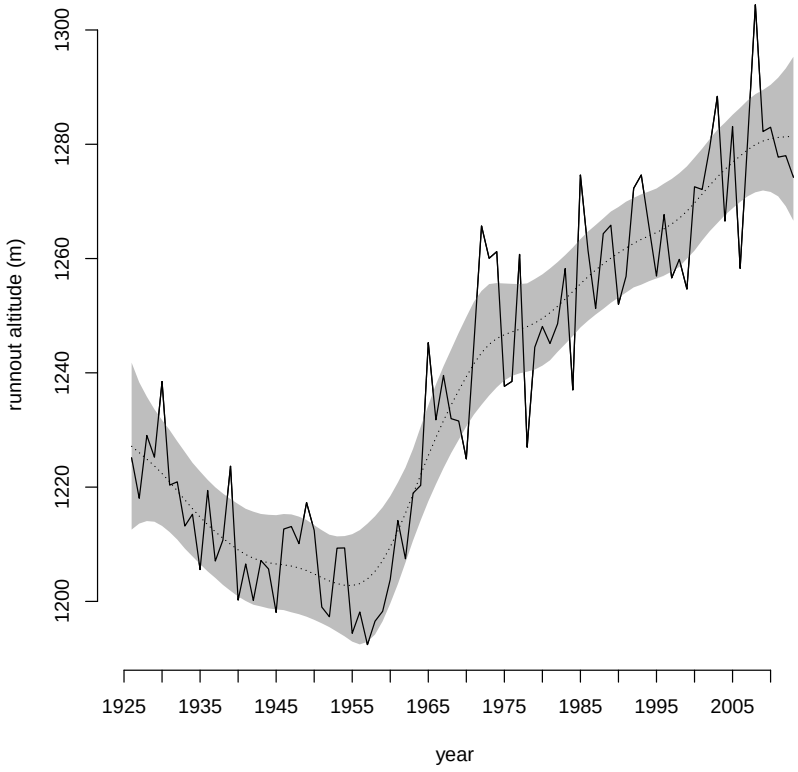
Carte du terme spatial $C_c$




| | portée $\phi$ | pallier $\tau^2$ | pépite $\rho^2$ | cte $a$ | alt. val. $b$ | exposition $c_S$ |
|---|---|---|---|---|---|---|
| esp. | 3037.61 m | 4644.89 $\mathrm{m}^2$ | 2426.96 $\mathrm{m}^2$ | -935.06 m | 0.83 | 7.02 m |
| éc. type | 1790.80 m | 1192.07 $\mathrm{m}^2$ | 616.54 $\mathrm{m}^2$ | 31.02 m | 0.03 | 9.40 m |

# Résultats-Temporel

- Terme spatial représente seulement 1.5% de la variabilité spatio-temporelle




| | Espérance | Ecart-type |
|---|---|---|
| $\delta_0$ | 137.7 $m^2$ | 33.1 $m^2$ |
| $\delta_1$ | 0.88 $m^2$ | 1.09 $m^2$ |

# Conclusion

- Identification d'un problème de convergence de l'algorithme de Gibbs
  - Troncature des données
  - Distribution des données presque discrètes à cause de l'erreur de mesure
- Solution : ajout de contraintes
  - Connaissance *a priori*
  - Peu restrictives
  - Impossible de mesure son effet sur les estimateurs dans ce cas

# Références

A'Hearn B (2004) A restricted maximum likelihood estimator for truncated height samples. Economics & Human Biology 2(1) :5–19

Cope EW (2011) Penalized likelihood estimators for truncated data. Journal of Statistical Planning and Inference 141(1) :345–358

Griffiths W (2004) A Gibbs' sampler for the parameters of a truncated multivariate normal distribution. In : Becker R, Hurn S (eds) Contemporary issues in economics and econometrics : theory and application, Edward Elgar Pub, pp 75–91

Rodriguez-Yam G, Davis R, Scharf L (2004) Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. Unpublished Manuscript