Stoch Environ Res Risk Assess (2005) 19: 173–183
DOI 10.1007/s00477-004-0215-3

ORIGINAL PAPER

J. van de Kassteele · A. L. M. Dekkers · A. Stein
G. J. M. Velders

Model-based geostatistical interpolation of the annual number of ozone exceedance days in the Netherlands

Published online: 26 February 2005

Abstract This paper discusses two model-based geostatistical methods for spatial interpolation of the number of days that ground level ozone exceeds a threshold level. The first method assumes counts to approximately follow a Poisson distribution, while the second method assumes a log-Normal distribution. First, these methods were compared using an extensive data set covering the Netherlands, Belgium and Germany. Second, the focus was placed on only the Netherlands, where only a small data set was used. Bayesian techniques were used for parameter estimation and interpolation. Parameter estimates are comparable due to the log-link in both models. Incorporating data from adjacent countries improves parameter estimation. The Poisson model predicts more accurately (maximum kriging standard deviation of 2.16 compared to 2.69) but shows smoother surfaces than the log-Normal model. The log-Normal approach ensures a better representation of the observations and gives more realistic patterns (an RMSE of 2.26 compared to 2.44). Model-based geostatistical procedures are useful to interpolate limited data sets of counts of ozone exceedance days. Spatial risk estimates using existing prior information can be made relating health effects to environmental thresholds.

Keywords Model-based geostatistics · Bayesian inference · Count data · Ozone · Exceedance days

Abbreviations MCMC: Markov chain Monte Carlo

J. van de Kassteele · A. Stein
Mathematical and Statistical Methods Group,
Biometris, Wageningen University, 100,
6700 AC, Wageningen, The Netherlands

J. van de Kassteele (✉) · A. L. M. Dekkers · G. J. M. Velders
Netherlands Environmental Assessment Agency - RIVM,
1, 3720 BA, Bilthoven, The Netherlands
E-mail: Jan.vandeKassteele@wur.nl
Tel.: +31-317-482384
Fax: +31-317-483554

Introduction

Ground level (tropospheric) ozone is a major air pollutant in Western Europe. Tropospheric ozone results from photochemical reactions with ozone precursors, volatile organic compounds, nitrogen oxides, carbon monoxide and methane in the atmosphere. Environmental focus on ozone concentrations has increased as a result of the possible inflammatory responses and reduction in lung function caused when humans are exposed to periods of several days‘ high ozone concentration. Ozone can also affect ecosystems, mainly through damage to leaves and other parts of plants (WHO 1996; UNECE 1996).

As a protection instrument for human health, the European Commission has set several targets and objectives for ozone levels in the atmosphere. The indicator applied in this study is the number of days per year in which an 8-h moving average ozone concentration exceeds 120 μg/m^3 (EC 2002).

Currently, rural ozone concentrations in the Netherlands are measured hourly within the Netherlands Air Quality Monitoring Network at 23 stations, spread across the country (van Elzakker 2001). Each station registers the annual number of exceedance days. EU regulations (EC 2002) require the number of exceedance days to be reported at the measuring sites. Interpolation of the exceedance days to produce maps for the Netherlands are a basis for assessment studies related to public health and environmental effects (e.g., see EEA 1998).

Here we analyse the use of geostatistical interpolation of annual ozone count data. So far, no attention has been paid in the literature to geostatistical interpolation of counts for ozone exceedance days. In geostatistics, spatial data are assumed to be a realisation of a random field, and often without the assumption of any stochastic model being declared. Usually, normality is implicitly assumed (Christakos 1992). The data analysed in this paper, however, are positively valued count data, without constant variance and normally distributed errors. In this case, the normality assumption may no longer be appropriate. Count data require a different approach. The question addressed in this paper is then which interpolation procedure will be most appropriate and practically applicable for environmental scientists.

The aim of this study is to investigate the applicability of either a Poisson procedure or a log-Normal model-based geostatistical procedure (Diggle et al. 1998; Ribeiro and Diggle 1999) to interpolate the number of exceedances of the 120 μg/m^3 threshold. A complication is sparseness of the data. Therefore, data from 2000 measured at 120 rural ozone monitoring stations in the Netherlands, Belgium and Germany were analysed first. Then there is a focus on the small subset of 23 stations in the Netherlands.

Materials and methods

Data

Verified hourly data for 2000 were collected from the Airbase database (ETC-ACC 2003) at 120 rural background ozone stations for the Netherlands, Belgium and Germany. These data were then aggregated, first, by calculating 8-h moving averages, and, second, by taking the daily maxima. Finally, the days on which these maxima exceeded the threshold of 120 μg/m^3 were summed to obtain the annual number of exceedance days. The data were aggregated according to the guidelines of the European Commission for missing data (EC 2002). Nine stations had therefore to be excluded. It was assumed that small differences between measurement techniques in the monitoring networks had not influenced the annual number of exceedances, since all observations of ozone had to satisfy the same quality control specifications (EC 2002).

Figure 1 shows the 111 observations over the whole study region. The coordinates were obtained by transforming the geographical coordinates with an azimuthal equidistant projection centred on 51° north and 9° east. This projection preserves a correct absolute distance between the stations in the region considered. The number of exceedances was lower near the North Sea coast and higher in the south-east of the region. High ozone concentrations are caused by photochemical reactions during warm and sunny days, while the strong dependence on these meteorological conditions caused the number of exceedance days to fluctuate sharply from year to year (Feister and Balzer 1991). In and near large cities, ozone concentrations are usually lower than in rural areas (see e.g., Gregg et al. 2003). Since this effect introduces local non-stationarities, only stations in rural areas were considered in this study.

Next to showing the observations in the data set of 111 stations, the study focused on analysing data from the national air quality monitoring network of the Netherlands only. This network consists of 23 rural ozone stations (van Elzakker 2001), but one station had to be excluded due to missing data. The reason for analysing this small subset only is practical: i.e. the Netherlands Environmental Assessment Agency needs to report the number of exceedance days to the European Commission as soon as the data has become available.

