Article:
Gyarmati-Szabo, J, Bogachev, L and Chen, H (2011) Modelling threshold exceedances of air pollution concentrations via non-homogeneous Poisson process with multiple change-points. Atmospheric Environment, 45 (31). 5493 - 5503 . ISSN: 1352-2310

https://doi.org/10.1016/j.atmosenv.2011.06.049

Modelling threshold exceedances of air pollution concentrations via non-homogeneous Poisson process with multiple change-points

János Gyarmati-Szabó [a,*], Leonid V. Bogachev [b], Haibo Chen [a]

[a] *Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK*
[b] *Department of Statistics, University of Leeds, Leeds LS2 9JT, UK*

A B S T R A C T

A Bayesian multiple change-point model is proposed to analyse violations of air quality standards by pollutants such as nitrogen oxides (NO_2 and NO) and carbon monoxide (CO). Threshold exceedance occurrences are modelled by a step rate Poisson process fitted after short-range correlations in the exceedance data are removed via declusterisation. The change-points are identified, and the rate function is estimated, using an MCMC algorithm adapted from Green (1995). This technique is applied to the daily concentration data collected in Leeds, UK (1993–2009). Results are validated by running the MCMC estimator on the posterior-replicated data. Findings are discussed in the context of the past environmental actions and events. The proposed methodology may be useful for the air quality management by providing quantitative means to measure the efficacy of pollution control programmes.

Keywords: Air quality action plan; Urban air pollution; Threshold exceedances; Non-homogeneous Poisson process; Change-points; MCMC

1. Introduction

Due to harmful effects of pollution on human health and the environment, air quality in urban areas is a matter of worldwide concern amongst scientists, policy makers and public alike. Significant health problems (e.g., respiratory and cardiovascular) can be caused or worsened by exposure to the air pollution, from mild effects such as increased use of inhalers by asthmatics to more serious including advancement of death (Brimblecombe, 1998; Ayres, 2002).

The UK National Air Quality Strategy, adopted in 1997 (see AQS, 2007), determined statutory standards for eight main air pollutants, including nitrogen dioxide (NO_2), nitrogen monoxide (NO) and carbon monoxide (CO). In particular, UK local authorities are required to identify Air Quality Management Areas (AQMA) where appropriate, along with Air Quality Action Plans in pursuit of the air quality objectives (AQMA, 1997).

It is well documented that the pollutant emission to the atmosphere is dominated by road transport, estimated to contribute 54% of CO and 32% of NO_x emissions in

* Corresponding author. Tel.: +44 (0)113 343 1788; fax: +44 (0)113 343 5334.
E-mail address: tsjs@leeds.ac.uk (János Gyarmati-Szabó).

the UK (NAEI, 2010). In zones close to roadside, road traffic may be responsible for over 90% of the total emission of harmful pollutants such as NO_2 (AQAP, 2004). The UK Traffic Management Act 2004 (TMA, 2004) has put in place legislation aiming to promote a systematic integration of air quality management with transport planning. As a result of these measures, including tough regulations for industry and tightened emission standards for vehicles (EUR, 2003), an improved air quality has been achieved in the UK, even though elevated levels of pollution still occur (AQS, 2007; Faulkner and Russell, 2010).

Leeds City Council is taking a proactive stance in monitoring and managing air quality by deploying the state-of-the-art instrumentations in the city (with the data accessible to transportation and environmental researchers and modellers) and by implementing traffic control schemes. Two AQMAs were declared in small areas of the inner city (AQAP, 2004), with the action plans including measures such as traffic demand management, improvements to the highways network, reducing emissions from industrial and domestic sources, and raising awareness.

In order to monitor and assess the efficacy of these and future policies, it is important to develop adequate statistical methods to measure the impact of the regulations on the dynamics of various pollutants, especially with regard to the set standards. Clearly, any such methods should be linked to exceedance episodes, when the concentrations of one or more pollutants violate certain thresholds. This underpins the importance of developing reliable tools for statistical modelling of exceedances, including their prediction and control. Progress in this direction would provide the necessary feedback helping to assess and improve the current environmental policies, thus avoiding a grossly inefficient and costly "trial and error" approach. Furthermore, in addition to the environmental and health benefits, advancement of research methodologies in the statistical modelling of pollutant concentration extremes could have a significant economic value, potentially leading to massive savings in the public spending related to the government's environmental policies.

