

Latent process models for temporal extremes with an application to rainfall data

*Modelli a processi latenti per valori estremi temporali con
un'applicazione a dati di pioggia*

Paola Bortot and Carlo Gaetan

Abstract The paper reviews a hierarchical modelling procedure for extreme values of stationary series proposed by Bortot and Gaetan (2013). A modification of the model to allow for different types of marginal tail behaviour is presented. The two models are applied to the analysis of extreme hourly and daily rainfalls and their performances compared.

Abstract *Nel lavoro viene presentata una modifica di un modello gerarchico per i valori estremi di serie stazionarie sviluppato da Bortot e Gaetan (2013) che permette di includere vari tipi di code della distribuzione univariata marginale. I due modelli vengono applicati all'analisi di estremi di pioggia orari e giornalieri e le loro capacità di adattamento poste a confronto.*

Key words: Extremal dependence, Generalized Pareto distribution, Hierarchical models, Rainfall data.

1 Introduction

Statistical analysis of extreme values has played an important role in environmental sciences and related fields (for a review, see Jonathan and Ewans, 2013). A widespread approach for inferring the extremal behaviour of a phenomenon of interest is the so-called Peaks over Threshold (POT) approach whereby the recorded exceedances over a sufficiently high threshold are modelled using the Generalized Pareto distribution (GPD) (Pickands, 1975). This class of procedures has an asymp-

Paola Bortot

Università di Bologna, Dipartimento di Scienze Statistiche, Via Belle Arti, 41, e-mail: paola.bortot@unibo.it

Carlo Gaetan

Università Ca' Foscari – Venezia, Dipartimento di Scienze Ambientali, Informatica e Statistica, San Giobbe, Cannaregio 873, e-mail: gaetan@unive.it

totic justification since the GPD arises as the limiting distribution of an exceedance as the threshold tends to the upper endpoint of the underlying distribution. However, in the application of the POT approach care must be taken of the temporal dependence, as this might persist even at very high levels, causing the exceedances to occur in clusters. A simple solution is declustering, i.e. filtering the exceedances so that the resulting sequence is made of approximately independent observations (Ferro and Segers, 2003). A downside of filtering is a loss of information that leads to reduced estimation precision. Another possibility, that overcomes the waste of data, but at the cost of a greater modelling effort, is to infer the joint distribution of all the exceedances. This approach has the additional benefit of allowing inference of general functionals of exceedances, for example the aggregate exceedance within an extremal event. Various attempts to model the joint distribution of exceedances can be found in the literature. One example is the method of Smith *et al.* (1997) which assumes Markovianity for successive exceedances and uses multivariate extreme values distributions to model the chain transitions above the threshold. Another example can be found in Reich *et al.* (2013) which consists of modelling the series of exceedances as a realization of a censored max-stable process (de Haan, 1984). Alternatively, Bortot and Gaetan (2013) suggest a hierarchical procedure based on a latent process to capture serial dependence at high levels. Their proposal allows for the strength of extremal dependence to decrease when moving further into the tail of the series, a feature which is exhibited by many environmental time series. Two drawbacks of this procedure are that it is limited to long-tailed margins and there is no separation between model parameters controlling marginal and temporal aspects, respectively, which, in some cases, constrains the range of dependence that the model can capture. In the next section the model of Bortot and Gaetan (2013) is briefly reviewed and a modification that aims at overcoming the aforementioned limitations is proposed. A comparison of the two models is carried out in Section 3 through an application to rainfall data.

2 Hierarchical models for stationary series

A way of capturing temporal dependence in stationary series is through hierarchical structures. In extreme value studies this is often implemented by letting the parameters of a standard extreme value distribution be driven by a latent stochastic process, and assuming conditional independence of extremes given the latent process (see, for example, Gaetan e Grigoletto, 2004 and Huerta and Sansó, 2007). Typically, however, the resulting marginal distributions are no longer of an extreme value type and the induced temporal dependence is weak, in the sense that it decreases as the threshold increases eventually converging to independence at asymptotically high levels. The hierarchical model proposed by Bortot and Gaetan (2013) aims to overcome both limitations.