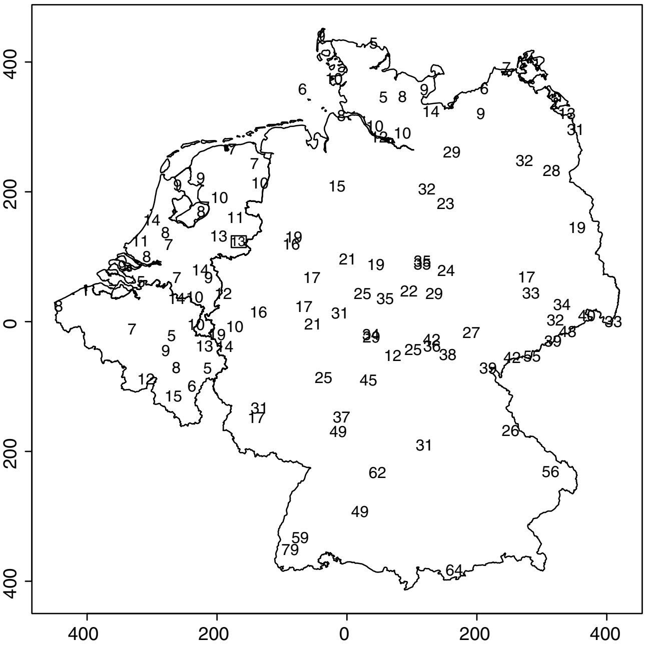

Fig. 1 The annual number of days where the daily maximum 8-h moving average ozone concentration exceeds the 120 μg/m^3 threshold value in 2000. The *rectangle* indicates the location of Eibergen station

Figure 1 shows the number of exceedance days to increase from the north-west to the south-east. The spatial variability increases along with this trend, making the random field a spatial non-stationary process, typical behaviour of a Poisson-like process. To gain more insight into the distribution, the count data were analysed and simulated at one single point. At other locations a similar process may occur. Observations of daily maximum 8-h moving average ozone concentration from 1991 to 2000 are shown in Fig. 2, at one particular station, Eibergen, a village in the eastern part of the Netherlands. Meteorological conditions and human activities contribute to fluctuations in the concentration. The graph shows extreme concentrations during the spring/summer season. Circles indicate the days that the concentration exceeds the threshold of 120 μg/m^3. One particular difficulty in assigning any statistical distribution to these data is the clustering of the exceedances. Such behaviour is typical for extreme events. This can be described with an extreme value model (Smith 1989). Shively (1991) models the sequence of exceedances as a non-homogeneous Poisson process. A simpler and more straightforward approach assumes the exceedance days to follow a Poisson process in the limit, i.e. that the dependence between daily maxima separated

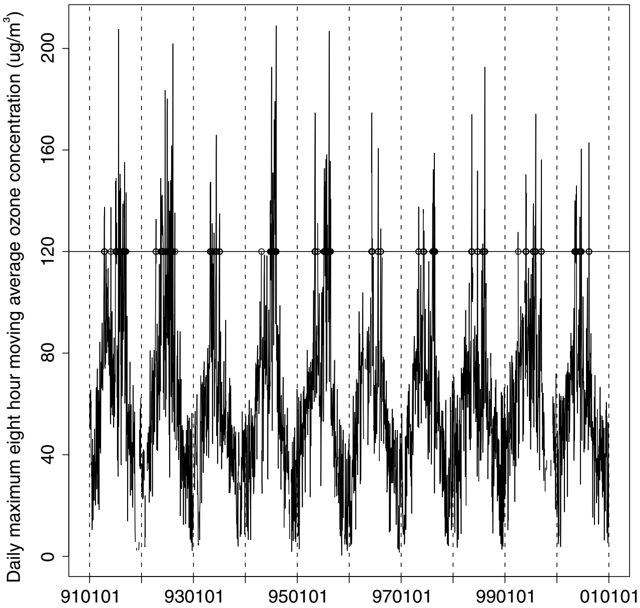

Fig. 2 Daily maximum 8-h moving average ozone concentration (μg/m^3) at Eibergen station for 1991 – 2000. Days that this concentration exceeds the threshold of 120 μg/m^3 are indicated by *circles*

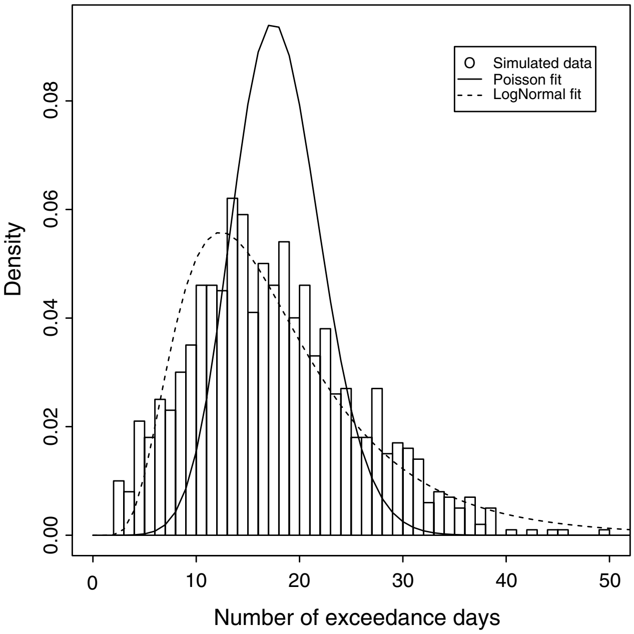

Fig. 3 Simulations of the annual number of exceedance days (bars) at Eibergen station with a fitted Poisson distribution (*solid line*) and log-normal distribution (*dashed line*)

by a given number of days decreases sufficiently fast as the separation increases.

As an experiment, we simulated the annual number of exceedance days. We modelled the occurrence of exceedance days over one ozone season at the Eibergen station by sampling from the Bernoulli distribution. To account for temporal dependence, the probabilities were conditional on the outcome of the previous day. These conditional probabilities were estimated from the 10 years of observations shown in Fig. 1. Summing the resulting sequences of zeros and ones yielded the annual numbers of exceedance days. The distribution of simulated exceedance days is presented as a histogram in Fig. 3, to which a Poisson distribution was fitted. It describes the average well (17.9), and is suitable for handling count data. It shows a smaller variance (17.9 for the Poisson distribution) than the simulated data (64.4), however. It overestimates the top and underestimates the tails. Also the log-Normal density function was fitted, which better accounts for the tails and the mean (18.2) of the simulations. The function however handles the data as being continuous and overestimates (105.6) the variance of the simulated data (64.4).

This exploring analysis showed that neither a Poisson distribution or a log-Normal distribution can describe ozone exceedance count data very well, but both distributions have properties that do fit the data. On the other hand, the Bernoulli simulation of the occurrence of exceedance days might not have been correct, since it was only a very simple model for the real situation (see the discussion section). It should also be realised that comparison with only one observed datum could be performed, since only one realisation was available.

The Poisson model

Assuming counts to be spatially independent Poisson distributed, they could be analysed with a generalised linear model (McCullagh and Nelder 1989). Generalised linear models allow data to follow any distribution of the exponential family, accommodating both continuous and discrete non-Normal distributions. Generalised linear-mixed models (Breslow and Clayton 1993) allow for correlated data. Diggle et al. (1998) embedded kriging into the framework of generalised linear models, providing a way to analyse spatially correlated Poisson data. This model was applied in this study.

