A default Bayesian approach for regression on extremes

Stefano Cabras [a,1]

[a] *Department of Mathematics, University of Cagliari, Via Ospedale 72, 09124 Cagliari (Italy).*

Maria Eugenia Castellanos Nueda [b,2]

[b] *Department of Statistics and Operations Research, Rey Juan Carlos University, c/ Tulipán, 28933, Móstoles (Spain)*

Dani Gamerman [c,3]

[c] *Department of Mathematics, Universidade Federal do Rio de Janeiro Rio de Janeiro (Brazil).*

Abstract

A default Bayesian approach to predict extreme events in the presence of explanatory variables is presented. The approach is based on the non-homogenous Poisson-Generalized Pareto Distribution point process and allows for variation of the tail behaviour according to explanatory variables. The prior proposed is based on a similar Jeffreys' rule for the regression parameters. Inference is performed approximately via MCMC and the posterior turns out to be relatively easy to be computed. The model is applied to two real data sets from meteorological applications.

Key words: Extreme values, Jeffreys's prior, Peaks Over the Threshold, non-homogenous Poisson-GPD

1 Introduction

In the past two decades there has been an increasing interest for statistical methods that model rare events in numerous disciplines as Environmental Sciences, Finance and Insurance, among others (see for instance Coles (2001) and Smith (2003)). We consider the class of models arising from extreme value theory (Gnedenko (1943) and Pickands (1975)) with particular attention to model exceedances of a random variable Z over a fixed high threshold, u (Davison and Smith (1990)). Let $Z \sim H(z)$ represents the quantity of interest (i.e. rainfall levels or returns of a financial asset) and $Y = Z - u > 0$ the exceedance over u. Pickands (1975) showed that under regularity conditions on $H(z)$, the limiting distribution of Y, when $u \to \infty$, is the Generalized Pareto Distribution (GPD in the sequel) with shape ξ and scale σ

$$f(y \mid \xi, \sigma) = \begin{cases} \sigma^{-1} (1 + y\xi/\sigma)^{-(1+\xi)/\xi}, & (1 + y\xi/\sigma) > 0 \\ \\ \sigma^{-1} \exp(-y/\sigma), & \xi = 0 \end{cases} . \qquad (1)$$

Email addresses: s.cabras@unica.it (Stefano Cabras),
maria.castellanos@urjc.es (Maria Eugenia Castellanos Nueda),
dani@im.ufrj.br (Dani Gamerman).

[1] Supported by Ministero dell'Istruzione, dell'Univesità e della Ricerca of Italy
[2] Supported by the Spanish Ministry of Science and Technology, under Grants MTM2007-61554 and HBP2005-040.
[3] Supported by grants from CNPq-Brazil and FAPERJ-Brazil.

Note that the support of Y depends on distribution parameters: $Y > 0$ for $\xi \geq 0$ and $0 \leq Y \leq -\sigma/\xi$ for $\xi < 0$. This means that Y is upper bounded for negative values of the shape parameter ξ and the model does not satisfy usual regularity conditions for the Maximum Likelihood Estimator (MLE) (Smith, 1984).

It is often the case that the probability of hydrological extreme events depends on the period of the year, or the characteristics of the geographical area. We therefore relate the parameters of the GPD to available covariates $\mathbf{X}$ through regression functions as in Chavez-Demoulin and Davison (2005). However, we do this in a context of a default Bayesian approach, by using linear regression functions on a reparametrization of GPD parameters. Their prior distribution is derived from a formal rule similar to the Jeffreys's one. The main idea is to extend the Jeffreys's rule employed in Castellanos and Cabras (2006) to regress ξ and σ on $\mathbf{X}$.

Our proposal for reparametrization, link functions and prior distribution over the regression parameters are presented in the following Section. In Section 3, the posterior distribution for regression parameters is derived along with the algorithm to simulate from it. Two applications to real data sets appear in Section 4. Concluding remarks and comments are left for Section 5.

2 Prior distributions and link functions

The model and the corresponding prior distributions will be constructed in steps. We start from the Generalized Pareto Distribution and then successively incorporate elements from Poisson point processes. These are shown to be

required in order to predict future events.

2.1 The Generalized Pareto Distribution

Suppose a random sample $\mathbf{y} = (y_1, \ldots, y_n)$ of observations from (1) is available. Castellanos and Cabras (2006) proposed the following prior distribution derived from Jeffreys' rule

$$\pi(\xi, \sigma) \propto \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2}, \quad \xi > -1/2, \sigma > 0,$$

leading to a proper posterior distribution (Castellanos and Cabras, 2006, Theorem 1). They also showed that Jeffreys rule produces more accurate inference than other bayesian estimators, as $\pi(\xi, \sigma)$ mitigates the odd behavior of the likelihood function derived from (1).

