

A multivariate extreme value model for short term ozone concentrations forecasting.

Un modello multivariato per valori estremi per la previsione a breve termine della concentrazione di ozono.

Francesco Pauli

Dipartimento di Scienze Economiche e Statistiche
Università degli Studi di Trieste
francesco.pauli@econ.univ.trieste.it

Riassunto: Obiettivo di questa analisi è di costruire uno strumento previsivo a breve termine per la concentrazione di ozono a livello del terreno. In particolare, vogliamo esplorare le capacità di previsione di un modello bivariato per valori estremi recentemente proposto in Coles and Pauli (2001), che riteniamo particolarmente versatile per la struttura di dipendenza estrema. Una volta adattato il modello a coppie di osservazioni in giorni consecutivi utilizziamo i risultati per prevedere la concentrazione al giorno t condizionatamente alla concentrazione osservata il giorno precedente.

Keywords: Extreme values; Asymptotic dependence; Ozone concentrations

1. Introduction

Due to its bad sanitary and economic effects, high concentration of ozone at ground level is considered in air quality standards by all legislation. Italian law, in particular, fixes a level for the daily maximum of ozone which should not be exceeded for more than three consecutive days. Within this framework, we will explore the possibility of forecasting ozone concentration one day ahead by modeling concentrations in consecutive days using a multivariate extreme value distribution. We focus on the issue of extremal dependence since we believe that, being extreme ozone concentrations the event we are especially interested in, the extremal dependence structure is of overwhelming importance. It is well known that a fundamental issue in extreme dependence analysis is to determine whether variables are asymptotic dependent (ASD) or independent (ASI) (see Ledford and Tawn (1997)). Extreme value distributions are classified according to the value of $\chi = \lim_{u \rightarrow 1} P(Y > F_Y^{-1}(u) | X > F_X^{-1}(u))$ and $\bar{\chi}$ (the last is a transformation of the more commonly used coefficient of tail dependence defined in Ledford and Tawn (1997), in fact $\bar{\chi} = 2 \times CTD - 1$, we use $\bar{\chi}$ to ease comparison with non parametric estimates). The value of χ is 0 for ASI variables, positive for ASD variables; the coefficient $\bar{\chi}$ is 0 for independent variables, 1 for ASD variables. We use the model recently proposed in Coles and Pauli (2001), which is either ASD and ASI as the parameter varies. The model is defined by

$$P(X > x, Y > y) = (S(x)^{1-k} + S(y)^{1-k} - 1)^{-\frac{p}{k-1}} (S(x)S(y))^{1-p} \quad (1)$$

where $p \in [0, 1]$, $k > 1$ and $S(\cdot)$ is the marginal survivor function. The $\bar{\chi}$ for this distribution is $p/(2-p)$ and $\chi = 2^{1/(1-k)}$ if $p = 1$ and 0 otherwise. Model (1) has symmetric dependence structure, to assess whether this is reasonable for our dataset we

also considered an asymmetric model from Coles and Pauli (2001) given by

$$P(X > x, Y > y) = (S(x)^{p(1-k)} + S(y)^{q(1-k)} - 1)^{-\frac{1}{k-1}} S(x)^{1-p} S(y)^{1-q} \quad (2)$$

where $p, q \in [0, 1]$, $k > 1$ and $S(\cdot)$ is the marginal survivor function. The $\bar{\chi}$ for this distribution is $\min\{p, q\}$ and $\chi = 2^{1/(1-k)}$ if $p = q = 1$ and 0 otherwise. Note that model (2) is not a generalization of (1), the main advantages of model (2) with respect to (1) are the asymmetry in the dependence structure and its easy extension to more than two variables.

The two models have been fitted to data recorded in Padua where Y represents ozone concentrations in t and X is ozone concentrations in $t - 1$.

In order to check the results of this model, we also used a non parametric estimation technique proposed by Coles et al. (2000). The authors defined two functions $\chi(u)$ and $\bar{\chi}(u)$ which can be estimated non parametrically and whose limits, as u tends to the final point of the distribution, are equal to χ and $\bar{\chi}$, respectively.

2. Data

We examined hourly observations of ozone concentrations recorded from March, 1st 1994 to October, 31st 1998 in Padua city center.

Ozone is a secondary pollutant and comes as the result of a chemical reaction in the atmosphere driven by solar radiation. High ozone concentrations are observed only during high temperature periods. We discard winter periods, when ozone concentration is stable at low level, and model periods from March, 1st to October, 31st.

The first problem to deal with is that of missing data. The object of our analysis is to model daily maxima of ozone concentration. If we examine the hourly concentration of ozone we easily see that there is a daily cycle, where the high values correspond to high temperature daytime. This means that, if observations of a day are partially missing but the missing hours occur during the night, we can be quite confident that the maximum was, in fact, observed. Following this idea we estimate the probability of the maximum occurring in each hour of the day and compute for each partially missing day the probability of the maximum being unobserved. Finally, we retain those values for which the probability of the maximum being actually observed is higher than 0.8 (figure 1).

By looking figure 1 we can spot a seasonal structure with a period of one year. This is not surprising, in fact we already pointed out that high temperature is a condition for ozone production and so ozone concentration broadly follows the pattern of temperature. We examined the seasonal pattern by assuming an extreme value model for the daily maxima (Y_t): $P(Y_i \leq y) = \exp \left\{ - \left(1 + \xi \frac{y-\mu}{\sigma} \right)^{-1/\xi} \right\}$. We choose this distribution because of its wide range of tail behaviours, this choice is independent of the choice of model (1) or (2).

We estimated by maximum likelihood a GEV model with month-specific position parameter (μ), and constant scale and shape parameters, finding a significant seasonal effect (estimation of a GEV model with month-specific position and scale parameters (σ) did not show evidence of a seasonal effect on the variability of the phenomenon).

