Multivariate and Spatial Extremes for the Analysis of Air Quality Data [1]

Simone A. Padoan and Alessandro Fassó
Department of Information Technology and Mathematical Methods,
University of Bergamo, Viale Marconi 5, 24044 Dalmine, Bergamo, Italy
email: simone.padoan@unibg.

Abstract: In recent years statistical analyses for monitoring the environment are increasingly in demand in different areas such as epidemiology, engineering, economy, etc. An example is the statistical monitoring of air quality, which makes it possible to statistically quantify the amount of certain pollutants in the lower troposphere. For a better understanding of the stochastic behavior of pollutants we focus on describing their extreme responses, because excessively extreme levels in the air may have implications in the environment and on human health. We then consider multivariate extreme value models and the class of maxstable processes in order to asses the frequencies of several extreme pollutant levels in central Europe and their spatial dependence structure.

Keywords: max-stable processes, multivariate extreme value distributions, generalized extreme value distribution, extremal coefficient, correlation function, Fréchet distribution, pollution.

1 Introduction

Nowadays in many disciplines such as epidemiology, engineering, economy, etc, are in great demand the statistical analyses for monitoring the environment. Specifically, it is very important to statistically quantify the amount of certain pollutants in the lower troposphere and this is possible thanks to the statistical monitoring of the air quality. A main aspect of environmental processes is their natural spatial domain, presupposing a statistical spatial analysis approach. One of the primary aims of the latter is to asses the dependence structure of the underlying process. In this case it is important to determine the degree of dependence of the pollutants' levels among the monitoring stations. There are a number of generic approaches to spatial modeling that to date have already been widely applied (e.g., Diggle and Ribeiro, 2007). But these are suitable for modeling the mean process levels, therefore they are inappropriate for handling extremal aspects. For a better understanding of the stochastic behavior of pollutants we focus on describing their extreme responses, be-

[1] This research is part of Project EN17, "Methods for the integration of different renewable energy sources and impact monitoring with satellite data", funded by Lombardy Region under "Frame Agreement 2009".

cause excessively extreme levels in the air may have implications in the environment and on human health.

With this work we aim to describe the extreme values of certain pollutants, such as fine particulate matters, sulphure, nitrogen dioxides, etc. recorded in central Europe. Each pollutant is recorded at $s = 1, \ldots, S$ locations, within a continuous region, for n-temporal observation with $n = 1, 2, \ldots$. At each site we compute the maximum with respect to a block of N temporal observations. For example, for hourly observations, we set $N = 24 \times 366$ and this implies that we focus on annual maxima of the process. Thus, we derive a temporal series of componentwise maxima of process measurements denoted by $\{y_t(s)\}$ with $t = 1, \ldots, T$ the sample of block maxima. In order to perform the analyses of the pollutants' extreme levels we consider the classes of multivariate extreme value models and of maxstable processes (see e.g. Chapters 6, 9 of de Haan and Ferreira, 2006). These families provide a quite general framework, with similar asymptotic motivations to the univariate case, suitable to model extreme processes incorporating temporal or spatial dependence. Statistical methods for max-stable processes and data analyses of practical problems are discussed by Padoan et al. (2010).

2 Methods

A suitable setting for addressing spatial problems in the extreme values context is provided by max-stable processes.

Let $\{Y(x)\}_{x \in \mathcal{X}}$ be a stochastic process defined on $\mathcal{X} \subseteq \mathbb{R}^q$, $q \in \mathbb{N}$, with continuous sample path. Assume that n independent and identically distributed (iid) copies of it, Y_i with $i = 1, \ldots, n$, are available, and hence focus on the limit of the rescaled process $\{M_n(x)\}_{x \in \mathcal{X}}$. Specifically, if there exist continuous positive functions $a_n(x)$ and real functions $b_n(x)$, with $n \in \mathbb{N}$ such that

$$Z(x) = \lim_{n \to \infty} \left\{ \frac{M_n(x) - b_n(x)}{a_n(x)} \right\}_{x \in \mathcal{X}} \tag{1}$$

is not a trivial limit, that is the normalized sequence $M_n(x)$ converges in distribution to a process $Z(x)$ with non-degenerate marginals for all $x \in \mathcal{X}$, then we call Z an extreme value process. Observe, that the limiting process Z posses three important proprieties: a) it is a *max-stable* process; b) all its univariate marginal distributions belong to *the generalized extreme value* class of distributions; c) all its finite p-dimensional distributions, with $p \geq 2$, are characterized to be *multivariate extreme value distributions* (see e.g. Chapters 1, 6 of de Haan and Ferreria, 2006).

Correlation coefficients and correlation functions are typically used in order to describe pairwise dependence, under Gaussianity assumption, respectively for high dimensional and spatial analysis. Similarly extremal coefficients and the extremal coefficient functions describe the dependence for extremes. Specifically, given Z_i, $i = 1, \ldots, n$, iid copies of a component-wise random vector $Z = (Z_1, \ldots, Z_p) \in \mathbb{R}^p_+$

with common unit Fréchet margins, then from the following relation

$$\mathbb{P}\left\{\max(Z_1,\ldots,Z_p) \leq z\right\} = \mathbb{P}\left\{Z_1 \leq z\right\}^{\theta} = \exp(-\theta/z), \quad z > 0,$$

