Science of the Total Environment 409 (2011) 1123–1133

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Analysis of high level ozone concentrations using nonparametric methods

Alejandro Quintela-del-Río, Mario Francisco-Fernández *

University of A Coruña. Faculty of Computer Science, Campus de Eviña, s/n, A Coruña 15071, Spain

A R T I C L E I N F O

Article history:
Received 10 June 2010
Received in revised form 18 October 2010
Accepted 29 November 2010
Available online 31 December 2010

MSC:
62G32
62G08
62P12

Keywords:
Nonparametric estimation
Air-pollution
Ozone
Extreme values

A B S T R A C T

Controlling emissions of air pollutants and establishing air quality objectives to improve and protect ambient air quality are very important tasks of Governments. Ozone (O_3), one of those pollutants of concern, is not emitted directly into the atmosphere, but is a secondary pollutant produced by reaction between nitrogen dioxide (NO_2), hydrocarbons and sunlight. High levels of ozone can produce harmful effects on human health and the environment in general. Therefore, the study of extreme values of ozone represents an important topic of research in environmental problems. Classical extreme value theory has been usually used in air-pollution studies. It consists of fitting a parametric generalized extreme value (GEV) distribution to a data set of extreme values and using the estimated distribution to compute quantities like the probability of exceedance, the quantiles, the return levels or the mean return periods. In this paper, we propose nonparametric methods to estimate those quantities. Additionally, nonparametric estimators of the trends of very high values of ozone are proposed. The nonparametric estimators are applied to real samples of maximum ozone values obtained from several monitoring stations belonging to the Automatic Urban and Rural Network (AURN) from the UK. Results show that nonparametric estimators work satisfactorily, generally outperforming the behaviour of classical parametric estimators.

1. Introduction

In the last years, numerous studies on air-pollution problems using statistical methods have been published. Generally, in these studies, statistical techniques such as time series analysis, regression methods, multivariate statistical analysis or spatial statistics are used to deal with problems like forecasting high levels of a certain pollutant, identifying trends in high levels of this pollutant, or mapping the spatial distribution of this element in a region. One of these pollutants of concern is ozone (O_3). Ozone is a natural component of the troposphere, produced by the photochemical reactions of nitrogen ($NO_x = NO + NO_2$) and volatile organic compounds (VOCs). Among other things, these reactions depend on meteorological conditions, like sunlight, temperature, wind speed or wind direction, producing complex seasonal patterns and trends in ozone levels. High ozone levels are taken as indicative of high pollution, representing a risk factor to human life, vegetation or materials. Therefore, due to these risks, controlling those levels as well as other harmful sources which can cause global warming and serious environmental problems is an important task of Governments.

Standards for air pollution are concentrations over a given time period that are considered to be acceptable in the light of what is scientifically known about the effects of each pollutant on health and on the environment. Regarding ozone, the World Health Organization, in the 2005 global update of its quality guidelines (World Health Organization, 2006), reduced the guideline given in its second edition (World Health Organization, 2000) from 120 μg m^{-3} (8-h daily average) to 100 μg m^{-3} for a daily maximum 8-h mean. It is considered that ozone levels higher than this value can produce health problems. These problems depend on the sensitivity of each individual and the type of exposure, and go from slight disabilities to permanent damages. However, some countries have developed their own regulations or air quality objectives for protection of human health or protection of vegetation and ecosystems. Usually, these regulations are based on the number of exceedances of an established value (threshold) in a period of time. Therefore, it is obvious that the importance of carrying out statistical analyses to estimate the probability of obtaining an ozone level higher than a threshold, or to estimate the mean number of times that a threshold could be exceeded in a certain period of time. In the present paper, we focus on those and other related estimation problems, using nonparametric methods. Furthermore, we use nonparametric regression techniques to analyse the trends of very high values of ozone. These methodologies can help environmental agencies give out public health warnings or evaluate the effectiveness of their regulation programs, for example. While in some papers statistical analysis of environmental ozone data are tackled using time series analysis (Dueñas et al., 2005; Kumar and Jain, 2010; Liu, 2009; Prybutok et al., 2000; Slini et al., 2002), in the present work a combination of nonparametric methods and extreme value theory is used.

* Corresponding author. Tel.: +34 981167000x1222; fax: +34 981167160.
E-mail address: mariofr@udc.es (M. Francisco-Fernández).

0048-9697/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2010.11.032

The statistics of extremes (Gumbel, 1958; Leadbetter et al., 1983) plays a very important role to deal with some of the problems described at the end of the previous paragraph. This methodology has been usually used in many fields of environmental studies such as climatology (Elsner et al., 2006; Perrin et al., 2006; Rajabi and Modarres, 2008), hydrology (Katz et al., 2002), agricultural management (Gomes et al., 2003) and many others. There are also interesting papers in the analysis of ground-level ozone using extreme value theory (see, e.g. Huerta and Sansó, 2007; Küchenhoff and Thamerus, 1996; Reyes et al., 2009; Smith, 1989). In those papers, classical parametric methods to model extreme values are used. Extreme value theory relies on asymptotic arguments for a sample of an observed data set of extreme values. The basis behind parametric methods in the previous papers is that presented in (Leadbetter et al., 1983), where it is shown that, under general conditions, the distribution of extreme values in stationary processes corresponds to the type of the named generalized extreme value (GEV) distribution. Basically, the GEV distribution is characterized by three values, shape, scale and location. Once these parameters are estimated, some other important quantities, such as the probability of exceedance, the quantiles or return levels, or the mean return period, can be estimated. Alternatively, a different kind of ideas based on nonparametric tools can be used in the analysis of these last quantities.

