Structurer une couche latente à l'aide de la géostatistique

Application aux fréquences d'avalanches dans les Alpes

Aurore Lavigne, Liliane Bel, Eric Parent et Nicolas Eckert

ANR project: MOPERA

AppliBUGS, 17 juin 2011

Introduction

FIGURE: Avalanche in Montroc 1999 (Cemagref)

Risk management :

- **Forecast** Avalanche activity in the next days
- **Predetermination** Long time avalanche activity (frequency and magnitude)

Frequency :

- Evolution over the past 60 years $\Rightarrow$ Stationarity
- Spatial repartition in the French Alps $\Rightarrow$ Dependence

Data

Data come from the EPA database :

EPA Enquête Permanente sur les Avalanches Managed by the CEMAGREF, it is a chronicle describing the avalanche events on approximately 3900 paths in the Alps and the Pyrenees. On each path, information are intended to be the most complete as possible.

We use avalanche counts by year and by township :

- The study period is 1946-2008
- We prefer the township scale to the path scale because of path localization problems by past.

A quality study has been performed by the CEMAGREF and 40% of missing data have been identified.

With these data,

- Eckert et al (2007) identify a spatial dependance in Savoie at the scale of the township, using the Besag, York and Mollié's model.
- Eckert et al (2009) show that the temporal effect account for 17% of the avalanche frequency variability.

We pursue this work, and use hierarchical model under the Bayesian paradigm to

- define a temporal evolution using smoothing spline.
- model a spatial structure using spatial dependance :
  - based on neighborhood relationships
  - based on distance between sites

Outline

Poisson distribution for counts data

Let Y_{ct} be the number of avalanches the year t in the township c

$$Y_{ct} \propto \frac{E_c RR_{ct}^{Y_{ct}}}{Y_{ct}!} \exp(-E_c RR_{ct})$$

We assume the $Y_{ct}|E_c RR_{ct}$ for $c \in \{1, \ldots, N\}$ and for $t \in \{1, \ldots, T\}$ are independent.

- E_c is the expected number of avalanches depending on the number of paths.
- RR_{ct} is the relative risk, it depends on the year and the township and is modeled on a latent layer.

A separable spatio-temporal model with 2 symmetrical structures

$$RR_{ct} = \alpha + \eta_t^T + \eta_c^S$$

The temporal term $\boldsymbol{\eta}^T = (\eta_1^T, \dots, \eta_T^T)'$	The spatial term $\boldsymbol{\eta}^S = (\eta_1^S, \dots, \eta_N^S)'$
$\boldsymbol{\eta}^T = \mathbf{g} + \boldsymbol{\epsilon}$	$\boldsymbol{\eta}^S = \mathbf{u} + \mathbf{v}$
$\boldsymbol{\epsilon} \sim N(0, \delta_0 I_T)$ $\mathbf{g} \sim N$ precision $\frac{1}{\delta_1} A$	$\mathbf{v} \sim N(0, \sigma_v^2 I_N)$ $\mathbf{u} \sim N$ precision $\frac{1}{\tau^2} P$
$\delta_0, \delta_1 InvGamma(0.1, 0.1)$	$\sigma_v^2, \tau^2 InvGamma(0.1, 0.1)$

A and P are semi-definite positive matrix. For identifiability purposes we set $\sum \eta_t^T = 0$ and $\sum \eta_c^S = 0$. α has a constant prior.

The temporal evolution as a smooth process

Eckert et al 2009 model the avalanche frequencies evolution with jumps between different levels.
We propose to model a cubic smoothing spline

- It is flexible
- It is a non-parametric model, and does not need any covariates
- It is a continuous process

We use a random walk of order 2 :

$$[\mathbf{g}] \propto \frac{|A|_+^{\frac{1}{2}}}{\delta_1^{\frac{1}{2}(T-2)}} exp(-\frac{1}{2\delta_1}\mathbf{g}' A \mathbf{g})$$

with $|A|_+$ the product of the non-nul eigen values of A
Speckman and Sun 2003 show that using proper prior on δ_0 and δ_1 we make the distribution of $\mathbf{g}|\boldsymbol{\eta}^T$ proper.

