Analysis of high level ozone concentrations using nonparametric methods

Alejandro Quintela-del-Río[a], Mario Francisco-Fernández[a,*]

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

Abstract

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 in 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

[*] Corresponding author. Phone: +34 981167000 (ext. 1222). Fax: +34 981167160.
Email address: `mariofr@udc.es` (Mario Francisco-Fernández)

Preprint submitted to Science of the Total Environment *October 18, 2010*

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.

Keywords: Nonparametric estimation, air-pollution, ozone, extreme values
2000 MSC: 62G32,
2000 MSC: 62G08,
2000 MSC: 62P12

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-hour daily average) to 100 μg m^{-3} for a daily maximum 8-hour 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 a established value (threshold) in a period of time. Therefore, it is obvious 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 to give out public health warnings or to 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 (Prybutok et al., 2000; Slini et al., 2002; Dueñas et al., 2005; Liu, 2009; Kumar and Jain, 2010), in the present work a combination of nonparametric methods and extreme value theory is used.

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. Küchenhoff and Thamerus, 1996; Huerta and Sansó, 2007; 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); 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 realiable 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 $X_1, ..., X_n$ 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-hours, a month, a year,...). An important function in this context is the function that, for a 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

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

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 $X_1, ..., X_n$ 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), given by

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

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

and

$$RT_{\hat{\theta}}(c) = \frac{1}{1 - F_{\hat{\theta}}(c)}, \tag{8}$$

are immediately calculated. When a particular sample of extreme ozone values, $x_1, ..., x_n$, is observed, numeric estimates of those functions are computed by substituting the sample values in the expressions of the corresponding estimators.

2.2. Nonparametric estimators

As it was indicated in the Introduction Section, nonparametric curve estimation methods do not assume a prespecified functional form for the curve to be estimated. This enables the estimator to fit a wide range of possible curves. Different nonparametric estimators, depending on the curve to be estimated, have been developed in the last decades. Next, nonparametric kernel estimators of the density function, the distribution function and the regression function are briefly described. These estimators allow us to obtain nonparametric estimators of (1), (3) and (4), and of the trends of high ozone levels.

Let X be a continuous random variable, with density function f and distribution function F. Given a random sample $X_1, ..., X_n$, each X_i having the same distribution as X, the Parzen-Rosenblatt nonparametric kernel estimator (Parzen, 1962) of $f(\cdot)$ is defined by

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{9}$$

In this expression, K is a kernel function and $h = h(n) \in \mathbb{R}^+$ is the smoothing parameter, or bandwidth. While the election of the function K has not play an important role in the fitting of the estimation (normally, K is a density function with some regularity conditions), the election of the bandwidth is more crucial, because the shape of the resulting estimator varies greatly according to its value. If the value of h is small, an undersmooth estimator, with high variability, will be obtained. On the contrary, if the value of h is big, the resulting estimator will be very smooth and farther from the function that we seek to estimate (see, e.g. Silverman, 1986).

From the relation between a density function and a distribution function, it is also possible to construct a nonparametric kernel estimator of the distribution function, given by the expression

$$F_h(x) = \int_{-\infty}^{x} f_h(t)dt = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right), \tag{10}$$

where $H(u) = \int_{-\infty}^{x} K(t)dt$.

From (10), nonparametric estimators of the probabilities of exceedance, the return levels and the mean return period, defined in (1), (2) and (4), respectively, can be immediately obtained. Their expressions are,

$$R_h(c) = 1 - F_h(c), \tag{11}$$

$$RL_h(T) = F_h^{-1}\left(1 - \frac{1}{T}\right) \tag{12}$$

and

$$RT_h(c) = \frac{1}{1 - F_h(c)}. \tag{13}$$

Theoretical motivations for this kind of approaches can be found in Youndjé and Vieu (2006).

As it was explained previously, an important step to compute (11), (12) and (13) is to select a bandwidth for the nonparametric estimator of the distribution function. A popular technique to select that bandwidth is the least-squares *cross-validation* method (Sarda, 1993).

A different approximation is related to considering some type of quadratic error, such as the Mean Integrated Squared Error (MISE),

$$MISE_F(h) = \int (F_h(x) - F(x))^2 dx, \tag{14}$$

and then to select the bandwidth minimizing an asymptotic approximation of this error. This bandwidth is given by:

$$h_{AMISE}(F_h) = Cn^{-1/3} = \left(\frac{0.5\int V_F^2(x)w(x)f(x)dx}{\int B_F^2(x)w(x)f(x)dx}\right)^{1/5} \cdot n^{-1/3}, \tag{15}$$

where $B_F = \frac{1}{2}\left(f''(x)\right)^2\left(\int x^2 K(x)dx\right)$ and $V_F^2(x) = 2f(x)\int xK(x)H(x)dx$ (see, e.g. Quintela, 2007). Constant C in (15) depends on the kernel function and the theoretical (unknown in practice) distribution function of the data (Reiss, 1981). A *plug-in* approximation of (15) considers the bandwidth

$$\hat{h} = \hat{C}\ n^{-1/3}, \tag{16}$$

where $\hat{C}$ is a preliminary nonparametric approximation of the unknown values (Altman and Leger, 1995; Polanski and Baker, 2000).

The simulation study in Bowman et al. (1998) compares the cross-validation technique with the plug-in method proposed in Altman and Leger (1995), showing that better results can be obtained with the cross-validation criterion. A disadvantage of the cross-validation method when it is compared with plug-in methods is its worse performance, in terms of computing time. Certainly, cross-validation involves an integration term, that is needed to be computed numerically. Besides, a sum of n terms, including in each term a new estimator like (10), but computed with a subsample of size $(n-1)$, has to be calculated. In the present work, we use the iterative method presented in Polanski and Baker (2000) to obtain the bandwidth for the nonparametric distribution function estimators. There are mainly two reasons for this election: the fast computation (here, we are working with large sample sizes) and the best results obtained with this method in our own simulation studies. See Section 3 of Polanski and Baker (2000) for a description of the way to obtain $\hat{C}$ in (16).

An interesting problem in air-pollution studies is to investigate the behaviour of very high values (represented by the quantiles (2)) of a pollutant of interest through time. Tracing seasonal trends in the level of ozone is essential for predicting high-level periods, observing long-term trends, and discovering potential changes in pollution. Traditional methods for modelling seasonal effects are based on the conditional mean of ozone concentration; however, the upper conditional quantiles are more critical from a public health perspective. For $0 < p < 1$, let us denote by $\hat{z}_p^k$ the estimated quantiles of order $1-p$ at consecutive times (years, for example) t_k, with $k = 1, 2, \ldots, n$. A regression problem can be formulated to estimate the trend of those quantiles. Following these lines, considering the sample $\left\{\left(t_k, \hat{z}_p^k\right)\right\}_{k=1}^{n}$, the regression model

$$\hat{z}_p^k = m(t_k) + \varepsilon_k, \quad 1 \leq k \leq n, \tag{17}$$

where ε_k are random errors, can be used to estimate the trend $m(\cdot)$ of the high values of ozone.

