A Dirichlet process mixture model for Bayesian analysis of extreme avalanches

Ophélie Guin[1], Nicolas Eckert[2], Liliane Bel[3], Eric Parent[3]

[1] Université de Lille
[2] UR ETNA, INRAE Grenoble,
[3] UMR MIA-Paris-Saclay, AgroParisTech, INRAE, université Paris-Saclay

AppliBUGS - 19 décembre 2023

Extreme avalanches

Avalanches with important runout distances / low stopping altitude,

$\Rightarrow$ provoke human and material losses.

Need methods to predict such avalanches.

Preliminary investigate :

- Change over time (e.g. impact of climate change),
- Stopping altitude distribution.

$\Rightarrow$ **Modeling stopping altitudes for extreme avalanches**.

EPA dataset

Available data = Enquête Permanente des Avalanches (EPA).

id_evt	dept	nom_dept	massif	id_massif	insee	commune	num_site_epa	id_site_epa	nom_site	observation_s	type_observat	alt_site
33344	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33345	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33346	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33347	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33348	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33349	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33350	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33351	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33352	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33353	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33354	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33355	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33356	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33357	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33358	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33359	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33360	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33361	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33362	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33363	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33364	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33365	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33366	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33367	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33368	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33369	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33370	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33371	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33372	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33373	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33374	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33375	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33376	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33377	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33339	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33340	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33341	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33342	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
33343	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
77979	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
77980	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
77981	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
79544	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	3756	RESSEC	Site observé	Observation	
81562	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	5912	RESSEC	Site observé	Observation	
81563	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	5912	RESSEC	Site observé	Observation	
81897	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	5912	RESSEC	Site observé	Observation	
82154	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	5912	RESSEC	Site observé	Observation	
83883	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	5912	RESSEC	Site observé	Observation	
85640	73	SAVOIE	HAUTE-MAUR	10	73144	LANSLEVILLARD	2	5912	RESSEC	Site observé	Observation	

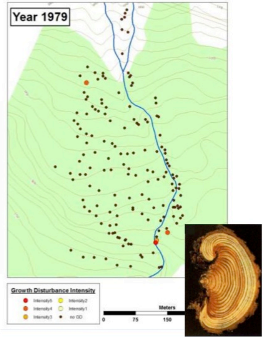
Extrait base EPA pour le site de Ressec

- Managed and developed by Inrae.
- Collection of data on avalanches (dates, snow cover, departure and arrival altitudes, type of avalanche, etc.).
- 3900 sites in 11 departments.
- More than 90,000 events available.

Time-limited database that only begins in 1900 + unclear at first.

More data

How to retrieve more data for better estimations ?

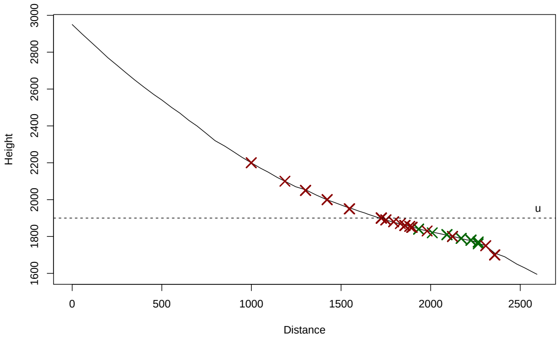


Dendrochronological data.

1. Sample trees in an avalanche corridor,
2. detect impacted trees (in tree-rings) for each year,
3. determine avalanches years with spatial impact repartition,
4. set runout distance as the distance of the last impacted tree.

Difficulty : **censored data**.

Data for Ressec site (Savoie).



Goal : probabilistic modeling of extreme avalanches as a function of time and stopping altitude.

Prerequisites :

- bivariate modeling,
- flexible modeling,
- taking into account that data may be censored,
- and non-stationary.

Peak over threshold (POT) modeling for the stationary case (1)

Let $X_1, X_2, ..., X_n \overset{\text{iid}}{\sim} F$, X_j occur at regular time intervals.
u a threshold level, we focus on the distribution $F_u = \mathbb{P}(X - u \leq z | X > u)$.

- F_u converges in distribution to the Generalized Pareto's Distribution (GPD) :

$$\lim_{u \to u_\infty} F_u(z) = 1 - \left(1 + \xi \frac{z}{\sigma}\right)^{-1/\xi}.$$

  [Pickands, 1975].

- Number of exceedances follows a Poisson's distribution.

Peak over threshod (POT) modeling for the stationary case (2)

Relabel sub-sample $(T_i, u + Z_i)$, with $i = 1, ..., r$, from ordered pairs (j, X_j) such that X_{T_i} exceeds u.

After re-normalization on $\mathcal{A} = [0, 1] \times \mathbb{R}^+$, (T_i, Z_i) converges towards homogeneous Poisson process on $\mathbb{R}^2$ with a separable intensity :

- homogeneous in time,
- GPD in the second component.