To date, a number of statistical methods have been developed to model violations of the air quality standards. The common approach is based on extreme value theory (see review papers by Thompson et al., 2001; Smith, 2004, and further references therein). Exceedances of air pollution concentrations can also be modelled using non-homogeneous Poisson processes (see Raftery, 1989; Smith, 1989; Shively, 1991; Smith and Shively, 1995; Achcar et al., 2009). This approach is motivated by the asymptotic theory of stationary point processes stating that the series of occurrences of high threshold exceedances can be approximated by a Poisson process (Leadbetter et al., 1983). For nonstationary data, the approximating Poisson process has to be non-homogeneous, with the intensity rate depending on time. Estimation of the unknown rate function $\lambda(t)$ is greatly facilitated by choosing certain parametric classes, such as $\lambda(t) = \alpha t^{\beta-1}$, $\lambda(t) = \beta/(t+\alpha)$, $\lambda(t) = \alpha e^{-\beta t}$ ($\alpha, \beta > 0$), etc. (Achcar et al., 2009). Note that these and other conventional formulas are restricted by a monotone behaviour of $\lambda(t)$. More complex models are obtained by incorporating the dependence of parameters on "external" factors (covariates) such as meteorological

conditions (Smith and Shively, 1995). However, such models are not suitable if the rate may change abruptly (e.g., following an updated air quality action plan). To capture this type of variation in the exceedance dynamics, Achcar et al. (2009) proposed a parametric non-homogeneous Poisson model allowing to handle up to three change-points using complicated iterative updates of the prior information.

In the present paper, we use a non-homogeneous Poisson model with no prior restriction on the number of change-points. For simplicity, the rate $\lambda(t)$ is assumed to be a step function (i.e., constant between adjacent change-points). Note that statistics of threshold exceedances is likely to involve short-range correlations due to patches of elevated concentration levels. Therefore, to ensure a better fit of the Poisson model it is essential to apply a suitable thinning (declusterisation) of the threshold exceedance data (Shively, 1991). An efficient MCMC estimator is then used to automatically determine the number and locations of change-points, as well as the respective values (heights) of the rate function. These techniques are applied to NO_2, NO and CO data collected in Leeds, UK, as daily concentration maxima over nearly 17 years (1993–2009).

The paper is organised as follows. In Section 2, we review basic properties of non-homogeneous Poisson processes and set out the step rate model; formal statistical tests to compare different Poisson models are also outlined there. Section 3 deals with the model specification and preprocessing, including thresholding, imputing missing values, deseasonalisation and declusterisation. The results of fitting a step rate Poisson model to the NO_2, NO and CO threshold exceedance data, using a suitable MCMC estimator, are reported in Section 4, followed by the model verification and validation. Section 5 contains the discussion, including some preliminary interpretation of the results in terms of past air quality related actions and events. Finally, a brief summary of the work is given in Section 6.

2. Poisson model

2.1. Non-homogeneous Poisson process

Poisson process (Cox and Lewis, 1966) is a statistical model describing a series of uncorrelated random events (“arrivals”), whereby new arrivals are statistically independent of any past arrivals. Rate function $\lambda(t)$ of the Poisson process determines its local intensity. The number $N(t)$ of arrivals on $[0,t]$ is distributed according to a Poisson law,

$$P\{N(t)=n\} = \frac{\Lambda(t)^n}{n!} e^{-\Lambda(t)}, \qquad n = 0,1,2,\ldots, \tag{1}$$

where $\Lambda(t) = \int_0^t \lambda(u)\,du$ is the expected number of arrivals. More generally, the increments $N(t)-N(s)$ $(t>s)$ have Poisson distribution with parameter $\Lambda(t)-\Lambda(s)$. The waiting time τ to first arrival after time s satisfies

$$P\{\tau > t - s\} = P\{N(t) - N(s) = 0\} = e^{-\Lambda(t)+\Lambda(s)} \qquad (t > s).$$