2.1 A hierarchical model for exceedances

The basic idea of the modelling procedure of Bortot and Gaetan (2013) is using a representation of the GPD as a mixture of an Exponential distribution with Gamma distributed rate parameter and inducing serial dependence among exceedances by specifying a latent process with Gamma margins for the rate parameter. More precisely, classical extreme value results (Pickands, 1975) justify modelling the distribution of the exceedances of a random variable X over a high threshold u , i.e.

$$F_u(x) = P(X \leq x + u | X > u)$$

with the GPD whose density function is given by

$$f(x; \xi, \sigma) = \frac{1}{\sigma} \left(1 + \xi \frac{x}{\sigma}\right)_+^{-(1/\xi+1)}, \quad x \geq 0, \quad (1)$$

where $(x)_+ = \max(0, x)$, ξ is a real shape parameter and σ is a positive scale parameter. With the reparametrization $\alpha = 1/\xi$ and $\beta = \sigma/\xi$, for $\alpha > 0$, the density of the GPD can be represented as

$$f(y; \alpha, \beta) = \int_0^\infty \lambda e^{-y\lambda} g(\lambda; \alpha, \beta) d\lambda, \quad (2)$$

where

$$g(\lambda; \alpha, \beta) = [\beta^\alpha / \Gamma(\alpha)] \lambda^{\alpha-1} e^{-\beta\lambda}$$

is the density of a Gamma(α, β) distribution (Reiss and Thomas, 2007). In other words, when $\alpha > 0$, which corresponds to a heavy right tail, the GPD can be characterized as a Gamma mixture of Exponential distributions.

Starting from representation (2), if $\{X_t\}$ is a stationary random sequence, its behaviour above u is modelled through a hierarchical structure which, at the first stage, assumes that conditionally on a latent process Λ_t , the X_t 's are independent with

$$X_t - u | (\Lambda_t, X_t > u) \sim \text{Exp}(\Lambda_t), \quad (3)$$

where $\text{Exp}(a)$ denotes the Exponential distribution with mean $1/a$. To model the occurrence of $X_t > u$, the authors propose

$$\text{pr}(X_t > u | \Lambda_t) = \exp(-\kappa \Lambda_t),$$

where κ is a positive constant controlling the percentage of crossings of the threshold u . At the second stage, temporal dependence above the threshold u is incorporated by choosing a stochastic process with Gamma(α, β) margins for $\{\Lambda_t\}$. Two choices are considered: the Warren process (WP) introduced by Warren (1992) and the Gaver and Lewis process (GLP) developed by Gaver and Lewis (1980) (see also Walker, 2000, for an autoregressive representation of the process). The two specifications of $\{\Lambda_t\}$ are first-order Markov and have the same autocorrelation function

which depends on a parameter $\rho \in (0, 1)$. However, when embedded in the hierarchical structure they lead to different forms of extremal serial dependence. Broadly speaking, under WP, the strength of dependence of the exceedances of u^* , with $u^* > u$, decreases as u^* increases, eventually reaching independence. Under GLP, the dependence also gets weaker as u^* increases, but in the limit the exceedances still exhibit dependence and occur in clusters. Thus, under WP there is asymptotic independence of exceedances, while under GLP exceedances are asymptotically dependent.

Combining the first and second stages, marginally with respect to $\{\Lambda_t\}$, the distribution of $X_t - u | X_t > u$ is still GPD and

$$\text{pr}(X_t > u) = \left(\frac{\beta}{\beta + \kappa} \right)^\alpha.$$

Inference for this hierarchical model can be carried out using a censored likelihood. As likelihood contributions are computationally intractable, a pairwise approximation of the full likelihood is used.

2.2 An alternative hierarchical formulation

In Bortot and Gaetan (2013) the performance of the model of Section 2.1 is assessed through a simulation study and an application to real data sets. In both cases, it shows a good flexibility in capturing different types of tail behaviour. However, some drawbacks are also present. Perhaps the most serious is the constraint $\alpha > 0$, which limits the application of the method to heavy tailed data. Many series arising in the financial and insurance areas fall within this class, but in the environmental context, this constraint could be rather severe as many phenomena, such as temperature series, are known to be lighter tailed. A less evident limitation is related to the double role of α as a parameter of the marginal GPD of the exceedances and of the latent process controlling temporal dependence. In particular, under both specifications of $\{\Lambda_t\}$, larger values of α , which correspond to lighter tails, imply weaker dependence of exceedances. Both WP and GLP have an additional dependence parameter ρ ; for WP, ρ can compensate for this effect, but for GLP this is only partly true. It can be shown that, regardless of the value of ρ , for sufficiently large α , the degree asymptotic dependence allowed by GLP is rather low.

