NISS

A Point Process Approach to Modeling Trends in Tropospheric Ozone Based on Exceedances of a High Threshold

Richard L. Smith and Thomas S. Shively

Technical Report Number 16
May, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Although the information in this document has been funded wholly or in part by the United States Environmental Protection Agency under assistance agreement #CR819638-01-0 to the National Institute of Statistical Sciences, it may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

A POINT PROCESS APPROACH TO MODELING TRENDS IN TROPOSPHERIC OZONE BASED ON EXCEEDANCES OF A HIGH THRESHOLD

by

Richard L. Smith
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27599-3260

and

Thomas S. Shively
CBA 5.202
Department of Management Science
and Information Systems
University of Texas at Austin
Austin, Texas 78712

May, 1994

Abstract: A major issue with the analysis of data on tropospheric ozone is to establish whether observed trends in the data are real, meaning that they could be attributed to actual changes in the emissions of toxic gases into the atmosphere, or whether they are the result of meteorological changes affecting the conditions under which ozone is generated. One way of investigating this question is to construct a regression model in which the level of ozone is represented as a function of both meteorological variables and time, in order to determine the significance of the time component when the meteorological variables are taken into account. However, the conventional methods of regression analysis do not make any distinction between low and high levels of the series, whereas with ozone it is largely the high levels that are of interest and concern. This paper proposes a method of regression analysis that is based entirely on the exceedances over a high threshold, and applies the method to data from the Houston area.

Keywords: Diagnostic model testing; Generalized Pareto distributions; Meteorological conditions; Nonhomogeneous Poisson process.

Please send all correspondence to:

Thomas S. Shively
CBA 5.202
Department of Management Science and Information Systems
University of Texas at Austin
Austin, Texas 78712

1. Introduction

The problem considered in this paper is to construct a statistical model for the exceedances of tropospheric ozone over a high threshold. According to the National Ambient Air Quality Standard (NAAQS) for ozone established by the U.S. Environmental Protection Agency, each monitoring station for ozone is assessed by the exceedances of daily maximum ozone over a threshold level of 12 parts per hundred million (pphm). Each monitoring station is supposed to have not more than three excesses in any three-year period. Therefore, a natural starting point for the analysis of ozone data is to consider all exceedances of this or some nearby threshold level. A major issue of current interest is to analyze such series for trend, while taking into account the meteorological factors which are known to influence the formation of ozone.

Shively (1991) presented one analysis of this nature, in which the point process consisting of all the times at which the threshold is exceeded is modeled as a nonhomogeneous Poisson process, with the parameters depending on both time and a number of meteorological variables. By modeling the dependence on time as a function of meteorology, he was able to separate out the two effects. This improved on earlier models for threshold exceedances, which did not take meteorology into account.

However, one disadvantage of Shively's analysis is that it uses the data only to the extent of determining at what times the threshold was exceeded — a violation of 13 pphm is treated in exactly the same way as one of 30 pphm. An analysis that took into account the magnitudes of the exceedances should be of greater interest to regulatory and public health agencies, and also of greater value in terms of drawing scientific inferences.

The principles behind an analysis of this sort were developed, in the context of tropospheric ozone data, by Smith (1989). Smith's analysis took into account the actual levels of exceedances as well as the times at which the exceedances occurred. However, as presented in that paper, the method could only be used to detect and measure trends in the data and not to take account of other factors such as meteorology.

The purpose of the present paper is to combine these two methodologies. We are able to extend the methodology of Smith (1989) to model the two-dimensional point process of exceedance times and exceedance levels as a nonhomogeneous Poisson process whose parameters depend on both time and meteorology. Then, using the models of Shively (1991) as a starting point, we develop specific models for the meteorological influence on the ozone exceedances. In addition, we explicitly allow for the effects of missing values in the estimated parameters of the model. Over the ten year period 1983-1992 for which we have data, approximately 20% of the observations are missing due to the monitoring equipment being out of service, so it is important to account for the effects of missing data on the model. The results of our analysis reinforce Shively's earlier conclusions, which essentially showed that meteorology does have a major influence on the assessment of a trend, and also suggest a number of approaches for more detailed studies.

2. Modeling the Time and Size of the Exceedances

In this section we develop the two-dimensional nonhomogeneous Poisson process used to the model the time and size of the exceedances of a specified threshold level. Consider first the hypothetical case in which the series of daily ozone maxima form a stationary time series, unaffected by either a time trend or by changes in meteorology. (There is inevitably a seasonal trend, the ozone problem in most places being confined to the summer months. However, in the analysis that follows we consider only the summer data, so that the seasonal effect can be ignored.) By confining attention to exceedances of a high threshold, we obtain a two-dimensional point process of exceedances, where the two dimensions represent the times at which the exceedances occur, and their levels. When both the threshold and the total time span of the observations are large, the point process may be approximated by a two-dimensional Poisson process. This intuitive picture is supported by a very large body of probabilistic theory about the limiting

properties of extremes in stationary stochastic process, as represented for example by the monograph of Leadbetter, Lindgren and Rootzén (1983), and further reviewed by Smith (1989).

The purpose of our model is to allow for the relationship between meteorological conditions and the frequency and size of the exceedances. To develop our model, we begin by letting

$$\Psi_t(y) = \begin{cases} \Pr\{Y > y \text{ on day } t\} & \text{if day } t \text{ is not missing} \\ 0 & \text{if day} t \text{ is missing} \end{cases} \tag{2.1}$$

i.e. if day t is not missing, $1-\Psi_t(y)$ is the probability distribution of Y on day t. Also, let

$$\psi_t(y) = -\frac{d}{dy}[\Psi_t(y)]$$

Suppose the process is observed over a time period $(0,T^*)$ and the peaks over the threshold u are represented by $\{(T_i,Y_i),\ 1 \le i \le N\}$. That is, the ith peak occurs at time T_i and takes the value $Y_i \ge u$. The total number of peaks N is itself a random variable. Then the approximate joint density of the data is

$$\left[\prod_{i=1}^{N} \frac{\psi_{t_i}(y_i)}{\Psi_{t_i}(u)}\right]\left[\left(\prod_{i=1}^{N} \Psi_{t_i}(u)\right)\exp\left\{-\int_0^{T^*} \Psi_t(u)dt\right\}\right]. \tag{2.2}$$

The term $\Psi_{t_i}(u)$ is left in the denominator of the first term in square brackets and in the second term in square brackets to facilitate the interpretation discussed below of each of these terms.

An important point to make at this stage is that we are explicitly allowing for the possibility of missing data on day t. As we noted in the Introduction, over the ten year period which we are analyzing, approximately 20% of the observations are missing. Therefore, it is clearly important to account for the effects of missing data. If we assume

that the data are missing at random, then using the expression for $\Psi_t(y)$ in (2.1) and the density function in (2.2) properly accounts for the effects of missing values.

We now interpret each term in square brackets individually. Consider the product in the first set of square brackets. The *i*th term in this product is the density function of y_i given $y_i \geq u$, i.e. it is the density function of y_i given that an exceedance of the threshold level u occurred. The second term in square brackets in (2.2) is the density function for a nonhomogeneous Poisson process with frequency function $\Psi_t(u)$. Thus, one can interpret the first term as modeling the size of the exceedance on day t given that an exceedance occurred, while the second term models the times of the exceedances of the threshold level u.