This kind of analysis was used in Reyes et al. (2009) to study the trends of very high values of tropospheric ozone in some monitoring stations in Mexico City during the period from 1986 to 2005. They used a parametric polynomial regression model to estimate the trend functions. Taking into account the advantages of nonparametric function estimators, in the present paper, we propose to estimate the trends by using nonparametric estimators of the regression function. Specifically, the nonparametric kernel estimator, called the local polynomial regression (LPR) estimator, is used in our analyses. This estimator consists in locally fitting a q-degree polynomial to the observed sample, $\left\{\left(t_k, \hat{z}_p^k\right)\right\}_{k=1}^{n}$, and it can be written as:

$$\hat{m}_h(t) = \mathrm{e}_1' \left(\mathrm{X}_t' \mathrm{W}_t \mathrm{X}_t\right)^{-1} \mathrm{X}_t' \mathrm{W}_t \mathrm{Y} \tag{18}$$

where $\mathrm{e}_1 = (1, 0, \ldots, 0)'$, $\mathrm{Y} = (\hat{z}_p^1, \ldots, \hat{z}_p^n)'$,

$$\mathrm{X}_t = \begin{bmatrix} 1 & t_1 - t & \ldots & (t_1 - t)^q \\ \vdots & \vdots & & \vdots \\ 1 & t_n - t & \ldots & (t_n - t)^q \end{bmatrix},$$

and $\mathrm{W}_t = \mathrm{diag}\{K((t_1 - t)/h), \ldots, K((t_n - t)/h)\}$, with $K(\cdot)$ a kernel function, q the degree of the local polynomial and h the bandwidth. Similarly to the estimators given in (9) and (10), the bandwidth h is a very important parameter to be selected by the user in order to obtain reliable estimators. Once again, cross-validation and plug-in methods are the most used here. Significant references on the LPR estimator, including some guidelines on how to select the bandwidth, are Fan and Gijbels (1996) or Wand and Jones (1995), for example.

A different approach to tackle this problem would be to use quantile regression (Koenker, 2005; Baur et al., 2004; Sousa et al., 2009). Specifically, a nonparametric quantile curve could be directly fitted to the original data set (Koenker, 2005, chap. 7). This would be a more direct way to deal with this problem. However, model (17) would be in line with a regression model with repeated measurements, the original high ozone level concentrations in each year. Once the quantiles per year are estimated using the nonparametric estimator, a nonparametric regression estimator would represent a good way to obtain the trend of these quantiles through the years. Given the good theoretical and practical properties of these nonparametric estimators, this approach seems to be reasonable.

3. Application to real data

In this Section the nonparametric estimators presented in the previous Section are applied to ozone real data in the UK. Moreover, the results are compared with those obtained by using the classical parametric estimators described in Subsection 2.1.

3.1. The data

Ozone concentration measurements are obtained from the UK Air Quality Archive, available at http://www.airquality.co.uk. This web page provides measurements of several pollutants monitored in different networks across the UK. There are two major types - automatic and non-automatic networks. There are currently 4 automatic networks and 11 non-automatic networks, funded by The Department for Environment, Food and Rural Affairs (Defra) and the Devolved Administrations, across the UK. The largest automatic monitoring network in the UK is the Automatic Urban and Rural Network (AURN). Currently, AURN includes automatic air quality monitoring stations measuring oxides of nitrogen ($\mathrm{NO_x}$), sulphur dioxide ($\mathrm{SO_2}$), ozone ($\mathrm{O_3}$), carbon monoxide (CO) and particles ($\mathrm{PM_{10}}$). In 2007, there were 133 operating sites in this network covering urban and rural areas. Rural and urban monitoring sites are organized by site type. Additionally, AURN stations are also grouped according to their locations in different regions. Regarding ozone, the pollutant of concern in this paper, there were 91 automatic urban and rural sites monitoring $\mathrm{O_3}$ using UV absorption analysers all around the UK in 2007. For more details about AURN or other monitoring networks in the UK, we refer the readers to the UK Air

Quality Archive web site (http://www.airquality.co.uk) or the Defra web site (http://www.defra.gov.uk/).

It is important to note that whereas nitrogen dioxide (NO_2) acts as a source of ozone, nitrogen oxide (NO) destroys ozone acting as a local sink. For this reason, ozone levels are not as high in urban areas (where high levels of NO are emitted from vehicles) as in rural areas. Furthermore, there are several meteorological and geographical variables, like number of sunlight hours, temperature, wind, altitude, etc. which influence ozone production (see Coyle et al., 2002). Taking this into account and in order to carry out a very general study of the nonparametric estimators presented in Section 2.2, we select 6 rural monitoring sites and 6 urban sites, belonging to the AURN network. These stations are selected trying to cover different regions and different type of sites. Table 1 shows the monitoring sites considered in the present study and Figure 1 their locations. All these stations are currently active and they started collecting ozone measurements more than 10 years ago (regarding ozone, the oldest of these 12 stations began working in 1986). To obtain the results presented in the following Subsection, measurements of hourly ozone concentrations in each monitoring site listed in Table 1 are considered. The data collected in each station go from the specific starting date of that station until the end of 2009.

[Table 1 about here.]

[Figure 1 about here.]

3.2. Results

Next, we present the main results of our study summarized in several tables and figures that are representative of the comprehensive research performed. The behaviour of the nonparametric estimators introduced in Section 2.2 is studied and they are compared with the classical parametric estimators presented in Section 2.1. To carry out this comparison, ozone data until 2008 are used to calculate (6), (7), (8), (11), (12) and (13), defined in Section 2. From them, parametric and nonparametric predictions of some important values in the year 2009 can be obtained and checked with the real values in that year. An important issue in order to apply classical parametric extreme value methods (Section 2.1) is that of independence of extreme data. To get this assumption, the maxima of daily concentration measurements are considered in our analysis. Smith (1989) worked with a similar assumption. In that paper, cluster intervals ranging from 24 hours to 72 hours were selected and no strong sensitivity was found to the election of this value. Similar approaches were followed in Küchenhoff and Thamerus (1996) and Huerta and Sansó (2007). Anyway, it should be noted that independence is not necessary for correct estimation of the parameters of (5). It is sufficient that dependence decreases suitable fast with increasing time separation (Perrin et al., 2006). However, nonparametric estimators can be correctly applied in this field and have good theoretical properties, although the assumption of independence is not strictly fulfilled (Youndjé and Vieu, 2006).

In Figure 2, box-plots of the daily maxima of hourly measurements in each location are shown. This descriptive plot gives useful information about the different levels of high values of ozone in each one of the considered monitoring stations, as well as the global difference between rural and urban sites.

[Figure 2 about here.]

Given a monitoring station and once a sample of daily maxima of hourly measurements is calculated in the period of time under consideration, the first step in our study is to estimate the distribution function of that sample of extreme values. Following the lines indicated in Section 2, this can be done by means of parametric or nonparametric methods. In the first case, location (μ), scale (σ) and shape (γ) parameters must be estimated from the observed sample, while to compute the nonparametric kernel estimator of the distribution function, a kernel function and a bandwidth parameter must be selected. The Gaussian kernel, $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$, $-\infty < u < \infty$ is used here. Table 2 shows the parameters calculated for each one of the locations considered, using the maximum likelihood method to estimate GEV parameters (in brackets, bootstrap standard deviations of these parameters) and the plug-in bandwidth method proposed in Polanski and Baker (2000) to select h. The free statistical software R (R Development Core Team, 2008) was used to implement the estimators computed in the present paper. For the parametric versions of these estimators, we use the evir package (S original (EVIS) by Alexander McNeil and R port by Alec Stephenson, 2008).

[Table 2 about here.]

Once the distribution function F is estimated, estimates for the probability of exceedance (1) above different daily maxima of ozone levels are readily obtained. As an example, in Figure 3, the nonparametric and the parametric estimations of this function for four stations, -Aston Hill, Strach Vaich, London N. Kensington and Thurrock- (2 rural and 2 urban) are shown. These estimates are computed with the parameters presented in Table 2. Similar graphs are obtained for all the monitoring sites.

