

Hierarchical modeling of extremes for regional data

Modelli gerarchici per valori estremi regionali

Daniela Cocchi, Michele Scagliarini¹

Dipartimento di Scienze Statistiche

Università di Bologna

cocchi@stat.unibo.it , scagliar@stat.unibo.it

Riassunto: Nel lavoro si propone una modellazione statistica per l'analisi di valori estremi rilevati su aree geografiche diversificate. Si sviluppa una struttura gerarchica stratificata che consente di verificare la presenza di differenze tra gruppi di dati.

Keywords: environmental risk, extreme values, generalized extreme value distributions, hierarchical models, sub-regional models.

1. Introduction

Many real situations require to estimate the extremal behavior of an environmental process over a specified geographical region. A typical application is the study of extremal rainfall behavior within a catchment region (Coles and Tawn, 1996; Cocchi and Scagliarini, 2001), useful for the evaluation and prevention of hydrological risks. Researchers are often interested in a model holding for the whole region, though in practice only repeated pointwise rainfall measurements from a network of sites over the region are available. This paper proposes a modeling strategy for the extremal behavior of the aggregated process, starting from the pointwise data in the presence of auxiliary information at the site level. Situations occur typically where, besides the pointwise time series data, also cross section information about the different sites is available. The whole area is not considered as homogeneous and the existence of different sub-areas is hypothesized. This structure of information suggests a hierarchical formulation of the statistical model.

2. The models

At the single site level, an extreme value analysis can be performed by fitting the maxima, recorded under a fixed time interval, to the GEV distribution

$$G(x; \mu_i, \sigma_i, \gamma_i) = \exp \left\{ - \left[1 + \gamma_i \left(\frac{x - \mu_i}{\sigma_i} \right) \right]^{-\frac{1}{\gamma_i}} \right\}, \quad i = 1, \dots, s \quad (1)$$

where i denotes the monitoring site.

Stratified hierarchical modeling of extremes

Covariates available at the site level can be employed to estimate a model for the whole area. The following model is built for situations where there are several regions and the aim is to test if the parameters differ among regions. The definition of several

¹Research supported by a 2000 grant (project n.26, sector: Economics and Statistics) for researches of national interest by the Italian Ministry of the University and Scientific and Technological Research.

regions is allowed by the knowledge of a categorical covariate able to classify the sites in different groups. In the presence of r sub-regions, r different $s_k \times p$ matrices of covariates $\mathbf{Y}_k$, ($k = 1, 2, \dots, r$) are available, where s_k is the number of sites for each region and p is the number of covariates. Each matrix $\mathbf{Y}_k$ is represented as: $\mathbf{Y}_k = [y_{ij}]_k$, where $i = 1, 2, \dots, s_k$ and $j = 1, 2, \dots, p$. We assume that the single site parameters in each sub-region are functions of the covariates according to a second level model. A functional dependence, suitable for rainfall data in the case of a regional model without strata (Coles and Tawn, 1996), is

$$\begin{aligned}\boldsymbol{\mu}_k &= \mathbf{Y}_k \boldsymbol{\mu}_k^* \\ \boldsymbol{\sigma}_k &= \exp(\mathbf{Y}_k \boldsymbol{\sigma}_k^*), \quad k = 1, \dots, r \\ \gamma_k &= \mathbf{e}_{s_k} \gamma_k^*\end{aligned}\quad (2)$$

where $\boldsymbol{\mu}_k^*$ e $\boldsymbol{\sigma}_k^*$ are p -dimensional vectors of hyperparameters, γ_k^* is a scalar and $\mathbf{e}_{s_k}$ is a s_k -dimensional vector of units. The elements of the s_k -dimensional vectors $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ are $\mu_{ik} = [\mathbf{y}'_i]_k \boldsymbol{\mu}_k^*$ and $\sigma_{ik} = \exp([\mathbf{y}'_i]_k \boldsymbol{\sigma}_k^*)$ respectively, where $[\mathbf{y}'_i]_k$ denotes the row i of the matrix $\mathbf{Y}_k$. The likelihood of the complete model, under the assumption of independence among sub-regions, can be written as

$$L(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*, \boldsymbol{\gamma}^*; x, y) = \prod_{k=1}^r \prod_{i=1}^{s_k} L_{ik}([\mathbf{y}'_i]_k \boldsymbol{\mu}_k^*, \exp([\mathbf{y}'_i]_k \boldsymbol{\sigma}_k^*), \gamma_k^*; x, y) \quad (3)$$

where $\boldsymbol{\mu}^*$, $\boldsymbol{\sigma}^*$ and $\boldsymbol{\gamma}^*$ are of the form