We deseasonalized the series using this maximum likelihood estimates in order to analyse the dependence structure. Furthermore, according to the same estimates, data were transformed to get uniform marginals, this makes easier the implementation of the bayesian MCMC algorithm.

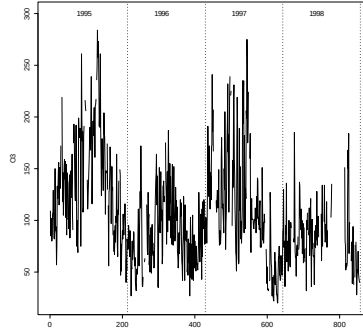


Figure 1: Time series of daily maxima for summer periods. Maxima were computed on days for which the empirical probability of observing the day maximum during the missing hours is less than 0.2.

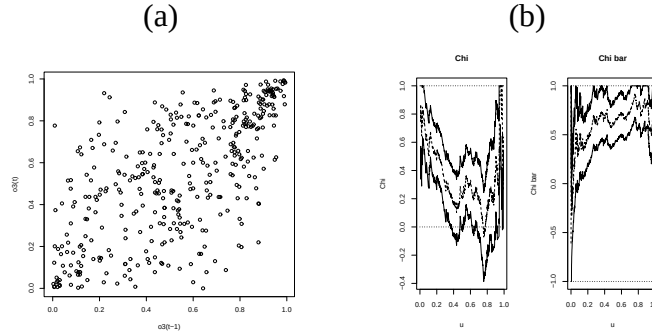


Figure 2: Scatterplot of ozone concentrations at day t versus concentrations at day $t - 1$ for 1995 and 1996 after transformation to get uniform marginals (a) and non parametric estimates for tail dependence (b).

3. Dependence analysis

After a transformation in order to get uniform marginals, we analyzed the dependence structure of (Y_t, Y_{t-1}) . Both the scatter and non parametric estimates in figure 2 suggest a relatively strong extremal dependence structure: this is encouraging for our forecasting aims. In order to check forecasts the model was estimated using only the data for years 1995 and 1996, and checked against data for 1997 and 1998.

We implemented an MCMC algorithm in order to obtain a bayesian estimate of the parameters. Posterior distributions from our analysis are shown in figure 3 for $\bar{\chi}$ and k in the two models, as can be seen, the estimates are reasonably similar (the mean values of the two posteriors are respectively 0.86 and 0.91). Also, both estimates seem reasonable if compared to non parametric estimates in figure 2.

Eventually, we used the model for (Y_t, Y_{t-1}) to forecast ozone. Conditioning on the value of Y_{t-1} , we obtained the predictive distribution for Y_t . In order to check the forecast capability of our model we compared quantiles from the above predictive distribution with empirical quantiles (from the empirical distribution of ozon concentrations in t conditional on the value in $t - 1$). Some preliminary results are shown in figure 4.

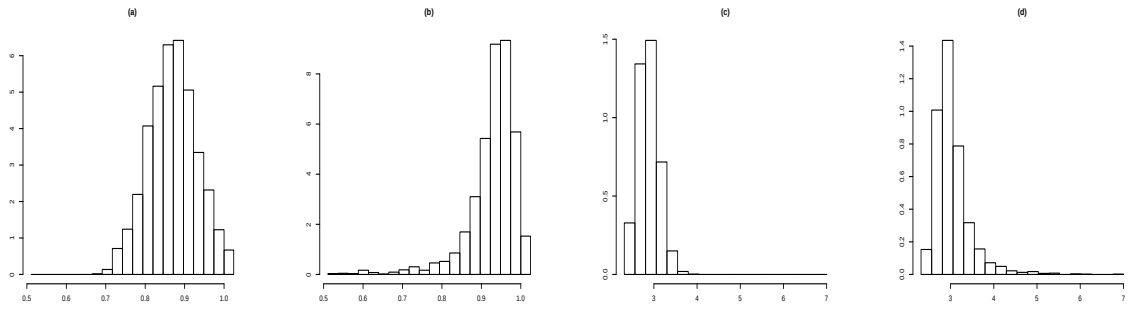


Figure 3: Bayesian estimates: posteriors for $\bar{\chi}$ for model (1) (a) and (2) (b), last bar represents probability of $\bar{\chi} = 1$; posteriors for k for model (1) (c) and (2) (d)

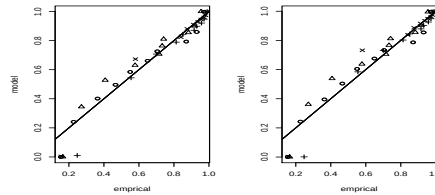


Figure 4: Quantile plots of empirical versus model quantiles for model (1) (left) and model (2) (right) computed for conditioning event $y_{t-1} = 0.7$ ($\diamond$); $y_{t-1} = 0.8$ ($\triangle$); $y_{t-1} = 0.9$ ($+$); $y_{t-1} = 0.95$ ($\times$).

4. Discussion

An important issue which is not addressed in this analysis but is the plan for future work is that of estimating simultaneously the dependence structure, the marginals and the seasonal effect.

Moreover, the time dependence may extend to more than one day, so another direction to improve the analysis consist in using the multivariate version of model (2) for the d -pla $(Y_t, Y_{t-1}, \dots, Y_{t-d})$.

References

- Coles, S. G. and Pauli, F. (2001). Models and inference for uncertainty in extremal dependence, *Biometrika* (to appear).
- Coles, S. G., Heffernan, J. E. and Tawn, J. A. (2000). Dependence measures for extreme value analyses, *Extremes* **2**, 339–365.
- Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions, *JRSS B* **59**, 475–499.