Hence, the joint density of n consecutive arrivals restricted to $[0,T]$ reads

$$f(t_1,\ldots,t_n) = e^{-\Lambda(T)+\Lambda(t_n)} \prod_{i=1}^{n} \lambda(t_i)\, e^{-\Lambda(t_i)+\Lambda(t_{i-1})} = e^{-\Lambda(T)} \prod_{i=1}^{n} \lambda(t_i) \qquad (2)$$

$$(0 = t_0 < t_1 < \cdots < t_n < T)$$

with the corresponding log-likelihood

$$\log L(t_1,\ldots,t_n) \equiv \log f(t_1,\ldots,t_n) = \sum_{i=1}^{n} \log \lambda(t_i) - \Lambda(T) \qquad (0 < t_1 < \cdots < t_n < T). \qquad (3)$$

Formulas (1), (2) imply that the joint probability density of arrivals conditioned on their total number $N(T) = n$ is given by

$$f_n(t_1,\ldots,t_n) = \frac{n!}{\Lambda(T)^n} \prod_{i=1}^{n} \lambda(t_i) \qquad (0 < t_1 < \cdots < t_n < T), \qquad (4)$$

which coincides with the density of the increasing order statistics for n independent points thrown in the interval $[0,T]$ with the density $\lambda(t)/\Lambda(T)$ each (e.g., for a constant rate this amounts to the uniform distribution on $[0,T]$). In turn, formula (4) implies that the values $t_i' = \Lambda(t_i)/\Lambda(T)$ have the same distribution as the uniform order statistics on $[0,T]$ (Lewis and Shedler, 1976), which allows one to construct approximate goodness-of-fit tests for particular classes of the rate functions using estimates of the model parameters (see Section 2.2 below).

In the present paper, the rate $\lambda(t)$ is assumed to be a step function,

$$\lambda(t) = h_j \quad \text{if} \quad s_j \le t < s_{j+1} \quad j = 0,1,\ldots,k \quad (0 = s_0 < s_1 < \cdots < s_k < s_{k+1} = T), \qquad (5)$$

where the number k of change-points, as well as their positions s_j and heights h_j of the corresponding steps are not predetermined and therefore have to be statistically estimated.

2.2. Statistical tests for the Poisson model

To assess the utility of the change-point Poisson process as a fitting model, one can carry out statistical tests to compare three different models:

- M_0: homogeneous Poisson process, $\lambda(t) \equiv \text{const}$;
- M_1: Poisson process with a log-linear rate, $\lambda(t) = \alpha \exp(-\beta t)$;
- M_2: Poisson step rate model with one change-point.

For observed threshold exceedance data $X: 0 < t_1 < t_2 < \cdots < t_n < T$, denote

$$S_n = \sum_{i=1}^{n} u_i\,, \qquad u_i = t_i / T \qquad (i = 1, \ldots, n)\,. \tag{6}$$

A classical goodness-of-fit test for M_0 is based on the uniform distribution of arrivals (Section 2.1), implying that S_n is approximately normal with mean $n/2$ and variance $n/12$ (Cox and Lewis, 1966). A test for M_0 against M_2 (Raftery, 1989) is based on the statistic (Raftery, 1989)

$$\Delta_n = \frac{1}{\sqrt{n}} \max\left\{ \left| g(i-1,\, u_i) \right|, \left| g(i,\, u_i) \right| \right\}, \tag{7}$$

where the maximum is taken over all $u_i \in [0.01,\, 0.99]$, and

$$g(i,u) = i\sqrt{(1-u)/u} - (n-i)\sqrt{u/(1-u)} \qquad (i = 1, \ldots, n, \quad 0 < u < 1)\,.$$

Approximate critical values can be computed from the formula (Akman and Raftery, 1986)

$$P(\Delta_n > z) \approx \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \left(\xi z - \xi z^{-1} + z^{-1} \right), \qquad \xi = \log \frac{0.99}{0.01} = 4.595\ldots \tag{8}$$

For the model M_1, the coefficient α is a nuisance parameter eliminated by switching to the conditional distribution with $n > 0$ observed arrivals in $[0, T]$. The slope parameter β is estimated from the maximum likelihood equation $\frac{\partial \log L_n}{\partial \beta}(\hat{\beta}) = 0$, in view of (3) reduced to (Cox and Lewis, 1966)

$$\frac{1}{\hat{\beta} T} - \frac{e^{-\hat{\beta} T}}{1 - e^{-\hat{\beta} T}} = \frac{S_n}{n}\,. \tag{9}$$

As was explained in Section 2.1, the modified values

$$t_1' = \frac{\hat{\Lambda}(t_1)}{\hat{\Lambda}(T)}, \quad \ldots, \quad t_n' = \frac{\hat{\Lambda}(t_n)}{\hat{\Lambda}(T)}\,, \tag{10}$$

with $\hat{\Lambda}(t) = \int_0^t \hat{\lambda}(u)\, du = \alpha \hat{\beta}^{-1} \left(1 - \exp(-\hat{\beta} t) \right)$, are approximately distributed as the uniform order statistics, so the uniformity test described above may be applied (Lewis and Shedler, 1976).