where the rightmost term is a Fréchet(θ) distribution, the parameter $1 \leq \theta \leq p$ defines the extremal coefficient. When $\theta = 1$ indicates complete dependence, whereas $\theta = p$ corresponds to full independence. The extremal dependence of stochastic processes has a similar definition. If now we consider a stationary max-stable process $Z(x)$ with univariate unit Fréchet margins then, for any pair of locations x_1, $x_2 \in \mathcal{X}$ separated by $h = x_2 - x_1$, from the following relation

$$\mathbb{P}\left\{\max(Z(h), Z(o)) \leq z\right\} = \exp(-\theta(h)/z), \quad z > 0,$$

the real-valued function $\theta(h)$ defines the pairwise extremal coefficient function, where o denotes the origin (e.g. Schlather and Tawn, 2003). From a practical point of view we consider, for modeling the extremes of the pollutants, two specific families of max-stable processes such as the Brown-Resnick process (e.g. Kabluchko et al., 2009) and the Extremal Gausssian process (e.g. Schlather, 2002) and the class of multivariate extreme value distributions named the Extremal-t model (e.g. Nikoloulopoulos et al., 2009). We can easily fit these models to the pollutants data using the maximum composite likelihood estimation method (e.g. Padoan et al., 2010) and then to compare the different results. Moreover, for these models the closed form of the extremal coefficients is known so that we can, after the fitting step, assess the dependence structure and estimate the frequencies with which different high levels of the pollutants occur.

3 Data

The dataset considered for the analysis consists of hourly measurements of some pollutants of central Europe (see left panel of Figure 1) available on the Internet at the website: `http://www.eea.europa.eu`. Specifically, we took into account a time-period of 13 years, from January 1996 to December 2008, and we selected a region of approximately 341.000 km^2. The right panel of Figure 1 shows the area where the monitoring weather stations are located and the numbers from 1 to 3 denote the locations for the different pollutants. In particular, the 139 stations indicated with the number 1 monitor the benzene ($COH6$), carbon monoxide (CO) and nitrogen dioxide ($NO2$), the 126 stations indicated by the number 2 monitor the ozone ($O3$) and the 68 stations indicated by the number 3 monitor the particulate matters ($PM10$), sulphure ($SO2$). For each pollutant there are some missing data but the percentage is small, we can account on average (between sites) 2 % of missing values. Given that in the analysis we focus on block maxima of pollutants, where the blocks are formed by 24 × 366 temporal observations leading to sequences of annual maxima, the small percentage of missing data should not have an impact on the description of the extremes.

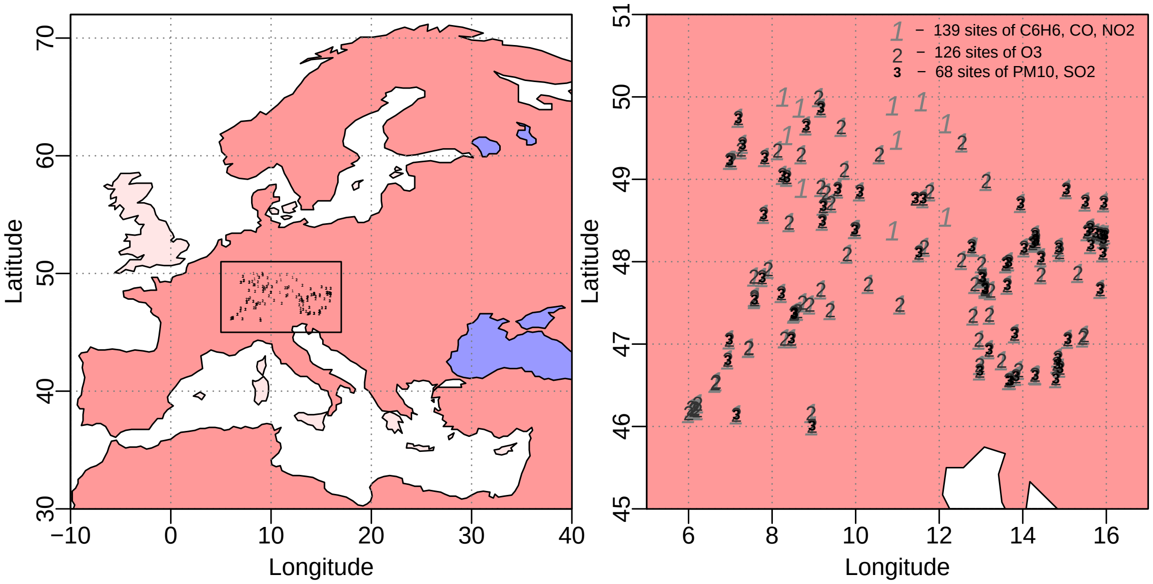

Figure 1: *Air quality data: the left panel reports the European map and the rectangle displays the central part where the monitoring weather stations are located. The right panel shows the expanded zone marked by the rectangle of the left panel and displays with the numbers the locations of the monitoring stations.*

References

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory An Introduction*. New York: Springer.

Kabluchko, Z., Schlather, M. and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *The Annals of Probability*, 37, 2042–2065.

Nikoloulopoulos, A. K., Joe H. and Li H. (2009). Extreme value properties of multivariate t copulas. *Extremes*, 12, 129–148.

Padoan, S. A., Ribatet, M. and Sisson S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association, Theory & Methods*, 105, 263–277.

Diggle, P. J. and Ribeiro P. J. (2007) *Model–Based Geostatistics*. London: Springer.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5, 33–44.

Schlather, M. and J. A. Tawn (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90, 139–154.