In the multivariate inference case with many parameters, as the one here considered, an orthodox application of the Jeffreys's rule may be not a suitable choice. This point is discussed in general in (Box and Tiao, 1992, Section 1.3.6), and in particular in our Section 5 for case of the GPD model. We then start to reparametrize density (1) in terms of orthogonal parameters ξ and ν according to Chavez-Demoulin and Davison (2005) where $\nu = \sigma(1+\xi)$. The density (1) becomes

$$f(y|\xi, \nu) = \begin{cases} \nu^{-1}(1+\xi)(1+y\xi(1+\xi)/\nu)^{-(1+\xi)/\xi}, & (1+y\xi(1+\xi)/\nu) > 0 \\ \nu^{-1}\exp(-y/\nu), & \xi = 0 \end{cases} \tag{2}$$

Note that in this parametrization the support complicates as $Y > 0$ for $\xi \geq 0$ and $(1 + Y\xi(1+\xi)/\nu) > 0$ for $\xi < 0$.

Along the paper we assume that each observation y_i comes from a GPD with different parameters ν_i and ξ_i, and we assume that they can be modeled using fixed design matrix $\mathbf{X}_{n\times p}$ through link functions $\nu_i = \nu(\mathbf{x}_i^T\boldsymbol{\beta})$ and $\xi_i = \xi(\mathbf{x}_i^T\boldsymbol{\gamma})$, where $\mathbf{x}_i$ indicates row i of $\mathbf{X}$, $i = 1, \ldots, n$. Generalization to different sets of covariates for ξ and ν is straightforward and would not be pursued here.

In this parametrization the two blocks of p parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, are orthogonal and this simplifies the derivation of the joint prior $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The diagonal elements of the Fisher information matrix for model (2) as appears in Smith (1984) when $\xi > -0.5$ are

$$i_{\xi\xi} = \frac{1}{(1+\xi)^2}, \quad i_{\nu\nu} = \frac{1}{\nu^2(1+2\xi)}. \tag{3}$$

Therefore, if we consider these parameters to be independent a priori, as in page 57 of Box and Tiao (1992), the link functions that makes the Jeffreys' priors for ξ and ν locally uniform can easily be obtained. In particular, following this rule, one can consider the marginal prior for each parameter ξ and ν when only one of them is unknown. The prior for ξ, considering known ν is $\pi(\xi) \propto 1/(1+\xi)$ therefore $\pi(\log(1+\xi)) \propto 1$, while the prior for ν with ξ known is $\pi(\log(\nu)) \propto 1$.

Thus, incorporation of the effect of covariates into the model can be made in two steps: initally, linear predictors $\eta_\xi = \mathbf{x}^T\boldsymbol{\gamma}$ for ξ and $\eta_\nu = \mathbf{x}^T\boldsymbol{\beta}$ for ν are constructed; then, these linear predictors are related to ξ and ν via $\eta_\xi = \log(1+\xi)$ and $\eta_\nu = \log(\nu)$. This approach reproduces the regression strategy adopted for many standard models such as log-linear regression and logistic regression. Based on this fact, we propose the inverse of the above transformations as link functions in order to obtain uniform priors. More explicitly, for each $(y_i, \mathbf{x}_i)$ we consider a GPD with parameters (ξ_i, ν_i) defined as linear combinations of

covariates through the link functions:

$$\xi_i = \exp(\mathbf{x}_i^T \boldsymbol{\gamma}) - 1,\ \nu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}). \tag{4}$$

This allow us to pose a flat prior for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$,

$$\pi(\boldsymbol{\theta}) \propto \mathbf{1},\ \text{ for } \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{2p}.$$

Prior and likelihood support Θ depends on $\mathbf{X}$ and parameter constraints in model (2), as $\boldsymbol{\theta}$ is a point satisfying the following joint restrictions for all i

$$\begin{cases} \exp(2\mathbf{x}_i^T \boldsymbol{\gamma}) - \exp(\mathbf{x}_i^T \boldsymbol{\gamma}) > -y_i^{-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\ \mathbf{x}_i^T \boldsymbol{\gamma} > -\log(2) \end{cases}. \tag{5}$$

The last restriction is needed for the existence of Fisher information (Smith, 1984). These restrictions can only be verified numerically.

Motivations for (4) and $\pi(\boldsymbol{\theta})$ stays in the fact that the linear predictors η_ν and η_ξ have a uniform prior mass on their respective support. This amounts to assume the joint improper prior on $\boldsymbol{\theta}$ given by the product of independent Jeffreys priors for ν_i and ξ_i, obtained when only one of the two parameter is unknown.

2.2 The homogenous Poisson-GPD process

In many situations we are interested not only in the intensity of the exceedances, but also in the number of exceedances that occur by unit time. A possible way to model this setting is by using the Poisson-GPD point process. This model can be used when the number of exceedances N over u by unit

time (i.e. a year or a season) follows the Poisson distribution, with intensity λ, and conditionally on $N \geq 1$, the exceedances $Y_1, \ldots, Y_N$ are independently drawn from (2). The joint probability density function of n exceedances, $\mathbf{y}$, observed in a unit time, for $n \geq 1$ is:

$$\Pr(N = n) f(\mathbf{y} \mid N = n) = \frac{\lambda^n}{n!} \exp(-\lambda) \prod_{i=1}^{n} f(y_i \mid \xi_i, \nu_i), \tag{6}$$

while the probability of observing $N = 0$ exceedances is $e^{-\lambda}$.

