Küchenhoff, Thamerus:

Extreme value analysis of Munich airpollution data

Sonderforschungsbereich 386, Paper 4 (1995)

Online unter: http://epub.ub.uni-muenchen.de/

Extreme value analysis of Munich air pollution data

Helmut Küchenhoff and Markus Thamerus

Abstract

We present three different approaches to model extreme values of daily air pollution data. We fitted a generalized extreme value distribution to the monthly maxima of daily concentration measures. For the exceedances of a high threshold depending on the data the parameters of the generalized Pareto distribution were estimated. Accounting for autocorrelation clusters of exceedances were used. To get information about the relationship of the exceedance of the air quality standard and possible predictors we applied logistic regression. Results and their interpretation are given for daily average concentrations of O_3 and of NO_2 at two monitoring sites within the city of Munich.

Keywords: air pollution, extreme values, generalized extreme value distribution, generalized Pareto distribution, logistic regression.

1 Introduction

Peak concentrations of single air pollutants are of particular interest in the determination of the air quality. Extreme concentrations of harmful atmospheric substances support the greenhouse effect and in addition to that their occurrence represents a risk to people's health and can lead to other environmental damages as a consequence. Thus the presence of extreme concentrations of ozone (O_3) or of nitrogen oxides (NO_x) always causes some sort of distress for the people living in the affected area. The direct effects to one's health range from slight disabilities to permanent damages. The prediction of these extreme concentrations and the assessment of their contribution to atmospheric pollution are subjects of strong environmental concern. Therefore extreme value concepts have recently been used in the monitoring of environmental data such as ground-level ozone (Smith, 1989) or as a tool in modelling trends in the urban ozone air quality (Rao et al., 1992).

During the summer season of the last few years German monitoring sites have registered, almost daily, the exceedances of the required national air quality standard for

ozone. The frequency of these exceedances has caused a political debate about new emissions control strategies. The discussion focuses on steps to reduce the emissions in case that extreme concentrations can be expected. The effect of traffic limitations for such situations is tested at present in field experiments at different locations in Germany. The surveillance of air quality is regulated by a federal law and therefore long-term data bases are available.

In this article we analyse extreme values of daily averages of ozone and nitrogen dioxide concentrations measured at two monitoring sites in Munich. A detailed description of our data is given in Section 2. In the third section we review the different statistical methods we use and give the results of our analysis. At first we model the distribution of extreme values by fitting the generalized extreme value (GEV) distribution to the monthly maxima of our series. The second model is based on the exceedances of high thresholds where we use the generalized Pareto distribution as a stable distribution for the excesses. To cope with the time dependent structure of the data exceedances occurring on subsequent days were linked to a cluster. Most definitions of air pollution standards only consider the fact that a certain limit is exceeded or the corresponding frequency of it. Therefore we fit logistic regression models with the binary response variable of limit exceedance. As predictors we include lagged concentrations, meteorologic variables and a weekend dummy in the model. The results indicate that these models can provide useful information about the occurrence of limit exceedances of air pollution and can be used to make short term predictions about limit exceedances.

2 The data

The data consist of daily averages of the concentration of ozone (O_3) and of nitrogen dioxide (NO_2) measured at two monitoring sites within the city of Munich (M-Stachus and M-Lothstr). The samples were obtained from the Bavarian State Office for Environmental Protection through a period from January 80 to October 92. The summary statistics of the observed data including the 95% and 99% quantiles are

given in Table 1 in more detail. Further we give an overview of the number of observed exceedances of the official air pollution standards for daily averages from 1989 of the German Engineering Association. The standard air quality level for ozone is 0.05 mg/m^3(about 25 ppb), for nitrogen dioxide it is 0.1 mg/m^3(about 53 ppb). The 95% quantile (0.1207 mg/m^3) of the NO_2 concentrations at the site M-Stachus exceeds its standard level and the 95% quantile (0.0816 mg/m^3) of the O_3 concentrations measured at M-Lothstr passes the standard as well.

	M-Stachus		M-Lothstr	
	O_3	NO_2	O_3	NO_2
time period	1/80 - 9/92	1/84 - 10/92	3/89 - 9/92	1/84 - 10/92
sample size	3871	2984	1271	2896
mean	0.0149	0.0718	0.0354	0.0514
s.e.	0.0124	0.0274	0.0252	0.0214
q_{50}	0.0106	0.0677	0.0326	0.0486
q_{95}	0.0404	0.1207	0.0816	0.0857
q_{99}	0.0555	0.1615	0.0976	0.1313
exceedances	77	392	358	74
rate of exc.	0.0199	0.1314	0.2817	0.0256

Table 1. Summary statistics of the daily averages of concentrations including the median (q_{50}), the 95% (q_{95}) and 99% (q_{99}) quantiles. The exceedances are the number of data above the standard air quality level (0.05 mg/m^3 for O_3 and 0.1 mg/m^3 for NO_2).

Different traffic conditions (high traffic at M-Stachus, low traffic at M-Lothstr) cause different exceedance rates of O_3 and NO_2 at both sites. Since our main interest

concerns the extreme values, we confine our analysis on NO_2 at the site M-Stachus and O_3 at the site M-Lothstr. The plots of these time series, including moving averages, are given in Figures 1 and 2. After trying MA filters of different orders we present the O_3-series with the MA(16) and the NO_2-series with the MA(20). Some of the difficulties typical for the application of extreme value concepts to samples of daily air pollution data can be seen. All observed concentrations show a positive short range correlation and in the case of ozone a pronounced seasonal variation is found.