[Coles, 2001]

Extending for non-stationary case

Assumptions of the POT model :

- a sufficiently high threshold,
- independence for peaks over an acceptably high level.

$\Rightarrow$ Stationary behavior of the series.

Problem : stationary assumption in statistical avalanche analysis not realistic.

Modeling

The point process (T_i, Z_i) are realizations such that :

- inhomogeneous in time,
  $\Rightarrow$ relax marginal Uniform distribution for a Beta distribution.
- frequency and severity of avalanches are not independent,
  $\Rightarrow$ in bayesian context components are independent *a priori* but not *a posteriori*.

More flexibility : take benefit from the classical POT model (with Beta x Pareto kernel) to generate weighted components of mixture model.

Model interpretation : study model uncertainty around the baseline of the classical POT approach for extremes.

Intensity density of Poisson process

Main question : estimating the intensity density $\lambda(t, z)$ of Poisson process.

Let the decomposition :

$$\lambda(.) = \gamma f(.),$$

$\gamma = \int_{\mathcal{A}} \lambda(t, z) dtdz$ the total intensity and $f(.)$ a density function.

Poisson process likelihood function :

$$L(\gamma, f(.); \{(t_i, z_i) : i = 1, ..., r\}) \propto \exp(-\gamma)\gamma^r \prod_{i=1}^{r} f(t_i, z_i),$$

Estimation can be broken down into two independent problems :

- $\gamma \rightarrow$ easy in Bayesian context.
- $f(.) \rightarrow$ difficult.

Kernel specification (1)

Specialists do not have much idea about density form in avalanches context.

$\Rightarrow$ Flexible model by relying on a nonparametric mixture :

$$f(t,z) = f(t,z;G) = \int_{\Theta} k(t,z|\boldsymbol{\theta}) dG(\boldsymbol{\theta}),$$

with $k(t,z|\theta)$ a parametric density with parameter θ and G a random mixing distribution.

Specification of bivariate kernel :

$$k(t,z|\boldsymbol{\theta}) = k(t,z|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = k_1(t|\boldsymbol{\theta}_1) k_2(z|\boldsymbol{\theta}2),$$

where k_1 and k_2 independent before mixing.

Kernel specification (2)

Beta distribution for kernel component over time :

$$k_1(t|\boldsymbol{\theta}_1) = \frac{\Gamma(\tau)}{\Gamma(\tau\kappa)\Gamma(\tau(1-\kappa))} t^{\tau\kappa-1}(1-t)^{\tau(1-\kappa)-1},$$

with $\kappa \in (0,1)$ the mean and $\tau > 0$ a scale parameter.

Generalized Pareto Distribution (GPD) for the exceedances :

$$k_2(z|\boldsymbol{\theta}_2) = \frac{1}{\sigma}\left(1 + \frac{\xi(z-u)}{\sigma}\right)^{-1/\xi-1}, z \geq u,$$

with $\boldsymbol{\theta}_2 = (\sigma, \xi)$ with $\sigma > 0$ and $\xi > 0$.

Dendrochonological records considered as censored data.

$\Rightarrow$ $k_2(z|\boldsymbol{\theta}_2)$ will be replaced by $\mathbb{P}(Z > z) = 1 - K_2(z|\boldsymbol{\theta}_2)$ for these data.

Bayesian Hierarchical Model

Dirichlet process prior $DP(\alpha, G_0)$ is one of the most widely used Bayesian nonparametric priors.

[Ferguson, 1973]

G_0 the base distribution and α controls how close the realization G is to G_0.

$\Rightarrow$ **Hierarchical model** :

$$\lambda(t,z) \equiv \lambda(t,z;G,\gamma) = \gamma f(t,z;G) = \gamma \int_{\Theta} k(t,z|\boldsymbol{\theta}) dG(\boldsymbol{\theta})$$

$$G|\alpha,\boldsymbol{\theta} \sim DP(\alpha, G_0),$$

γ inference

Marginal prior for γ : $p(\gamma) \propto \gamma^{-1} 1_{\gamma>0}$.
[Kottas and Behseta, 2010].

$\Rightarrow$ Proper posterior distribution $p(\gamma|t, z)$ is a gamma$(n, 1)$.

Dirichlet process algorithm (1)

Dirichlet process realizations are discrete with probability one,

$\Rightarrow$ model can be viewed as infinite mixtures *[Ferguson, 1983]*.

Equivalent model obtained by taking the limit as L goes to infinity of finite mixture models with L components :

$$(t_i, z_i)|\kappa_{L_i}, \tau_{L_i}, \sigma_{L_i}, \xi_{L_i} \sim k_1(t_i|\kappa_{L_i}, \tau_{L_i})k_2(z_i|\sigma_{L_i}, \xi_{L_i}),\ i = 1, ..., r$$
$$L_i|\boldsymbol{p} \sim \text{Discrete}(p_1, ..., p_L)$$
$$\boldsymbol{p}|\alpha \sim \text{Dirichlet}(\alpha/L, ..., \alpha/L)$$
$$\theta_l = (\kappa_l, \tau_l, \sigma_l, \xi_l) \sim G_0(\theta_l|\psi),\ l = 1, ..., L,$$

L_i represent à “latent class” associated with observation (t_i, y_i).