The most common feature of interest for the Poisson-GPD process is the K-year return level, L_K, defined as the level which is exceeded on average once in K-years, assuming the year as the time unit. For the homogeneous Poisson-GPD process, L_K is derived using the fact that the mean crossing rate for the event $Z > u$ in a year is $\lambda \left[1 + (z-u)\xi(1+\xi)/\nu\right]^{-1/\xi}$. Setting this quantity equal to $1/K$ and isolating z gives

$$L_K = u + \frac{\nu}{\xi(1+\xi)} \left[(\lambda K)^{\xi} - 1\right].$$

2.3 The non-homogenous Poisson-GPD process

As mentioned before, the extremal processes usually depend on the time span and on the geographical region. Thus, the process is no longer homogenous. As interest lies in the posterior distribution of L_K in a certain domain $\mathcal{D}$, and this quantity is described by $\mathbf{X}$ (eg. seasonality and geographical locations), we consider the *non-homegeneous* Poisson-GPD process also described in Leadbetter et al. (1983) and Resnick (1987). The intensity λ in (6) now depends on time and space, that is the domain is $\mathcal{D} \equiv [0, T] \times S$, and will be denoted by $\lambda(t, s)$ for $(t, s) \in \mathcal{D}$ to make this dependence explicit. Then for a measurable subset $A \subset \mathcal{D}$, if we denote $N(A)$ the number of exceedances in

A, then $N(A)$ has a Poisson distribution with mean

$$\Lambda(A) = \int_A \lambda(t,s)dtds.$$

In the special case considered in our applications, $\lambda(t,s)$ is assumed to have a constant value λ_{qr} along season T_q, $q = 1, \ldots, p_T$, and region, S_r, $r = 1, \ldots, p_S$, with $p_T + p_S = p$. Then, the probability of having n_{qr} exceedances over u in region S_r and season T_q along T years is

$$\frac{\left(\sum_1^T \int_{T_q} \int_{S_r} \lambda_{qr} dtds\right)^{n_{qr}}}{n_{qr}!} \exp\left\{-\sum_1^T \int_{T_q} \int_{S_r} \lambda_{qr} dtds\right\} = \tag{7}$$
$$\frac{\left(T\mu(T_q)\mu(S_r)\lambda_{qr}\right)^{n_{qr}}}{n_{qr}!} \exp\left\{-T\mu(T_q)\mu(S_r)\lambda_{qr}\right\}.$$

where $\mu(T_q)$, $\mu(S_r)$ are the measures of seasons and regions.

We are mainly interested in cases where seasons (and regions) have the same time (and space) length. Nevertheless this can be easily extended to different situations, for example when observations are not referred to the same time length and are not points in space. If the time span is divided in p_T equal interval times and observation space in p_S equal area subregions, the probability of having n_{qr} over seasons $q = 1, \ldots, p_T$ in T years and regions $r = 1, \ldots, p_S$ is:

$$\prod_{q=1}^{p_T} \prod_{r=1}^{p_S} \left[\frac{\left(T/(p_T p_S)\lambda_{qr}\right)^{n_{qr}}}{n_{qr}!} \exp\left\{-\frac{T}{p_T p_S}\lambda_{qr}\right\}\right].$$

Now, assume that intensity λ_{qr} also depends on explanatory variables $\mathbf{x}_i$ and denote it by λ_i. Moreover, we assume the following link function:

$$\lambda_i = \left(\mathbf{x}_i^T \boldsymbol{\alpha}\right)^2, i = 1, ..., n,$$

where $\mathbf{x}_i$ has p indicators of seasons and regions. Using as link function the

square of the linear predictor $\mathbf{x}_i^T\boldsymbol{\alpha}$ and the Jeffreys prior for the rate of a Poisson distribution, that is $\pi(\lambda_i) \propto \lambda_i^{-1/2}$, we can assume the flat prior

$$\pi(\boldsymbol{\alpha}) \propto \mathbf{1}, \text{ for } \boldsymbol{\alpha} \in \Lambda \subset \mathbb{R}^p,$$

in analogy to $\pi(\boldsymbol{\theta})$. Since the λ_i's are orthogonal to $\boldsymbol{\theta}$ then the joint prior distribution is also assumed in the independent form

$$\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\alpha}). \tag{8}$$

3 Likelihood and posterior distribution for the Poisson-GPD process

For a random sample $\mathbf{y}$ and a corresponding matrix $\mathbf{X}$ the likelihood (6) for the Poisson-GPD process factorizes into two parts. Applied to the regression setting considered here, and considering logarithms, these parts are $l_{Poisson}(\boldsymbol{\alpha})$ involving only the counting part and $l_{GPD}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ involving only the exceedances $\mathbf{y}$. They are respectively given by

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = l_{Poisson}(\boldsymbol{\alpha}) + l_{GPD}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where

$$l_{Poisson}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} 2\log(\mathbf{x}_i^T\boldsymbol{\alpha}) - \left(\frac{T}{p_T p_R}\mathbf{x}_i^T\boldsymbol{\alpha}\right)^2,$$

and

$$l_{GPD}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log f\left(y_i | \xi(\boldsymbol{\beta}^T\mathbf{x}_i), \nu(\boldsymbol{\gamma}^T x_i)\right),$$

with f according to density (2).