Considering spatial observations $y(\mathbf{x})$ as realisations of a random field process $Y(\mathbf{x})$, where $\mathbf{x} \in \Re^2$, the random field process for spatial correlated Poisson data is written as

$$Y(\mathbf{x})|S(\mathbf{x}) \sim \text{ Poisson}[M(\mathbf{x})]. \tag{1}$$

The interpretation is that conditional on an underlying surface $S(\mathbf{x})$, $Y(\mathbf{x})$ is an independent Poisson distributed spatial variable with the conditional expectation $M(\mathbf{x})$. $M(\mathbf{x})$ is a stochastic variable containing the deterministic large-scale trend, $\mu(\mathbf{x})$, and the underlying surface $S(\mathbf{x})$. For Poisson data the relation between $M(\mathbf{x})$, $\mu(\mathbf{x})$, and $S(\mathbf{x})$ is attenuated by the log link function

$$\log[M(\mathbf{x})] = \mu(\mathbf{x}) + S(\mathbf{x}). \tag{2}$$

The trend $\mu(\mathbf{x})$ is a linear function $d(\mathbf{x})^{\mathrm{T}}\beta$ of known functions of covariates $d(\mathbf{x})$, which, in our study, only depends on the location variable $\mathbf{x}$, and unknown regression or trend parameters, β. The underlying surface $S(\mathbf{x})$ is modelled by a zero mean second-order stationary Gaussian process with covariance matrix Σ

$$S(\mathbf{x}) \sim \mathrm{N}\left(0, \sum\right). \tag{3}$$

The elements of Σ depend upon the distance vector h_{ij} between two locations, $\mathbf{x}_i$ and $\mathbf{x}_j$, by means of a permissible correlation function ρ with parameters σ^2 (the variance) and a range parameter ϕ. The Poisson model predicts the intensity in space $M(\mathbf{x})$. For the Poisson model, intensity and variance are equal. Hence, the predicted intensity field will always be smoothed because it can explain deviations from the intensity value by its corresponding Poisson variance.

The log-Normal model

We considered the log-Normal model (Cressie 1993) as an alternative method. We now assume the logarithm of the random field followed a Normal distribution. The log-Normal model can be written in an equivalent model-based formulation of a linear mixed model (Ribeiro and Diggle 1999; Pinheiro and Bates 2000). Conditional on the underlying surface $S(\mathbf{x})$, the log$[Y(\mathbf{x})]$ are independently normally distributed, with conditional expectation $M(\mathbf{x})$ and variance τ^2

$$\log[Y(\mathbf{x})]|S(\mathbf{x}) \sim \mathrm{N}\left[M(\mathbf{x}), \tau^2\right]. \tag{4}$$

Note that here log$[Y(\mathbf{x})]$ is in fact a noisy version of $M(\mathbf{x})$, with residual variance τ^2. The relationship between $M(\mathbf{x})$, $\mu(\mathbf{x})$, and $S(\mathbf{x})$ is the identity link for a Gaussian model so

$$M(\mathbf{x}) = \mu(\mathbf{x}) + S(\mathbf{x}). \tag{5}$$

Interpretation of $\mu(\mathbf{x})$ and $S(\mathbf{x})$ remains unchanged in comparison to the Poisson model. In conventional geostatistics, the variance τ^2 is called the nugget, $\sigma^2 + \tau^2$, the sill, and σ^2, the partial sill. The parameter τ^2 can be considered to resemble variations that cannot be attributed to spatial correlation and thus introduces smoothing. Finally, the log-Normal model makes spatial predictions of the expected number of exceedance days. This is a major difference with the Poisson model.

Parameter estimation and spatial prediction

Parameters were estimated using Bayesian inference (Gelman et al 1995), in particular using Markov Chain Monte Carlo (MCMC) methods (Gilks et al. 1996) based upon the Langevin–Hastings algorithm (Besag 1994; Papaspilliopoulus et al. 2003). This is a Metropolis-Hastings algorithm in which the proposal distribution uses gradient information from the log-posterior distribution. The algorithm iteratively generates a chain, where in each step a proposal is generated for an update of the current state of the chain. The update is then accepted or rejected according to a certain acceptance probability. Proposal variances for σ^2 and ϕ have to be found manually in such a way that approximately 60% of the proposals is accepted (Christensen and Ribeiro 2002). The predictive distribution is obtained by first sampling from the posterior distributions, and then taking, for each, samples from the multivariate Gaussian distribution of $S(\mathbf{x})$. This procedure automatically incorporates parameter uncertainty in the predictions. For the mathematical formulation of the above process we refer to Diggle et al. (1998) and Gelman et al. (1995).

A re-parameterisation of the nugget τ^2 as a relative nugget $\tau^2{}_{\mathrm{rel}} = \tau^2 / \sigma^2$ was carried out to still be able to write the covariance matrix Σ as a product between σ^2 and the correlation matrix. Discrete intervals for ϕ and$\tau^2{}_{\mathrm{rel}}$ had to be taken, because their posteriors cannot be written as a standard statistical distribution. (Ribeiro and Diggle 1999; Christensen and Waagepetersen 2002).

Prior specification and setup of the MCMC algorithm

Bayesian inference needs a specification of prior distributions of the parameters. Prior knowledge was available (see Fig. 1). To allow modelling of the trend towards the south-east, we included covariates$d(\mathbf{x}) = (1, \mathbf{x}_1, \mathbf{x}_2)^{\mathrm{T}}$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the coordinates in the east-west and north-south directions, respectively, and associated regression parameter $\beta = (\beta_0,\beta_1,\beta_2)^{\mathrm{T}}$. The variance σ^2 of the log-data is positive, approximately equal to 0.1, and the correlation distance ϕ a few hundred kilometres, which is typical for ozone concentrations found in previous years. An exponential variogram model $\rho(u) = \exp(-u)$ was chosen for the covariance structure. The resulting priors for both the Poisson and log-Normal models are:

$$\begin{aligned}
&\pi(\beta_0) \sim \mathrm{N}(3, 0.5)\\
&\pi(\beta_1) \sim \mathrm{N}(0.001, 0.001)\\
&\pi(\beta_2) \sim \mathrm{N}(-0.002, 0.001)\\
&\pi\left(\sigma^2\right) \sim \chi^2_{\mathrm{inv}}(1, 0.1)\\
&\pi(\phi) \sim \exp(1/100)\\
&\pi\left(\tau^2_{\mathrm{rel}}\right) \propto 1
\end{aligned}$$

We chose Gaussian priors for the trend parameters, an inverse-χ^2 distribution with one degree of freedom and a scale parameter 0.1 forσ^2 and an exponential prior with an expectation of 100 km for the range parameter. The relative nugget $\tau^2{}_{\mathrm{rel}} = \tau^2 / \sigma^2$, only used in the log-Normal model, was given a uniform prior. The Gaussian distributions and inverse-χ^2 distribution are conjugate priors for the trend and sill parameters, respectively. The exponential distribution for the range parameter leads to more equally spaced correlations at a fixed distance (Ribeiro and Diggle 1999).

The variances of the trend parameters seem rather strict. They are not however, because these variances are scaled by the partial sill parameter. In combination with the fact that coordinates are given in kilometers, these priors are relatively flat.