Dirichlet process algorithm (2)

Posterior inference for this type of model based on the Chinese Restaurant Process sampler *[Neal, 2000]*.

Here G_0 base measure is a non-conjugate prior for $\boldsymbol{\theta}$,
$\Rightarrow$ more difficulties and use of numerical techniques.

Algorithm used for inference is Algorithm 8 of *[Neal, 2000]*.

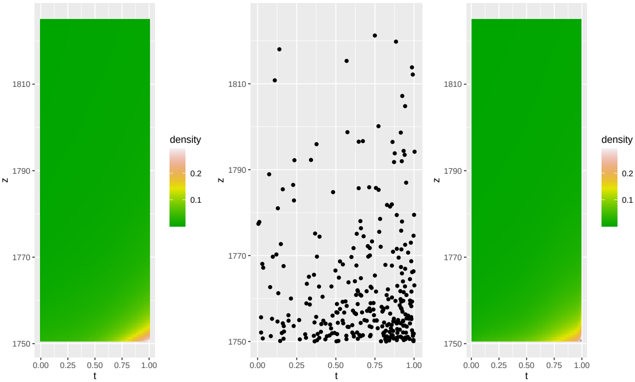
Base distribution

The different base distribution components are *a priori* independent :

$$G_0(\boldsymbol{\theta}) = G_0(\kappa, \tau, \sigma, \xi) = G_0^{\kappa}(\kappa) G_0^{\tau}(\tau) G_0^{\sigma}(\sigma) G_0^{\xi}(\xi)$$

- G_0^{κ} is an uniform distribution,
- G_0^{τ} and G_0^{σ} are inverse-gamma distributions with fixed shape parameter,
- G_0^{ξ} is an exponential distribution.

Simulations



Datasets

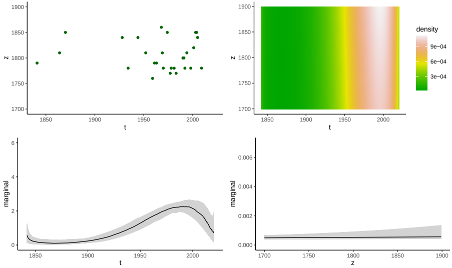
Two types of data for the Ressec site :

- 67 EPA data between 1901 and 2022,
- 28 dendrochronological data from 1840.

Data pre-treatment : select only the avalanche from the EPA base (more precise) when avalanches appear in the two bases.

Threshold $u = 1900$.

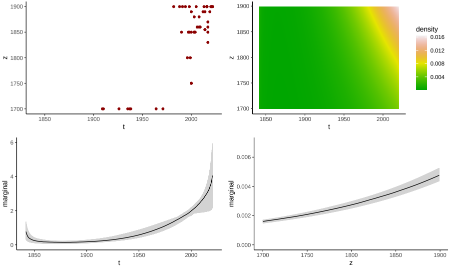
Model apply to dendrochronological data



- Increase in avalanches over time.
- Unclear results for stopping altitudes.

  ⚠ Only worked with censored data.

Model apply to EPA data

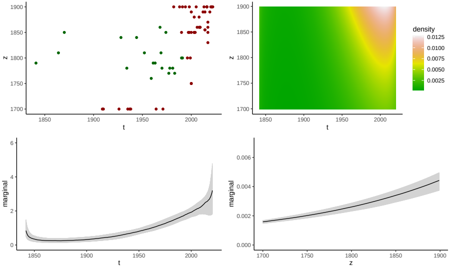


- Probability of extreme avalanches increases with time (with acceleration in the 2000s).

  ⚠ Few data before the 1980s.

- Stopping altitudes probability increases for higher altitude.

Model apply to EPA and dendrochronological data



- Dendrochronological data provide information (before 1980 and for stopping altitudes greater than 1700m).
- Probability of extreme avalanches increases with time.
- Stopping altitudes probability increases but acceleration in the 1950s.

Conclusion and perspectives

Non-parametric modeling of extreme non-stationary and censored data.

Perspectives :

1. Here, restriction to the case $\xi > 0$,

   $\Rightarrow$ extend to all values of ξ.

2. Return period estimations.

References

Coles, S. (2001)
An Introduction to Statistical Modeling of Extreme Values
Springer

Ferguson, T. S. (1973)
A bayesian analysis of some nonparametric problems.
The Annals of Statistics, 1 :209–230.

Neal, R. (2000).
Markov chain sampling methods for dirichlet process mixture models.
Journal of Computational and Graphical Statistics, 9(2) :249–265.

Pickands, J. (1975).
Statistical inference using extreme order statistics.
The Annals of Statistics, 3(1) :119–131.