We aim :

- To check that η^S is spatially dependent
- To identify the role of massifs. Townships are embedded in massifs, which have homogeneous climatic characteristics.

We construct 2 models based on the popular BYM model :

model 1 Using the townships' networks

model 2 Using the massifs and townships' networks

Intrinsic *CAR* (Mollié et al 1991) with W the weight matrix. $W_{cc'} = 1$ if c and c' share a common boundary, else $W_{cc'} = 0$, $W_{c+} = \sum_{c' \neq c} W_{cc'}$.

Conditional distributions, for $c \in \{1, \ldots, N\}$

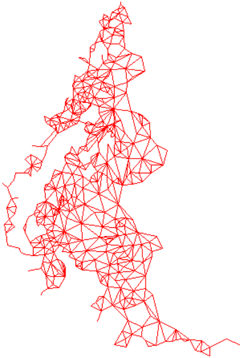
$$[u_c | \mathbf{u_{-c}}] \propto exp(-\frac{w_{c+}}{2\tau^2}(u_c - \sum_{c' \neq c} \frac{w_{cc'} u_{c'}}{w_{c+}})^2)$$

with $\mathbf{u_{-c}} = (u_1, \ldots, u_{c-1}, u_{c+1}, \ldots, u_N)$

Full distribution :

$$[\mathbf{u}] \propto exp(-\frac{1}{2\tau^2}\mathbf{u}'(D_w - W)\mathbf{u})$$

With D_w the diagonal matrix, and $(D_w)_{cc} = W_{c+}$.



Neighboring graph

Each township belongs to one massifs : $u_c = u_{cm}$

We decompose u_c in two terms

$$u_{cm} = u_m + u_{c|m}$$

- u_m is relative to the massif m, M the number of massifs, $\mathbf{u}^M = (u_1, \ldots, u_M) \sim$ intrinsic CAR with variance τ_M^2
- $u_c|m$ is relative to the township c in the massif m, for $m \in \{1, \ldots, M\}$ $(u_{1|m}, \ldots, u_{n_m|m}) \sim$ intrinsic CAR with variance τ^2, with n_m the number of townships in the massif m.

*N*eighboring Graph :
In black the massifs
In red the townships.

Outline

Inference is realized on the 63 years series of avalanche counts.
The last year (2009) is let to prediction.
We write MCMC algorithm using Gibbs sampling on the R software.
For each model, 40,000 iterations are performed, and the 10,000 first one are deleted.

To compare models, we compute the DIC criterion (Spiegelhalter et al, 2002).
Let θ be the parameters and $D(\theta)$ be the deviance :

$$p_D = E[D(\theta)] - D[E(\theta)]$$
$$DIC = p_D + E[D(\theta)]$$

p_D is the effective number of parameters of the model.

Models Comparison

Estimates :

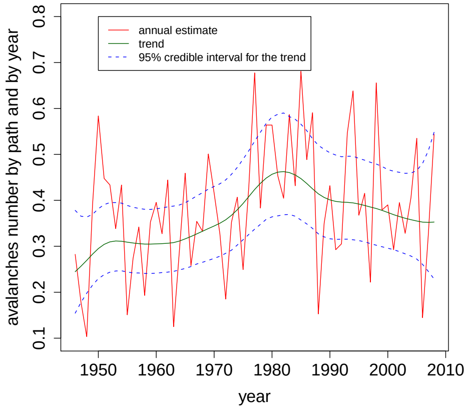
Model 1	Model 2
α -1.03 (0.06)	α -1.05 (0.06)
δ_0 0.16 (0.04)	δ_0 0.18 (0.04)
δ_1 9.8510^{-4} (9.5210^{-4})	δ_1 9.9610^{-4} (9.4310^{-4})
τ^2 2.48 (0.67)	τ_M^2 0.62 (0.25)
	τ_C^2 1.28 (0.35)
σ_v^2 0.18 (0.09)	σ_v^2 0.31 (0.08)