The joint density function in (2.2) is an approximation to the actual density function because we are modeling the times of the exceedances as a continuous time stochastic process while the "true" model is a discrete time stochastic process (by definition there is at most one exceedance per day because we are working with daily maxima). However, we believe that the density in (2.2) is a reasonable approximation and this is supported empirically by the results of the diagnostic tests discussed in Section 3.

We now consider the *i*th term in the first set of square brackets $\psi_{t_i}(y_i)/\Psi_{t_i}(u)$. For notational purposes the subscript i is dropped for the remainder of this paragraph. If we know the underlying distribution function $\Psi_t(y)$, then an expression for $\psi_t(y)/\Psi_t(u)$ can be obtained directly. However, we can avoid making any explicit assumptions about the parametric form of $\Psi_t(y)$ by using a result due to Pickands (1975) and Davison and Smith (1990). In our context this results states that if $\Psi_t(y)$ is in the domain attraction of one of the extreme value distributions, then the limiting distribution of $X = Y - u$, given $Y \geq u$, as $u \to \infty$ is the generalized Pareto distribution

$$G(x;\beta(t),\xi(t)) = 1-\left(1+\xi(t)\beta(t)x\right)^{-1/\xi(t)}$$

where $x = y - u$. The limiting density function for x is

$$\lim_{u\to\infty}\frac{\psi_t(y)}{\Psi_t(u)} = g(x;\beta(t),\xi(t)) = \beta(t)(1+\xi(t)\beta(t)x)^{-(1/\xi(t)+1)} \tag{2.3}$$

Since the domain of attraction of the extreme value distributions contains all the distributions that are likely to arise in practice, no assumptions are required regarding the underlying parametric form of the distribution function $\Psi_t(y)$.

To complete the specification of the model for the exceedances sizes we need to specify a functional form for $\beta(t)$ and $\xi(t)$. We will assume

$$\beta(t) = \beta_0 + \beta_1 s(t) + \sum_{j=2}^{p}\beta_j w_j(t) \tag{2.4}$$

where $s(t)$ is the year in which observation t occurred and $w_j(t)$, $j = 2, \ldots, p$, is a vector of p-1 meteorological variables for each time t. We will assume that $\xi(t) = \xi$ does not depend on t. Note that

$$\lim_{\xi\to 0} g(x;\beta(t),\xi) = \beta(t)\exp\{-\beta(t)x\} \tag{2.5}$$

so in this case $1/\beta(t)$ can be interpreted to be the expected size of an exceedance, given that an exceedance occurred, conditional on the meteorological variables on day t. The empirical evidence discussed in Section 5 suggests that $\xi \approx 0$, and therefore that the exponential distribution is the appropriate distribution for the exceedance sizes.

With regard to the second term in brackets in (2.2), we will assume

$$\Psi_t(u) = \text{pr}\{Y \geq u \text{ on day } t\} = \exp\{\alpha(t)\} \tag{2.6}$$

where

$$\alpha(t) = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p}\alpha_j w_j(t) \tag{2.7}$$

with $s(t)$ and $w_j(t)$, $j = 2, \ldots, p$, defined as before. This assumption implies that the functional form of the probability of an exceedance of the threshold level u is known. This is a considerably weaker assumption than assuming that the entire distribution function $\Psi_t(y)$ is known. The assumption in (2.6) implies knowledge only about the tail of the distribution of Y (assuming u is sufficiently large) rather than about the entire distribution.

Analyses of ozone data have been done by Smith (1989) and Shively (1991). Smith (1989) modeled the joint density of $\{(T_i, Y_i),\ 1 \le i \le N\}$ using a linear trend model for $\alpha(t)$, so $\alpha(t) = \alpha_1 + \alpha_2 t$, and assuming that $\beta(t)$ and $\xi(t)$ were constants. However, his model did not allow for meteorological variables. Shively (1991), in contrast, constructed a model in which all the meteorological variables were taken into account, but the analysis was based only on the exceedance times $\{T_i\}$ and not the values $\{Y_i\}$. The model in (2.2) combines the best of the methodologies previously proposed by Smith (1989) and Shively (1991).

3. Diagnostic Testing of the Model

This section outlines the diagnostic testing procedures we use to check that the assumptions used to derive the model in (2.2) are satisfied. These assumptions are listed below along with the methods we use to check that they are satisfied for our data.

(1) The exceedance times T_i, $i = 1, \ldots, n$, are modeled by a nonhomogeneous Poisson process with rate function given by (2.6) and (2.7).

(2) The density function

$$g(x;\beta(t)) = \lim_{u\to\infty} \frac{\psi_t(x+u)}{\Psi_t(u)} = \beta(t)\exp\{-\beta(t)x\}$$

is the appropriate model for the exceedance sizes $X_i = Y_i - u$, given that $Y_i \ge u$. By checking this assumption we are checking (i) the adequacy of the expression for $\beta(t)$ in (2.4); (ii) whether $u = 11.5$ pphm is a sufficiently high threshold level so that the

exponential density function $g(x;\beta(t))$ provides a good approximation to the actual density function for X; and (iii) whether (2.5) is the appropriate model for the exceedance sizes or if the more general model in (2.3) should be used.

(3) Exceedances that occur on days T_i and T_j are independent, i.e. we need to check that $X_i = Y_i - u$, given that $Y_i \geq u$, is independent of $X_j = Y_j - u$, given that $Y_j \geq u$, when $i \neq j$.

3.1 Assumptions for the nonhomogenous Poisson process

To check that the exceedance times T_i, $i = 1, \ldots, n$, are modeled by a nonhomogeneous Poisson process we check that (i) the inter-event times $S_i = T_i - T_{i-1}$ have the distribution function implied by a nonhomogeneous Poisson process; and (ii) the inter-event times are independent.

A probability plot is used to check that the inter-event times S_i have the distribution function implied by a nonhomogeneous Poisson process. To construct a probability plot the distribution function for the inter-event times, given the frequency function, is required. To begin, we write the frequency function $\Psi_r(u)$ in (2.6) as $\Psi(r)$. For the frequency function Ψ(r), and a given time t, the density function for the time to the next event (from t) is given by

$$f_t(s) = \Psi(t+s)e^{-h(s)} \tag{3.1}$$

where $h(s) = \int_t^{t+s} \Psi(r)dr$, and the distribution function is given by

$$F_t(s) = \int_0^s f_t(u)du \tag{3.2}$$

Due to the fact that the exceedance times for the ozone data are integer-valued (i.e. the first exceedance occurs on day t_1, the second exceedance occurs on day t_2, etc.), we only need to compute $F_t(s)$ for integer values of t and s. Taking this efficiency into account, it is relatively straightforward to show that the approximate distribution function is

$$F_t(s) = 1 - \exp\left\{-\sum_{k=1}^{s} \overline{\Psi}(k)\right\}$$

where $\overline{\Psi}(k) = \Psi(t+k)$.