Proposal variances for σ^2 and ϕ were found to be 0.002 and 100. To check on convergence and mixing, we considered trace plots of the individual samples and their corresponding auto-correlation functions. The samples preferably show stationarity with low auto-correlation. The chain's burn-in time was set at 10,000 iterations and it was sampled every 200th iteration to reduce the auto-correlation.

The prior specification for the subset of 22 observations for the Netherlands only was based on information on the full set. Only the intercept parameter was given a lower value, and no prior trend was specified. The resulting priors for the subset are:

$$\pi(\beta_0) \sim \mathrm{N}(2, 0.5)$$
$$\pi(\beta_1) \sim \mathrm{N}(0, 0.001)$$
$$\pi(\beta_2) \sim \mathrm{N}(0, 0.001)$$
$$\pi(\sigma^2) \sim \chi^2_{\mathrm{inv}}(0.1, 1)$$
$$\pi(\phi) \sim \exp(1/100)$$
$$\pi(\tau^2_{\mathrm{rel}}) \propto 1$$

The proposal variances for σ^2 and ϕ were found to be 0.01 and 300, respectively. The chain's burn-in time and thinning remained unchanged.

Validation

A cross validation by "leaving one out" was carried out to see which interpolation method performs better. The root-mean-squared error (RMSE) was chosen as the error measure. The two models do not predict the same quantity. Therefore, results have to be interpreted with care.

Results

The data sets were analysed with the software packages *geoR* (Ribeiro and Diggle 2001) and its extension, *geoRglm* (Christensen and Ribeiro 2002). Both packages run under the programming environment of *R* (Ihaka and Gentleman 1996). The *geoR* package contains several functions for handling (log-)Normal spatial data; *geoRglm*can deal with spatial Poisson data. *R* and both packages are available free of charge on the Internet.

The results are presented under three headings: (1) parameter estimation and interpolation using the full data set, (2) interpolation results of 1, focusing on the Netherlands, and (3) parameter estimation and interpolation using the subset of the Netherlands only. Interpolation was done on a 15×15 km grid for the Netherlands, Belgium and Germany, while for the Netherlands a 5×5 km grid has been taken.

Part 1: analysis and interpolation using the full set

Posterior densities of the six model parameters are shown in Fig. 4. Since the Poisson model does not contain a nugget effect, no posterior is shown. Values of the modes and standard deviations are given in Table 1. Since we work with a number of days, sill and nugget have no units.

The posterior densities of the trend parameter vector β of both models are practically identical. This is not surprising since both models estimate the trend on a log-scale. The modes are $\hat{\beta} = (3.01, 0.0015, -0.0024)^{\mathrm{T}}$ and $\hat{\beta} = (3.00, 0.0013, -0.0025)^{\mathrm{T}}$ for the Poisson model and log-Normal model, respectively.

The partial sill σ^2 of the Poisson model is smaller than that of the log-Normal model. Their modes are 0.093 and 0.12, respectively. Although they have the same order of magnitude due to the log-scale, we can understand the difference from distributional assumptions of both models. The Poisson model predicts the intensity field. The Poisson model can describe the variation in the original data by its corresponding Poisson variance. For this reason, σ^2 may be smaller than for the log-Normal model. If simulations of equal probable fields were made, these fields would be close to the original data.

For the posterior range distribution for the Poisson model we have a mode $\hat{\phi} =$ 102 km and for the log-Normal model $\hat{\phi} =$ 68 km. The effective correlation distance is three times larger, because of the exponential correlation function. The Poisson posterior is more uncertain, shown by its smaller peak and wider tail (see also Table 1). Further, we found the range to be positively correlated with the partial sill.

The nugget effect is estimated only in the log-Normal model. Its mode is relatively small, $\hat{\tau}^2 =$ 0.028, but it will introduce smoothing in the interpolation.

Figure 5 shows the predicted spatial fields (left panels) and their corresponding standard deviations (right panels). Minimum and maximum values are given in Table 2. From the original number of exceedance days, the Poisson model (upper panels) predicts the Poisson intensity of the number of exceedance days, while the log-Normal model (lower panels) predicts the expected number of exceedance days. The patterns and values of both models are rather similar. Both models smoothed the original data.

The standard deviations clearly show the difference between what both models predict. Related to its lower sill and higher range, the Poisson model shows considerable lower values. This indicates the predicted intensity to be more certain than the expected number of exceedance days. In both models, the log-link yields larger standard deviations at those locations where predictions are larger.

Cross validation (Table 3) shows that the RMSE for the Poisson model was smaller (7.09) than that for log-Normal model (7.15).

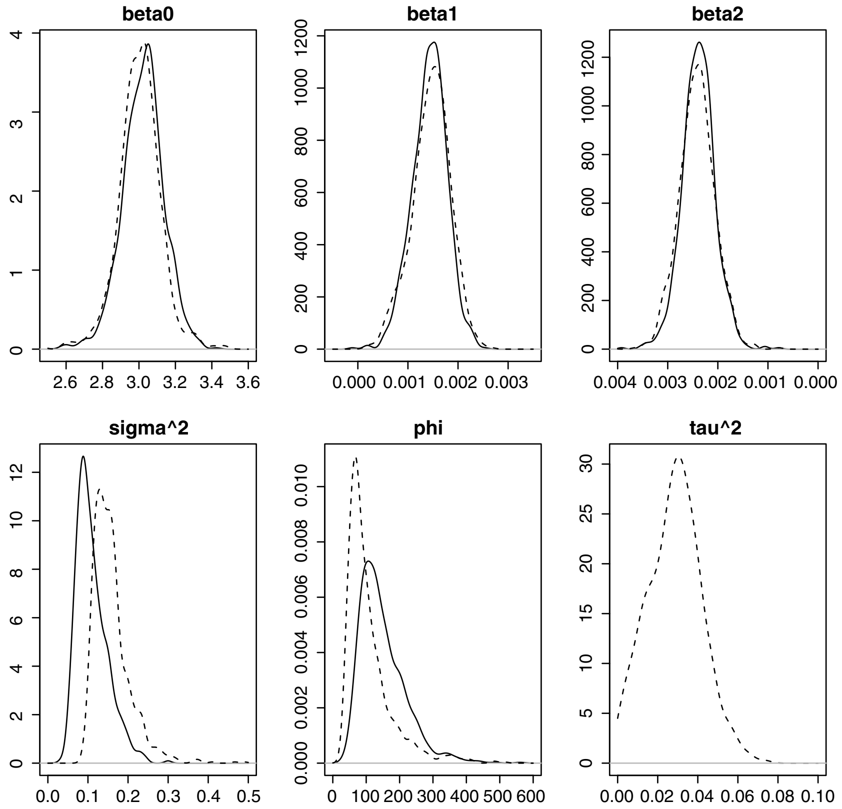

Fig. 4 Posterior distributions of the Poisson model parameters (*solid line*) and log-Normal model (*dashed line*) using the full data set