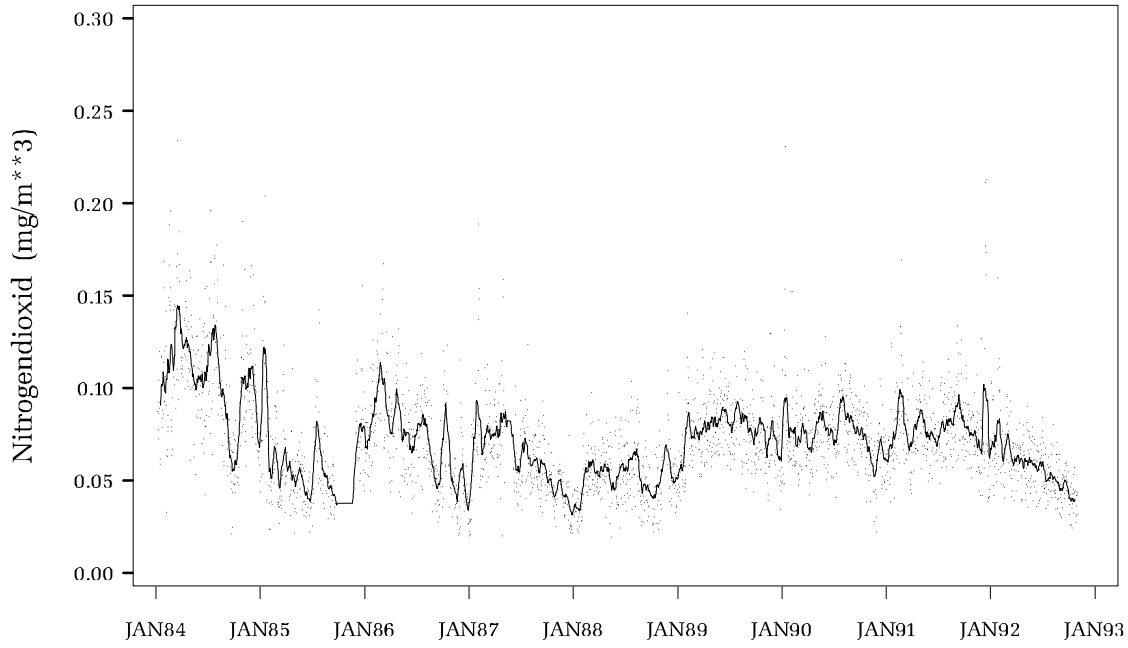

Figure 1. Time series plot and moving averages of order 20 for the NO_2 concentrations (mg/m^3) at the site M-Stachus from January 84 to October 92.

A balanced reaction of oxygen and nitrogen dioxide with the photochemical ozone and nitrogen monoxide takes place all the time in our atmosphere. This reaction is introduced by sunlight and therefore this chemical process is encouraged by the meteorological conditions during the summer season, where extreme values of ozone are exclusively observed. With the help of Box-and-Whisker plots of the yearly observed maxima for each month we find a different behavior of ozone for the months from April to August compared to the other months of the year. Therefore we split the data into these two seasons and fit two different extreme value models.

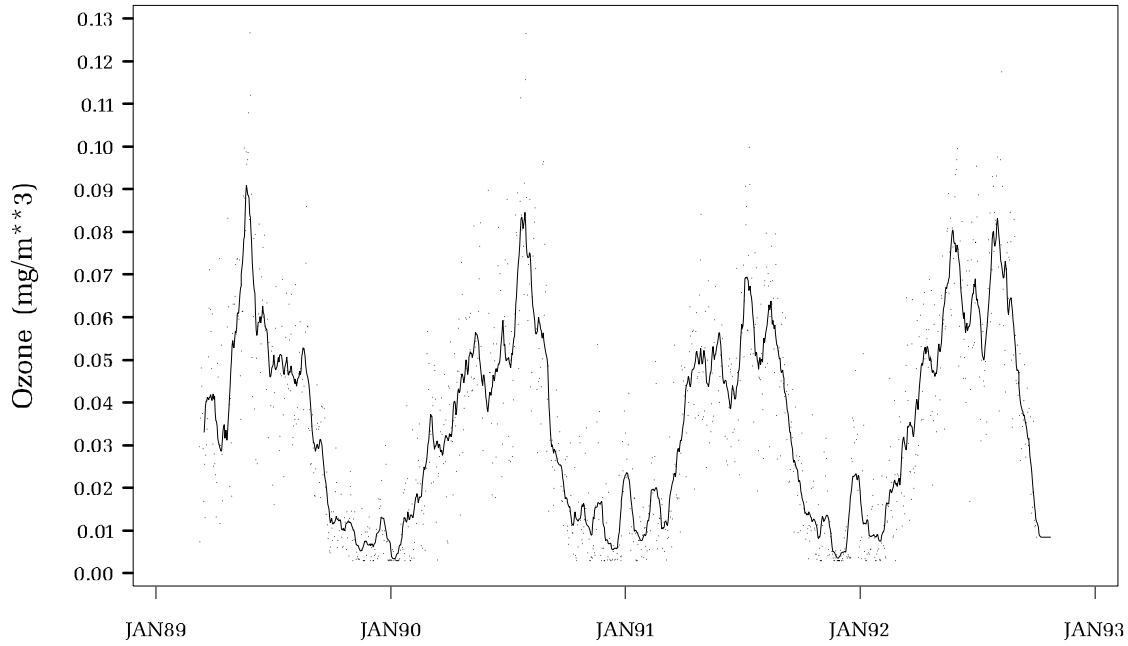

Figure 2. Time series plot and moving averages of order 16 for the O_3 concentrations (mg/m^3) at the site M-Lothstr from March 89 to September 92.

The ozone concentrations at the site M-Lothstr show that during summer time the official air pollution standard of 0.05 mg/m^3 is permanently exceeded. This is confirmed by the marked MA(16) process of the daily ozone averages which climbs over the limit as well during this time.

3 The models

3.1 The generalized extreme value distribution

Extreme value distributions are usually considered to correspond to one of three families. These families, first introduced by Fisher and Tippett (1928), were derived as the limiting forms of the distribution of maxima in samples of i.i.d. random variables. To each type of extreme value distribution belongs a set of parent distributions, called 'domains of attraction', for which necessary and sufficient conditions for the derivation of the limit are satisfied (see e.g. Resnick, 1987 or Leadbetter et al., 1983). For

statistical purpose it is convenient to combine the three types into the generalized extreme value (GEV) distribution. Its distribution function is given by

$$\mathrm{H}\,(x;\mu,\alpha,k) = \begin{cases} \exp\left\{-\left[1-\frac{k(x-\mu)}{\alpha}\right]^{\frac{1}{k}}\right\} & \text{for} \quad k \neq 0 \\ \exp\left\{-\exp\left[-\frac{(x-\mu)}{\alpha}\right]\right\} & \text{for} \quad k = 0 \end{cases} \tag{1}$$

with $x < \mu + \frac{\alpha}{k}$ for $k > 0$ and $\mu + \frac{\alpha}{k} < x$ for $k < 0$, respectively, and $-\infty < x < \infty$ for $k = 0$. The Fisher-Tippett-distributions of type I, II and III correspond to the cases $k = 0,\ k < 0,$ and $k > 0$.

The theory requires that the sample of extreme values $x = (x_1\, x_2 \ldots x_n)$ should each be drawn from single i.i.d. samples. Leadbetter et al. (1983) have shown that under certain conditions the distribution of extreme values in stationary stochastic processes corresponds to the same type of family as the distribution of extreme values in i.i.d. samples.

The most common method for the estimation of the parameters of the GEV distribution (1) is that of maximum likelihood. In small or moderate samples it sometimes appears that the loglikelihood function does not have a local maximum. Therefore we use the method of probability-weighted moments (PWM) proposed by Hosking et al. (1985) to estimate the GEV parameters . They have shown that the PWM estimators yield small sample properties ($n = 15,\ 25$) favourable to those estimators obtained by the method of maximum likelihood and that they have comparable standard deviations for moderate sample sizes ($n = 50,\ 100$).