[Figure 3 about here.]

Figure 4 shows the estimated mean return periods for different ozone levels (50, 75, 100, 120, 150 and 180). The nonparametric and the parametric estimates showed in this plot are computed in the same stations as those selected in Figure 3. Similarly, plots showing the T–return levels for different values of time could be obtained (not shown here for reasons of space).

[Figure 4 about here.]

In Figure 4, we also plot the real values obtained in the year 2009 by black point symbols. As the parametric and the nonparametric estimations are calculated using the data until the end of the year 2008, including these real values of the year 2009 can serve as a mean of comparison between the parametric and the nonparametric approaches.

A way to measure the uncertainty of the estimates presented in Figures 3 and 4 is computing confidence bands. We have designed a specific procedure combining bootstrap techniques with nonparametric methods to obtain simultaneous confidence bands (not only pointwise intervals) for the functions estimated in Figures 3 and 4. These bands are designed to contain the whole function with a prescribed high probability, typically 0.95.

Generally speaking, the method consists in, first, computing pointwise intervals with the pre-specified confidence level (95%, for instance) and with the confidence level calculated using a Bonferroni correction (Bonferroni, 1935). Then, applying an iterative algorithm, using the generated bootstrap curves, the proper confidence level for the simultaneous intervals is obtained. Finally, pointwise confidence intervals with this correct level of confidence are computed. A detailed description of this procedure is in Appendix A. As an example, Figure 5 shows the nonparametric estimator of the mean return period function in Aston Hill (shown in Figure 4) and the corresponding band computed using this new approach. For simplicity, to plot Figure 5, we selected the pilot bandwidth g with the same value as the bandwidth h used to obtain the estimator $RT_h(\cdot)$, given in (13) and also plotted in this figure (see Appendix A).

[Figure 5 about here.]

In addition to this, a good way to compare the parametric estimators with our nonparametric proposals is to estimate, using both approaches, the expected number of days with the daily maximum of ozone level larger than a threshold c in the year 2009. Then, these numbers can be compared with the real values of this variable. Formally, denoting by ND(c) the number of days in 2009 with the daily maximum of hourly ozone levels exceeding a threshold c, the expected number of this variable is given by

$$\mathrm{E}(ND(c)) = 365 \times R(c), \tag{19}$$

where $R(c)$ is given in (1). A natural estimator of (19) is

$$\mathrm{E}(\widehat{ND(c)}) = 365 \times \hat{R}(c), \tag{20}$$

where $\hat{R}(c)$ could be (11) or (6), respectively, depending on whether a nonparametric or a parametric approach is used. Figure 6 shows the parametric and nonparametric estimations and the real values of this variable for the monitoring sites previously used in Figures 3 and 4, and the thresholds previously used in Figure 4. Moreover, this information (ND, the parametric estimations, $\widehat{\mathrm{E}(ND)}_\theta$, and the nonparametric estimations, $\widehat{\mathrm{E}(ND)}_h$), is reported in Table 3 for the 6 rural monitoring sites and in Table 4 for the 6 urban monitoring sites for the different thresholds. For the sake of comparison, the estimations presented in Tables 3 and 4 are rounded to the nearest integer. Additionally, we include in these tables the mean squared error (MSE), given by

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathrm{ND}(c_k) - \mathrm{E}(\widehat{ND(c_k)})^*\right)^2,$$

where $\mathrm{E}(\widehat{ND(c_k)})^*$ can be $\mathrm{E}(\widehat{ND(c_k)})_\theta$ or $\mathrm{E}(\widehat{ND(c_k)})_h$ computed at the values $c_k = 50, 75, 100, 120, 150$ and 180.

[Figure 6 about here.]

[Table 3 about here.]

[Table 4 about here.]

We can observe in Tables 3 and 4 that nonparametric estimators give, in general, very good results, improving those obtained with the classical parametric estimators. This improvement is general in all stations, no matter the kind of station under consideration, location, etc.

The next part of our study is to analyse the trends of very high values of ozone in the stations under consideration. These high values are given by the quantiles (2) of order $1-p$ for small values of p. As explained in Section 2.2, for each monitoring site, this problem can be formulated as a regression problem, where the response variables are the estimated quantiles in each one of the years of consideration, and the explicative variables are the corresponding years. Considering the benefits, previously showed, of applying nonparametric methods in this field, we use nonparametric estimators of the quantiles and of the regression function. For each station and for each year (including the year 2009) quantiles of order 0.95 are estimated using (12) with $T = 20$. For each year, those estimations represent approximately the values such that only a 5% of the daily maxima of ozone concentrations are larger than them in that year. The nonparametric LPR estimator of the regression function, given in (18), with $q = 1$ (local linear estimator) is used in our research. The Gaussian kernel is also used here, while the bandwidths needed to compute the estimators are selected using the plug-in method proposed in Ruppert et al. (1995). In Figure 7, the estimations of the trends throughout the years, jointly with the confidence intervals computed with the row-wise method of Hannig and Marron (2006), in the rural monitoring sites are presented. Figure 8 shows the same information for the urban monitoring stations.

[Figure 7 about here.]

[Figure 8 about here.]

Figures 7 and 8 show different patterns in the trends of the different monitoring sites, revealing the different performance of maximum ozone levels throughout the years. In general, the estimated trends are not monotonic functions. That does not allow us to obtain clear conclusions of its behaviour. However, we can conclude that, unlike what it happened in the Mexico sites studied by Reyes et al. (2009), a parabolic regression fit would not model correctly the behaviour of the data (in several cases).

4. Concluding remarks

In this paper, we focus on the problem of analysing ozone extreme values by using nonparametric estimators. Based on the nonparametric kernel estimator of the distribution function, we develop a new methodology to obtain estimators of the probability of exceedance, the quantiles, the return levels and the mean return periods. These functions play an important role in environmental problems because they can be used to estimate or predict high ozone levels. Nonparametric estimation is also useful in showing the trend of the maximum ozone levels of each studied monitoring site over the years, allowing for more accurate and flexible models than those obtained by typical polynomial regression fits.

Although parametric and nonparametric techniques have been compared before, up to our knowledge, this is the first time in which a comparison is made in the context of ozone extreme values. We have been able to observe that nonparametric techniques yield highly accurate estimates of both the interest functions and parameters, and also to capture more complex patterns in the data that those allowed by the classical GEV fits. Perhaps the only disadvantage of the nonparametric methods can be the higher demanding in computational terms (requiring to obtain bandwidth parameters), but that, really, does not suppose a clear disadvantage nowadays.

We would like to stress the fact that the methodology developed here could be applied to other extreme value problems, for example, in hurricanes, rainfall or related problems in hydrology (Loaiciga and Leipnik, 1999; Koutsoyiannis and Baloutsos, 2000), insurance (Embrechts et al., 1997), reliability (Melchers, 1999) and many others.

5. Acknowledgements

The authors thank a referee for constructive comments that improved the presentation of this article. This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 and by Xunta de Galicia Grant PGIDIT07PXIB105259PR for the second author.

References

Altman N, Leger C. Bandwidth selection for kernel distribution function estimation. J Stat Plan Infer 1995;46:195–214.

Baur D, Saisana M, Schulze N. Modelling the effects of meteorological variables on ozone concentration - a quantile regression approach. Atmos Environ 2004;38:4689–99.

Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. In: Studi in Onore del Professore Salvatore Ortu Carboni. Rome; 1935. p. 13–60.