Combination of the above likelihood with the joint prior (8) by Bayes theorem

gives the posterior distribution. The kernel of posterior density is

$$\pi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}|\mathbf{y}) \propto \prod_{i=1}^{n} \frac{(\mathbf{x}_i^T\boldsymbol{\alpha})^2 \exp\left\{-\frac{T}{p_T p_R}(\mathbf{x}_i^T\boldsymbol{\alpha})^2\right\} \exp(\mathbf{x}_i^T\boldsymbol{\gamma} - \mathbf{x}_i^T\boldsymbol{\beta})}{(1 + y_i(\exp(\mathbf{x}_i^T\boldsymbol{\gamma}) - 1)\exp(\mathbf{x}_i^T\boldsymbol{\gamma} - \mathbf{x}_i^T\boldsymbol{\beta}))^{\exp(\mathbf{x}_i^T\boldsymbol{\gamma})/\left(\exp(\mathbf{x}_i^T\boldsymbol{\gamma})-1\right)}},$$

where the normalization constant does not an analytical form. The posterior support is given by the product space $\Theta \times \Lambda$, that can be described only numerically.

3.1 Approximation of posterior distribution

Approximation of posterior distribution $\pi(\boldsymbol{\theta},\boldsymbol{\alpha}|\mathbf{y})$ may be obtained using Markov Chain Monte Carlo methods (see Gamerman and Lopes (2006)). In particular, we used a Metropolis step nested in a Gibbs sampling algorithm. At iteration t of the chain we update in a random order all $3p$ parameters using as proposal distribution a univariate Normal density with mean given by the parameter at the previous iteration and variance fixed such that the acceptance rate for all parameters is between 30% and 40%.

Please note that in our applications $p \approx 35$ and the dimension of the space of parameters $\Theta \times \Lambda \subset \mathbb{R}^{3p}$ is sizable as $3p \approx 105$. Nevertheless, diagnostic plots (not shown here) seem to indicate that the chain has converged. We think that the important ideas to facilitate this are the proposed reparametrizations and corresponding link functions.

4 Application

We apply the Poisson-GPD model to two real data sets. The first application shows that extreme value analysis may be improved by accounting for

covariates. We use a data set also analyzed in Sisson et al. (2006) without covariates. The second application is a larger data set with meteorological measurements from Rio de Janeiro, Brazil. It shows the computational feasibility of the inference technique here proposed even for moderate to large data sets.

4.1 San Juan, Puerto Rico data

This data set relates to rainfall levels Z in San Juan, Puerto Rico, measured at Luis Muñoz Marin International Airport. Data have been analyzed in Sisson et al. (2006) without a regression approach, but using instead a Gumbel distribution and Generalized Extreme Value (GEV) model for maxima. They use all available annual maxima (from 1967 to 2001) with a Gumbel and GEV model in a Bayesian and maximum likelihood framework showing the superior predictive behavior of GEV over Gumbel.

We consider the Poisson-GPD model for rainfalls $Z > u$, where $u = 50\ mm$. In order to predict the rainfall of 17 September 1989 we use the $n = 94$ rainfall measurements above u showed in Figure 1. We use as covariate a season dummy variable. According to Rivera-Ramirez et al. (2002), San Juan has two seasons coded by $x_i = 0$ for dry season (January through April, June and July) and $x_i = 1$ for wet season (May, and August through December). The marginal posterior distribution of each parameter along with its normalized profile likelihood are shown in Figure 2.

It can be seen that the marginal posterior for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is likelihood dominated as its shape does not differ from that of the likelihood. Differences are appre-

ciable for $\boldsymbol{\gamma}$ as the shape ξ is near to its lower bound of $\xi = -0.5$ corresponding to $\mathbf{x}_i^T\boldsymbol{\gamma} = -\log 2$. This indicates that there exists an upper bound in rainfall levels exceedances Y. Also, observing Figure 2, the effect of the season is fairly significant for the regression parameters $\boldsymbol{\alpha}$ of process rate with the number of extreme events increasing in wet season. For the other Poisson-GPD model features there is not a clear effect.