To construct the probability plot, the inter-event time S_i is transformed to

$$U_i = F_{t(i-1)}(S_i)$$

where $t(i\text{-}1)$ is the time of $(i\text{-}1)$st event. If the inter-event times are independent with distribution function $F_t(s)$, then this transformation should reduce the inter-event time S_i to a random variable with a uniform distribution on (0,1). Following Smith (1986), we order the values U_i, and plot the ordered values against $i/(n+1)$, $i = 1, \ldots, n$. If the inter-event times have the density function given by (3.2), then this plot should be close to the straight line that forms a 45 degree angle with the horizontal axis.

To check that the inter-event times are independent we check that adjacent inter-event times are uncorrelated, i.e. that $S_i = T_i - T_{i-1}$ is uncorrelated with $S_{i-1} = T_{i-1} - T_{i-2}$. It is important to note that if random variables are uncorrelated, this does not necessarily imply that the variables are independent. However, from a practical point of view, if there is dependence between neighboring inter-event times, this will probably manifest itself with a strong serial correlation.

Given the inter-event times s_i, $i = 1, \ldots, n$, the correlation between adjacent inter-event times is computed using

$$r = \sum_{i=2}^{n} \left[\frac{s_i - E_{t(i-1)}(S_i)}{\sigma_{t(i-1)}(S_i)}\right]\left[\frac{s_{i-1} - E_{t(i-2)}(S_{i-1})}{\sigma_{t(i-2)}(S_{i-1})}\right]$$

where $E_{t(i-1)}(S_i)$ is the expected time between the $(i\text{-}1)$th and ith events, and $\sigma_{t(i-1)}(S_i)$ is the standard deviation of the distribution of the time between the $(i\text{-}1)$th and ith events.

A computationally efficient technique for computing $E_{t(i-1)}(S_i)$ and $\sigma_{t(i-1)}(S_i)$ is outlined below. For notational purposes, the i-1 and i subscripts are dropped. As before, we write the intensity function $\Psi_r(u)$ as $\Psi(r)$.

For the intensity function Ψ(r), and a given time t, the density function for the time to the next event $f_t(s)$ is given by (3.1). Then the expected time to the next event is

$$E_t(S) = \int_0^\infty s f_t(s) ds = \int_0^\infty s\Psi(t+s) \exp\left\{ \int_t^{t+s} \Psi(r) dr \right\} ds \tag{3.6}$$

While the integral in (3.6) is staightforward conceptually, it is computationally infeasible to integrate numerically using standard numerical integration techniques because of the integral in the exponential function. To make the integration computationally feasible, we divide the open interval $(0,\infty)$ into sections and show how to obtain an efficient technique for integrating the function over the individual sections of the real line. As with the distribution function for the inter-event times, t will be an integer.

Note that

$$E_t(S) = \int_0^\infty s f_t(s) ds = \sum_{k=0}^{\infty} \int_{k-1}^{k} s f_t(s) ds \tag{3.7}$$

For notational purposes, let $\overline{\Psi}(k) = \Psi(t+k)$. If day $t + k$ is not missing, then

$$\int_{k-1}^{k} s f_t(s) ds = \int_{k-1}^{k} s\Psi(t+s) e^{-h(s)} ds$$

$$= \int_{k-1}^{k} s\overline{\Psi}(k) \exp\left\{ -\left[\sum_{j=1}^{k-1} \overline{\Psi}(j) + (s-(k-1))\overline{\Psi}(k) \right] \right\} ds \tag{3.8}$$

because, for $s \in (k\text{-}1, k]$, $\Psi(t+s) = \overline{\Psi}(k)$ and

$$h(s) = \int_t^{t+s} \Psi(r) dr = \sum_{j=1}^{k-1} \overline{\Psi}(j) + (s-(k-1))\overline{\Psi}(k) \tag{3.9}$$

The expression for $h(s)$ in (3.9) follows immediately from the fact that $\Psi(r)$ is a step function. Integrating (3.8) by parts gives

$$\int_{k-1}^{k} s f_t(s)ds = \left[(k-1)+\frac{1}{\overline{\Psi}(k)}\right]\exp\left\{-\sum_{j=1}^{k-1}\Psi(j)\right\}+\left[k+\frac{1}{\overline{\Psi}(k)}\right]\exp\left\{-\sum_{j=1}^{k}\Psi(j)\right\} \qquad (3.10)$$

If day $t+k$ is missing, then $\int_{k-1}^{k} s f_t(s)ds = 0$

Combining (3.7) and (3.10) provides an expression for $E_t(S)$ that can be computed efficiently. In practice, we must truncate the infinite summation in (3.7) after a sufficiently large number of terms. We truncate the summation after 60 terms because the density $f_t(s)$ is very small after 60 terms, i.e. the probability that the inter-event times are greater than or equal to 60 days is negligible.

To compute the $\sigma_t(S)$, we use the expression

$$\sigma_t^2(S) = Var_t(S) = E_t(S^2) - [E_t(S)]^2$$

where

$$E_t(S^2) = \int_0^{\infty} s^2 f_t(s)ds = \int_0^{\infty} s^2 \Psi(t+s)\exp\left\{\int_t^{t+s}\Psi(r)dr\right\}ds.$$

$E_t(S^2)$ can be computed using a technique similar to the one outlined above to compute $E_t(S)$.

The standard error of the correlation coefficient is approximately $1/n^{1/2}$.

It is difficult to apply more sophisticated time series tests for the independence of inter-event times than the one we have used here because of the complicated nature of our model for exceedance times and inter-event times. The difficulties in applying the more sophisticated techniques to test for independence arise because (i) the inter-event times are not stationary; (ii) the inter-event times are not normally distributed; and (iii) we are using

a continuous time stochastic process to approximate the discrete time process of exceedance times. These problems combine to make it difficult to understand the properties of the tests for independence previously proposed in the literature when used in our problem.

3.2 Assumptions regarding the distribution and independence of the exceedance sizes

To check that the exponential distribution is the appropriate distribution function for the exceedance sizes, and that the correct expression for $\beta(t)$ is given by (2.4), we construct a probability plot. The technique for obtaining the probability plot is similar to the one outlined in Section 3.1 where a probability plot was constructed for the inter-event times.

To determine that the exceedance sizes are independent, we check whether exceedances that occur on consecutive days are related. This is accomplished by computing the correlation between the exceedance sizes that occur on consecutive days. There are two points to make with regard to this discussion. First, we compute the correlation only between exceedances that occur on consecutive days. The reasoning is that if these exceedances are unrelated, then exceedances more than one day apart will probably be unrelated. Second, as was noted in Section 3.1, if random variables are uncorrelated this does not necessarily imply independence. However, in practice if there is dependence between exceedances occurring on consecutive days it will probably manifest itself in a significant correlation coefficient.

Given the n_{adj} pairs of adjacent exceedance sizes, $\{s_{i-1}(t\text{-}1),\ s_i(t)\}$, where $s_i(t)$ reperesents the ith exceedance that occurs on day t, the correlation between adjacent exceedance sizes is

$$r = \sum \left[\frac{s_i(t) - E(S_i(t))}{\sigma(S_i(t))} \right] \left[\frac{s_{i-1}(t-1) - E(S_{i-1}(t-1))}{\sigma(S_{i-1}(t-1))} \right]$$

where the summation is over the n_{adj} pairs of adjacent exceedances, and $E(S_i(t)) = 1/\beta(t)$ and $\sigma(S_i(t)) = 1/\beta(t)$. The standard error of the correlation coefficient is approximately $1/n_{adj}^{1/2}$.