Bowman A, Hall P, Prvan T. Bandwidth selection for the smoothing of distribution functions. Biometrika 1998;85:799–808.

Cai Z, Wang X. Nonparametric estimation of conditional var and expected shortfal. J Econom 2008;147:120–30.

Coles S. An Introduction to Statistical Modeling of Extreme Values. London: Springer Verlag, 2001.

Coyle M, Smith RI, Stedman JR, Weston KJ, Fowler D. Quantifying the spatial distribution of surface ozone concentration in the UK. Atmos Environ 2002;36:1013–24.

Dueñas C, Fernández MC, Cañete S, Carretero J, Liger E. Stochastic model to forecast ground-level ozone concentration at urban and rural area. Chemosphere 2005;61:1379–89.

Elsner JB, Jagger TH, Tsonis AA. Estimated return periods for Hurricane Katrina. Geophys Res Lett 2006;33:L08704.1–. Doi: 10.1029/2005GL025452.

Embrechts P, Klüppelberg C, Mikosch T. Modelling Extremal Events for Insurance and Finance. Berlin: Springer, 1997.

Fan J, Gijbels I. Local Polynomial Modelling and its Applications. London: Chapman & Hall, 1996.

Fisher RA, Tippett LHC. Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proc Cambridge Philos Soc 1928;24:180–90.

Gomes L, Arrúe JL, López MV, Sterk G, Richard D. Wind erosion in a semiarid agricultural area of Spain: the WELSONS project. Catena 2003;52:235–56.

Gumbel EJ. Statistics of Extremes. New York: Columbia University Press, 1958.

Hannig J, Marron JS. Advanced distribution theory for SiZer. J Am Stat Assoc 2006;101:484–99.

Hosking JRM, Wallis JR, Wood EF. Estimation of the Generalized Exteme Value distribution by the method of probability-weighted moments. Technometrics 1985;27:231–62.

Huerta G, Sansó B. Time-varying models for extreme values. Environ Ecol Stat 2007;14:285–99.

Katz RW, Parlange MB, Naveau P. Statistics of extremes in hydrology. Adv Water Resour 2002;25:1287–304.

Koenker R. Quantile Regression. New York: Cambridge University Press, 2005.

Koutsoyiannis D, Baloutsos G. Analysis of a long record of annual maximum rainfall in athens, greece, and design rainfall inferences. Nat Hazards 2000;22:31–51.

Küchenhoff H, Thamerus M. Extreme value analysis of Munich air pollution data. Environ Ecol Stat 1996;3:127–41.

Kumar U, Jain VK. Arima forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). Stoch Environ Res Risk Assess 2010;24:751–60.

Leadbetter MR, Lindren G, Rootzén H. Extremes and related properties of random sequences and processes. New York: Springer, 1983.

Liu PG. Simulation of the daily average pm_{10} concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. Atmos Environ 2009;43:2104–13.

Loaiciga HA, Leipnik RB. Analysis of extreme hydrologic events with gumbel distributions: marginal and additive cases. Stoch Environ Res Risk Assess 1999;13:251–259.

Melchers RE. Structural Reliability Analysis and Prediction. New York: John Wiley & Sons, 1999.

Miller RG. Simultaneous Statistical Inference. New York, USA: Springer-Verlag, 1991.

Parzen E. On estimation of a probability density function and mode. Ann Math Statis 1962;32:1065–76.

Perrin O, Rootzén H, Taesler R. A discussion of statistical methods used to estimate extreme wind speeds. Theor Appl Climatol 2006;85:203–15.

Polanski A, Baker ER. Multistage plug-in bandwidth selection for kernel distribution function estimates. J Statist Comput Simul 2000;65:63–80.

Prybutok VR, Yi J, Mitchell D. Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. Eur J Oper Res 2000;122:31–40.

Quintela A. Plug-in bandwidth selection in kernel hazard estimation from dependent data. Comput Stat & Data Anal 2007;51:5800–12.

Quintela A. On non-parametric techniques for area-characteristic seismic hazard parameters. Geophys J Int 2010;180:339–46.

R Development Core Team . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2008. http://www.R-project.org.

Rajabi MR, Modarres R. Extreme value frequency analysis of wind data from Isfahan, Iran. J Wind Eng Ind Aerodyn 2008;96:78–82.

Reiss RD. Nonparametric estimation of smooth distribution functions. Scand J Statist 1981;8:116–9.

Reyes H, Vaquera H, Villaseñor JA. Estimation of trends in high urban ozone levels using the quantiles of (GEV). Environmetrics 2009;Doi: 10.1002/env.997.

Ruppert D, Sheather SJ, Wand MP. An effective bandwidth selector for local least squares regression. J Am Stat Assoc 1995;90:1257–70.

S original (EVIS) by Alexander McNeil and R port by Alec Stephenson . evir: Extreme Values in R; 2008. R package version 1.6.

Sarda P. Smoothing parameter selection for smooth distribution function. J Stat Plan Inference 1993;35:65–75.

Sharma A, Lall U, Tarboton DG. Kernel bandwidth selection for a first order nonparametric streamflow simulation model. Stoch Hydrol Hydraul 1998;12:33–52.

Sharma A, Tarboton , G. D, Lall U. Streamflow simulation: A nonparametric approach. Water Resour Res 1997;33:291–308.

Silverman BW. Density estimation for statistics and data analysis. London: Chapman & Hall, 1986.

Simonoff J. Smoothing Methods in Statistics. New York: Springer, 1996.

Slini T, Karatzas K, Moussiopoulos N. Statistical analysis of environmental data as the basis of forecasting: an air quality application. Sci Total Environ 2002;288:227–37.

Smith R. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Stat Sci 1989;4:367–93.

Sousa SIV, Pires JCM, Martins FG, Pereira MC, Alvim-Ferraz MCM. Potentialities of quantile regression to predict ozone concentrations. Environmetrics 2009;20:147–58.

Wand MP, Jones MC. Kernel Smoothing. London: Chapman & Hall, 1995.

World Health Organization . Air Quality Guidelines for Europe. 2nd ed. Number European series No. 91. Copenhagen, WHO regional publication, 2000.

World Health Organization . Air Quality Guidelines: Global Update 2005: Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide. Copenhagen, 2006. Avalible at: http://www.euro.who.int/Document/E87950.pdf.

Youndjé E, Vieu P. A note on quantile estimation for long dependent stochastic processes. Stat Probab Lett 2006;76:109–16.

Appendix A. Bootstrap simultaneous confidence bands

In this Appendix, the procedure to obtain the simultaneous confidence bands for the probability of exceedance function $R(\cdot)$, given in (1) is described in detail. A similar approach can be followed to obtain simultaneous confidence bands for the functions defined in (3) or (4).

First, given an initial confidence level, $1-\alpha$, for a small α ($\alpha = 0.01$ or 0.05, typically), we start by constructing individual $(1-\alpha)$-confidence intervals, (ℓ_j, u_j), for $\hat{R}(c_j)$, in a selected grid of ozone levels, c_j, $j = 1, \ldots, k$.

For every c_j, $j = 1, \ldots, k$, using the nonparametric estimator $R_h(c_j)$ given in (11), the sampling distribution of

$$D_n\left(c_j\right) = R_h(c_j) - R(c_j) \tag{A.1}$$

is approximated. This is done using its bootstrap distribution, from which it is easy to obtain pointwise $(1-\alpha)$-confidence intervals. The process is the following:

1. Obtain a bootstrap sample $X_1^*, \ldots, X_n^*$ from the original sample of ozone high levels, $X_1, \ldots, X_n$.

2. With this bootstrap sample, compute an approximated bootstrap version of (A.1), using

$$D_n^*(c_j) = R_h^*(c_j) - R_g(c_j), \tag{A.2}$$

where $R_g(c_j)$ is the nonparametric estimator of $R(c_j)$ with a pilot bandwidth g.