The probability-weighted moments (PWM) of a random variable X with distribution function F are given by

$$M_{(p,\,r,\,s)} = \mathrm{E}\,\left[X^p\{\mathrm{F}\,(X)\}^r\{1-\mathrm{F}\,(X)\}^s\right]$$

where $p,\ r,\ s$ are real numbers. The usual noncentral moments are the quantities $M_{(p,0,0)},\ p = 0,1,2,\ldots$. For estimating the parameters of the GEV distribution we only consider moments of the form

$$\beta_r = M_{(1,\,r,\,0)} = \mathrm{E}\,\left[X\{\mathrm{F}\,(X)\}^r\right]\ ,\ r = 0,1,2,\ldots$$

If F is chosen as the GEV distribution the probability-weighted moments for $k \neq 0$ result in

$$\beta_r = (r+1)^{-1} \left[\mu + \alpha \left\{ 1 - (r+1)^{-k} \, \Gamma(1+k) \right\} / \, k \right] \; ; \; k > -1. \tag{2}$$

For a random sample $x_1, x_2, \ldots, x_n$ of i.i.d. variables $X_i, i = 1, \ldots, n$, the moments are estimated using the order statistics $x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$. An unbiased estimator of β_r is given by

$$\hat{\beta}_r = n^{-1} \sum_{j=1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} \, x_{[j]}. \tag{3}$$

Relation (2) implies the following three equations:

$$\text{(i)} \qquad \beta_0 = \mu + \alpha \left\{ 1 - \Gamma(1+k) \right\} / \, k$$

$$\text{(ii)} \qquad 2\,\beta_1 - \beta_0 = \alpha \, \Gamma(1+k) \left(1 - 2^{-k} \right) / \, k$$

$$\text{(iii)} \qquad (3\,\beta_2 - \beta_0) / \, (2\,\beta_1 - \beta_0) = \left(1 - 3^{-k} \right) / \, \left(1 - 2^{-k} \right).$$

The estimators for the distribution parameters μ, α and k are obtained by the solution of the equations (i)-(iii) whereby the unknown parameters β_r are replaced with their estimators (3). For the estimation of the shape parameter k the equation (iii) is solved by using a polynomial approximation proposed by Hosking et al. (1985).

Our random samples are the monthly maxima of daily concentration measurements. Since the maxima are nearly always separated by a few days, the assumption of independence for the maxima $X_i, i = 1, \ldots, n$, seems reasonable. Smith (1989) established a similar assumption for the 24-hour maxima of hourly readings of ozone concentrations. Our results show that the presence of autocorrelation did not interfere with the application of extreme value models.

Extreme value distributions of type I ($k = 0$) are also called Gumbel distributions (Gumbel, 1958) and most extreme value applications refer to distributions of that type. Hosking et al. (1985) derived a statistical test for the hypothesis that the shape parameter k equals zero in the GEV distribution which is equivalent to the presence of a Gumbel distribution. The corresponding test statistic z is given by $z = \hat{k}\sqrt{n/0.5633}$, which is standard normally distributed under the null hypothesis.

Results

In Table 2 we give a summary of the sample maxima used for the analysis.

		n	$mean$	$s.e.$	q_{50}	q_{95}	max
NO_2 (M-Stachus)		105	0.1157	0.0388	0.1104	0.1959	0.2338
O_3 (M-Lothstr)	summer	20	0.0896	0.0178	0.0851	0.1266	0.1267
	winter	24	0.0439	0.0147	0.0466	0.0646	0.0712

Table 2. Summary statistics of the monthly maxima including the median (q_{50}) and the 95% (q_{95}) quantile. The summer season is the period April-August and the winter season September-March.

A histogram of the NO_2 maxima can be found in Figure 3. For illustration we add a kernel density estimator using a Gaussian kernel.

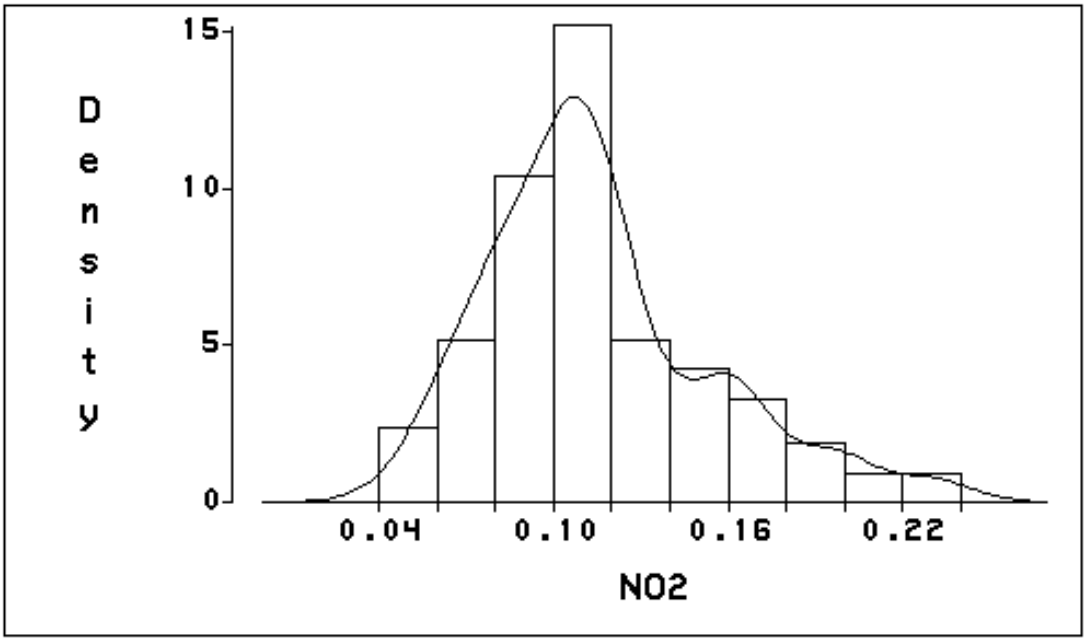

Figure 3. Histogram of the monthly observed NO_2 maxima at the site M-Stachus. The solid line represents the estimated Gaussian kernel density.

As indicated in Section 2 we apply two different models for the monthly maxima of the ozone concentrations. The histogram (Figure 4) includes the maxima of all months and clearly shows two peaks near the medians of the different seasons (q_{50}=0.0851 for the summer and q_{50}=0.0466 for the winter months).