4. The Data

The data we consider in this paper consist of ozone exceedances over 11.5 pphm at the Clinton monitoring site in Houston, Texas during the months May-October over the period 1983-1992, together with daily values of the meteorological variables given below. The months May-October are considered the "high ozone" season and is the time of the year when the majority of the high values of ozone occur.

The daily ozone value used in our analysis is the maximum of the 13 hourly ozone readings taken each hour from 6am to 6pm. If 7 or more hourly observations are missing during this 13 hour period on a given day, then the ozone reading for this day is considered missing.

The meteorological variables used in the analysis are:

Maximum temperature (*TMAX*): Maximum of the hourly temperature readings for the period 6am to 6pm. The expected effect of higher temperatures is to increase ozone levels.

Temperature range (*TRANGE*): Difference between the minimum and maximum hourly temperature readings for the period 6am to 6pm. *TRANGE* is considered to be a proxy for the amount of sunlight occurring during the day. (A direct measure of sunlight is not available at the Clinton monitoring site.) The expected effect of large temperature swings is to increase ozone levels.

Average wind speed (*WSAVG*): Average of the hourly wind speed readings for the period 6am to 6pm. The expected effect of increased wind speed is to reduce ozone levels because higher wind speed tends to disperse pollutants present in the ambient air.

Wind speed range (*WSRANGE*): Difference between the minimum and maximum hourly wind speed readings for the period 6am to 6pm. Experts have conflicting views as to the effect of this variable. The empirical results given below suggest that large values of *WSRANGE* are associated with an increase in the probability of an exceedance.

The hourly wind direction readings are broken into four classifications: NW-NE, NE-ESE, ESE-SSW and SSW-NW. These wind directions correspond roughly to the different types of frontal passages through the Houston area (see Shively, 1991). For each wind direction we compute the percentage of time during the day when the hourly wind direction fell into each of these categories.

NW/NE: Percentage of time from 6am to 6pm that the wind direction was between NW and NE. The expected effect initially is to reduce ozone levels as a new air mass moves into the area from the North.

NE/ESE: Percentage of time from 6am to 6pm that the wind direction was between NE and ESE. The expected effect is to increase ozone levels because wind blowing from the NE-ESE direction tends to blow in pollution from the Beaumont-Port Arthur area and the Houston Shipping Channel.

ESE/SSW: Percentage of time from 6am to 6pm that the wind direction was between ESE and SSW. The expected effect is to increase ozone levels as there are several large industrial sources of pollution to the south of Houston.

SSW/NW: Percentage of time from 6am to 6pm that the wind direction was between SSW and NW. The expected effect is to reduce ozone levels because there are no major sources of pollution to the west of Houston. During the months May-October, the wind direction is seldom SSW/NW.

The missing data convention for the meteorological variables is the following: If more than 7 hourly readings in the period 6am to 6pm are missing, the data are considered to be missing for the day.

If either the ozone or meteorological variables are missing on a given day, then the data for that day are considered to be missing. If there were no missing observations we would have 1840 days of data. However, due to missing ozone and meteorological data, our data set is reduced to 1478 observations, i.e. approximately 20% of the daily data are missing.

Some exploratory analysis of the data is contained in Figures 1 and 2. Figure 1 simply shows a histogram of each of the eight meteorological variables. This is useful in giving some idea of the practical ranges of these variables. Figure 2 shows plots of the high-level ozone values against each of the covariates. The left-hand figure for each covariate was constructed by computing the proportion of ozone exceedances and an approximate 95% confidence interval for each value of the covariate. The sample proportions and upper and lower confidence limits are plotted. These figures are useful in gaining some idea of which are the significant factors — for example, comparing Figures 2.a.2 and 2.a.4 shows that the variation in *WSAVG* is far more significant than that in *TMAX* — but beyond that it is difficult to make definitive conclusions. One difficulty in interpreting these plots is that the meteorological variables are themselves highly correlated, so there is a considerable amount of confounding in the effects of the different variables. Thus, although the plots are useful in gaining an idea of what is going on, they are of only limited help in deciding which models to fit. The plots on the right-hand side are of actual ozone values over 11.5 pphm against the covariate. This gives an indication of whether the actual level of ozone, as opposed to just the probability that it is over the threshold, depends on the covariate. Again, the confounding of meteorological variables makes it difficult to draw definitive conclusions from these figures, but it is clear in several plots that there is dependence between ozone level and the covariate, so a realistic model should take this into account.

5. Analytical Results

To model the exceedance times and exceedance sizes we need to first determine which variables should be included in the expressions for $\alpha(t)$ and $\beta(t)$ in (2.7) and (2.4), respectively. The variables in $\alpha(t)$ can be interpreted to be those variables that are related to the frequency through time of the exceedances of the threshold level 11.5 pphm. Of particular interest is the coefficient α_1 associated with the trend variable because it represents the trend through time in ozone levels holding the important meteorological conditions constant. The coefficients α_2, ..., α_p provide a measure of the relationship between meteorological conditions and the frequency of high values of ozone. Similarly, the variables in $\beta(t)$ are interpreted to be those variables that are related to the size of the exceedances of the threshold level 11.5 pphm. To interpret the individual coefficients of $\beta(t)$, assuming $\xi \approx 0$, note that the expected value of the size of the exceedance on day t, given that an exceedance occurred on day t, is $1/\beta(t)$. Therefore, a downward trend in the size of the exceedances, holding meteorological conditions constant, would be indicated by a positive value of β_1 in (2.4). The coefficients β_2, ..., β_p provide a measure of the relationship between meteorological conditions and the exceedance sizes.

An important point with regard to the variable selection process for $\alpha(t)$ and $\beta(t)$ is that there is no reason the same set of variables need to be included in the expressions for $\alpha(t)$ and $\beta(t)$. In fact, due to the factorization of the density function in (2.2), we can select the variables to be included in $\alpha(t)$ and $\beta(t)$ independently. One would expect that for the most part meteorological conditions that are related to the frequency of exceedances would also be related to the size of the exceedances. The results given below confirm this. They show that *TRANGE*, *WSAVG*, and *NW/NE* are related to both the frequency and size of the exceedances. Two additional variables, *WSRANGE* and *NE/ESE*, also appear in the expression for $\alpha(t)$ but do not appear in the expression for $\beta(t)$.