3. Repeat steps 1 and 2 a large number of times B ($B = 1000$ or 5000, for example). After steps 1-3, we have B values of the bootstrap distribution $D_n^*(c_j)$.

4. Compute the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the bootstrap distribution: $D_n^{*\left(\left\lceil \frac{\alpha}{2}B\right\rceil\right)}(c_j)$ and $D_n^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right)B\right\rceil\right)}(c_j)$, where $\lceil x \rceil$ denotes the integer part of x. They are the values that are in positions $\left\lceil \frac{\alpha}{2}B\right\rceil$ and $\left\lceil \left(1-\frac{\alpha}{2}\right)B\right\rceil$, when sorting the bootstrap resample in an increasing order.

5. Compute the confidence interval as $D_n^{*\left(\left\lceil \frac{\alpha}{2}B\right\rceil\right)}(c_j) \leq R_h(c_j) - R(c_j) \leq D_n^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right)B\right\rceil\right)}(c_j)$.

6. The final bootstrap confidence interval for $R(c_j)$ is

$$\left[R_h(c_j) - D_n^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right)B\right\rceil\right)}(c_j), R_h(y) - D_n^{*\left(\left\lceil \frac{\alpha}{2}B\right\rceil\right)}(c_j)\right]$$

or, considering the expression (A.2), equivalently,

$$\left[R_h(c_j) + R_g(c_j) - R_h^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right)B\right\rceil\right)}(c_j), R_h(c_j) + R_g(c_j) - R_h^{*\left(\left\lceil \frac{\alpha}{2}B\right\rceil\right)}(c_j)\right],$$

where $R_h^{*\left(\left\lceil \frac{\alpha}{2}B\right\rceil\right)}(c_j)$ and $R_h^{*\left(\left\lceil \left(1-\frac{\alpha}{2}\right)B\right\rceil\right)}(c_j)$ are the corresponding $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the bootstrap distribution $R_h^*(c_j)$.

The previous algorithm is repeated for every grid point c_j, with $j = 1, \ldots, k$ and, therefore, k pointwise intervals (ℓ_j, u_j), for each value $\hat{R}(c_j)$, $j =$

$1, \ldots, k$ are calculated. Individual confidence intervals have approximately the nominal coverage probability $(1-\alpha)$ when they are considered separately (for a particular grid point). However, the probability that the whole growth curve is included in the band depicted by the whole set of intervals is much smaller. This is known as the multiple range testing problem or the false discovery rate in high dimensional statistical problems (Miller, 1991).

A classical way to correct for multiple testing is the popular Bonferroni approach (Bonferroni, 1935). In a hypothesis testing context, the idea behind this approach is to consider a new significance level, $\alpha_{\text{Bonf}} = \frac{\alpha}{J}$, and compute individual tests using this new level. The resulting multiple test has a multiple level which is much closer to the desired α. However, it is well known that the Bonferroni approach is a conservative procedure. In our context, this means that the joint coverage probability of the confidence band would be larger than the desired $1-\alpha$.

Starting from the conservative Bonferroni approach and the anticonservative individual testing approach, the following algorithm finds an approximate $(1-\alpha)$-confidence interval, with a given approximation error δ (typically δ is small in comparison with the nominal α, for instance $\delta = \frac{\alpha}{10}$):

1. Fix $\alpha_{\text{low}}^{(0)} = \alpha_{\text{Bonf}} = \frac{\alpha}{J}$ and $\alpha_{\text{high}}^{(0)} = \alpha$. Fix the iteration number, $k = 0$.
2. Compute $\alpha_{\text{mean}}^{(k)} = \frac{\alpha_{\text{low}}^{(k)} + \alpha_{\text{high}}^{(k)}}{2}$.
3. Use the bootstrap resamples to compute individual confidence intervals with $1-\alpha_{\text{low}}^{(k)}$, $1-\alpha_{\text{mean}}^{(k)}$ and $1-\alpha_{\text{high}}^{(k)}$ confidence levels.
4. Compute with the same bootstrap resamples, the proportion of bootstrap curves that are included in each of these confidence bands. These proportions satisfy $p_{\text{low}}^{(k)} \geq p_{\text{mean}}^{(k)} \geq p_{\text{high}}^{(k)}$, $p_{\text{low}}^{(k)} \geq 1-\alpha \geq p_{\text{high}}^{(k)}$ and

$p_{\text{low}}^{(k)} > p_{\text{high}}^{(k)}$.

5. If $p_{\text{mean}}^{(k)} \geq 1 - \alpha$, then define $\alpha_{\text{low}}^{(k+1)} = \alpha_{\text{mean}}^{(k)}$ and $\alpha_{\text{high}}^{(k+1)} = \alpha_{\text{high}}^{(k)}$. Otherwise define $\alpha_{\text{low}}^{(k+1)} = \alpha_{\text{low}}^{(k)}$ and $\alpha_{\text{high}}^{(k+1)} = \alpha_{\text{mean}}^{(k)}$.
6. Stop at step k if $\left| p_{\text{mean}}^{(k)} - (1 - \alpha) \right| < \delta$. Otherwise increase k in one unit and repeat Steps 2-5.

The final approximate $(1 - \alpha)$ simultaneous confidence intervals are those obtained for level $1 - \alpha_{\text{mean}}^{(k)}$ in the last iteration.

Figure legends

Figure 1: O_3 monitoring sites (AH–Aston Hill, Es–Eskdalemuir, LN–Lough Navar, LH–Lullington Heath, SV–Strach Vaich, WF–Wicken Fen, LC–Leeds Centre, LK–London Kensington, MS–Manchester South, Md–Middlesbrough, NC–Nottingham Centre, Th–Thurrock).

Figure 2: Box-plots of the daily maxima of hourly measurements in each monitoring station.

Figure 3: Estimations of the probability of exceedance in Aston Hill, Strach Vaich, London Kensington and Thurrock. Parametric estimations in solid lines and nonparametric estimators in dashed lines.

Figure 4: Estimated mean return periods for different daily maxima of ozone levels in Aston Hill, Strach Vaich, London Kensington and Thurrock. Parametric estimations in solid lines, nonparametric estimations in dashed lines and real values in year 2009 in black point markers.

Figure 5: Estimated mean return periods for different daily maxima of ozone levels in Aston Hill and bootstrap simultaneous confidence band with $\alpha = 0.05$.

Figure 6: Estimated expected number of days in 2009 with the daily maximum of hourly ozone level exceeding different thresholds in Aston Hill, Strach Vaich, London Kensington and Thurrock. Parametric estimations in solid lines, nonparametric estimations in dashed lines and real values in year 2009

in black point markers.

Figure 7: Estimations of high-level ozone trends and confidence bands in rural monitoring sites.

Figure 8: Estimations of high-level ozone trends and confidence bands in urban monitoring sites.

Table 1: Monitoring sites with their geographical area and type.