Variability :

	Model 1	Model 2
$\frac{\sigma_v^2+var(E(\mathbf{u}))}{\delta_0+var(E(\mathbf{g}))+\sigma_v^2+var(E(\mathbf{u}))}$	87.26%	82.64%
$\frac{var(E(\mathbf{u}))}{var(E(\mathbf{u}))+\sigma_v^2}$	85.75%	68.55%

DIC :

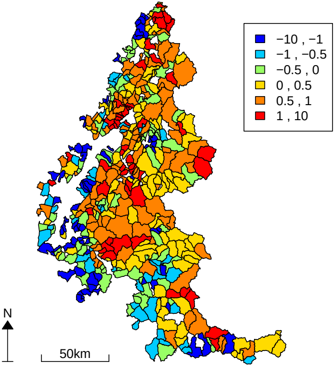
	Model 1	Model 2
DIC	68546	70030
P_D	689	605



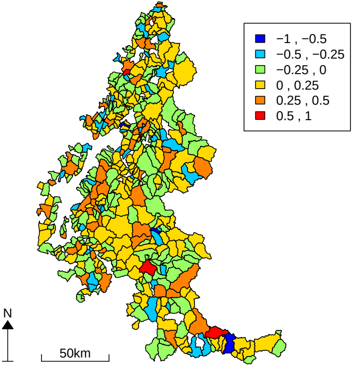
- The temporal term accounts for only 12% of the relative risk variability.
- The temporal trend $E[g|Y]$ shows few variations over the study period
- The annual avalanche frequency presents large variations from one winter to another

Spatial estimation - Model 1

Posterior mean of **u**



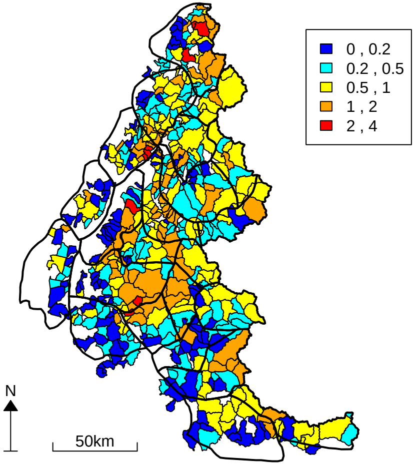
Posterior mean of **v**



- Spatially structure excess in the log relative risk in the eastern and northern regions.

- Posterior mean is rather small
- No spatially structured

Avalanche frequency estimate by path on a mean year - Model 1



- We found back the spatial structure
- Massifs does not correspond to spatially excess or deficit

Outline

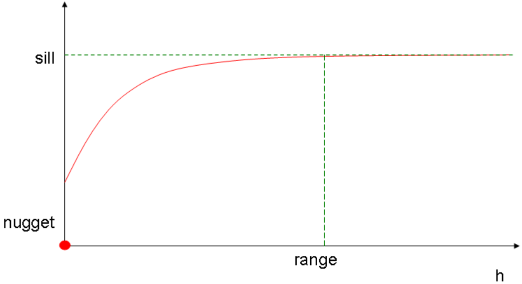
- Largely used to model spatial dependence on a continuous Gaussian field.
- Introduced parameters easily interpretable.

We suppose intrinsic stationarity :

$$\begin{cases} E[\boldsymbol{\eta}_c^S - \boldsymbol{\eta}_{c'}^S] = 0 \\ E[\eta_c^S - \eta_{c'}^S]^2 = 2\gamma(h_{cc'}) \end{cases}$$

with $h_{cc'}$ the distance between the centroids of townships c and c'

$$\gamma(h) = var(\eta_c^S - \eta_{c+h}^S)$$



$$cov(u_c, u_{c'}) = \tau^2 \exp(-\frac{h_{cc'}}{\phi})$$

with $h_{cc'}$ the distance between centroids of townships c and c' :

$$\mathbf{u} \sim N(\beta, \Sigma_\phi) \text{ with } \sum_c u_c = 0 \text{ and } [\beta] \propto cst$$