To select the variables in $\alpha(t)$, we model the exceedance times using a nonhomogeneous Poisson process. The likelihood function for a nonhomogeneous Poisson process is the second term in square brackets in (2.2). We begin by including a constant term, a trend variable, and all the meteorological variables listed in Section 3 (except for *SSW/NW*, for reasons discussed below) in the expression for $\alpha(t)$, and then eliminate variables that are not statistically significant. Using this procedure, the following expression for $\alpha(t)$ is obtained:

$$\begin{aligned}\alpha(t) = & \underset{(0.034)}{-0.149\, s(t)} + \underset{(0.016)}{0.072\, TRANGE(t)} - \underset{(0.080)}{0.926\, WSAVG(t)} \\ & + \underset{(0.051)}{0.223\, WSRANGE(t)} - \underset{(0.408)}{0.850\, NW/NE(t)} + \underset{(0.398)}{1.432\, NE/ESE(t)} \qquad (5.1)\end{aligned}$$

The approximate standard errors of the estimated coefficients are given in parentheses. These standard errors are obtained by computing the inverse of the Fisher information matrix at the estimated parameters. To avoid multicollinearity among the wind direction variables, *SSW/NW* was not included as an explanatory variable in the original expression for $\alpha(t)$. The procedure we used for eliminating variables from the expression is as follows: If the absolute values of the ratios of the estimated coefficients to their standard errors (i.e. the test statistics for each coefficient) were all above 2.0, no variables were eliminated and the variable selection process was complete. Otherwise, the variable with the smallest value for its test statistic (in absolute value) was eliminated. The frequency function was then reestimated with the remaining variables and the procedure was repeated until no tests statistics had a value less than 2.0. If the distributions of the test statistics are approximately normal, then the value 2.0 is the approximate 5% critical value for a two-sided test of the null hypothesis that a given coefficient is zero against the alternative that the coefficient is not zero.

To obtain an expression for $\beta(t)$, we fit the exponential distribution in (2.5) to the exceedances of the threshold level 11.5 pphm. Using a procedure simlar to the one outlined above the final model for $\beta(t)$ is

$$\beta(t) = \underset{(0.011)}{0.035\, s(t)} - \underset{(0.005)}{0.016\, TRANGE(t)} + \underset{(0.019)}{0.102\, WSAVG(t)} + \underset{(0.179)}{0.400\, NW/NE(t)} \qquad (5.2)$$

The approximate standard errors of the estimated coefficients are given in parentheses.

The diagnostic tests for the model in (2.2) using the expressions in (5.1) and (5.2) for $\alpha(t)$ and $\beta(t)$, respectively, appear to be relatively well-satisfied. The correlation coefficient for adjacent inter-event times is 0.13 with a standard error of 0.09 so it is within two standard errors of zero. The correlation between adjacent exceedance sizes is 0.25 with a standard error of 0.18 so this correlation is also within two standard errors of zero. The probability plot in Figure 3 for inter-event times shows that the distributions of inter-event times specified by a nonhomogeneous Poisson process with frequency function given by (2.6) and (5.1) provide a good fit to the inter-event times actually observed. The probability plot in Figure 4 for the exceedance sizes shows that the exponential distribution in (2.5) provides a good fit to the observed exceedances.

The more general model (2.3) can be used to model the exceedances. We will assume that the paramter $\xi(t)$ is a constant not depending on time or the meteorological variables. Using the same variables in the expression for $\beta(t)$ as we did above, the estimate of ξ is -0.054 with a standard error of 0.063 so ξ is not significantly different from 0. Therefore the exponential model in (2.5) for the exceedances is the appropriate model to use.

5.1 Interpretation

The coefficients in the expression for $\alpha(t)$ can be interpreted as follows. To begin, by combining (2.6) and (5.1), the probability of an exceedance on day t of the threshold level 11.5 pphm is

$$\Psi_t(12) = \exp\{-0.149s(t) + 0.072TRANGE\ (t) - 0.926WSAVG\ (t)$$

$$+ 0.223WSRANGE\ (t) - 0.850NW/NE(t) + 1.432NE/ESE(t)\}$$

The negative coefficient -0.149 associated with $s(t)$ implies that after allowing for the confounding effects of meterological conditions, there is strong evidence of a downward trend through time in the frequency of exceedances, i.e. for meteorological conditions held constant the probability of an exceedance of the threshold level 11.5 pphm is decreasing through time. The estimated coefficients associated with the meteorological variables are as we would expect. The positive coefficient 0.072 associated with *TRANGE* imples that the larger the range of temperatures during the day, the higher the probability of an exceedance. Similarly, the negative coefficient -0.926 for *WSAVG* implies that high wind speeds are associated with a low probability of an exceedance. When the wind blows from NW-NE, the probability of a high ozone value decreases, while if the wind blows from NE-ESE the probability of an exceedance increases. The latter result is an interesting finding because when the wind blows from the NE-ESE, it is blowing into Houston from the Beaumont-Port Arthur area and across the Houston Shipping Channel. A possible explanation for the increase in the high levels of ozone associated with this wind is that the wind blows the chemical precursors to ozone into Houston (where they react to form ozone) that are released into the air in the Beaumont-Port Arthur area and/or at the Houston Shipping Channel.

The coefficients in $\beta(t)$ are interpreted as follows. Because the expected size of an exceedance is inversely related to $\beta(t)$, the positive coefficient 0.047 associated with $s(t)$ implies that after allowing for the confounding effects of meterological conditions, there is strong evidence of a downward trend in the exceedance sizes. Similarly, as *TRANGE* increases, the expected exceedance size increases, as *WSAVG* increases the expected

exceedance size decreases and when the wind blows from NW-NE, the exceedance sizes tend to be smaller than when the wind blows from other directions.

An interesting comparison is to estimate the trend in the frequency and size of the exceedances with and without the meteorological variables in the model. This comparison indicates the importance of taking into account the effects of meteorology when estimating the trends. If we omit the meteorological variables the expression we obtain for $\alpha(t)$ is

$$\begin{aligned} \alpha(t) = \quad & -2.104 - 0.069s(t) \\ & (0.172) \quad (0.030) \end{aligned}$$

with the approximate standard errors of the estimated coefficients given in parentheses. The model for $\beta(t)$ is

$$\begin{aligned} \beta(t) = \quad & 0.291 + 0.018s(t) \\ & (0.057) \quad (0.011) \end{aligned} \tag{5.3}$$

In comparing these expressions to the ones in (5.1) and (5.2), where the meteorological variables are included, we see that the estimated trends are considerably smaller in the model that does not accout for meteorological conditions. In fact, for the expression for $\beta(t)$ in (5.3), the trend is less than two standard errors from zero. Comparing the two sets of results indicates the importance of allowing for the effects of meteorological conditions when estimating the trends in the frequency and size of the exceedances of the threshold level specified by the EPA.

6. Conclusion

We have identified a class of models that may be used to model both the exceedance times and the exceedance levels as functions of time and meteorological covariates. The empirical results indicate that after accounting for the effects of meteorological conditions

there is evidence of a downward trend through time in both the frequency of exceedances of the EPA's specified threshold level and in the size of these exceedances.

Acknowledgements

Richard Smith's research was partially supported by the U.S. Environmental Protection Agency under Cooperative Agreement CR819638-01-0 through the National Institute of Statistical Sciences, and by NSF grant DMS-9205112. Tom Shively's research was partially supported by the CBA/GSB Faculty Research Committee of the College of Business Administration, The University of Texas at Austin.

References

Davison, A.C. and Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc.*, **52**, 393-442.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Series*. Springer Verlag, New York.

Pickands, J. (1975), Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Shively, T.S. (1991), An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment* **25B**, 387-396.

Smith, R.L. (1986), Extreme value theory based on the r largest annual events. *J. Hydrology* **86**, 27-43.

Smith, R.L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science* **4**, 367-393.