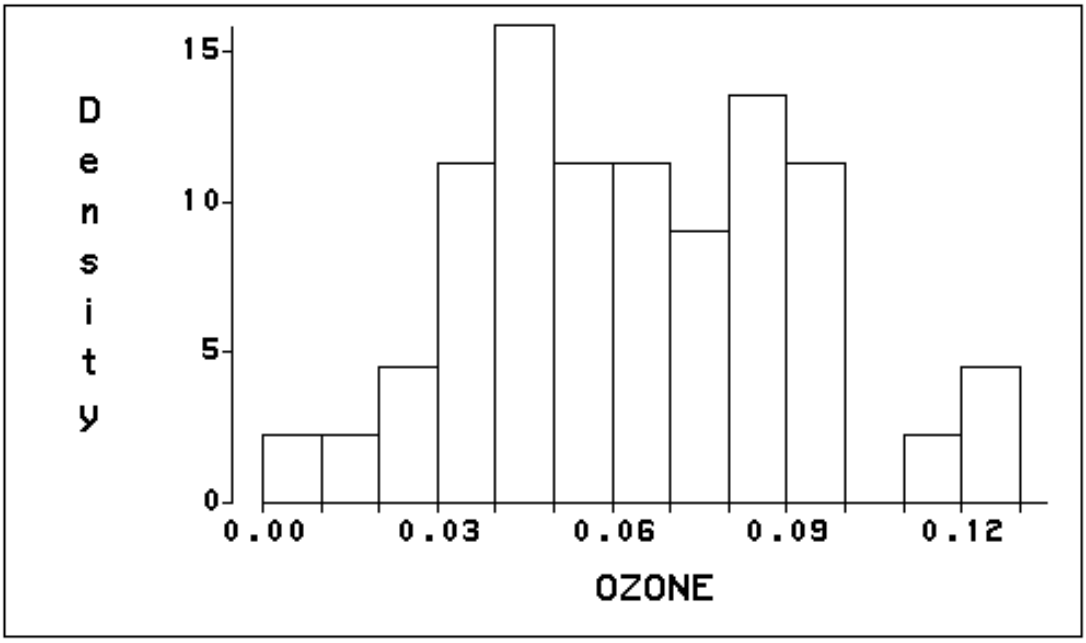

Figure 4. Histogram of the monthly observed O_3 maxima for both seasons at the site M-Lothstr.

For all data we fit the GEV distribution by the PWM method and estimators for standard errors of the parameters are obtained by using the proposed weights for the elements of the asymptotic covariance matrix of the PWM estimators (see Hosking et al., 1985). The results are summarized in Table 3.

The estimated GEV distributions of the monthly ozone maxima show a different location for the two seasons comparable to those of the empirical distributions. Only for the winter model at the site M-Lothstr the hypothesis of the presence of a Gumbel distribution has to be rejected. At this site only small samples of ozone data are available which results in bigger standard errors of the parameter estimates.

		n	k	μ	α
NO_2 (M-Stachus)	all year	105	-0.0132	0.0979	0.0302
			(0.0732)	(0.0033)	(0.0025)
O_3 (M-Lothstr)	summer	20	-0.0499	0.0810	0.0137
			(0.1678)	(0.0035)	(0.0026)
	winter	24	0.4964	0.0402	0.0159
			(0.1565)	(0.0036)	(0.0026)

Table 3. GEV distribution parameter estimates and their standard errors.

The fit of the models can be checked by a quantile-quantile (qq-) plot. In all cases the estimated GEV distribution gives an acceptable fit for the monthly maxima of the concentrations. As an example we present the qq-plot for the NO_2 concentrations at the site M-Stachus. Figure 5 shows that the data points scatter around the expected straight line and that the GEV model gives a good fit for the distribution of the NO_2 maxima.

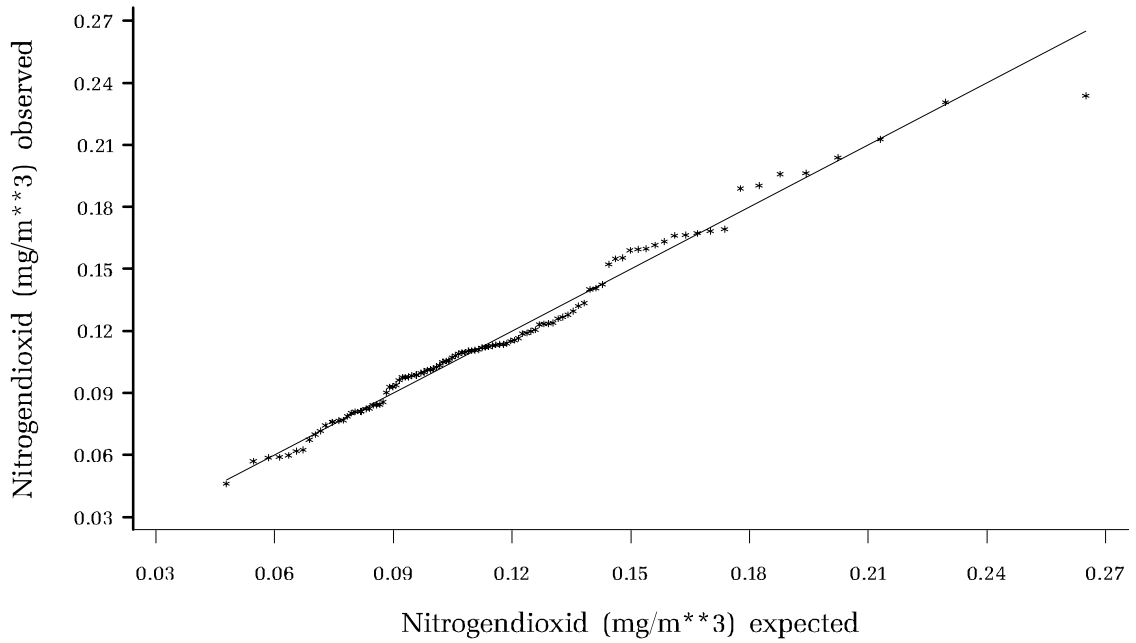

Figure 5. QQ-plot for the GEV model of the monthly NO_2 maxima at the site M-Stachus.

A quantity of interest in environmental studies is the N-month-return-level. It is defined as the value of the observed concentrations that can be expected to be once exceeded during a N month period. For each fitted distribution it is given by $\mathrm{F}^{-1}(1-\frac{1}{N})$; $N > 1$. Standard errors are derived by the δ-method using the estimated covariance matrix of the PWM estimates.

NO_2 (M-Stachus)		O_3 (M-Lothstr)		
N	all year model	N	summer model	winter model
2	0.1090 (0.0037)	2	0.0860 (0.0039)	0.0456 (0.0035)
4	0.1358 (0.0049)	4	0.0986 (0.0052)	0.0550 (0.0033)
6	0.1498 (0.0058)	6	0.1054 (0.0062)	0.0585 (0.0032)

Table 4. Estimated N-month-return levels (GEV) and their standard errors. The official standards for O_3 and NO_2 are 0.05 mg/m^3 and 0.1 mg/m^3.