Table 1 Modes and standard deviations (calculated as 1/4 of the 95% credible interval) of the posterior distributions (Figs. 4, 7)

		β_0 [–]	β_1[km^{-1}]	β_2[km^{-1}]	σ^2 [–]	ϕ[km]	τ^2 [–]
Full set	Poisson	3.01	0.0015	−0.0024	0.093	102	
		(0.11)	(0.00034)	(0.00033)	(0.037)	(68)	
	log-Normal	3.00	0.0013	−0.0025	0.12	68	0.028
		(0.12)	(0.00040)	(0.00037)	(0.044)	(74)	(0.014)
Subset	Poisson	2.07	−0.00010	0.000076	0.029	63	
		(0.14)	(0.00036)	(0.00036)	(0.039)	(113)	
	log-Normal	2.13	−0.000041	0.00011	0.073	56	0.013
		(0.16)	(0.00047)	(0.00048)	(0.052)	(76)	(0.011)

Part 2: interpolation in the Netherlands using the full set

In the previous section, we showed the most important properties of both models. Because the effective range was approximately a few hundred kilometres, we also incorporated data from the surrounding countries Belgium and Germany. In this section we zoom in on the interpolation results for the Netherlands only, while using the parameter estimates from the full set.

Figure 6 shows the predicted fields in the Netherlands (left panels) and their corresponding standard deviations (right panels). The presence of observations from Germany leads to higher values near the Netherlands–German border. The Poisson model (upper panels) shows more smoothing than the log-Normal model (lower panels), as is verified from the minimum and maximum predicted values (Table 2).

Standard deviations of the Poisson model are smaller and show less variation. This indicates that the predicted intensities are more certain than the expected number of exceedance days predicted by the log-Normal model. The RMSE of both models are practically equal, 2.28 and 2.27, respectively. The log-Normal model has shown less smoothing, but this has only a little effect on the RMSE.

Part 3: analysis and interpolation in the Netherlands using the subset

This section focuses on the Netherlands only. Parameter estimation and interpolation has been done using the subset of 22 observations. Posterior densities are shown in Fig. 7, with values of the modes and standard deviations given in Table 1. The posterior trend parameter vector of the Poisson model is again similar to that of the log-Normal model. Posterior modes are $\hat{\beta}=$ (2.07, −0.00010, 0.000076)$^{\mathrm{T}}$ and $\hat{\beta}=$ (2.13, −0.000041,

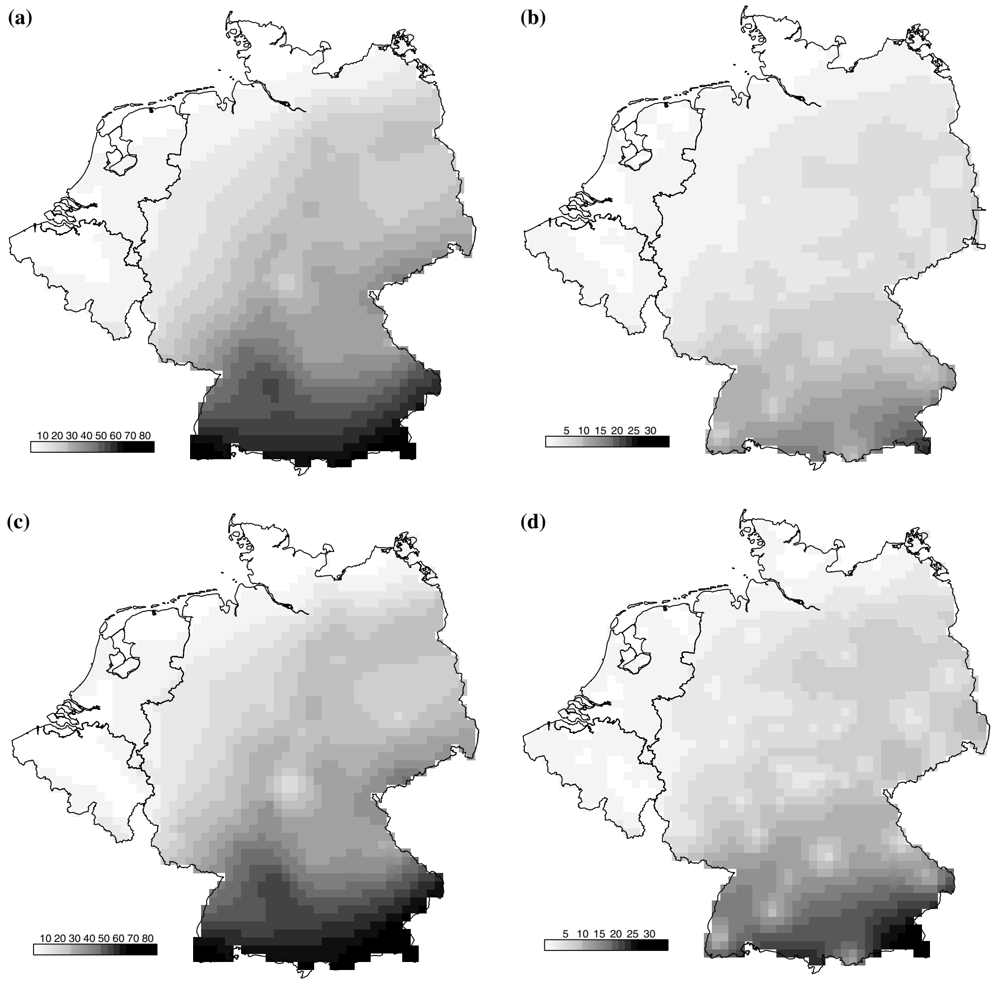

Fig. 5 Predicted number of ozone exceedance days (*left*) and corresponding kriging standard deviations (*right*) with the Poisson model (*top*) and log-Normal model (*bottom*) for 2000 using the full data set

Table 2 Minimum and maximum values of the data, model predictions and model standard deviations

	Full set NL-B-D		Full set NL		Subset NL	
	Min	Max	Min	Max	Min	Max
Poisson						
Observations	5	79	5	14	5	14
Predictions	6.60	75.84	7.23	13.28	8.11	10.03
Standard deviations	1.40	23.48	1.39	2.60	1.26	2.16
log-Normal						
Predictions	5.78	83.71	6.05	13.97	5.85	12.42
Standard deviations	1.20	34.47	1.00	3.56	0.85	2.69

$0.00011)^{T}$, respectively, indicating no significant trend in the data. The estimates for the Poisson model are more certain (Table 1).

The partial sill σ^2 in both models diminished in comparison to the values found using the full set. Posterior modes are 0.029 and 0.073, respectively. In particular for the Poisson model, the sill value has become very small, indicating that the Poisson model will only show little variation around its mean. As for the trend parameters, differences in uncertainty of σ^2 between the two models have grown.