The probability of September extreme event, conditionally on all extremes larger than $50mm$ and wet season is 2.8%. The same probability in the dry season is almost a half, indicating that season matters in predicting extreme rainfalls in San Juan. The posterior distributions of the mean return level L_K for wet season in $K = 50$ and $K = 100$ years are shown in Figure 3. We can see that the September 1989 rainstorm is in the 95% posterior credible intervals of L_{100}. Moreover, in wet season $\Pr(L > 200 \mid \mathbf{y}, \mathbf{x}, K = 50, season = wet) \approx$ 4.8%, while in $K = 100$ years it increases up to 80%. The latter probability decreases to only 35% for the dry season. Again, ignoring season effect may lead to a serious underestimation of these extreme events.

4.2 Rainfall in Rio de Janeiro

This data set consists of maximum rainfall levels cumulated in 15 minutes along a day observed during the 10 years from 1997 to 2007 ($n = 1221$) in 32 meteorological stations locations all around Rio de Janeiro city. These public domain data come from Fundação Instituto de Geotécnica do Municpio do Rio de Janeiro and are available at the web page `www.rio.rj.gov.br/alertario`. We consider as extremes all rainfall levels above threshold $u = 50mm$.

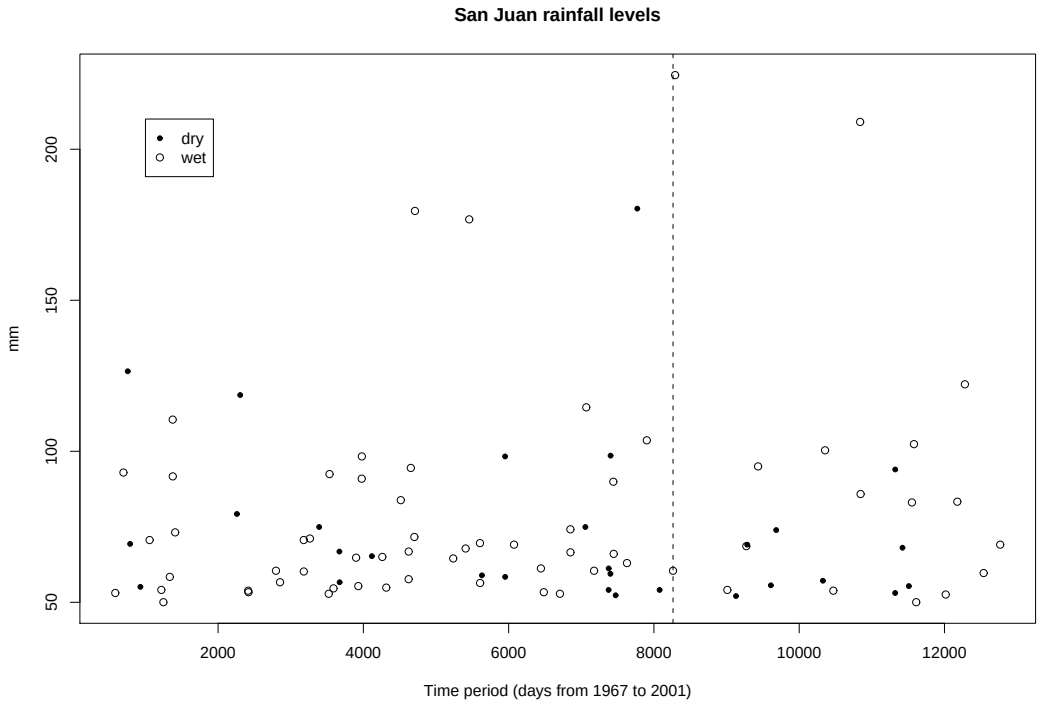

Fig. 1. *Rainfall levels in San Juan (Puerto Rico) measured at Luis Muñoz Marin International Airport.*

We want to show that in such large data set we can either evaluate the probability of high rainfall levels for a certain season / location at a relative lower computational cost. We consider as covariates the season (4 levels) and the meteorological station. These covariates enter as columns of dummy variables $\{0, 1\}$ in the $\mathbf{X}$ matrix. Thus $\mathbf{X}$ has 1221 rows and $3p = 105$ columns, where $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ have 35 elements each. This large amount of parameters is relative easy to manage with the algorithm explained in Section 3.1. Data are shown in Figure 4 where it seems that season and location are related to the underlying extremal process.

Posterior probability map of L_{100} for summer season (the most rainy season in Rio) is shown in Figure 5. It is clear that there are areas of Rio de Janeiro which have high probability of rain falls levels over 300 mm. These areas contain the