Rural		
Monitoring Sites	**Monitoring Zone**	**Site Type**
Aston Hill (AH)	West Midlands	Rural
Eskdalemuir (Es)	Scottish Borders	Rural
Lough Navar (LN)	Nothern Ireland	Rural-Remote
Lullington Heath (LH)	South East	Rural
Strach Vaich (SV)	Highlands	Rural-Remote
Wicken Fen (WF)	Eastern	Rural
Urban		
Monitoring Sites	**Monitoring Zone**	**Site Type**
Leeds Centre (LC)	Yorkshire/Humberside	Urban Centre
London Kensington (LK)	London Urban Area	Urban background
Manchester South (MS)	N. W. and Mereyside	Suburban
Middlesbrough (Md)	North East	Industrial Urban
Nottingham Centre (NC)	East Midlands	Urban Centre
Thurrock (Th)	Eastern	Urban Background

Table 2: Estimated location (μ), scale (σ) and shape (γ) parameters with bootstrap standard deviations in brackets, and bandwidths (h) needed to estimate F using parametric and nonparametric methods, respectively.

Monitoring	Parametric			Nonpar.
Sites	μ	σ	γ	h
Aston Hill (AH)	-0.09 (0.006)	21.33 (0.258)	68.43 (0.237)	1.07
Eskdalemuir (Es)	-0.10 (0.007)	18.88 (0.224)	61.24 (0.225)	1.08
Lough Navar (LN)	-0.11 (0.011)	18.82 (0.235)	59.22 (0.242)	1.15
Lullington Heath (LH)	-0.07 (0.007)	26.23 (0.295)	67.60 (0.322)	1.45
Strach Vaich (SV)	-0.11 (0.006)	15.62 (0.179)	74.29 (0.180)	1.24
Wicken Fen (WF)	-0.11 (0.013)	26.18 (0.392)	64.07 (0.464)	1.69
Leeds Centre (LC)	-0.15 (0.008)	22.99 (0.225)	47.36 (0.323)	1.64
London N. Kensington (LK)	-0.12 (0.008)	27.08 (0.276)	50.85 (0.414)	1.70
Manchester South (MS)	-0.12 (0.007)	19.16 (0.241)	44.04 (0.301)	1.31
Middlesbrough (Md)	-0.09 (0.015)	23.07 (0.322)	58.99 (0.379)	1.27
Nottingham Centre (NC)	-0.13 (0.008)	22.55 (0.228)	44.41 (0.360)	1.53
Thurrock (Th)	-0.12 (0.007)	26.29 (0.300)	54.21 (0.378)	1.69

Table 3: Number of days in the year 2009 above thresholds and estimations using parametric and nonparametric methods in rural monitoring sites.

Monitoring Sites		c						**MSE**
		50	**75**	**100**	**120**	**150**	**180**	
AH	ND	341	176	32	12	0	0	
	$\widehat{\mathrm{E}(ND)}_\theta$	172	129	92	69	42	25	6653.32
	$\widehat{\mathrm{E}(ND)}_h$	335	187	41	17.11	6	2	51.35
Es	ND	328	160	28	3	0	0	
	$\widehat{\mathrm{E}(ND)}_\theta$	161	114	77	54	31	16	6030.43
	$\widehat{\mathrm{E}(ND)}_h$	318	121	21	8	2	0	291.03
LN	ND	305	61	2	0	0	0	
	$\widehat{\mathrm{E}(ND)}_\theta$	151	105	69	479	259	13	5536.29
	$\widehat{\mathrm{E}(ND)}_h$	293	102	14	5	1	0	336.5
LH	ND	324	185	32	3	1	0	
	$\widehat{\mathrm{E}(ND)}_\theta$	175	132	95	72	45	28	6076.05
	$\widehat{\mathrm{E}(ND)}_h$	309	191	59	28	11	4	298.53
SV	ND	334	194	21	1	0	0	
	$\widehat{\mathrm{E}(ND)}_\theta$	158	120	87	66	42	25	7929.07
	$\widehat{\mathrm{E}(ND)}_h$	333	218	37	9	1	0	149.62
WF	ND	310	154	43	12	1	0	
	$\widehat{\mathrm{E}(ND)}_\theta$	173	127	88	64	38	21	4325.06
	$\widehat{\mathrm{E}(ND)}_h$	301	169	51	22	7	2	84.88

Table 4: Number of days in the year 2009 above thresholds and estimations using parametric and nonparametric methods in urban monitoring sites.

Monitoring Sites		c						**MSE**
		50	**75**	**100**	**120**	**150**	**180**	
LC	ND	232	50	1	0	0	0	
	$\widehat{\mathrm{E}(ND)}_{\theta}$	146	89	49	28	11	3	2031.74
	$\widehat{\mathrm{E}(ND)}_{h}$	223	64	13	5	1	0	73.47
LK	ND	257	101	18	1	0	0	
	$\widehat{\mathrm{E}(ND)}_{\theta}$	166	109	66	42	19	8	2127.63
	$\widehat{\mathrm{E}(ND)}_{h}$	250	105	29	13	4	1	57.04
MS	ND	148	25	3	0	0	0	
	$\widehat{\mathrm{E}(ND)}_{\theta}$	136	79	42	23	9	3	875.49
	$\widehat{\mathrm{E}(ND)}_{h}$	1909	399	9	30	0	0	338.62
Md	ND	254	90	6	0	0	0	
	$\widehat{\mathrm{E}(ND)}_{\theta}$	144	101	68	47	26	14	3196.97
	$\widehat{\mathrm{E}(ND)}_{h}$	260	112	19	6	2	1	129.21
NC	ND	215	77	11	1	0	0	
	$\widehat{\mathrm{E}(ND)}_{\theta}$	147	87	46	26	10	3	1125.4
	$\widehat{\mathrm{E}(ND)}_{h}$	208	58	12	5	1	0	74.55
Th	ND	341	176	32	12	0	0	
	$\widehat{\mathrm{E}(ND)}_{\theta}$	172	129	92	69	42	25	6653.10
	$\widehat{\mathrm{E}(ND)}_{h}$	335	187	41	17	6	2	51.57

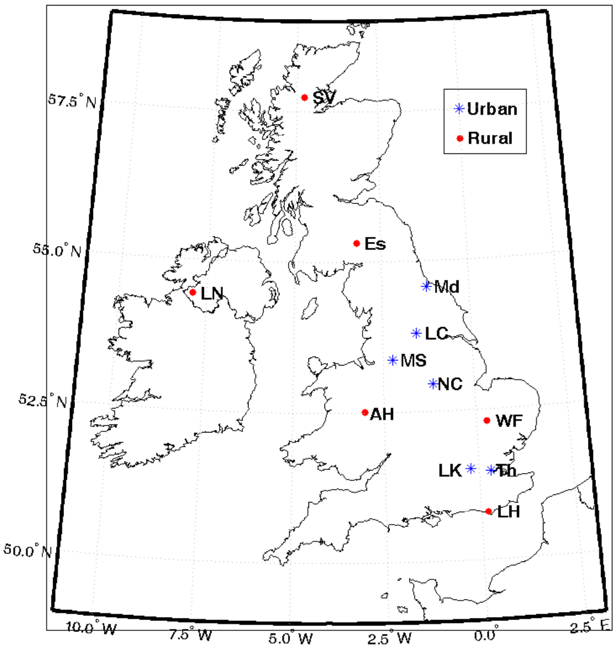

Figure 1: O_3 monitoring sites (AH–Aston Hill, Es–Eskdalemuir, LN–Lough Navar, LH–Lullington Heath, SV–Strach Vaich, WF–Wicken Fen, LC–Leeds Centre, LK–London Kensington, MS–Manchester South, Md–Middlesbrough, NC–Nottingham Centre, Th–Thurrock).