Table 3 Root mean-squared error values of the cross validation

	Full set NL-B-D	Full set NL	Subset NL
Poisson	7.09	2.28	2.44
log-Normal	7.15	2.27	2.26

The posterior range parameter ϕ has also become smaller, with modes of 63 and 56 km for the Poisson model and log-Normal model, respectively. The posterior range of the Poisson model (Fig. 7) has become more uncertain than the range in Fig. 4 (Table 1). It appeared to strongly depend on its prior.

The nugget of the log-Normal model has also reduced ($\hat{\tau}^2 = 0.013$). Compared to the estimates using the full

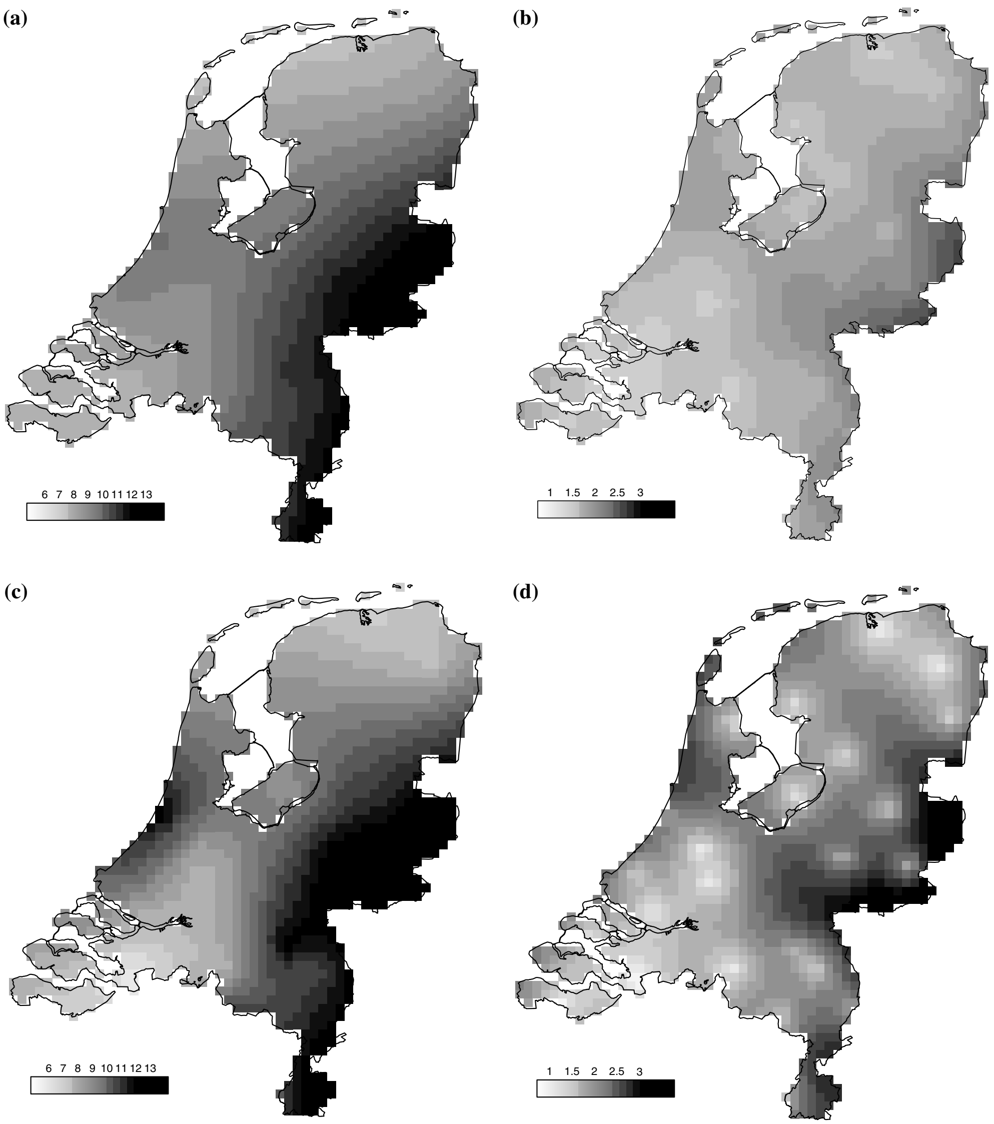

Fig. 6 Predicted number of ozone exceedance days (*left*) and corresponding kriging standard deviations (*right*) with the Poisson model (*top*) and log-Normal model (*bottom*) for the Netherlands in 2000 using the full data set

set, the standard deviations of all parameters, except the nugget, have increased (Table 1).

Figure 8 shows the predictions (left panels) and corresponding standard deviations (right panels) of both models. Contrary to Fig. 6, three aspects can be clearly seen. First, the influence of the observations from the surrounding countries has disappeared, especially near the Netherlands–German border. Second, the Poisson model has larger smoothing, and third, the log-Normal model has less smoothed. Standard deviations for both models decreased. Minimum and maximum values are given in Table 2. Cross validation shows a lower RMSE for the log-Normal model (2.26) than for the Poisson model (2.44) (Table 3).

Discussion

The data used in this study represent the annual number of days in which ozone exceeds a threshold level. Observations were used from the Netherlands, Belgium and Germany. One may question the possibility of pooling data, since different countries may use different ozone measurement techniques. According to European quality control specifications (EC 2002) however, measuring was done in a standardised way with calibrated equipment, but an intercomparison study (Borowiak et al. 2000) showed that the Netherlands ozone concentrations were measured 4% lower than its surrounding countries. We performed a recalculation of the number of exceedance days in the Netherlands, and the number of exceedance days increased from 0 to 5 days, depending on the station, with an average of 2.05 days. We could have corrected the data in advance, but on the other hand, in our research we analysed data that were actually reported by the Netherlands Environmental Assessment Agency, without correcting them first. The correction should be done by the Agency before releasing the data.

The study showed the effective correlation distance to be approximately a few hundred kilometres. This