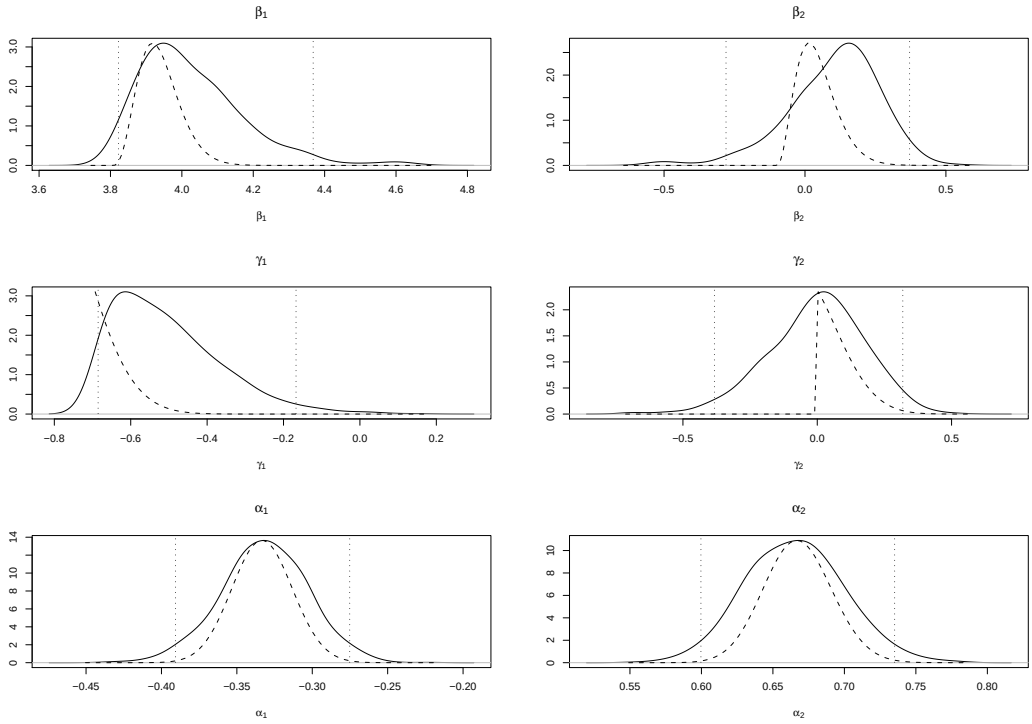

Fig. 2. *Posterior distribution for* $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}$ *parameters (solid) along with the corresponding profile likelihood (dashed). Vertical dotted lines represent 95% posterior credible intervals.*

meteorological stations that are mainly situated near Rio de Janeiro hills.

5 Remarks

We would like to conclude by remarking an undesirable effect of the application of the orthodox Jeffreys rule for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ parameters based on the $2p \times 2p$ information matrix (3). It can be noted that the $i_{\nu,\nu}$ element involves both ξ and ν, so the expression of elements $i_{\beta_s\beta_t}$ of the Fisher information when we consider the change $\nu_i = \nu(\mathbf{x}_i^T\boldsymbol{\beta})$ can be calculated as

$$i_{\beta_s\beta_t} = \sum_{i=1}^{n} x_{is}x_{it}\nu_i'^2 i_{\nu,\nu} = \sum_{i=1}^{n} x_{is}x_{it}\nu_i'^2\nu_i^{-2}(1+2\xi_i)^{-1},\ s,t \in (1,\ldots,p),$$

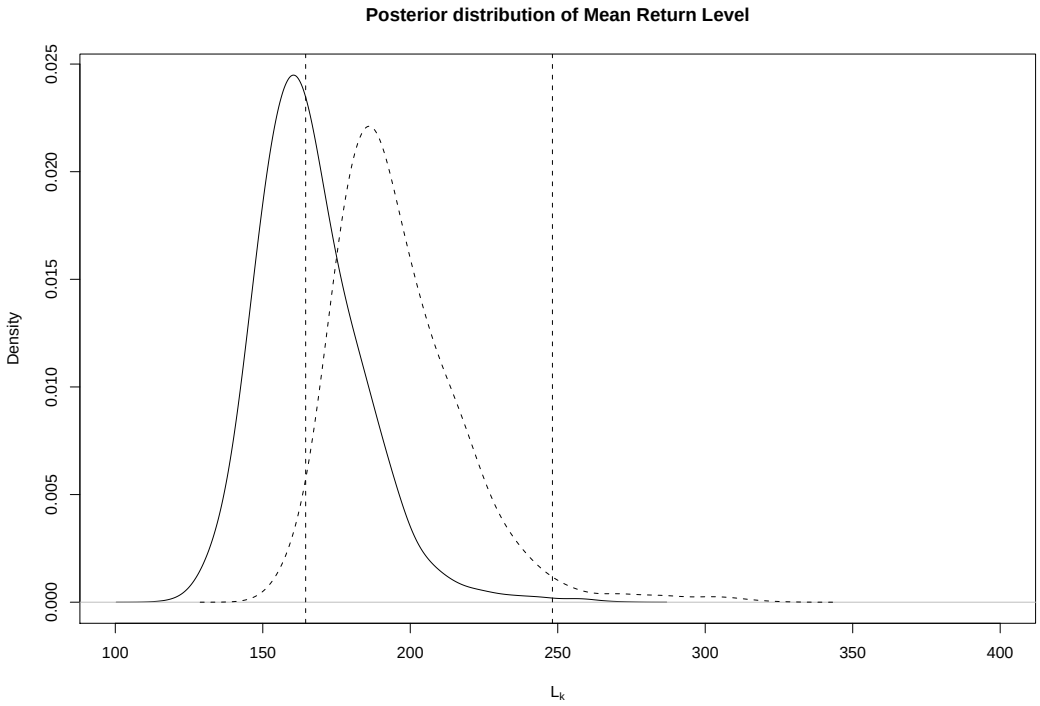

Fig. 3. *Posterior distribution for the K-year Mean Return Level, for $K = 50$ (solid) and $K = 100$ (dashed) years, for the wet season. Vertical dotted lines represent 95% posterior credible intervals for $K = 100$. In the observation period subsequent to that used to estimate this distribution, the largest rainstorm was about $220mm$.*