Parameters :

- range $-\phi log(0.05) \approx 3\phi$
- sill τ^2
- nugget model with the term v, σ_v^2

Outline

Introduction

A separable spatio-temporal model

Model description

Inference and results

Geostatistics' contribution

Model based exponential variogram

MCMC algorithme

Results

Gibbs' sampler Model 1

1. Initialisation
2. Sample δ_0 in $\delta_0|\boldsymbol{\epsilon} \sim IG(0.1 + \frac{T}{2}, 0.1 + \frac{1}{2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon})$
3. Sample δ_1 in $\delta_1|\mathbf{g} \sim IG(0.1 + \frac{T-2}{2}, 0.1 + \frac{1}{2}\mathbf{g}'A\mathbf{g})$
4. Sample σ_v^2 in $\sigma_v^2|\mathbf{v}, \delta_1 \sim IG(0.1 + \frac{N}{2}, 0.1 + \frac{1}{2}\mathbf{v}'\mathbf{v})$
5. Sample τ^2 in $\tau^2|\mathbf{u} \sim IG(0.1 + \frac{N-1}{2}, 0.1 + \frac{1}{2}\mathbf{u}'P\mathbf{u})$
6. Sample $\mathbf{g}$ in
$$\mathbf{g}|\boldsymbol{\eta}^T, \delta_1, \delta_0 \sim N(\mathbf{m}, \delta_0\Omega^{-1})$$
$$\text{avec} \quad \begin{array}{l} \mathbf{m} = \Omega^{-1}\boldsymbol{\eta}^T \\ \Omega = [I_{T\times T} + \frac{\delta_0}{\delta_1}A] \end{array}$$
7. Sample $\mathbf{u}$ in
$$\mathbf{u}|\mathbf{eta}^S, \tau^2, \sigma_v^2 \sim N(\mathbf{m}, \sigma_v^2\Omega^{-1})$$
$$\text{avec} \quad \begin{array}{l} \mathbf{m} = \Omega^{-1}\boldsymbol{\eta}^S \\ \Omega = [I_{N\times N} + \frac{\sigma_v^2}{\tau^2}P] \end{array}$$
8. Sample α in $\alpha|Y, E, \boldsymbol{\eta}^S, \boldsymbol{\eta}^T$
9. For $t = 1, \ldots, T$ sample ϵ_t in $\epsilon_t|Y, E, \alpha, \boldsymbol{\eta}^S, \mathbf{g}, \boldsymbol{\epsilon}_{-t}$
10. For $c = 1, \ldots, N$ sample v_c in $v_c|Y, E, \alpha, \boldsymbol{\eta}^T, \mathbf{u}, \mathbf{v}_{-c}$
11. Go back to 2

Model 1

$$\mathbf{u}|\boldsymbol{\eta}^S, \tau^2, \sigma_v^2 \sim N(\mathbf{m}, \sigma_v^2\Omega^{-1})$$

$$\text{avec} \quad \mathbf{m} = \Omega^{-1}\boldsymbol{\eta}^S, \quad \Omega = [I_{N\times N} + \frac{\sigma_v^2}{\tau^2}P]$$

- Ω depends on P
- Conditional distributions are easily obtained with the precision matrix Ω

Model 3

1. sample ϕ in $\phi|\mathbf{u}, \tau^2, \beta$
2. sample β in $\beta|\mathbf{u}, \tau^2, \phi$
3. sample $\mathbf{u}$ in

$$\mathbf{u}|\boldsymbol{\eta}^S, \tau^2, \sigma_v^2, \phi \sim N(\mathbf{m}, \sigma_v^2\Omega^{-1})$$

$$\text{avec} \quad \mathbf{m} = \Omega^{-1}(\boldsymbol{\eta}^S + \frac{\sigma_v^2}{\tau^2}\Sigma_\phi^{-1}\mathbf{1}\beta), \quad \Omega = [I_{N\times N} + \frac{\sigma_v^2}{\tau^2}\Sigma_\phi^{-1}]$$

- No known distribution for $\phi|\mathbf{u}, \tau^2$
- Inversion of Σ_ϕ at each iteration

Update of (ϕ, τ^2, β)

P.J. Ribeiro and P.J. Diggle, *Bayesian inference in Gaussian model-based geostatistics*, 2002, Technical Report.

