			Avalanches data	Conclusion
0000 0000	00000	00000	0000	00

A Dirichlet process mixture model for Bayesian analysis of extreme avalanches

Ophélie Guin¹, Nicolas Eckert², Liliane Bel³, Eric Parent³

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

AppliBUGS - 19 décembre 2023

Context		
0000		

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.

Context		
0000		

EPA dataset

Available data = Enquête Permanente des Avalanches (EPA).

	1000	1000,000	1441	a search and a search of the s	0.44		- Andrews	101000.004	10,000,004		1001, 100	shart and	the measure at one
ID+	(\cdots)	DOM 0	HISTORY,		2	TIDH	DOUGHTER		2	174	NEEDEC	Six aborriet	Descelor y
104	< - 1	IC SAUCE	100 Y.MA.		×	13344	14H5.F210.480		8	10	NPP 41	She allow with	Hoevener (
113-0	(DOM 0	HISTERAL)		20		UNITARY LINC		2	172	NUR	Big along of	Distriction (
1114	<u> </u>	9.9493	10070-0040			12344	100032012-000		6	100	RECEL	514 4000 VE	Monwhere a
TDA		NOVA P	NUTAR				UP/LTVU.MD		2	172	NUNC	Dis alsonal	Disarcation (
1114	<u> </u>	1 MOR	10070-0040				1003270.000			174	RECAL	Sta aboorwi	Steerather a
1179	<u> </u>	S SHOT	INCC.MA.				URINE AND			179	MOOT .	Steakerst	Boavaties (
LINC		DOM 0	10070-044.0		2		10-0.771.480			170	NEEDEC	Big along of	Overwhee a
110	<u> </u>	0.9908	INCOME.		*		DARK DOLLARS				NECHT	Station of	Monwhen a
LIF:		1 MOT	SUT MO				LEADING AND			172	1 M02T	Surgery .	Disposition is
III	\sim	DOM 0	10070-000				DOM: NO.			110	MERRIC	Six show'rd	Description a
1172		n'ssor	THEY ARE		· · · ·	TTM	LENGTHER AND			179	19221	Surgery .	Margarian 1
LIFE		DOM 0	10070-044.0		2		URB FYLIND		2	179	MERC	Site along of	Description a
1172		1000	100703001			11144	SHOP FOR ART.		*	1120	arrer	No shored	Manager 1
1172		DOM C	SUTAL				101501110				M225	Strainery!	Department of
1100		1,240.0	10070-000			TOH	DOM: NO.		2	119	HECHT	Site about we	Description of
TDS		N MOT	THEY MAN				LEATING AND			172	1 M02T	Surgery .	Discounting in
IIK		DOM 0	10070-044				UNITED IN			170	NEEDEC	Six show'rd	Description a
105		n'son	1017.000		· · · ·	TTM	LANS FOR ARE		X	119	19921	Strategy 1	Manager 1
IDC		TOMA C	SUTAL			TIME	UP& DYN ARD		2	172	M221	Stational	Departure 1
11 ho		IL MARKET	100120-004		· · · ·	TTM	AND THE ART.			174	arrar	No sheeted	Management of
102		1044 0	THE Y MAN		-	TIM	URINE COLUMN		2	172	M221	Summer of	Reporter 1
TTN		IN MACH	10070-044			TTM	MARK FROM AND		-	175	AFTER .	The sheeted	Secondary 1
106		1000	10071-004			11144	UNIT FUEL AND		2	109	and a second	No shared	Manual II.
ID9		DOM C	SUTAL				101501110		2	172	MINT	Straineyd	Department a
11160		1000	100110-0041		· · · ·	11144	MARK FOR ART.		*	179	arrer	We should	Manager at 1
IIC		TOMA OF	THE Y MAN		÷	TIME	1015101100		2	172	M221	Summer	Disposition in
1117		IN MACH	100120-004			TTM	MARK FROM AND		8	174	AFTAT	No sheets	Number 1
103		1000	100 1.000			TIM	UNITATION AND			109	19221	Strategy 1	Manufacture 1
THE		0.1007	100 77 144 1		-	TIME	100 Four set		-	110	MANAT	The should	Secondary 1
IIII		IL MARKET	100120-004		· · · ·	TTHE	AND THE ARE			174	arrar	No should	Management of
IIC		1044 0	100 1.000		-	TIM	URINE COLUMN			172	M221	Summer	Burnetin 1
TITL		IN MACH	10070-044			TTM	Mark Party and		÷	170	AFTAT.	The sheeted	Secondary 1
1005		IL SAVE	10071-004			11144	UNIT FOR ART.			109	and a second	No shared	Manufacture of
TITE		D MOOT	100 10 144			TTM	NAME AND ADDRESS		-	110	MANAT	The sheared	Secondary 1
104		DOM: NO	100 100 000 0			11.00	MARK FROM AND			100	arrer .	No should	Manual Inc.
1124		0.1807	100 10 144		-	TEAM	1848 Fair and		-	110	MANY	Excision 1	Management of the
100		IN MARCH	100 10.044				AND FAIL AND			100	HTTP:	The sheet of	Support of the local division of the local d
104		1042.0	1007.004			TIM	UPID FOR ARE			109	19221	Strategy 1	Manufact 1
110		0.1007	100 77 144 1		-	TIME	100 For 185		-	110	MANAT	The should	Secondary 1
1100		IN MARCH	100 10 100 1			11.00	MARK FROM AND			100	arrer -	No should	Theorem is a
These		0.1807	100 10 100 1		-	TRADE	THE PART AND		-	100	MONT	- Decidented	Management of the
10.0		N MACON	100 10 144		o		AND FAIL AND			100	ACC	The shoes of	Second in the
							COLUMN STATES		÷	-		and should be	Married Workshop
1100		N MOOT	1000 10044		-	100	NAME AND ADDRESS.		-	- 22	MANA	The sheet of	Second Second
1100		DOM: NO	100 100 000 0			11.00	MARK FROM AND			- 107	arrer .	No should	Manual Inc.
- 100		0.1007	1000 1000		-		THE OWNER.		-	- 22	MANY	The shore of	Street and a local division of the
1100		IN MARCH	100 10,044				AND FAIL AND			- 107	HTTP:	The sheet of	Support of the local division of the local d
							COLUMN TWO IS NOT		-	- 22	Marco		and the second s
					_								