Models can also be compared using the Bayes factor B_{ij} (Raftery, 1996), defined as the ratio of posterior to prior odds for M_i against M_j:

$$B_{ij} = \frac{P(X \mid M_i)}{P(X \mid M_j)}, \qquad X : t_1 < \cdots < t_n. \tag{11}$$

An indicative scale for interpreting B_{ij} is given in Raftery (1996, p. 165); e.g., values from 12 to 150 are classed as strong evidence in favour of the null hypothesis M_i. The Bayes factors for M_0 against M_1 and M_2 (Raftery, 1989) are given by

$$B_{01} = 0.6449(n-1)\left[\int_0^\infty e^{-S_n y}\left(\frac{y}{1-e^{-y}}\right)^{n-1} dy\right]^{-1}, \tag{12}$$

$$B_{02} = \frac{4\sqrt{\pi}\,\Gamma(n+1/2)}{\sum_{i=0}^{n} J_i \Gamma(i+1/2)\Gamma(n-i+1/2)}, \tag{13}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ and $J_i = \int_{u_i}^{u_{i+1}} x^{-(i+1/2)}(1-x)^{-(n-i+1/2)}dx$. Using equation (11), one also obtains

$$B_{12} = \frac{B_{02}}{B_{01}}. \tag{14}$$

3. Model preprocessing

3.1. Thresholding

To set a specific threshold in order to extract the statistics of exceedances, it might seem natural to use the air quality standards (AQS, 2007). However, there are several difficulties: (i) the standards are set out in terms of hourly or maximum daily rolling 8-hour averages, while we work with the daily data; (ii) the standards limit permitted frequencies of threshold violations, which causes the "moving window" type dependence in the exceedance data; (iii) the standard threshold values produce scarce statistics of the resulting exceedances (e.g., 0.11% for NO_2).

A simple alternative is to use the quantile thresholding at, say, a 90th empirical percentile, leading to the specific threshold values of 96 μg m^{-3} (NO_2), 185 μg m^{-3} (NO) and 2.1 mg m^{-3} (CO). Cumulative counts of the resulting exceedance data for all three pollutants are shown in Fig. 1.

3.2. Missing values

Our concentration level data contain a noticeable proportion of missing values (6.6%, 5.5% and 9.9% in the NO_2, NO and CO data, respectively), which considerably decreases the exceedance statistics available and therefore may adversely affect the estimation of the unknown rate function $\lambda(t)$.

One can improve estimation by compensating for the omissions (Cox, 1981). In this work, we impute the missing values using a (two-sided) moving average estimator with the window size ± 65 days. We first impute raw concentration values drawn independently according to their estimated (univariate) distribution, and then apply thresholding. Even though such interpolation neglects possible short-range correla-

tions in the concentration time series, it may be expected to work well due to a subsequent high-level thresholding (Leadbetter et al., 1983).

<Figure 1>

<Table 1>

The imputed exceedances for each pollutant are indicated in Fig. 1. Table 1 shows the MCMC results for NO_2 obtained before and after imputation. In both cases, the number of change-points identified by MCMC was found to be $k = 1$, with almost the same locations; however, the step heights of the rate function noticeably increased (by about 5%) as a result of more exceedances.

3.3. Deseasonalisation

It has to be taken into account that the concentration levels of all three pollutants under study (NO_2, NO and CO) are affected by meteorological conditions, e,g. inducing yearly oscillations (seasonality) in time series. For nitrogen oxides NO_2 and NO, this may be explained by photochemical reactions between NO_2 and NO involving ozone O_3 (Clapp and Jenkin, 2001), leading to decreased levels of NO_x during the summer due to higher solar radiation. Carbon monoxide CO is relatively nonreactive, but low temperatures in winter contribute to higher CO concentrations due to incomplete combustion in vehicle engines (CCME, 2002). Yearly cycles in the concentration time series are clearly visible in the daily data plots (Fig. 2, top), confirmed by computing the autocorrelation at lag 365 days (0.17 for NO_2 and NO, 0.33 for CO). Furthermore, analyzing the estimated spectral density of the observed time series (Venables, 2002), seasonality was found to be significant.