In Table 4 we give the return levels for the NO_2 data at the site M-Stachus and for the O_3 data at the site M-Lothstr including their standard errors. The table shows that the standard for NO_2 (0.1 mg/m^3) can be expected to be once exceeded every two months. The N-month-return-levels for the ozone concentrations at the site M-Lothstr indicate that even during a period of three winter months the limit of 0.05 mg/m^3 is once passed on average. For days with peaks of ozone in the summertime very high concentrations can be expected, e.g. in four months the double limit value is once exceeded on average.

3.2 The generalized Pareto distribution

An alternative approach to the problem of analysing extreme values is to model the exceedances over high thresholds. It is based on using the generalized Pareto distribution (GPD), first proposed by Pickands (1975). The first step in the application of threshold models consists in the selection of an appropriate threshold, then the

differences between the observations above the threshold and the threshold itself can be fitted by the GPD (see Davison and Smith, 1990).

The generalized Pareto distribution of a random variable Y is defined as

$$\mathrm{G}\left(y;\alpha,k\right)=\begin{cases}1-\left(1-k\frac{y}{\alpha}\right)^{\frac{1}{k}} & \text{for} \quad k\neq 0\\ \\ 1-\exp\left(-\frac{y}{\alpha}\right) & \text{for} \quad k=0,\end{cases} \tag{4}$$

where $\alpha>0$ and the range of y is $0<y<\infty$ if $k\leq 0$, in the case of $k>0$ the range is determined by $0<y<\frac{\alpha}{k}$. The motivation for equation (4) is the following: our aim is the derivation of the distribution of a random variable $W-u$ under the condition that $W\geq u$, where the values of W are the observations and u is the threshold. Let F denote the unknown distribution function of the random variable W and F_u the conditional distribution of $Y=W-u$ given $W\geq u$,

$$\mathrm{F}_u\left(y\right)=\frac{\mathrm{F}\left(u+y\right)-\mathrm{F}\left(u\right)}{1-\mathrm{F}\left(u\right)} \quad \text{for} \quad 0\leq y\leq w_o,$$

where $w_o\leq\infty$ is the upper boundary of the distribution of W. For $u\rightarrow w_o$ the distribution F_u can be approximated by the GPD (see e.g. Pickands, 1975 or Smith, 1984).

Since Hosking and Wallis (1987) have shown that for samples with less than 500 observations the estimators derived by the method of probability-weighted moments are more reliable than the ML-estimators for the parameters of the generalized Pareto distribution we apply the PWM-method here as well. For the estimation of the GPD parameters we use the probability-weighted moments

$$\alpha_s=M_{(1,0,s)}=\mathrm{E}\left[Y\{1-\mathrm{F}\left(Y\right)\}^s\right]=\frac{\alpha}{\left(s+1\right)\left(s+1+k\right)} \tag{5}$$

of a random variable Y, which exist for $k>-1$. The GPD parameters α and k can be estimated with the help of the moments α_0 and α_1 by the following equations

$$\alpha=\frac{2\,\alpha_0\,\alpha_1}{\alpha_0-2\,\alpha_1}\;, \quad k=\frac{\alpha_0}{\alpha_0-2\,\alpha_1}-2\;\;. \tag{6}$$

For an ordered sample $y_{[1]}\leq y_{[2]}\leq\ldots\leq y_{[n]}$ of Y an unbiased estimator $\hat{\alpha}_s$ for α_s is given by

$$\hat{\alpha}_s=n^{-1}\sum_{j=1}^{n}\frac{\left(n-j\right)\left(n-j-1\right)\ldots\left(n-j-s+1\right)}{\left(n-1\right)\left(n-2\right)\ldots\left(n-s\right)}y_{[j]}\;. \tag{7}$$

The GPD parameter estimates (6) are obtained by replacing the α_s in (5) with their estimators $\hat{\alpha}_s$ (7), which are asymptotically normally distributed for $k > -0.5$ (see Hosking and Wallis, 1987).

The selection of an appropriate threshold has been discussed by Davison and Smith (1990). A useful device for the choice of possible threshold levels for the models is a mean residual life plot where the mean excess of a threshold u is plotted against the threshold u. If the excess Y of a high threshold u_0 follows a GPD with parameters α and k, the plot should approximate a straight line with slope $-k/\,1+k$ for $u \geq u_0$. Therefore the starting points of straight lines in the mean residual life plot are selected as possible thresholds.

Davison and Smith also suggested estimators for further quantities: the crossing rate λ of the threshold and the N-month-return level. The exceedance process over a fixed threshold u is assumed to be Poisson with rate λ, where the size of the exceedances are assumed to be independent of λ. Then an unbiased estimator for λ is $\hat{\lambda} = n/\,l$ with estimated variance $n/\,l^2$, where n is the number of exceedances and l is the number of observed intervals (we use monthly intervals). The N-month-return level is obtained by $q_N = u + \alpha\frac{1-(\lambda N)^{-k}}{k}$, which can be estimated by replacing the parameters with their estimators. Standard errors of the estimated return levels are computed by the δ-method using the asymptotic covariance matrix for the parameter estimates.

Results

The threshold models are fitted for the O_3 concentrations at the site M-Lothstr and for NO_2 at M-Stachus. In both cases the estimated threshold models give a good fit for the distribution of the exceedances if the threshold is chosen high enough. The use of the official limits as threshold levels was not practicable because the rate of exceedances would have been too high. For the ozone data threshold models are applied only to the concentration measures of the months May - August. The frequency of extreme values recorded in April was too low, so the data of this month are not included in the models. During the four months of the summer period the estimated crossing rate of the official air pollution standard for ozone is 18.75 per

month, the median of all ozone concentrations is 0.057 mg/m^3, a value that already exceeds the standard. For both air pollutants we use several threshold levels suggested by the mean residual life plot of the exceedances. The mean residual life plot for the NO_2 data (Figure 6) suggests two thresholds (0.13 and 0.16 mg/m^3), the change of slope indicates different distribution parameters for the exceedances. The irregular behavior of the plot at his right hand side is due to the small number of exceedances over the highest threshold levels.