A way to overcome these limitations, while working within the same hierarchical modelling framework, exploits a copula type of technique. The basic idea is to transform marginally the process defined in Section 2.1 so to have marginal GPD with unconstrained shape parameter. In greater detail, in the definition of the hierarchical model of Section 2.1, we set α and β equal to 1. Let $\{X_t^*\}$ be the series obtained from the model under this setting. An integral transformation is applied marginally to each of the X_t^* , with $X_t^* > u$ to derive

$$Y_t = \frac{\sigma}{\xi} \left\{ \left(1 + \frac{X_t^* - u}{\kappa + 1} \right)^\xi - 1 \right\}$$

having GPD with parameters $\xi \in R$ and $\sigma > 0$ as in (1). Finally, by letting $X_t = Y_t + u$ a new hierarchical model for the series of exceedances of u is obtained. Inference for this model can again be carried out through a pairwise likelihood).

The modification proposed in this section is characterized by a separation of parameters governing marginal behaviour, ξ and σ , and parameters controlling temporal dependence, ρ . The advantage of this separation is that the GPD shape parameter ξ is no longer restricted to be positive and different types of marginal tail behaviour can be modelled. In addition, the model can be easily adapted to account for covariates through the inclusion of regression terms in the GPD parameters. This extension was precluded in the previous version as any transformation of α would affect both marginal and dependence aspects. On the other hand, ρ remains the only dependence parameter, so that a greater flexibility in marginal behaviour might be offset by an increased rigidity in the temporal structure. The choice between the two modelling frameworks is thus not so clear cut and investigation of the relative merits of each of the two models is required.

3 Application to rainfall extremes

When analyzing rainfalls, accounting for serial dependence at high levels is fundamental as functionals of extreme values are of interest, such as the average length of a storm or the aggregate exceedance during a storm. Typically, a decreasing strength of extremal dependence with threshold is observed for rainfall data which often give evidence of asymptotic independence. In addition, rainfall recordings are known to be heavy tailed, so this is an area of application where the two versions of the hierarchical model can be compared and their relative performances assessed.

The data analyzed are series of hourly and daily rainfalls in Camborne, west Cornwall, England. The recordings, obtained from the British Atmospheric Data Centre, cover the period 01/01/1980–31/12/2012, but for the whole of September 1994 both hourly and daily measurements are missing. As the model of Section 2.1 cannot be easily adjusted to capture non-stationarity, only the summer season from June to September was analyzed. This season, within which the data seem to be stationary, produces the most extreme events. The results of the study can be summarized as follows. As might be expected there is stronger dependence at extreme levels for hourly than daily rainfalls. However, both series support the hypothesis of independence in the limit. Various diagnostics have been carried out to assess performance of the two models at different thresholds u^* , above the model threshold u . These focussed on probabilities of consecutive uncrossing of u^* , distribution of aggregate exceedances and average length of extremal events. It was found that for daily data, the differences between the two versions of the model, with both WP and GLP specifications, are negligible and that they both provide a good fit. On the other

hand, for hourly data the new model, under WP, largely outperforms the previous one for any choice of the latent process and describes well the extremal behaviour of the series and the way this changes when raising the threshold.

References

- Bortot, P., Gaetan, C.: A latent process model for temporal extremes, *Scandinavian Journal of Statistics*, to appear.
- de Haan, L.: A spectral representation for max-stable processes, *The Annals of Probability*, **12**, 1194–1204 (1984).
- Ferro, C. A.T., Segers, J.: Inference for clusters of extreme values, *Journal of the Royal Statistical Society, Series B*, **65**, 545–556 (2003).
- Gaetan, G., Grigoletto, M.: Smoothing sample extremes with dynamic models, *Extremes*, **7**, 221–236 (2004).
- Gaver, D., Lewis, P.: First-order autoregressive Gamma sequences and point processes, *Advances in Applied Probability*, **12**, 727–745 (1980).
- Huerta, G., Sansò, B.: Time-varying models for extreme values, *Environmental and Ecological Statistics*, **14**, 285–299 (2007).
- Jonathan, P., Ewans, K.: Statistical modelling of extreme ocean environments for marine design: A review, *Ocean Engineering*, **62**, 91–109 (2013).
- Pickands, J.: Statistical inference using extreme order statistics, *Annals of Statistics*, **3**, 119–131 (1975).
- Reich, B.J., Shaby, B.A., Cooley, D.: A hierarchical model for serially-dependent extremes: a study of heat waves in the Western US, *Journal of Agricultural, Biological, and Environmental Statistics*, **18**, 1–17 (2013).
- Smith, R., Tawn, J.A., Coles, S.: Markov chain models for threshold exceedances, *Biometrika*, **84**, 249–268 (1997).
- Walker, S.: A note on the innovation distribution of a Gamma distributed autoregressive process, *Scandinavian Journal of Statistics*, **27**, 575–576 (2000).
- Warren, D.: A multivariate Gamma distribution arising from a Markov model, *Stochastic Hydrology and Hydraulics*, **6**, 183–190 (1992).