Beginning with Parzen (1962), extensive statistical literature exists inside what has been called nonparametric curve estimation. This methodology is a flexible and potent tool used to describe the behaviour of univariate and multivariate data sets, because it does not need the specification of a concrete model to work with (such as the normal distribution, or a linear relation). The nonparametric statistical techniques are, in some cases, a supplement for parametric models, because parametric models are usually well suited only to a sequence of events that have similar causes. Moreover, parametric models can be insensitive to anomalous events, since these models tend to be formulated through experience of relatively conventional activity. The reader can find a wide discussion about the use of smoothing ideas in many statistical problems through the following books: Silverman (1986); Simonoff (1996); and Wand and Jones (1995). Nonparametric methods have been also used in applied extreme problems more recently, for example in hydrologic studies (Sharma et al., 1997, 1998), modelling the earthquake risk (Quintela, 2010), or in econometric risk analysis (Cai and Wang, 2008). In those papers, the distribution of annual maxima (AM), partial duration or annual minimum are estimated via nonparametric estimators.

In the present work, nonparametric estimation methods are considered in the study of maximum ozone concentrations. On one hand, nonparametric estimators of the probability of exceedance, the return levels and the mean return period are proposed. On the other hand, nonparametric regression estimators of the trend of the return levels are used to investigate the behaviour of these quantities through time. Similar analyses to estimate these trends were carried out in Reyes et al. (2009), although they used polynomial regression methods. As nonparametric methods do not assume a prespecified functional form (as linear, quadratic or logistic) for the trend and let the data speak by themselves, more reliable estimates are obtained with our proposal. In that part of the study, we focus on the 95% quantiles. Note that, although nonparametric kernel methods could be notoriously unstable for extreme quantiles, the 95% quantiles are in the range where kernel techniques should provide reliable results. We apply these estimators to ozone real data from the UK.

The content of the paper is as follows. In Section 2, quantities of interest such as the probability of exceedance, the return levels, or the mean return period are defined. Moreover, we briefly describe the classical parametric estimators and our nonparametric proposals to estimate those values. In Section 3, we apply and compare both kinds of techniques (parametric GEV and nonparametric) to ozone data from the UK. Finally, Section 4 collects the main conclusions.

2. Statistical methods

Suppose $X_1,\ldots,X_n$ a sequence of extreme values with common distribution function F. In the setting addressed in this paper, these variables can represent the maximum ozone concentrations measured in a specific period of time (24 h, a month, a year,…). An important function in this context is the function that, for an ozone level c, gives the probability of obtaining a maximum ozone concentration larger than c (per unit of time); that is, the function returning the probabilities of exceedance. It is defined as

$$R(c) = P(X > c). \tag{1}$$

Related with (1), the following quantities can be defined (Coles, 2001). For $0<p<1$, the quantile of order $1-p$ of F is defined as the value z_p such that

$$1-p = P\left(X \le z_p\right) = F\left(z_p\right) \Leftrightarrow z_p = F^{-1}(1-p). \tag{2}$$

Thus, the T-return level is defined as the value of the observed concentrations that can be expected to be once exceeded during a T-period of time. It is given by

$$RL(T) = F^{-1}\left(1-\frac{1}{T}\right) = z_{1/T}. \tag{3}$$

The mean return period or recurrence interval of a concrete level c is an estimator of the interval of time between events of level c. It can be defined as the inverse of the probability that a level c will be exceeded in one period of time:

$$RT(c) = \frac{1}{P(X>c)} = \frac{1}{1-F(c)}. \tag{4}$$

As it was pointed out in the Introduction, it is very important to obtain reliable estimators of these values when ozone is the pollutant under consideration. Next subsections describe two ways to face these estimation problems, the classical parametric approach based on the GEV distribution and the nonparametric approach.

2.1. Parametric estimators. The GEV distribution

Classical extreme value theory uses the idea that, under certain regularity conditions (Fisher and Tippett, 1928), the limit of the distribution function F of the maximum is the GEV distribution function. This function is considered to correspond to one of the following three families,

$$F_\theta(x) = \begin{cases} \exp\left\{-[1+\gamma(x-\mu)/\sigma]^{-1/\gamma}\right\}, \\ 1+\gamma(x-\mu)/\sigma > 0, \gamma \neq 0, \\ \exp\{-\exp[-(x-\mu)/\sigma]\}, \gamma = 0 \end{cases} \tag{5}$$

with $\theta=(\mu,\sigma,\gamma)$. Here, μ is the location parameter, $\sigma>0$ is the scale parameter and γ is the shape parameter. The case of $\gamma=0$ is named the Gumbel distribution.

Based on a random sample $X_1,\ldots,X_n$ of extreme values, an estimator $\hat{\theta}$ for θ can be obtained. This can be done using, for example, the probability weighted moments method (Hosking et al., 1985) or by maximum likelihood related techniques (Coles, 2001). As soon as we get $\hat{\theta}$, an estimator $F_{\hat{\theta}}$ for F is derived. Using $F_{\hat{\theta}}$, parametric estimators for (1), (3) and (4), the given by

$$R_{\hat{\theta}}(c) = 1-F_{\hat{\theta}}(c), \tag{6}$$

$$RL_{\hat{\theta}}(T) = F_{\hat{\theta}}^{-1}\left(1-\frac{1}{T}\right) \tag{7}$$