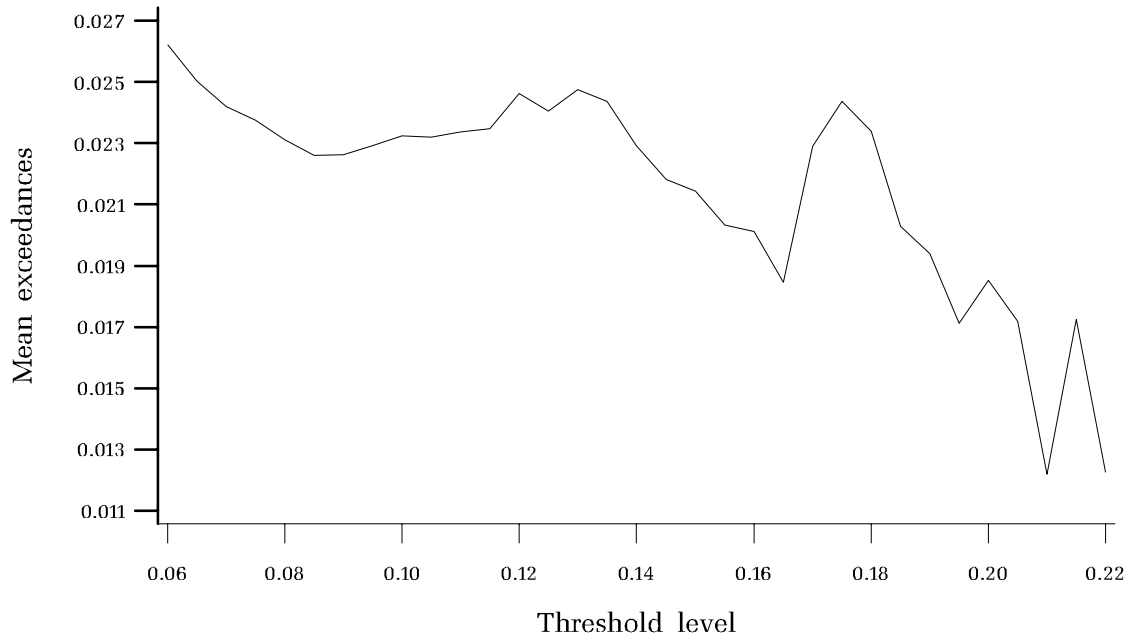

Figure 6. Mean residual life plot of the mean exceedances of the threshold u of the NO_2 concentrations (mg/m^3) at the site M-Stachus against u (mg/m^3).

The mean residual life plot of the ozone data indicates to take thresholds at 0.07 and 0.08 mg/m^3. For estimating the GPD parameters the dependence structure of the exceedances has to be taken into account. One way to deal with that problem is the assumption of a clustered Poisson process for the exceedances (see e.g. Smith, 1984). Since we find that the dependence structure of our series is similar to that of an AR(1), we link subsequent days of exceedances to clusters and use the maxima of these clusters which can be assumed as independent for our analysis. We estimate the GPD parameters for the different models by the PWM method and give the results

in Table 5. The standard errors of the estimators are obtained using the asymptotic covariance matrix of the parameter estimators (see Hosking and Wallis, 1987).

		O_3 (M-Lothstr)		NO_2 (M-Stachus)	
Threshold	u	0.07	0.08	0.13	0.16
Exceedances	n	54	31	53	21
Crossing rate	$\hat{\lambda}$	3.3750	1.9375	0.5000	0.1981
		(0.4593)	(0.3480)	(0.0687)	(0.0432)
Parameters	$\hat{k}$	0.0852	-0.0783	0.1273	0.0151
		(0.1609)	(0.2068)	(0.1653)	(0.2527)
	$\hat{\alpha}$	0.0150	0.0114	0.0324	0.0245
		(0.0031)	(0.0031)	(0.0068)	(0.0082)

Table 5. GPD parameter estimates and standard errors for the cluster maxima of the exceedances.

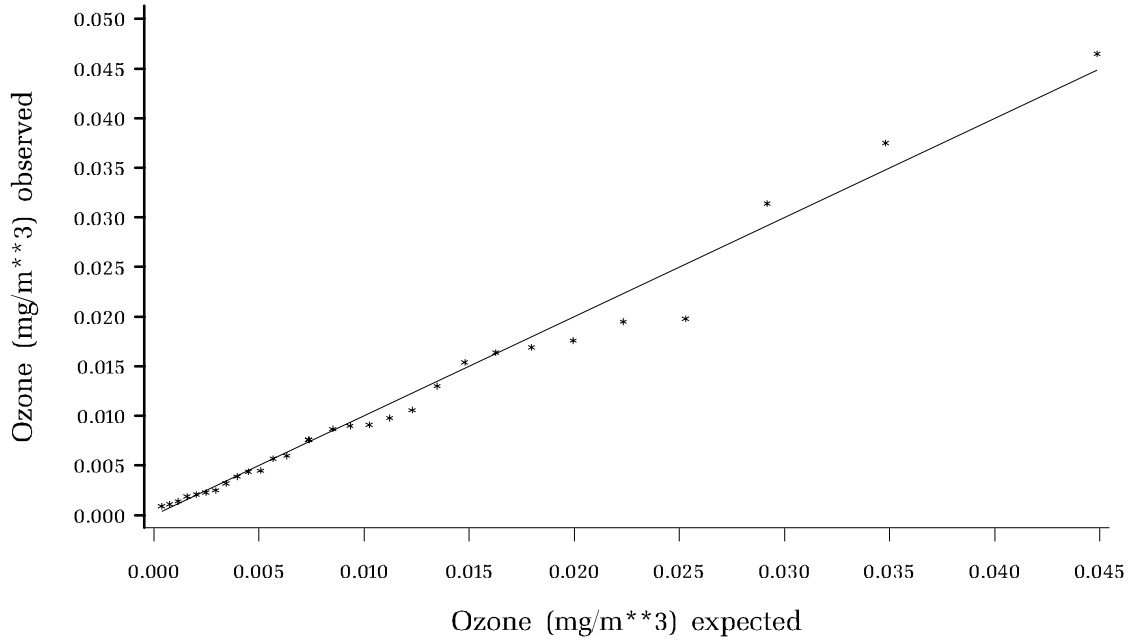

Figure 7. QQ-plot for the threshold model of the O_3 concentrations with u = 0.08 mg/m^3 at the site M-Lothstr.

Since unclustered data have been used to find possible thresholds by the graphical method mentioned above, the model fit has to be further checked. The qq-plots for the threshold models of ozone (Figure 7) show satisfactory results, the model with threshold at 0.08 mg/m^3 appears to fit the data best. A common feature of all estimated models is that the bulk of the data fit the model quite well and that the scattering of the data points around the line increases at the right end of the qq-plot. The effect that the largest order statistics were underestimated is found in both models, it is lowest for the model with the threshold taken at $u = 0.08$ mg/m^3. The plot also shows that these observations are separated from the other cluster maxima by their high values. Davison and Smith (1990) discussed the sensitivity of the GPD parameter estimators to huge observations and the counteraction of more robust methods and the loss of information. Point estimators for the quantiles of the distribution of the maximum cluster exceedances are obtained using the inverse of the estimated distribution function. For the threshold model with u=0.08 mg/m^3 the estimated quantiles are $q_{10} = 0.0012$ (0.0019), q_{25} =0.0033 (0.0032), $q_{50} = 0.0081$ (0.0087), q_{75} =0.0167 (0.0169), q_{90} =0.0287 (0.0314), where the values in parentheses are the quantiles of the empirical distribution function.