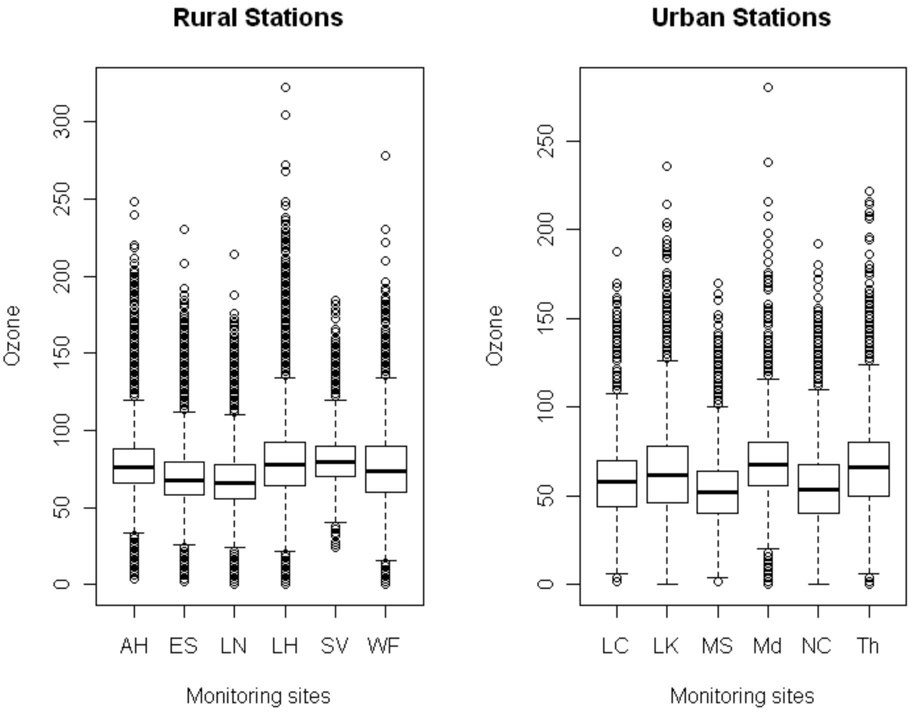

Figure 2: Box-plots of the daily maxima of hourly measurements in each monitoring station.

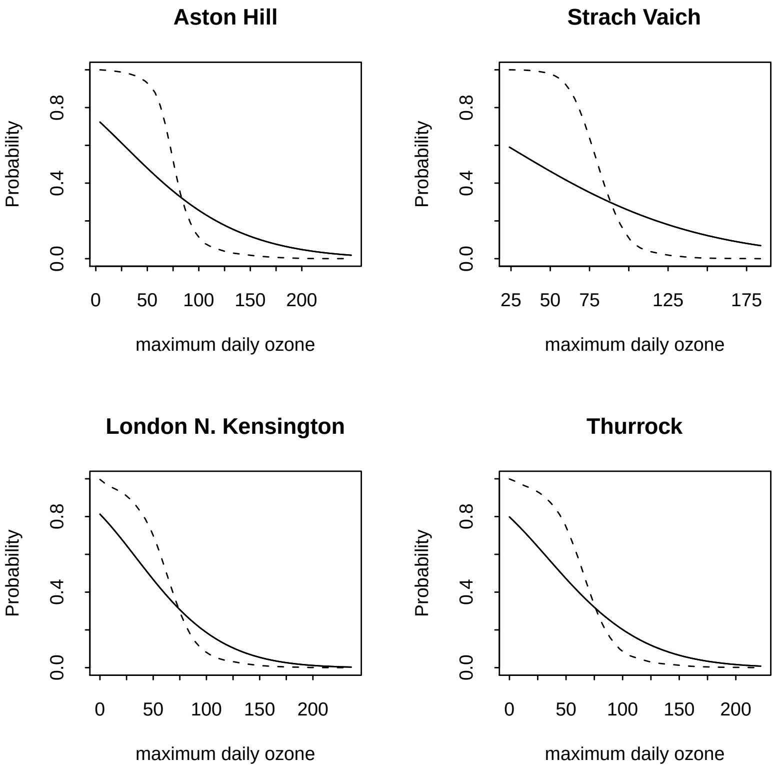

Figure 3: Estimations of the probability of exceedance in Aston Hill, Strach Vaich, London Kensington and Thurrock. Parametric estimations in solid lines and nonparametric estimators in dashed lines.

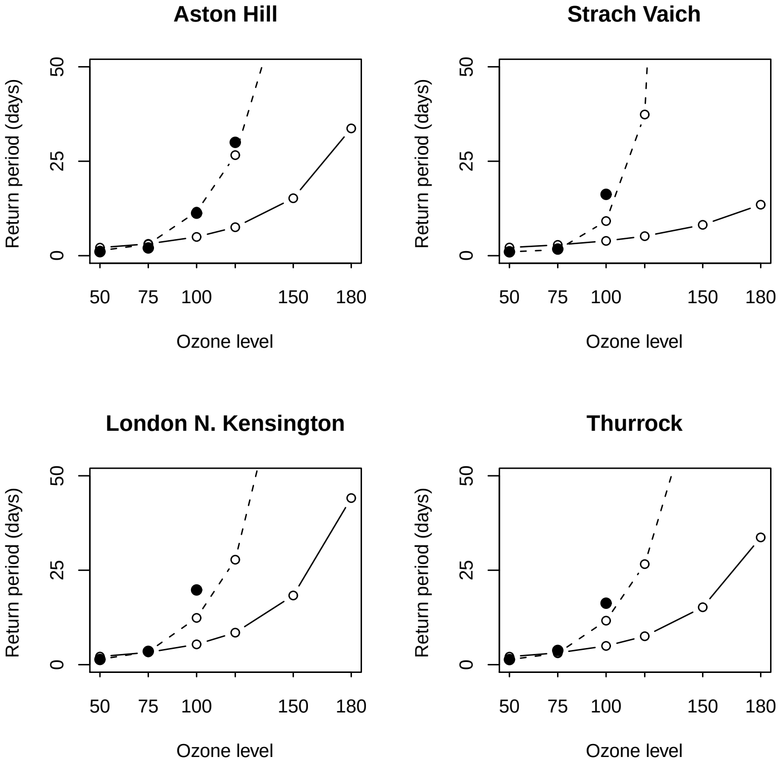

Figure 4: Estimated mean return periods for different daily maxima of ozone levels in Aston Hill, Strach Vaich, London Kensington and Thurrock. Parametric estimations in solid lines, nonparametric estimations in dashed lines and real values in year 2009 in black point markers.

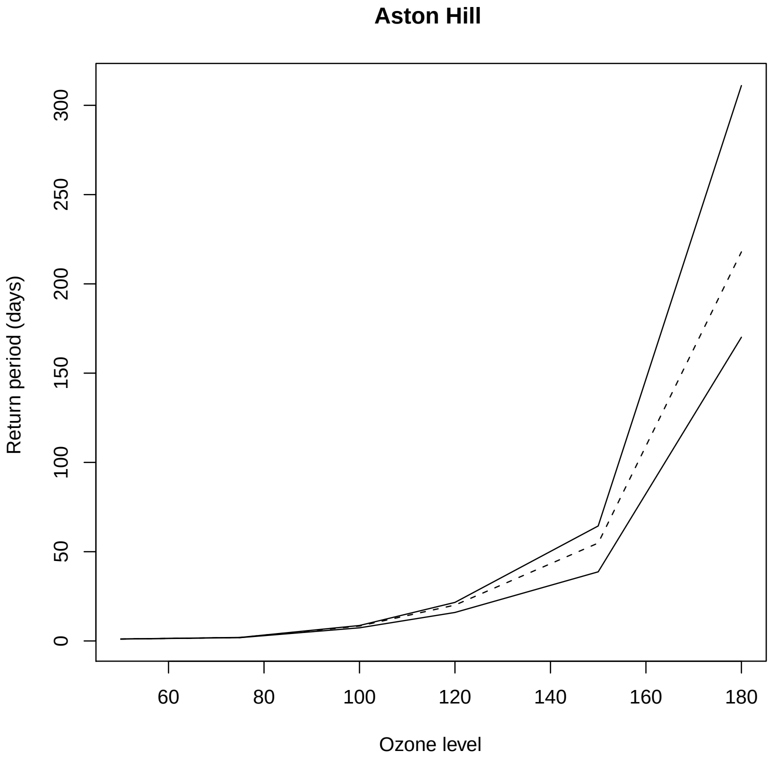

Figure 5: Estimated mean return periods for different daily maxima of ozone levels in Aston Hill and bootstrap simultaneous confidence band with $\alpha = 0.05$.