Deseasonalisation may be based on fitting a regression model for the log-transformed data (Cox and Lewis, 1966)

$$\log \tilde{X}(t) = \log X(t) - a\cos\omega t - b\sin\omega t - c\,, \qquad \omega = \frac{2\pi}{365}\,(\text{days}^{-1})\,, \tag{15}$$

Possible deviations from stationarity in the concentration data are likely to occur on a scale much slower than the annual variability, hence the regression estimation may be expected to give satisfactory results (cf. Cox and Lewis, 1966). Periodic trends filtered out from the raw data are shown in Fig. 2 (top). Deseasonalised data $\tilde{X}(t)$ are plotted in Fig. 2 (bottom) showing a much improved scattering, with no visible periodic pattern

<Figure 2>

Alternative methods of deseasonalisation may be based on moving-average, kernel type local smoothing or spectral techniques (Cox and Lewis, 1966; Lewis and Shedler, 1976). One could also take into account weekly periodicities due to traffic patterns caused by peak hours and weekday–weekend oscillations. However, weekly oscillations occur much more frequently as compared to the annual scale, so are better averaged out by fitting a step rate Poisson model; in addition, the weekly patterns are likely to be rather stable across decades.

3.4. Declusterisation

The time series of threshold exceedances by air pollution concentrations is likely to involve dependence (autocorrelation) manifested in clustering of consecutive exceedances, which may render Poisson model unsuitable. However, thanks to a relatively short range of such correlations, the dependence problem can be rectified by a suitable thinning, or "declusterisation" (Leadbetter et al., 1983; Shively, 1991), usually achieved by retaining one point from each cluster (merged with the subsequent point below the threshold to remove the tie).

Significance of dependence before and after declusterisation may be tested using the sample correlation coefficient for gaps between exceedances or by estimating serial correlation at various lags (Venables, 2002). We deployed the nonparametric Wald–Wolfowitz runs test of independence (Gibbons and Chakraborti, 2003) applied to a two-valued data sequence representing threshold exceedances versus non-exceedances. Note that the runs test applied to the entire declustered sequence may still reject H_0 due to nonstationarity, so it should be used on separate intervals between the change-points, which need to be estimated in advance.

In this study, declusterisation resulted in a reduction of the data size by about 30–35%. The runs test accepted independence on all intervals between the estimated change-points, with the p-values above 0.10.

4. Results for the air pollution data

In this section, we report the results of estimation of the unknown step rate function $\lambda(t)$ to fit a non-homogeneous Poisson process to threshold exceedance data obtained from the NO_2, NO and CO concentration level measurements. The data were collected in the Leeds city centre by a permanent monitoring station located about 30 metres from a busy 4-lane inner-city road and 150 metres from an urban motorway, with traffic flow of approximately 21,500 and 93,500 vehicles per day, respectively (AURN, 2010). The data used in the study (AQA, 2010) correspond to the daily maxima for each of the three pollutants, observed from 4 January 1993 to 31 December 2009 (i.e., for $T = 6206$ days).

4.1. Posterior estimates of the rate function

From the cumulative plot of exceedance frequencies for NO_2 (Fig. 1), it is evident that the initial part of the graph up to $t \approx 2500$ is close to a linear function (with the slope about $280/2500 = 0.112$), followed by a noticeable drop of the slope to ap-

proximately $(300-280)/(3600-2500)=0.018$ until $t\approx 3600$. One may also suspect one or two less significant change-points further on (e.g., near $t\approx 4500$). Similarly, the cumulative graphs for NO and CO in Fig. 1 suggest about four to five hypothetical change-points each. Hence, as a prior distribution for the unknown number k of change-points it is reasonable to choose a Poisson law with mean 4.5.