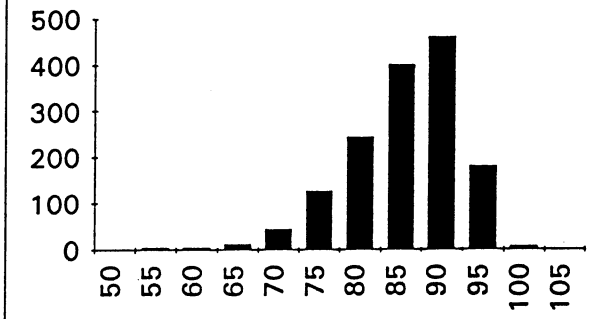

Figure 1.a, Histogram for TMAX

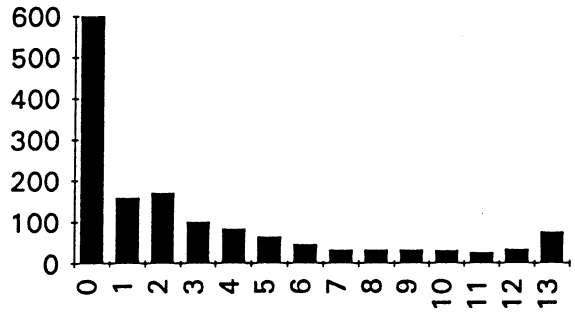

Figure 1.e, Histogram for NW/NE

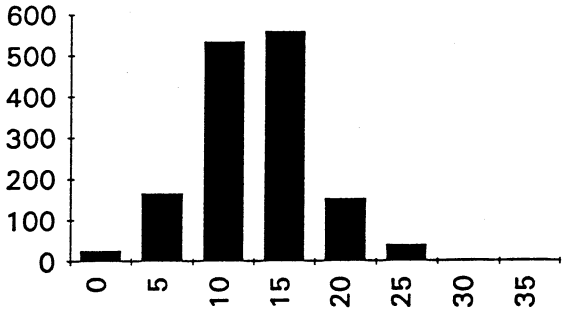

Figure 1.b, Histogram for TRANGE

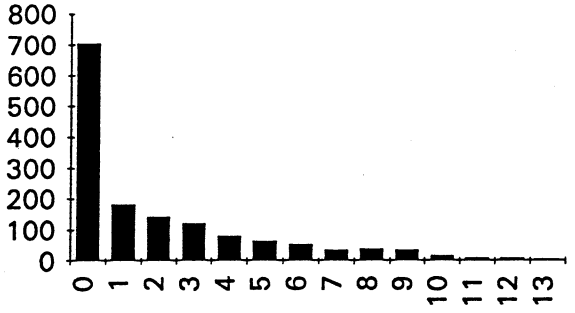

Figure 1.f, Histogram for NE/ESE

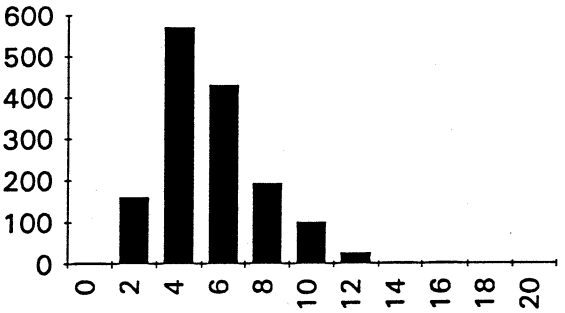

Figure 1.c, Histogram for WSAVG

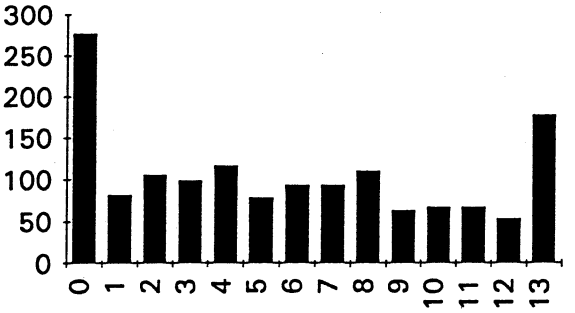

Figure 1.g, Histogram for ESE/SSW

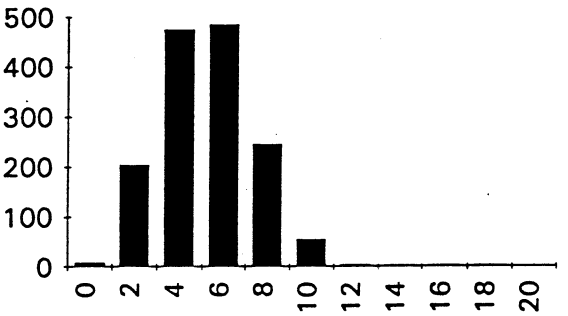

Figure 1.d, Histogram for WSRANGE

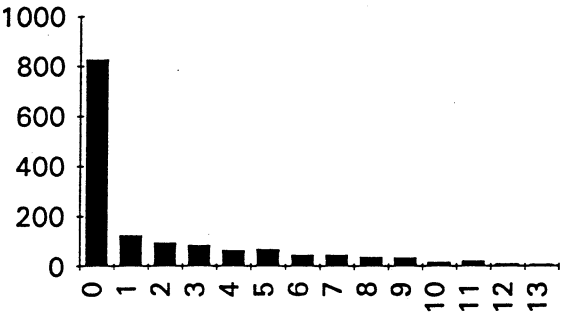

Figure 1.h, Histogram for SSW/NE

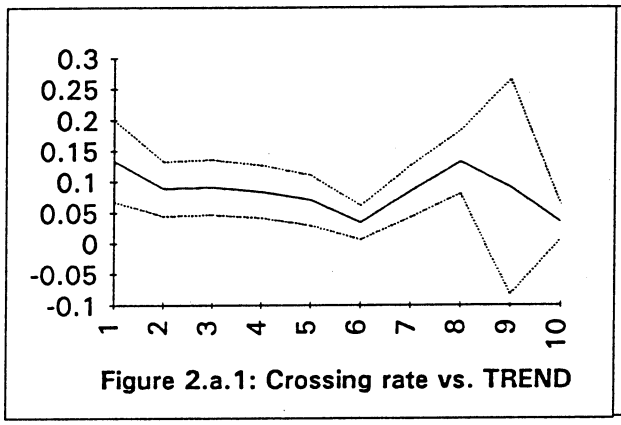

Figure 2.a.1: Crossing rate vs. TREND

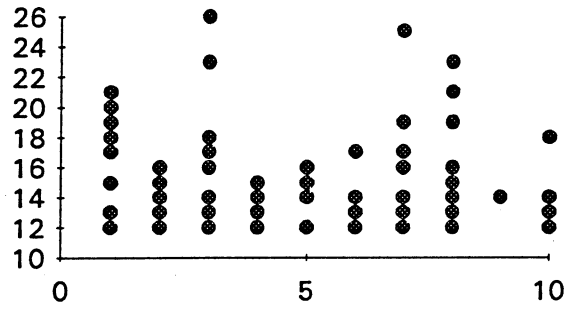