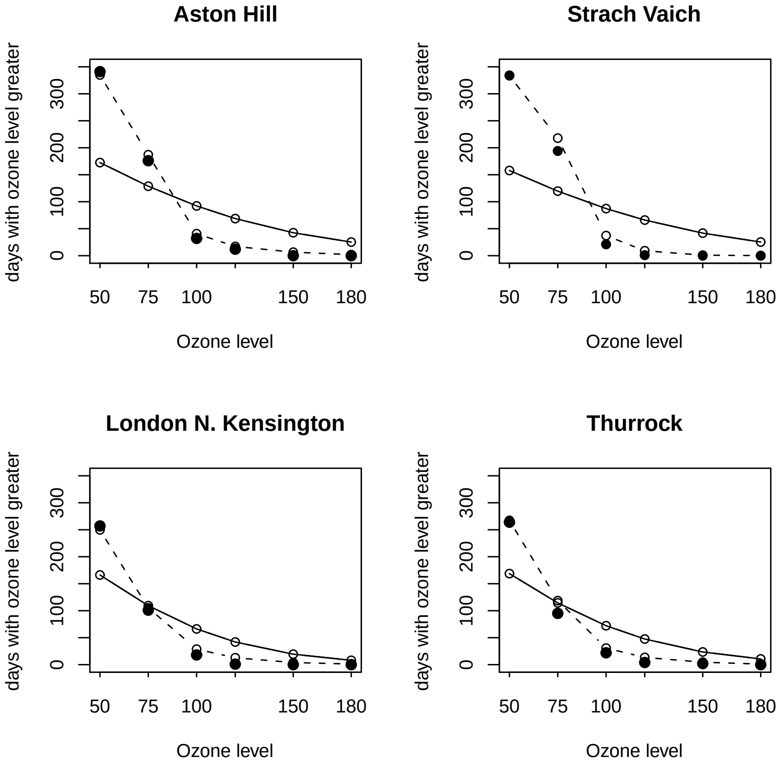

Figure 6: Estimated expected number of days in 2009 with the daily maximum of hourly ozone level exceeding different thresholds in Aston Hill, Strach Vaich, London Kensington and Thurrock. Parametric estimations in solid lines, nonparametric estimations in dashed lines and real values in year 2009 in black point markers.

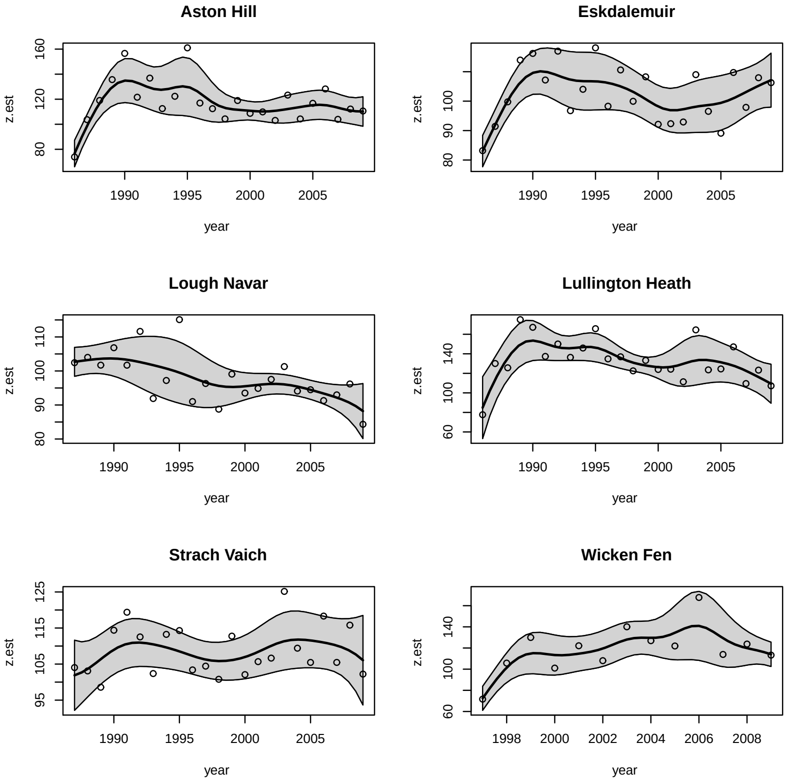

Figure 7: Estimations of high-level ozone trends and confidence bands in rural monitoring sites.

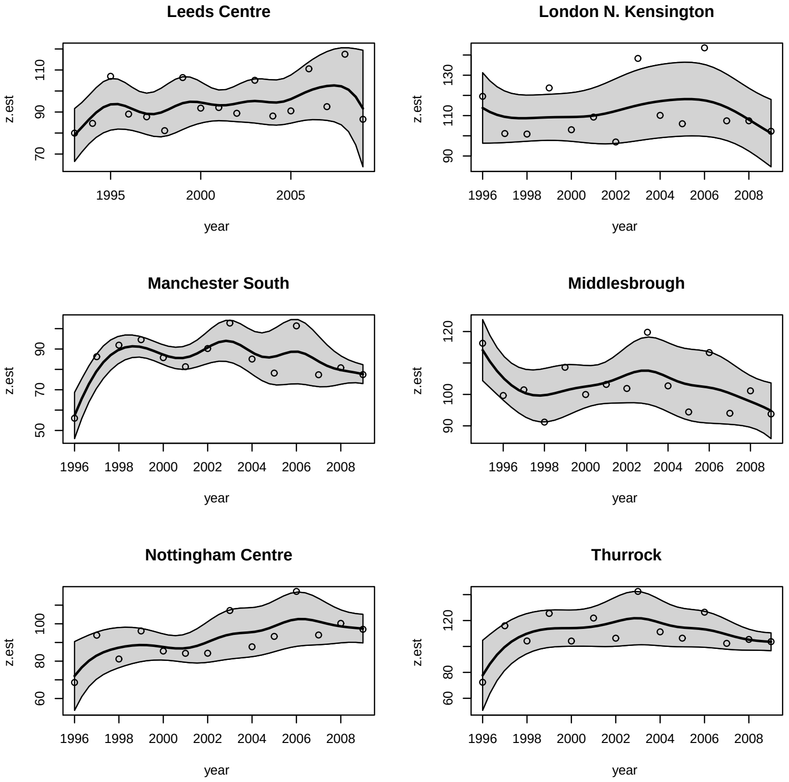

Figure 8: Estimations of high-level ozone trends and confidence bands in urban monitoring sites.