1. Choose a discrete uniform prior for ϕ.
2. Compute the posterior probabilities in this support set : $\pi(\phi|u)$ We obtain $\hat{\pi}(\phi|u)$, an approximation of $\pi(\phi|u)$.
3. Sample a value of ϕ in $\hat{\pi}(\phi|u)$.
4. Sample a value of τ^2 in $\pi(\tau^2|u, \phi)$, with ϕ sampled previously.
5. Sample a value of β in $\beta|u, \phi, \tau^2$.

- $$\beta|u,\phi,\tau^2 \sim N(SS^{-1}SP, \tau^2 SS^{-1})$$

- $$\begin{aligned}
[\tau^2|\mathbf{u},\phi)] &= \int \pi(\tau^2|\mathbf{u},\phi,\beta)\pi(\beta)d\beta \\
&\propto (\frac{1}{\tau^2})^{0.1+\frac{N-1}{2}+1}\exp(-\frac{1}{\tau^2}(0.1-\frac{1}{2}\frac{SP^2}{SS}+\frac{1}{2}PP))
\end{aligned}$$

- $$\begin{aligned}
[\phi|u] &\propto \frac{\pi(\phi,\tau^2,\beta)\pi(u|\phi,\tau^2,\beta)}{\pi(\tau^2|u,\phi)\pi(\beta|\phi,u,\tau^2)} \\
&\propto |\Sigma_\phi|^{-\frac{1}{2}} SS^{\frac{1}{2}}(0.1-\frac{1}{2}\frac{SP^2}{SS}+\frac{1}{2}PP)^{-(0.1+\frac{N-1}{2})}
\end{aligned}$$

with $SS = \mathbf{1}'\Sigma_\phi^{-1}\mathbf{1}$, $SP = \mathbf{u}'\Sigma_\phi^{-1}\mathbf{1}$ and $PP = \mathbf{u}'\Sigma_\phi^{-1}\mathbf{u}$

Outline

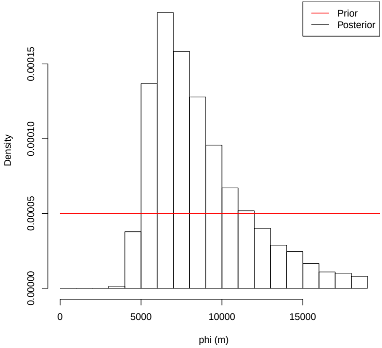
Models Comparison

Estimates :

Model 1	Model 3
α -1.03 (0.06) δ_0 0.16 (0.04) δ_1 9.8510^{-4} (9.5210^{-4})	α -0.98 (0.09) δ_0 0.15 (0.04) δ_1 9.5310^{-4} (8.6210^{-4})

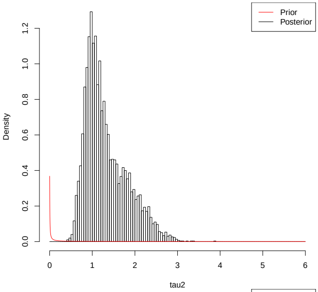
DIC :