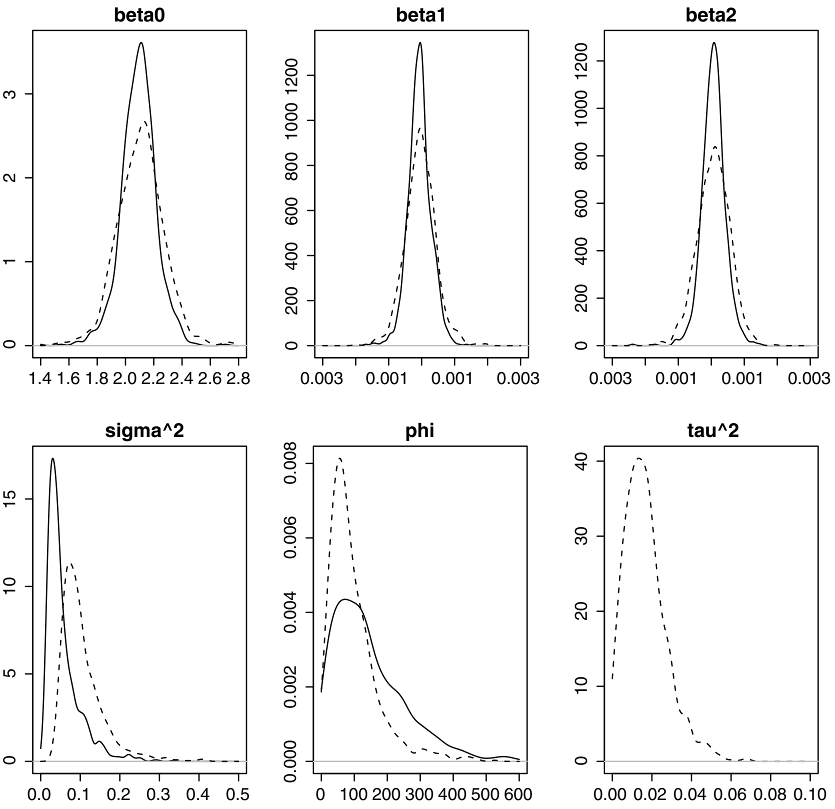

Fig. 7 Posterior distributions of the Poisson model parameters (*solid line*) and log-Normal model (*dashed line*) using the subset

satisfies analysis of the extensive data set covering the three countries. It further implies that when interpolating for the Netherlands only, data from surrounding countries have to be taken into account. One practical issue remains important as well: the Netherlands Environmental Assessment Agency needs to report the number of exceedance days as soon as the data has become available. Since data from other agencies can arrive late, analysing only data from the Netherlands is then the ultimate possibility, but on the other hand, information from previous years can be used as prior information.

In the study we chose an explicit model-based geostatistical approach to interpolate the annual number of exceedance days. First, we assumed an approximation by a Poisson distribution, and second, a log-Normal distribution. The log-link in both models made model and parameter comparison easier. The advantage of using a Poisson model was that data could be analysed as count data, with corresponding properties. This was indicated by increasing variance with increasing mean (Fig. 1) as well as by the simulation study (Fig. 3). Its disadvantage was that it did not properly fit the simulation study. The variance of the Poisson distribution was too small as compared to the variance of the simulation. Occurrences of exceedance days cluster in time, which weakens the assumption of a Poisson process, and it predicted the intensity of the annual number of exceedance days, as such complicating direct comparison with observations. The log-Normal model better fitted (Fig. 3). It also showed increasing variance with an increasing mean and it predicted the expected number of exceedance days. The disadvantage is that it handled data as continuous, which was not the case.

The most important difference between the models was that the Poisson model predicts an intensity field, whereas the log-Normal model predicted expected values. For this reason, σ^2 was smaller for the Poisson model (Table 2). As kriging standard deviations were smaller, the spatial predictions by the Poisson model seemed more accurate. Since expectation and variance are equal for a Poisson distribution, the Poisson model could describe more variation in the original data. Therefore the Poisson model described the original data by a smoothed intensity field that seemed more accurate.

The nugget of the log-Normal model can be considered similar to the Poisson variance and has a comparable effect to the smoothing properties of the Poisson model. This became clear for the full data set. For the subset, the nugget was lower, resulting in a less smoothed surface. The log-Normal model described most variation in the data by the underlying surface. The Poisson model, on the contrary, described this variation with its Poisson variance and showed a smoothed surface.

The choice for the prior of the range parameter was important. Earlier estimates using non-informative priors resulted in poor convergence in the MCMC algorithm. The choice of an exponential prior was an appropriate choice because it has the property that the correlation at a fixed distance was more uniformly

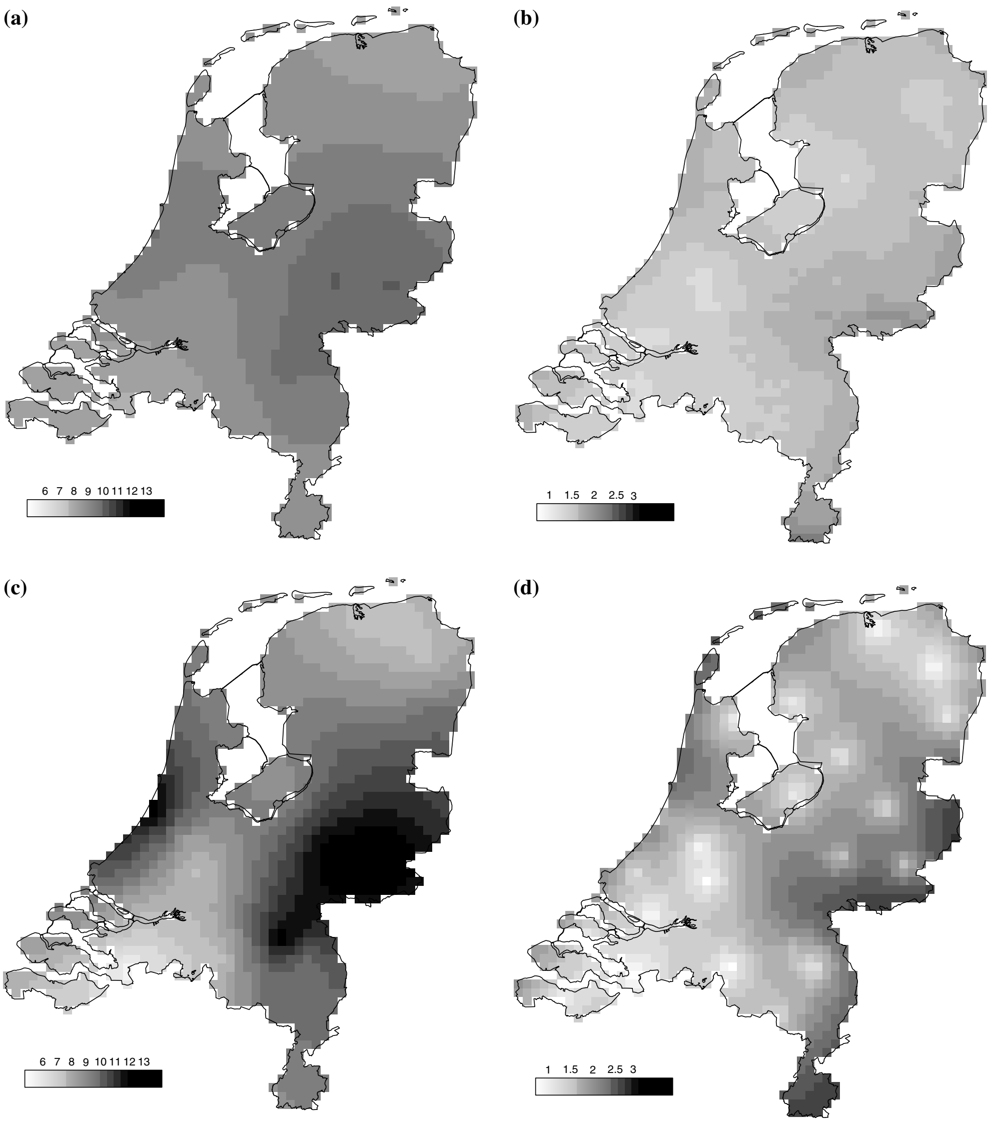

Fig. 8 Predicted number of ozone exceedance days (*left*) and corresponding kriging standard deviations (*right*) with the Poisson model (*top*) and log-Normal model (*bottom*) for the Netherlands in 2000 using the subset

distributed (Ribeiro and Diggle 1999). The other parameters seemed less sensitive and the chains always converged to reasonable values given our priors. The priors could be more specified if data from past years were analysed.