$$\begin{aligned}\boldsymbol{\mu}^* &= [\boldsymbol{\mu}_1^{*'} \quad \dots \quad \boldsymbol{\mu}_r^{*'}] \\ \boldsymbol{\sigma}^* &= [\boldsymbol{\sigma}_1^{*'} \quad \dots \quad \boldsymbol{\sigma}_r^{*'}] \\ \boldsymbol{\gamma}^* &= [\gamma_1^* \quad \dots \quad \gamma_r^*]\end{aligned}$$

and each L_{ik} is the likelihood of a GEV distribution. The log-likelihood can be written as

$$\begin{aligned}l(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*, \boldsymbol{\gamma}^*; x, y) &= \sum_{k=1}^r \sum_{j=1}^{s_k} \sum_{u=1}^{n_{ik}} \left\{ -[\mathbf{y}'_i]_k \boldsymbol{\sigma}_k^* - \left(1 + \frac{1}{\gamma_k^*}\right) \log \left[1 + \gamma_k^* \left(\frac{x_{uik} - [\mathbf{y}'_i]_k \boldsymbol{\mu}_k^*}{\exp([\mathbf{y}'_i]_k \boldsymbol{\sigma}_k^*)} \right) \right] \right. \\ &\quad \left. - \left[1 + \gamma_k^* \left(\frac{x_{uik} - [\mathbf{y}'_i]_k \boldsymbol{\mu}_k^*}{\exp([\mathbf{y}'_i]_k \boldsymbol{\sigma}_k^*)} \right) \right]^{-\frac{1}{\gamma_k^*}} \right\} \quad (4)\end{aligned}$$

where x_{uik} is observation u , in site i , of sub-region k ($u = 1, 2, \dots, n_{ik}$; $i = 1, 2, \dots, s_k$; $k = 1, 2, \dots, r$). The model has three different data components: the regions, the sites within regions, the observations about extreme behavior within sites; the number of parameters to estimate is $r \times (2 \times p + 1)$. Two levels of hierarchy appear in this model: at the lowest level there is the time series information about rainfall for each site, at the highest level one has cross-section information typical of the site.

A multi-sites model for the whole region

The case of one region simplifies the model structure. In this case a unique matrix of covariates $\mathbf{Y}_{s \times p} = [\mathbf{y}'_i]$ is available, where each $\mathbf{y}'_i$ ($i = 1, 2, \dots, s$) is the p -dimensional row vector of covariates for the site i . In this case, the functional dependence on the

covariates expressed by (2) simplifies in

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{Y}\boldsymbol{\mu}^* \\ \boldsymbol{\sigma} &= \exp(\mathbf{Y}\boldsymbol{\sigma}^*) \\ \gamma &= \mathbf{e}_s\gamma^*\end{aligned}\tag{5}$$

where $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}^*$ are p -dimensional vectors of hyperparameters common to the whole region, γ^* is a scalar and $\mathbf{e}_s$ is a s -dimensional vector of ones. The elements of the s -dimensional vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are $\mu_i = \mathbf{y}'_i\boldsymbol{\mu}^*$ and $\sigma_i = \exp(\mathbf{y}'_i\boldsymbol{\sigma}^*)$ respectively. The likelihood of the regional model is a particular case of (3) with $r = 1$ and can be written as

$$L(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*, \gamma^*; x, y) = \prod_{i=1}^s L_i(\mathbf{y}'_i\boldsymbol{\mu}^*, \exp(\mathbf{y}'_i\boldsymbol{\sigma}^*), \gamma^*; x, y)\tag{6}$$

while the log-likelihood is:

$$\begin{aligned}l(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*, \gamma^*; x, y) &= \sum_{i=1}^s \sum_{u=1}^{n_i} \left\{ -\mathbf{y}'_i\boldsymbol{\sigma}^* - \left(1 + \frac{1}{\gamma^*}\right) \right. \\ &\quad \left. \log \left[1 + \gamma^* \left(\frac{x_{ui} - \mathbf{y}'_i\boldsymbol{\mu}^*}{\exp(\mathbf{y}'_i\boldsymbol{\sigma}^*)} \right) \right] - \left[1 + \gamma^* \left(\frac{x_{ui} - \mathbf{y}'_i\boldsymbol{\mu}^*}{\exp(\mathbf{y}'_i\boldsymbol{\sigma}^*)} \right) \right]^{-\frac{1}{\gamma^*}} \right\}\end{aligned}\tag{7}$$

where x_{ui} is the observation u at site i and n_i is the number of observations in the site i . In this case, the number of parameters to estimate reduces to $2 \times p + 1$.