The programming code for the Markov chain Monte Carlo (MCMC) algorithm adapted from Green (1995) was implemented in R using a standard laptop (AMD processor DualCore 2.2 GB, RAM 4GB). The sampler takes 1000 steps within 20 seconds. Manual programming was necessary since the standard MCMC engines (e.g., WinBUGS, http://www.mrc-bsu.cam.ac.uk/bugs/) are not yet suitable for models with variable dimension.

After a burn-in period of length 15,000–20,000, pilot runs established that the dynamics of the Markov chain have reached the stationary regime. As usual, convergence of the sampling algorithm was monitored by visual inspection of the output plots and also by simple convergence diagnostic tools based on the MCMC output for a single Markov chain (Gilks et al., 1996). In addition, reliability of the burn-in runs was verified by trying different initial values, also utilised to control the termination time of the MCMC algorithm. In the stationary regime, the MCMC was run for 500,000 updates. Examining the autocorrelation plots, it was determined that independence within the sample is ensured by selecting every 40th generated value, thus providing a posterior sample of size $500{,}000/40=12{,}500$.

<Figure 3>

As a result of the MCMC performance, it was concluded that the MCMC sampler spent most of its time in parametric states corresponding to $k=1$, $k=3$ and $k=4$ change-points for NO_2, NO and CO, respectively (Fig. 3). The change-point locations s_j, estimated via the modes of the posterior marginal densities, are shown in Fig. 4. Specific time values of the change-points, along with 50% credible intervals, are given in Table 2. The step heights h_j of the rate function $\lambda(t)$ were estimated via the median of the posterior marginal distributions (see Fig. 5 and Table 3). Note that the parametric dimension of the model is bounded by

$$\dim(k)+\dim(s_j)+\dim(h_j)\le 1+4+5=10,$$

furnishing $12{,}500/10=1250$ sampled values per parameter, which ensures good accuracy of the posterior estimation.

<Figure 4>

<Table 2>

<Figure 5>

<Table 3>

<Figure 6>

The graph of the step function $\hat{\lambda}(t)$ (with estimated change-points s_j and step heights h_j) is shown in Fig. 6 along with an "integral" estimate $\bar{\lambda}(t)$ of the posterior mean rate function calculated by averaging 5000 sample step rate functions drawn from the MCMC posterior distribution. The transition areas for $\bar{\lambda}(t)$ near the change-points are typically quite narrow, confirming a stepwise nature of the rate function $\lambda(t)$. Fig. 6 also shows the cumulative graphs simulated via Poisson processes with rates $\hat{\lambda}(t)$ and $\bar{\lambda}(t)$, contrasted with the observed cumulative plot (cf. Fig. 1). As could be anticipated, the integrated (cumulative) estimates demonstrate a much better fit to the data.

4.2. Verification of the Poisson change-point model

The fitted step rate Poisson model can be verified by checking uniform distribution of exceedance occurrences between the consecutive change-points (Section 2.1). Specifically, for NO_2 there are two step intervals separated by the estimated change-point $s_1 = 2490$ and containing $n_0 = 257$ and $n_1 = 133$ exceedances. Using the uniformity test based on (6), we obtain the approximate p-values 0.6326 and 0.4907, respectively, confirming good quality of the fitted model. For NO and CO, the results were similar.

Further verification is motivated by challenging the lack of change-points on some intervals. For instance, to compare two competing models M_0 and M_2 for NO_2 on the interval $[2490, 6206]$, we evaluate the test statistic (7) (with $n = n_1 = 133$) and use formula (8) to find the approximate p-value 0.1457, indicating that model M_0 is acceptable. Moreover, formula (13) gives the Bayes factor $B_{02} = 33.39$, which constitutes strong evidence in support of M_0. Similar calculations for NO and CO also confirmed the absence of "extra" change-points.

It is interesting to assess the stepwise variation of the rate function $\lambda(t)$ as opposed to a more gradual change, especially for NO and CO (Fig. 1), by testing model M_1 against the alternative M_2. Starting with the NO data, from equation (9) we estimated the slope $\hat{\beta} = 0.000313$ and tested the uniformilty of the modified exceedances (10). The statistic (6) gave the p-value 0.025, suggesting that M_1 should be rejected. This was confirmed by calculating the Bayes factor $B_{21} = 53.17$ using

equations (12), (13) and (14). Similar results were obtained for CO, confirming that model M_1 is less suitable than the fitted step rate model.