Figure 2.b.1: High-level Ozone vs. TREND

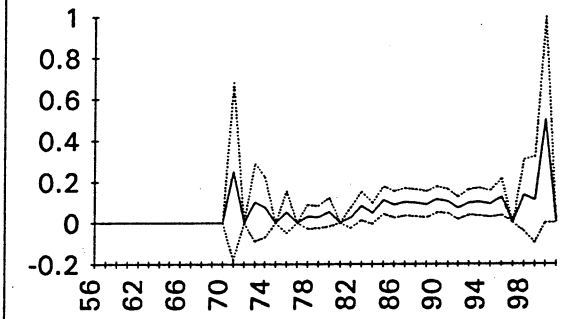

Figure 2.a.2: Crossing rate vs. TMAX

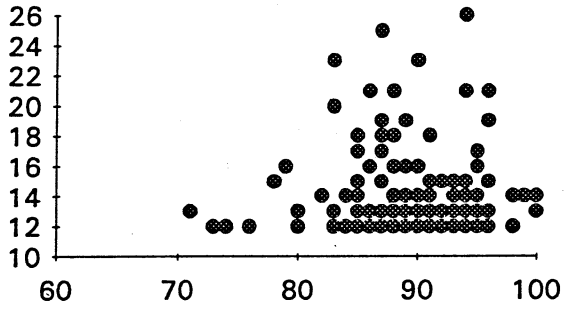

Figure 2.b.2: High-level Ozone vs. TMAX

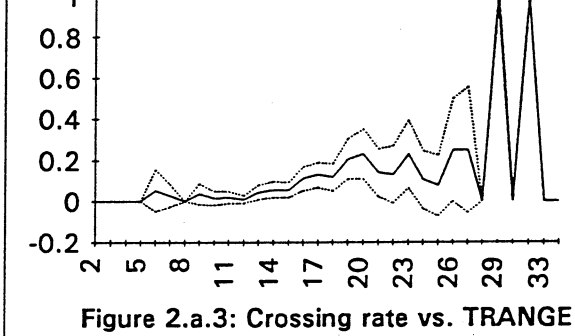

Figure 2.a.3: Crossing rate vs. TRANGE

Figure 2.b.3: High-level Ozone vs. TRANGE

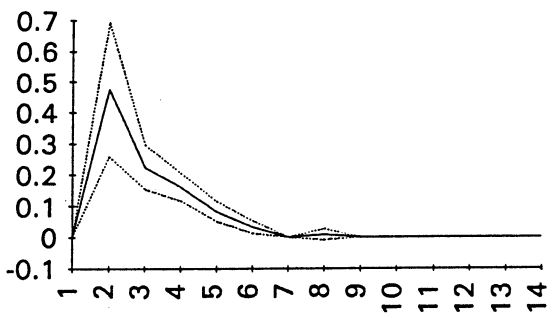

Figure 2.a.4: Crossing rate vs. WSAVG

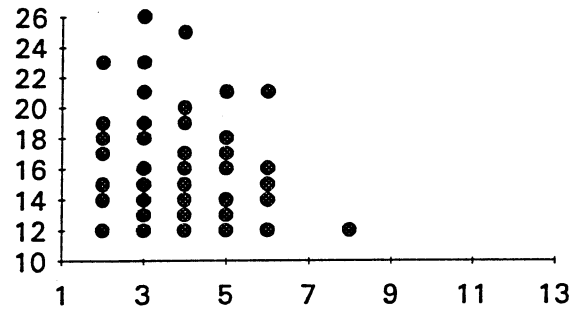

Figure 2.b.4: High-level Ozone vs. WSAVG

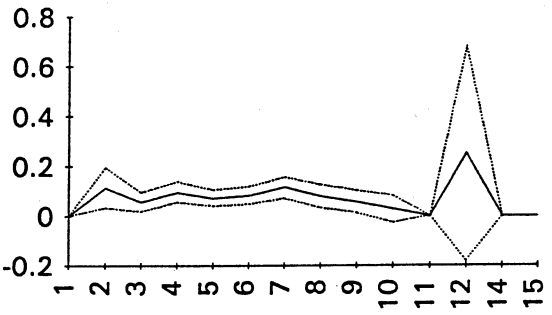

Figure 2.a.5: Crossing rate vs. WSRANGE

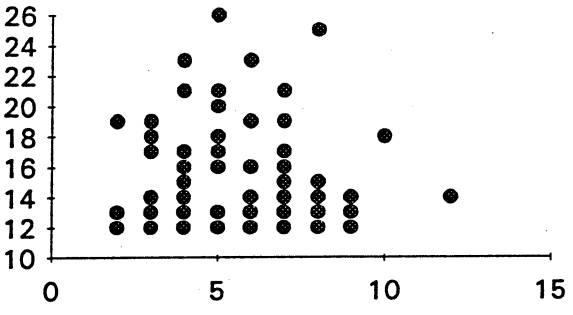

Figure 2.b.5: High-level Ozone vs. WSRANGE

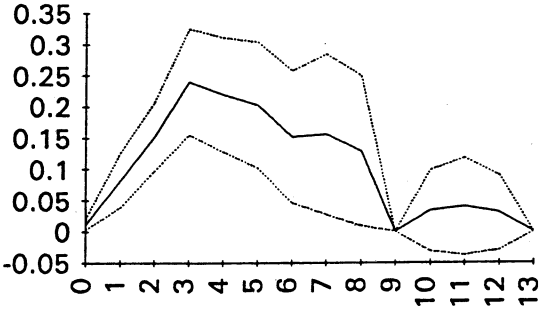

Figure 2.a.6: Crossing rate vs. NW/NE

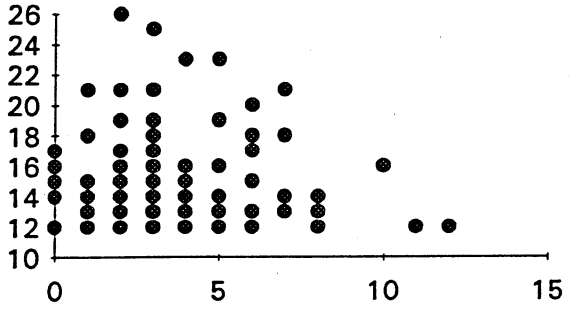