where ν_i' indicates the derivative of ν_i with respect to $\boldsymbol{\beta}$.

In this case parameter $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ will not be necessarily separated in the prior no matter which link functions is used. For example, using the link functions

$$\begin{cases} \nu_J^{-1}(\nu_i) \propto \int \nu_i^{-1} d\nu_i = \log(\nu_i) \\ \\ \xi_J^{-1}(\xi_i) \propto \int (1+\xi_i)^{-1}(1+2\xi_i)^{-1/2} d\xi_i = \arctan\sqrt{1+2\xi_i} \end{cases}$$

obtained by inverting the Jeffreys' prior in Castellanos and Cabras (2006), the marginal priors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are no longer flat.

With the approach here proposed we can model several extreme process fea-

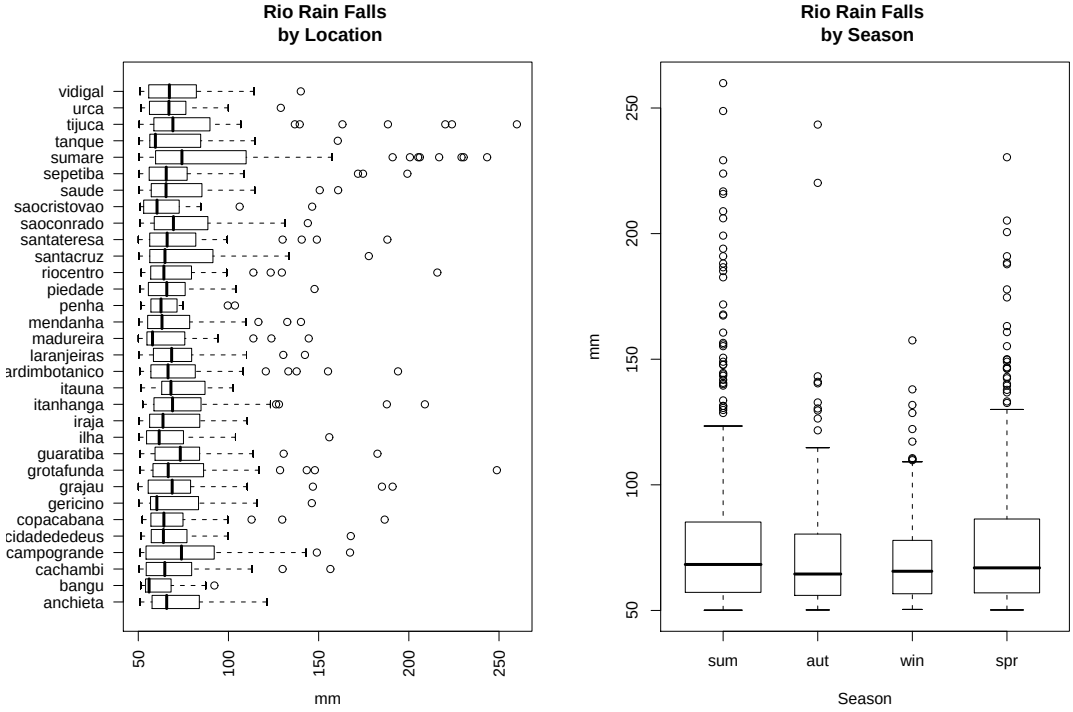

Fig. 4. *Rainfall levels in Rio de Janeiro grouped by location (left) and seasons (right). Boxplot width is proportional to the number of observations per group.*

tures by using a large set of available covariates at a relatively low computational cost. The choice of an appropriate threshold u has not been accounted explicitly in this work. In fact for San Juan data with $u = 40mm$ the probability of September extreme, given the wet season, decreases to 1% while for $u = 60$ it increases to 4.5%. These values are sensibly different from 2.8% obtained using $u = 50$.