4.3. *Validation of the fitted Poisson model*

To validate the employed MCMC sampler, we simulated data sets from non-homogeneous Poisson processes with the estimated step function $\hat{\lambda}(t)$ as a rate function and again applied MCMC to see if there were any significant differences in the estimates. The results (Fig. 7) show an excellent match of the posterior estimates for the simulated data sets with those obtained earlier for the observed data.

<Figure 7>

<Figure 8>

Following the method of *posterior predictive simulation* (Gilks et al., 1996), replicated the counting processes and compared them to the corresponding cumulative plots (Fig. 8). Due to random fluctuations, deviations of the replicated cumulative plots increase over time. The accuracy may be improved by simulating Poisson processes conditioned to have the same number of occurrences as in the real data. Conditional replicates are presented in Fig. 8, showing an excellent fit with the observed plot and also consistent with the sample mean plot for data sets simulated with a posterior estimate $\hat{\lambda}(t)$.

5. Discussion

5.1. General comments

Statistical analysis of observed pollutant concentrations developed in this paper is essentially based on the extreme value approach by utilising high threshold exceedances for statistical inference. We focused on the corresponding point process of exceedance occurrences in time, whereas the actual values of exceedances were not taken into consideration. This simplified model has an advantage of being less sensitive to deviations from the model assumptions, and hence more useful in real data applications. The algorithmic tractability of the model makes it suitable for an efficient MCMC multi-parametric estimation. As has been demonstrated in this work, MCMC provides a reliable identification of the unknown change-points in the rate function, including their number, location and the corresponding step heights.

A global downward trend in the behaviour of the rate functions for each pollutant (Fig. 6) indicates a general improvement in the air quality, arguably promoted by environmental actions (NAEI, 2010). It may also be attributable to the improvement of car engines (in particular, using the three-way catalytic converters as standard) and to the developing change in drivers' attitudes (Glaister, 2002). Since a step rate

model is unsuitable for capturing this kind of dynamics, it is interesting to consider a more general class

$$\lambda_\beta(t) = e^{-\beta t}\lambda(t), \tag{16}$$

where $\lambda(t)$ is a pure step function (5). Model (16) can be reduced to (5) by filtering out the log-linear term together with annual periodicity, using the regression equation (cf. (15))

$$\log \tilde{X}(t) = \log X(t) + \beta t - a\cos\omega t - b\sin\omega t - c. \tag{17}$$

Fig. 9 illustrates the results of fitting the combined trend (17) to the raw data; note a better fit as compared to a pure periodic trend (cf. Fig. 2).

Due to the computational complexity of the model and the distribution-oriented nature of Bayesian inference, specific conclusions of the MCMC-based analysis should be treated with caution. There is a need for an additional cross-examination and verification of these findings in co-operation with experts from other subject areas, including policy makers as well as environmental and transport modellers and practitioners. Special care is required when relating the conclusions of the threshold exceedance analysis with official statistics (see, e.g., NAEI, 2010; Faulkner and Russell, 2010) usually based on averaged observations rather than on extreme values. This underpins possible significant discrepancies in the interpretation, since extreme value statistics may appear more sensitive to traffic pollution changes than the typical, bulk values (Smith, 2004).

<Figure 9>

5.2. Mapping and interpretation

The following is a preliminary attempt of "mapping" of our results, i.e., relating the identified change-points with real transport and environment related circumstances.

5.2.1. High Occupancy Vehicle lane. Two change-points in Table 2 (29 October 1999, NO_2, and 11 November 1999, NO) may be associated with the High Occupancy Vehicle (HOV) lane scheme, designed to give priority to vehicles with more than one commuter (HOVL, 2004). The HOV lane in Leeds was opened on 11 May 1998 on the A647 Stanningley Road and Stanningley By-Pass, which form the principal radial route to the west of the city centre. By the official assessment (HOVL, 2004), a relatively small improvement of air quality was attributed "to reduced vehicle emissions rather than to the impact of the HOV lane". However, the reduced congestion on A647 could also affect the traffic in the city centre where our data were collected (AQA, 2010). In addition, due to predominantly westerly winds in North East England (MET-O, 2010), an improved air quality near A647 might positively contribute to the city centre (5 miles east of the HOV lane).