The log-Normal model described the variation in the original data well and incorporates it into the estimates of the spatial correlation parameters $\sigma^2$and ϕ. Consequently, the predictions passed practically through the observations at the monitoring stations. The Poisson model on the other hand avoided this. Apparently, the original data could be described by the predicted Poisson intensity parameter. When predicting exceedances near a critical level, e.g. the maximum allowed exceedance days per year, the log-Normal model approach was more appropriate.

As a further extension, the number of exceedance days may in fact follow a Negative Binomial distribution. This distribution can account for overdispersion and may fit the number of exceedance days better than the Poisson distribution or log-Normal distribution. In this case, the intensity $M(\mathbf{x})$ is Gamma distributed where the parameters vary in space.

The question remains how to interpolate this kind of count data exactly. The real situation is complex. The conceptual process is as follows: during smog days, the concentration in one area (a range of a about 100 km) increases, while in another area it does not. In the first area, an exceedance may occur, while in the other is does not. On another day, in the other area an exceedance may occur, while in the first area it does not. On average, there will be more exceedances in a certain area, in this case the southern part of Germany. The data are in fact a summation of different spatially correlated data over time. This may introduce large variability in space on small scales. To avoid interpolation of count data directly and using the bulk of information in hourly observations, one could imagine spatial-temporal interpolation. This can by done by interpolating hourly observed ozone concentrations (e.g., Guttorp et al. 1994) or daily maxima. In a second step, one can determine the number of days on every grid cell. Not only the inappropriate data assumptions or laborious MCMC parameter estimates can be avoided, but a more detailed map may also result. The primary interest is still the creation of an accurate national map showing the *actual* number of exceedance days at a certain location.

Conclusion

Two methods were discussed here for a model-based geostatistical interpolation of the annual number of exceedance days. The Poisson model was found to give a better representation of the random field process of the number of exceedance days. For environmental assessment applications, however, we concluded the log-Normal model to be the preferred method for interpolation, considering its capacity to predict the expected number of exceedance days instead of an intensity field.

When making interpolations for a small area such as the Netherlands, incorporating observations from surrounding countries in the analysis was beneficial since the effective correlation distance of the data was approximately 300 km. This means that predictions near the Netherlands border still depend on observations far in Germany. Furthermore, including more observations improved parameter estimation and led to predictions that were more accurate.

Use of prior information in the Bayesian inference procedures avoids problems with convergence of the MCMC algorithm, which kept on fluctuating if flat priors were used in the subset. Also, even use of a limited data set allowed us to map the number of exceedance days. These maps, including their uncertainties, might be used in the future to study environmental relations between ozone and risks for public health, like for example the economic consequences of environmental health policies.

References

Besag JE (1994) Discussion on the paper by Grenander and Miller. J R Stat Soc B 56:591–592

Borowiak A, Lagler F, Gerboles M, DeSaeger E (2000) EC harmonization programme for air quality measurements. Intercomparison exercises 1999/2000 for SO_2, CO, NO_2 and O_3. EC report EUR 19629. Joint Research Centre, Ispra Italy

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Ass 88:9–25

Christakos G (1992) Random field models in earth sciences. Academic, New York

Christensen OF, Ribeiro PJ Jr.(2002) GeoRglm—a package for generalised linear spatial models. R News 2(2):26–28

Christensen OF, Waagepetersen RP (2002) Bayesian prediction of spatial count data using generalised linear mixed models. Biometrics 58:280–286

Cressie NAC (1993) Statistics for spatial data, revised edition. Wiley, New York

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (with discussion). J R Stat Soc C 47:299–350

EC (2002) Directive 2002/3/EC of the European Parliament and of the Council of 12 February 2002 relating to ozone in ambient air. Official Journal L 067, 09/03/2002

EEA (1998) Europe's environment, The second assessment. European Environment Agency. ISBN 92-828-3351-8, Office for Official Publications of the European Communities, Luxembourg. Elsevier Science Ltd. Kidlington, UK, pp 94–108

Elzakker BG van (2001) Monitoring activities in the Dutch national air quality monitoring network in 2000 and 2001. Internal report. RIVM report 723101055, Bilthoven, The Netherlands

ETC/ACC (2003) Airbase air quality information system. European topic centre on air and climate change, Bilthoven, The Netherlands

Feister U, Balzer K (1991) Surface ozone and meteorological predictors on a sub-regional scale. Atmos Environ 25A:1781–1790

Gelman A, Carlin JC, Stern H, Rubin DB (1995) Bayesian data analysis. Chapman & Hall, New York

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London

Gregg JW, Jones CG, Dawson E (2003) Urbanization effects on tree growth in the vicinity of New York city. Nature 424:183–187

Guttorp P, Meiring W, Sampson P (1994) A space–time analysis of ground-level ozone data. Environmetrics 5:241–254

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comp Graph Stat 5:299–314

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Papaspilliopoulus O, Roberts GO, Skold M (2003) Non-centered parameterizations for hierarchical models and data augmentation. In: Bernardo JM, Bayarri S, Dawid JO, Heckerman D, Smith AFM, West M (eds) Bayesian statistics 7. Oxford University Press, Oxford

Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS. Springer, New York

Ribeiro PJ Jr, Diggle PJ (1999) Bayesian inference in Gaussian model-based geostatistics. Technical report ST-99-08. Department of Maths and Statistics, Lancaster University, Lancaster UK

Ribeiro PJ, Jr Diggle PJ (2001) GeoR: a package for geostatistical analysis. R N EWS 1(2):15–18

Shively TS (1991) An analysis of the trend in ground-level ozone using non-homogeneous Poisson processes. Atmos Environ 25B:387–395

Smith RL (1989) Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Stat Sci 4:367–393

UNECE (1996) Critical levels for ozone in Europe: test and finalising the concept. In: Kärenlampi L, Skärby L (eds) UN-ECE workshop report. University of Kuopio, Finland

WHO (1996) Update and revision of the WHO air quality guidelines for Europe. Classical air pollutants; ozone and other photochemical oxidants. European Centre for Environment and Health, Bilthoven, The Netherlands