In Figure 8 we present the qq-plot for the threshold model of the NO_2 concentrations with u=0.13 mg/m^3. The plot shows that the model fits the data sufficiently well. Point estimates for the quantiles of the distribution are $q_{10} = 0.0034$ (0.0032), q_{25} =0.0092 (0.0092), $q_{50} = 0.0215$ (0.0238), q_{75} =0.0411 (0.0373), q_{90} =0.0646 (0.0658), empirical quantiles given in parantheses. For both air pollutants the different threshold models produce almost identical N-month-return levels. The return levels, summarized in Table 6, are higher than those based on the estimated GEV distribution parameters. As mentioned, the cluster peaks used for threshold models of the ozone data are all from the summer season, so that the model can be compared to the GEV summer model.

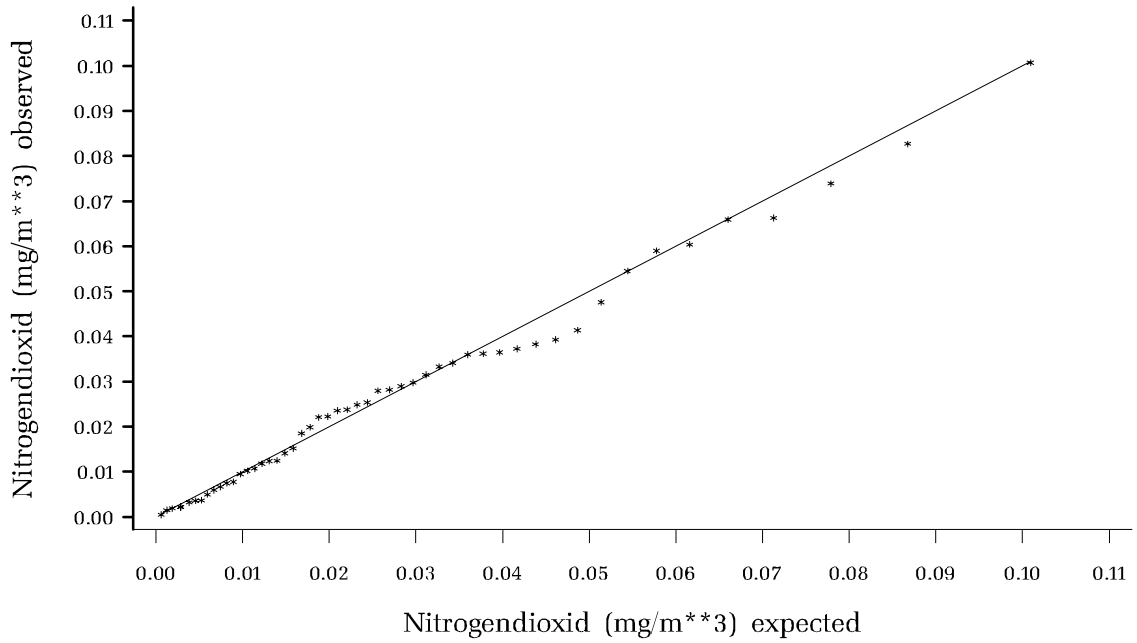

Figure 8. QQ-plot for the threshold model of the NO_2 concentrations with $u = 0.13$ mg/m^3 at the site M-Stachus.

	O_3 (M-Lothstr.)				NO_2 (M-Stachus)			
N	$u = 0.07$		$u = 0.08$		$u = 0.13$		$u = 0.16$	
1	0.0874	(0.0025)	0.0877	(0.0018)	0.1065	(0.0061)	0.1198	(0.0204)
2	0.0965	(0.0033)	0.0963	(0.0033)	0.1300	(0.0000)	0.1371	(0.0098)
4	0.1051	(0.0046)	0.1053	(0.0049)	0.1515	(0.0036)	0.1543	(0.0020)
6	0.1099	(0.0056)	0.1108	(0.0063)	0.1632	(0.0048)	0.1642	(0.0013)

Table 6. Estimated N-month-return levels (GPD) and their standard errors. The official standards for O_3 and NO_2 are 0.05 mg/m^3 and 0.1 mg/m^3.

The values of the return levels are alarmingly high, for ozone we can expect one cluster peak per month where the daily average of the ozone concentrations exceeds 0.0877 mg/m^3, within two months we can expect a cluster maximum which is almost twice

the standard. As the return levels are estimates for daily means of concentration measures the short time afternoon peaks of the ozone concentrations at one of these days could be enormously high.

3.3 The logistic regression model

It would be desirable to have a statistical tool to predict the exceedances of the ozone or nitrogen dioxide concentrations over their official air pollution standards. The limit L is usually fixed by law and if it is exceeded certain measures like smog warnings have to be taken. It defines that concentration level of an air pollutant above which the health of the people could be endangered. Thus it is of special interest to relate the probability of exceeding this standard L to relevant covariates. This can be done by a logistic regression model

$$\mathrm{P}\left(Z_L = 1 \mid H = h\right) = g\left(h^{'}\beta\right). \tag{8}$$

Let W denote the concentration of an air pollutant and H the vector of possible variables of influence on Z_L whereby the binary response variable Z_L is defined as

$$Z_L = \begin{cases} 1 & \text{for} \quad W > L \\ \\ 0 & \text{for} \quad W \leq L. \end{cases}$$

The inverse link function g is $g\left(t\right) = \left(1 + \exp(-t)\right)^{-1}$. The parameters of the logistic regression model are estimated by the method of maximum likelihood.

Results

As possible covariates we include several meteorological variables and other air pollutants that are supposed to have an influence on the limit exceedances of the O_3 and NO_2 concentrations. We also examine the coherence of the binary outcomes with lagged concentration measures of the same pollutant which have to be taken into account as regressors as well. At last a weekend dummy variable can be identified as an explanatory variable. We use an algorithm for a forward selection with subsequently backward elimination of the possible explanatory variables (see Hosmer and

Lemeshow, 1989). The meteorological variables are daily measurements of temperature (TE) in °C , wind velocity (WV) in m/sec and atmospheric humidity (HU). The lagged concentration measures are abbreviated with NO_2(x) and O_3(x) where x is the number of lags. The weekend dummy is denoted by WE. For the ozone data we took nitrogen monoxide (NO) into the model. The unit of the concentration measures is 100 mg/m^3. In Table 7 and 8 the final results of fitting the logistic regression model of the ozone data at M-Lothstr and of the nitrogen dioxide data at M-Stachus are presented. There the parameter estimates β_i with their standard errors $s(\hat{\beta}_i)$ and the p-values p_i for the hypotheses $H_0^i: \beta_i = 0$ versus $H_1^i: \beta_i \neq 0$ are listed. The last column contains the odds ratios $OR_i = \exp(\beta_i)$ of the variables.