Figure 2.b.6: High-level Ozone vs. NW/NE

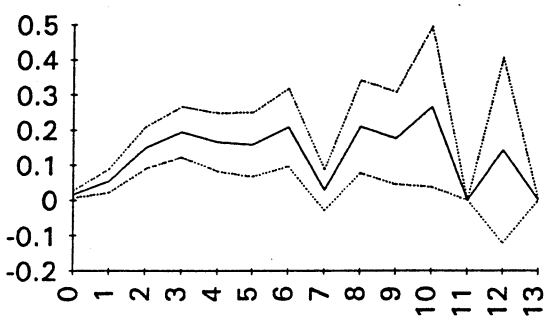

Figure 2.a.7: Crossing rate vs. NE/ESE

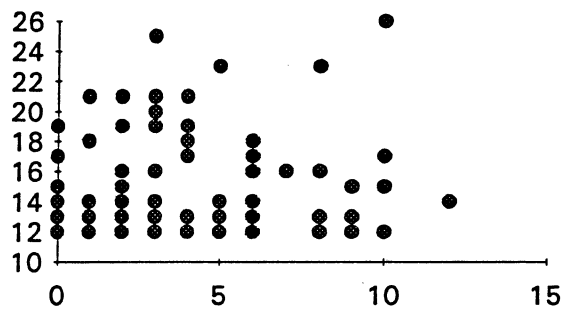

Figure 2.b.7: High-level Ozone vs. NE/ESE

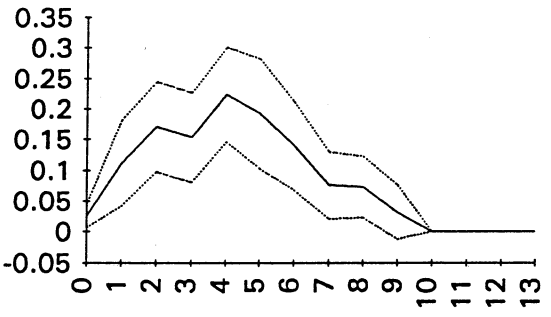

Figure 2.a.8: Crossing rate vs. ESE/SSW

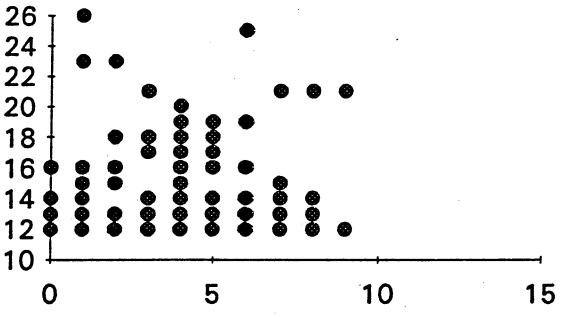

Figure 2.b.8: High-level Ozone vs. ESE/SSW

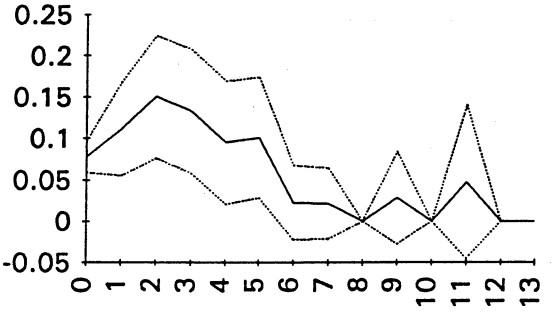

Figure 2.a.9: Crossing rate vs. SSW/NW

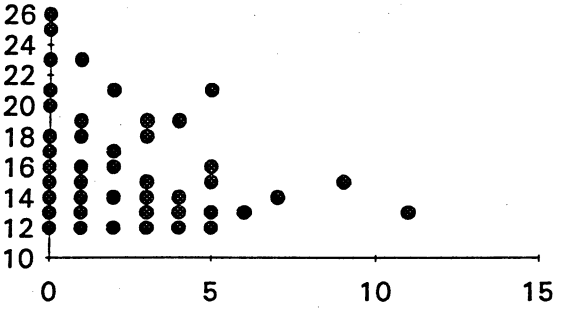

Figure 2.b.9: High-level Ozone vs. SSW/NW

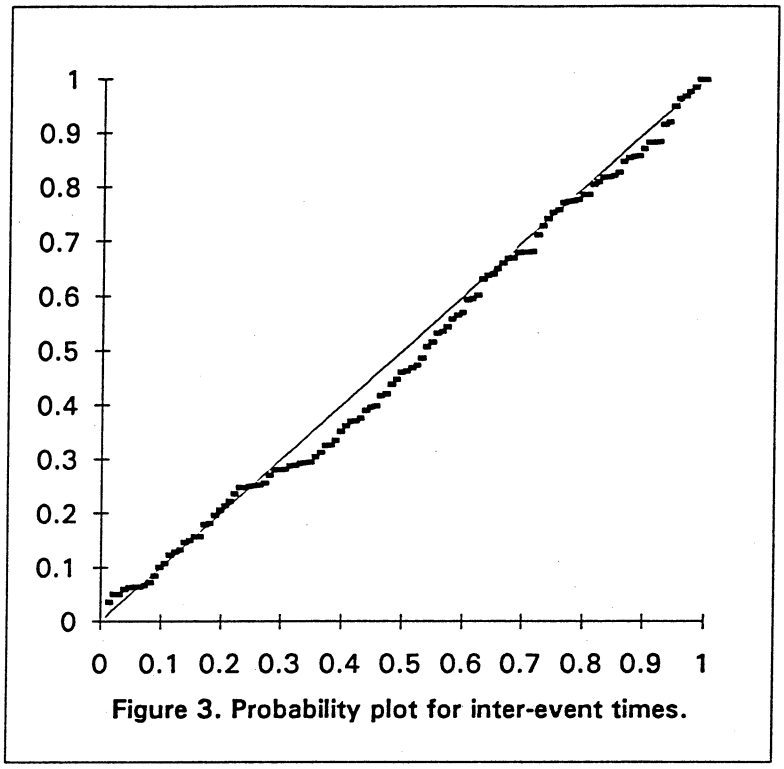

Figure 3. Probability plot for inter-event times.

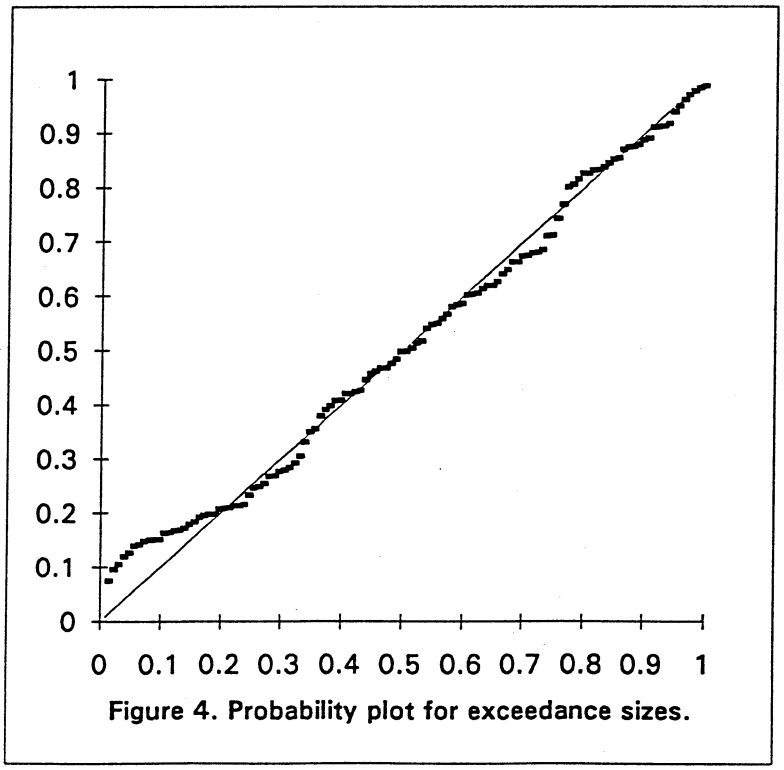

Figure 4. Probability plot for exceedance sizes.