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.

Context		
0000		

More data

How to retrieve more data for better estimations?



Dendrochronological data.

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

Difficulty : censored data.

Context		
0000		

Data for Ressec site (Savoie).



	Modelisation			
0000	•0000000	00000	0000	00

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.

Modelisation		
00000000		

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

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

• *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.

Modelisation		
00000000		

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]

Modelisation		
00000000		

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.

	Modelisation			
0000	00000000	00000	0000	00

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,
 ⇒ 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.

Modelisation		
000000000		

Intensity density of Poisson process

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

Let the decomposition :

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

 $\gamma = \int_{\mathcal{A}} \lambda(t, z) dt dz$ 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.
- $f(.) \rightarrow \text{difficult}$.

Modelisation		
000000000		

Kernel specification (1)

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

 \Rightarrow Flexible model by relying on a nonparametric mixture :

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

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

Specification of bivariate kernel :

$$k(t, z|\theta) = k(t, z|\theta_1, \theta_2) = k_1(t|\theta_1)k_2(z|\theta_2),$$

where k_1 and k_2 independent before mixing.

Modelisation		
000000000		

Kernel specification (2)

Beta distribution for kernel component over time :

$$k_1(t|\theta_1) = \frac{\Gamma(\tau)}{\Gamma(\tau\kappa)\Gamma(\tau(1-\kappa))} t^{\tau\kappa-1} (1-t)^{\tau(1-\kappa)-1},$$

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

Generalized Pareto Distribution (GPD) for the exceedances :

$$k_2(z|\theta_2) = \frac{1}{\sigma} \left(1 + \frac{\xi(z-u)}{\sigma}\right)^{-1/\xi-1}, z \ge u,$$

with $\theta_2 = (\sigma, \xi)$ with $\sigma > 0$ and $\xi > 0$.

Dendrochonological records considered as censored data.

 $\Rightarrow k_2(z|\theta_2)$ will be replaced by $\mathbb{P}(Z > z) = 1 - K_2(z|\theta_2)$ for these data.

Modelisation		
00000000		

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 :

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

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

	Inference	
	00000	

γ inference

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

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

	Inference	
	0000	

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 :

$$\begin{split} (t_i, z_i) | \kappa_{L_i}, \tau_{L_i}, \sigma_{L_i}, \xi_{L_i} \sim k_1(t_i | \kappa_{L_i}, \tau_{L_i}) k_2(z_i | \sigma_{L_i}, \xi_{L_i}), \, i = 1, ..., r \\ L_i | \boldsymbol{p} \sim \text{Discrete}(\boldsymbol{p}_1, ..., \boldsymbol{p}_L) \\ \boldsymbol{p} | \alpha \sim \text{Dirichlet}(\alpha/L, ..., \alpha/L) \\ \theta_l = (\kappa_l, \tau_l, \sigma_l, \xi_l) \sim G_0(\theta_l | \psi), \, l = 1, ..., L, \end{split}$$

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

	Inference	
	00000	

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 θ , \Rightarrow more difficulties and use of numerical techniques.

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

	Inference	
	00000	

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.

	Inference	
	00000	

Simulations



	Avalanches data	
	•000	

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.

	Avalanches data	
	0000	

Model apply to dendrochronological data



- Increase in avalanches over time.
- Unclear results for stopping altitudes.
 A Only worked with censored data.

	Avalanches data	
	0000	

Model apply to EPA data



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

 \triangle Few data before the 1980s.

 Stopping altitudes probability increases for higher altitude.

	Avalanches data	
	0000	

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
		00

Conclusion and perspectives

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

Perspectives :

Where, restriction to the case ξ > 0,
 ⇒ extend to all values of ξ.

2 Return period estimations.

		00

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.