Response: Ozone (M-Lothstr)				
Variable H_i	$\hat{\beta}_i$	$s(\hat{\beta}_i)$	p_i	$\widehat{OR}_i$
INTERCEPT	4.300	1.1624	0.0049	
TE	0.051	0.0203	0.0113	1.053
HU	-0.102	0.0118	0.0001	0.903
WE	0.580	0.2527	0.0217	1.786
O_3(1)	0.735	0.0734	0.0001	2.086
NO	-0.700	0.0988	0.0001	0.497

Table 7. Results of logistic regression analysis for O_3.

For O_3 we find a positive effect of temperature and of the weekend dummy, which means that the exceedances are more likely to occur on weekends and on days with high temperature. Most of the seasonal variation is thus explained by the temperature. We also observe the well-known fact that higher NO concentrations locally reduce the O_3 concentration, as the effect of NO is negative in our model.

Response: NO_2 (M-Stachus)				
Variable H_i	$\hat{\beta}_i$	$s(\hat{\beta}_i)$	p_i	$\widehat{OR}_i$
INTERCEPT	-3.678	0.868	0.0064	
TE	-0.065	0.010	0.0001	0.937
HU	-0.068	0.007	0.0001	0.934
WV	-0.996	0.099	0.0001	0.369
WE	-1.851	0.232	0.0001	0.157
$NO_2(1)$	0.609	0.046	0.0001	1.840
$NO_2(2)$	0.112	0.039	0.0043	1.119

Table 8. Results of logistic regression analysis for NO_2.

For NO_2 we get a negative weekend effect due to less traffic, and negative effects of temperature, humidity and wind velocity. In both models we have a strong effect of the lagged concentration of order 1 indicating the autoregressive structure of data. The probability of an exceedance given the covariates can be estimated by

$$\begin{aligned} \hat{\pi} &= g\left(h'\hat{\beta}\right) \quad \text{with} \\ h'\hat{\beta} &= 4.300 + 0.051\ \text{TE} - 0.102\ \text{HU} + 0.580\ \text{WE} \\ &\quad +0.735\ \text{O}_3(1) - 0.700\ \text{NO} \end{aligned}$$

in the O_3 model. Note that the model parameters are estimated for one site. Fitting the same model with data from other stations – not reported here – leads to slightly different results.

To predict exceedances one day in advance the weather forecast can be used for the weather variables and a regression model for NO proposed by Küchenhoff and Wolf (1995).

We examine the goodness of fit of the model of the O_3-concentration exceedances by calculating the model response probability of every single data point and classify them (see Table 9), where the used intervals can roughly be interpreted as categories of the risk for observing a limit exceedance.

Response: O_3 (M-Lothstr)				
Classification		Observed		
Exp. Risk	Interval	positive	total	prop
very low	[0.0, 0.1)	8	669	0.012
low	[0.1, 0.3)	31	141	0.220
moderate	[0.3, 0.5)	33	90	0.367
high	[0.5, 0.7)	50	76	0.658
very high	[0.7, 0.9)	81	103	0.786
extreme	[0.9, 1.0]	148	155	0.955
Total		351	1234	0.284

Table 9. Goodness of fit for the logistic regression model for the limit exceedances of O_3 at the site M-Lothstr.

If we assume that all estimated response probabilities in the interval (0.9,1.0] would lead to the prediction that an ozone concentration above the standard would be observed ($\hat{Z}_L = 1$), this would result in 155 positive response predictions. The observed sensitivity of the model would be $148/351 = 42.17\%$. The observed positive correctness of the model is very high, the ratio is $148/155 = 95.5\%$. The same assumption for all estimated response probabilities in the interval (0.7,1.0] results in 65.2 % for the observed sensitivity and 88.7 % for the observed positive correctness. Note that these values overestimate the goodness of prediction of the model. The assessment of the predictive power should be done with new data.

4 Discussion and further analysis

We analysed extreme values of daily air pollution data with different statistical methods. We modelled the distribution of monthly maxima and the distribution of maximum cluster exceedances of a suitable threshold. For this purpose the generalized form of the extreme value distribution and the Pareto distribution were used. Both distribution families were shown to be useful tools to get insights into the behavior of extreme air pollution concentrations and to estimate important attributes like its N-month-return levels. For a further research the time series structure of the data and covariates like weather variables could be included to try to improve the results. But since these models are rather complicated and many parameters have to be included, the results could be much more difficult to be interpreted than in our simple models. We also applied logistic regression to model the probability of the concentrations measures to exceed the official air quality standard. There for O_3 and NO_2 significant effects of weather variables and other pollutants were found. These models can be used for forecasts of exceedances and should be examined with further data.

Acknowledgements

The authors want to thank the Bavarian State Office for Environmental Protection for providing the data and for useful discussions. They are also grateful to the referees for their valuable comments and helpful suggestions, which considerably improved the paper.

5 References

Davison, A.C. and Smith, R.L. (1990) Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society.* **B 52**, 393-442.

Fisher, R.A. and Tippett, L.H.C. (1928) Limiting Forms of the Frequency Distributions of the Largest or Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society.* **24**, 180-190.

Gumbel, E.J. (1958) *Statistics of Extremes.* Columbia University Press, New York.

Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985) Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics.* **27**, 251-261.

Hosking, J.R.M. and Wallis, J.R. (1987) Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics.* **29**, 339-349.

Hosmer, D. W. and Lemeshow, S. (1989) *Applied Logistic Regression.* Wiley Series in Probability and Mathematical Statistics, New York.

Küchenhoff, H. and Wolf, R. (1995) Time Series Analysis of Daily Air Pollution Data. *Technical Report*, University of Munich.

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

Pickands, J. (1975) Statistical Inference using Extreme Order Statistics. *Annals of Statistics.* **3**, 119-131.

Rao, S.T., Sistla, G. and Henry, R. (1992) Statistical Analysis of Trends in Urban Ozone Air Quality. *Journal of the Air Waste Management Association.***42**, 1204-1211.

Resnick, S.I. (1987) *Extreme Values, Regular Variation and Point Process.* Applied Probability. Springer, New York.

Smith, R.L. (1984) Threshold Methods for Sample Extremes. in *Statistical Extremes and Applications,* ed. J. Tiago de Oliveira. Dordrecht: D.Reidel, 621-638.

Smith, R.L. (1989) Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science.* **4**, 367-393.