	Model 1	Model 3
DIC	68546	69771.2
P_D	689	884.7

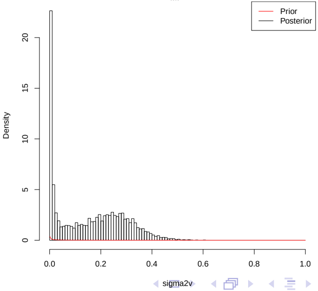


$\phi|\mathbf{Y}$

- Mixture of 2 distribution : $\sigma^2 \approx 0$ or $\sigma^2 \approx 0.3$



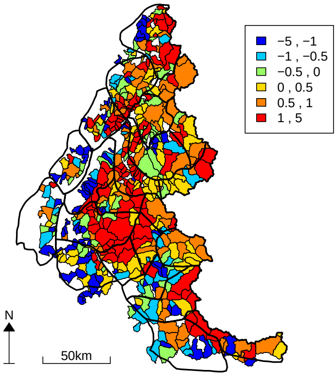
$\tau^2|\mathbf{Y}$



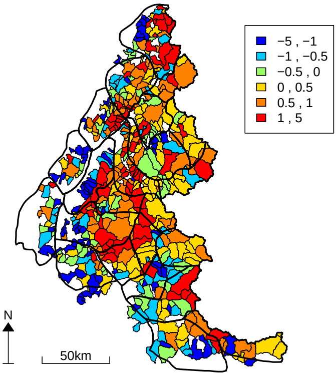
$\sigma_v^2|\mathbf{Y}$

Mean posterior of $\mathbf{u} + \mathbf{v}$

States when σ_v^2 sampled < 0.05

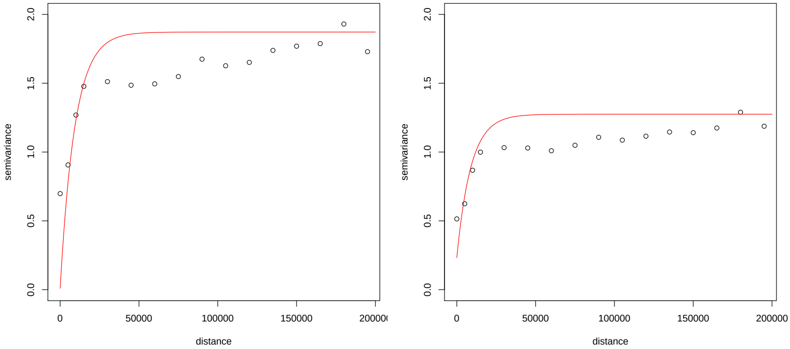


States when σ_v^2 sampled > 0.05



Empirical variogram of η^S

States when σ_v^2 sampled < 0.5	States when σ_v^2 sampled > 0.5



- 2 slope at short and mean range. Where is the sill ?
- 2 variograms are similar, but the scale of variance is different. Is our model too flexible ?

Outcome

- Using geostatistic to model spatial dependence brings flexibility.
- The range, a key parameter to characterize spatial dependence of avalanche frequencies is estimated.
- Empirical variograms seems to bring out 2 ranges.
- Inference is time consuming.
- The model seems too flexible (DIC and the mixture in the distribution a posteriori).

Conclusion

- Avalanche frequencies are spatially dependant. The range is estimated to 30km.
- Massifs seems useless to explain the number of avalanches.
- The evolution of avalanches frequency does not show monotonous trend.

Concerning the predetermination of the avalanches risk these results suggests that we can take benefits of the spatial structure. The separability between time and space assumption is discutable. To better understand the process at the origin of the avalanche frequency, we propose to remove the hypothesis of separability.

Bibliography

- Eckert, N., Parent, E., Bélanger L. et Garcia S. (2007) *Hierarchical Bayesian modelling for spatial analysis of the number of avalanche occurences at the scale of the township*, Cold Regions Science and technology, 50, 97-112.
- Eckert, N., Parent, E., Kies, R. et Baya, H. (2009) *A Spatio-temporal modelling framework for assessing the fluctuations of avalanche occurrence resulting from climate change : application to 60 years of data in the northern French Alps*, Climatic Change, 101, 515-553.
- Mollié, A. et Richardson, S. (1991) *Empirical Bayes estimate of cancer mortality rate using spatial models*, Statist. Med., 10, 95-112.
- Wahba, G (1978) *Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression*, Journal of the Royal Statistical Society. Series B (Methodological), 40, 364-372.