3. Testing differences among strata

In the stratified situation of the previous section, it is crucial to check whether extremes actually behave differently among sub-regions. The r vectors of hyperparameters should be, in this case, different among sub-regions. In the opposite situation, the r sub-regions ought to be considered as an area where the behavior of extremes is homogenous. In this case the functional dependence actually has the form (5) with $i = 1, 2, \dots, s_t$, where s_t is the sum of the number of sites across the r subregions: $s_t = \sum_{k=1}^r s_k$. The test of equality of the r sets of parameters can be performed by stating the two alternative hypotheses

$$\begin{aligned}H_0 &: \boldsymbol{\mu}_k^* = \boldsymbol{\mu}^*; \boldsymbol{\sigma}_k^* = \boldsymbol{\sigma}^*; \gamma_k^* = \gamma^* \\ H_1 &: \boldsymbol{\mu}_k^* \neq \boldsymbol{\mu}^*; \boldsymbol{\sigma}_k^* \neq \boldsymbol{\sigma}^*; \gamma_k^* \neq \gamma^*\end{aligned}$$

Under H_0 , the dimension of the parameter space is $2 \times p + 1$, while under H_1 it is $r \times (2 \times p + 1)$. The likelihood ratio test is defined as

$$\lambda(X) = \frac{L(H_0)}{L(H_1)} = \frac{\prod_{i=1}^{s_t} L_i(\mathbf{y}'_i\boldsymbol{\mu}^*, \exp(\mathbf{y}'_i\boldsymbol{\sigma}^*), \gamma^*; x, y)}{\prod_{k=1}^r \prod_{i=1}^{s_k} L_{jk}([\mathbf{y}'_i]_k \boldsymbol{\mu}_k^*, \exp([\mathbf{y}'_i]_k \boldsymbol{\sigma}_k^*), \gamma_k^*; x, y)}$$

and, under H_0 , the following asymptotic result holds:

$$-2 \ln [\lambda(X)] \longrightarrow \chi_{2 \times p + 1}^2.$$

The rejection of H_0 indicates significant differences in the behavior of extremes among sub-regions.

4. Empirical results and further developments

For our empirical analysis we consider the Reno River catchment (Italy), where daily data on rainfall accumulation have been collected in 18 monitoring sites and 458 monthly maxima are available for each site (the covered time interval goes from 1951 to 1989). The sites are assigned to two different subregions according to their altitude, setting the critical value at 450 meters: 8 sites fall into the highest stratum and 10 into the lowest one.

After defining the monthly rainfall as the sum of daily rainfall values for each month, the average monthly rainfall over the whole period considered, AMR , is used as a site covariate. The components of the alternative parameterizations are

$$\mathbf{y}'_i = \begin{bmatrix} 1 & AMR_j \end{bmatrix},$$

therefore

$$\begin{aligned} \boldsymbol{\mu}_k^{*/I} &= \begin{bmatrix} \mu_{0k}^* & \mu_{1k}^* \end{bmatrix}_{k=1,2} \\ \boldsymbol{\sigma}_k^{*/I} &= \begin{bmatrix} \sigma_{0k}^* & \sigma_{1k}^* \end{bmatrix}_{k=1,2} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\mu}^{*/I} &= \begin{bmatrix} \mu_0^* & \mu_1^* \end{bmatrix} \\ \boldsymbol{\sigma}^{*/I} &= \begin{bmatrix} \sigma_0^* & \sigma_1^* \end{bmatrix}. \end{aligned}$$

The fit of monthly maxima to the model has been verified by means of the usual diagnostic tools (see Cocchi and Scagliarini, 2001) and can be considered good. The log-likelihood under the null hypothesis of a unique region is $\ln(L(H_0)) = -73156.085$, while the log-likelihood under the alternative hypothesis of two subregions is $\ln(L(H_1)) = -73149.964$. The test statistics is $-2 \ln[\lambda(X)] = 12.242$ and $\Pr(\chi_{2 \times p+1}^2 \geq -2 \ln[\lambda(X)]) = 0.269$. The null hypothesis that a stratification is not relevant is accepted and the behavior of extreme rainfall can be considered as homogeneous within the whole catchment. The developments of this work consist in testing the model for a wider region where different river catchments can be considered as strata. A simulation study for checking the accuracy of the test approximation for moderate sample dimensions will be a parallel step.

References

- Coles S.G., Tawn J.A. (1996) Modelling Extremes of the Areal Rainfall Process, *Journal of the Royal Statistical Society B*, **58**(2) 329-347.
- Cocchi D., Scagliarini M. (2001) "Extreme Value Analysis Assisted by a Model for Spatial Variability: an application to rainfall data" Atti del convegno SCO2001 Modelli Complessi e Metodi Computazionali Intensivi per la Stima e Previsione, Bressanone 24-26 Settembre 2001, CLEUP EDITORE, Padova 223-228.