Threshold changes also affects the effect of season on scale ν and form ξ. In fact, the posterior distribution for the effect of season on these parameters are less concentrated around 0 with $u = 40$ then with $u = 50$ or $u = 60$ and the effect of season on λ is robust with respect to u. The threshold selection is an open direction for further research because available proposals, as in Behrens et al. (2004) or in Cabras and Morales (2006), do not account for the regression

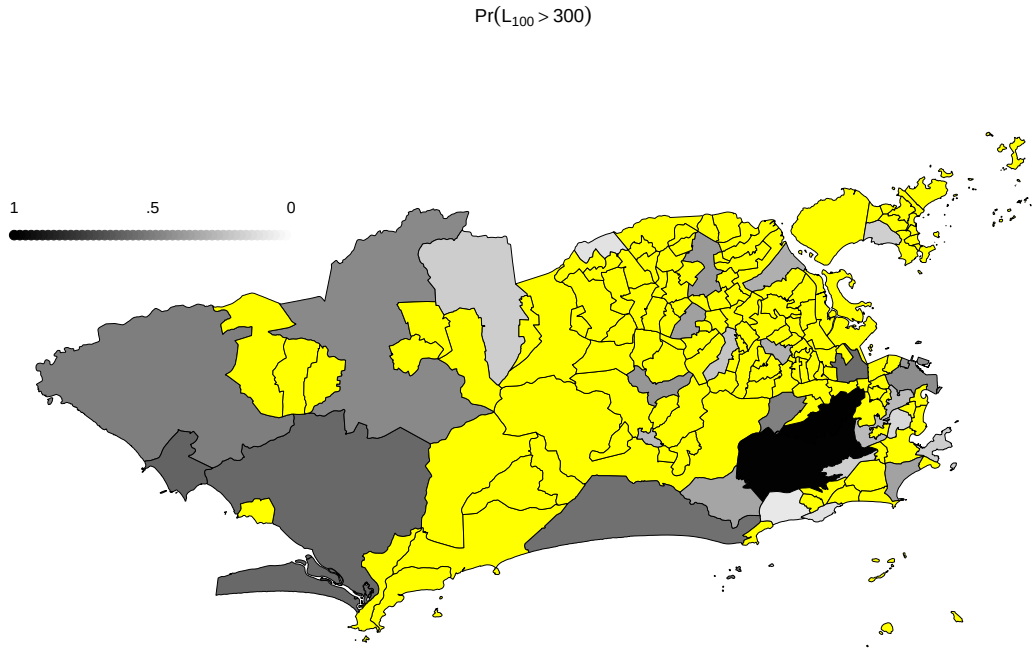

Fig. 5. *Posterior probability map of a mean return level greater than 300mm in 100 years in the administrative regions of Rio de Janeiro city. Each region is coloured according to its meteorological station. Yellow color indicates that no data were available.*

framework here discussed.

The analyses performed here illustrate the feasibility of the proposed methodology for a variety of practical settings. Appropriate models can be constructed and meaningful answeres can be drawn. Directions for further work include incorporation of threshold selection and of the central part of the distribution below the exceedances and incorporation of structured random effects in the regression framework to accommodate for additional unexplained heterogeneity.

References

Behrens, C. N., Lopes, H. F., Gamerman, D., 2004. Bayesian analysis of extreme events with threshold estimation. Statistical Modelling 4, 227–244.

Box, G., Tiao, G. C., 1992. Bayesian Inference in Statistical Analysis. Wiley Classics Library.

Cabras, S., Morales, J., 2006. Extreme value analysis within a parametric outlier detection framework. Applied Stochastic Models in Business and Industry 23(2), 157–164.

Castellanos, M., Cabras, S., 2006. A default bayesian pro cedure for the generalized pareto distribution. Journal of Statistical Planning and Inference 137 (2), 473–483.

Chavez-Demoulin, V., Davison, A., 2005. Generalized additive modelling of sample extremes. Applied Statistics 54, 207–222.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, New York.

Davison, A. C., Smith, R. L., 1990. Models for exceedance over high threshold. J. Roy. Stat. Soc. B 52 (3), 393–442.

Gamerman, D., Lopes, H. F., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. 2nd Edition. Chapman and Hall/CRC Press, Boca Raton.

Gnedenko, B., 1943. Sur la distribution limite du terme maximum d'une serie aleatoire. Ann. Mathematics 44, 423–453.

Leadbetter, M., Lindgren, G., Roorzén, H., 1983. Extremes and Related Properties of Random Sequences and Series. Springer-Verlag.

Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Statist. 3 (199-131).

Resnick, S. I., 1987. Extreme Value, Regular Variation, and Point Processes. Springer-Verlag, New York.

Rivera-Ramirez, H., Warner, G., Scatena, F., 2002. Prediction of master recession curves and baseflow recessions in the luquillo mountains of puerto rico. Journal of the American Water Resources Association 38, 693–704.

Sisson, S. A., Pericchi, L. R., Coles, S., 2006. A case for a reassessment of the risk of extreme hydrological hazards in the caribbean. Journal of Stochastic Environmental Research and Risk Assessment 20, 296–306.

Smith, R., 2003. Extreme Values in Finance, Telecommunications and the Environment. Chapman and Hall/CRC Press,, Ch. Chapter 1. Statistics of extremes, with applications in environment, insurance and finance.

Smith, R. L., 1984. Statistical extremes and Applications. Reidel, Dordrecht, pp. 621–638.