5.2.2. East Leeds Link Road. Leeds City Council initiated a number of projects to improve the highway network, including the East Leeds Link Road (ELLR) designed to reduce through-traffic in the city centre (ELLR, 2010). The new 4 km dual carriageway link between junction 45 of the M1 and the Leeds Inner Ring Road was opened on 10 February 2009, expected to provide "indirect" improvements in the air quality, with low to moderate impact (AQAP, 2004). Our model has not detected any change-points around February 2009 or after, but using more data beyond December 2009 might make it different.

5.2.3. Leeds Free City Bus. This zero-fare service (3.4 miles long loop in the city centre) began on 30 January 2006. It is estimated that the scheme has reduced the numbers of cars entering the city centre (LCCA, 2010) and hence contributed towards meeting a target of 10% reduction by 2011 in NO_2 (LCCA, 2010). The scheme's impact was not strong enough to be picked up by our model, perhaps being masked by the general downward trend.

5.2.4. Christmas 2004. Two change-points in Table 2 (1 December 2004, NO, and 8 January 2005, CO), with overlapping 50%-credible intervals, are both close to the Christmas and New Year holidays when the traffic drops dramatically. It is unclear if this alone has affected the threshold exceedance data so much as to lead to a statistically significant change-point for NO and CO. Incidentally, note a strong drop in the raw time series for NO_2 around 16 December 2001 (Fig. 2); however, no change-points were detected anywhere near.

5.2.5. *Construction works.* There were construction works near the monitoring station in the city centre, from mid-January 2005 until end of June 2009 (AURN, 2010). There were some concerns that the associated transportation and engineering activities might be responsible for sporadic elevated pollution levels in the area. From Table 2, the latest identified change-point was 8 December 2005 (for CO), with the 50% credible interval covering the commencement date of the construction. However, causal association seems unlikely due to a decrease of the estimated rate, suggesting that any contribution to the air pollution in NO_2, NO and CO was relatively mild as compared to a presumably dominant input from intensive traffic on the nearby roads (AURN, 2010). On the other hand, contributions to other pollutants might have been more prominent (e.g., particulate matter due to dust) and hence could be picked up by our methods; indeed, PM_{10} concentration data (not reported here) reveal substantially elevated levels from April to December 2005.

5.2.6. *Fuel crisis 2000*. The UK "fuel crisis", lasting one week in September 2000, was caused by the increase of the fuel prices and subsequent public protests including blockade of oil facilities, which led to disruption of fuel deliveries and hence to significant reductions in traffic flows (Hathaway, 2001). According to AQAP (2004), this resulted in a dramatic improvement of the air quality in the city (e.g., by 30–40% for NO_2). Our model has not picked up any noticeable change near this point in time, which is not surprising in view of a very short-lived impact of this episode.

5.2.7. *European emission standards.* These standards determined acceptable exhaust emission limits for new vehicles sold within the European Union (EUR, 2003); e.g., for passenger cars the staged inception dates were: July 1992 (Euro 1), January 1996 (Euro 2), January 2000 (Euro 3), January 2005 (Euro 4) and September 2009 (Euro 5). Some of these dates may be related to the identified change-points (Table 2). In particular, 29 October 1999 (NO_2) and 11 November 1999 (NO) are close to Euro 3 (covered by the 50% credible intervals), while the Christmas 2004 change-points (Section 5.2.4) may be linked with Euro 4. Furthermore, 18 May 1995 (NO) and 5 April 1995 (CO) may be attributed (even though with somewhat lower credibility) to Euro 2 which significantly tightened up the CO standard for diesel cars and reduced the NO_x permitted levels by 25–45% (EUR, 2003).

6. Conclusions

In this paper, we have used a non-homogeneous Poisson model to fit the observed series of threshold exceedance occurrences in time, extracted from the daily pollution concentration data obtained in Leeds (UK) in 1993–2009. The time dependence of the unknown Poisson rate was assumed to be a step function, with an arbitrary number of possible change-points and no restrictions on its values (in particular, with no *a priori* requirement of monotonicity). The principal aim of the modelling was to identify statistically significant change-points in the rate function, signalling notable changes in the pollution dynamics patterns. Statistical estimation of the unknown rate function was carried out using Bayesian posterior samples obtained by adapting an MCMC algorithm proposed by Green (1995). Our results have demonstrated the computational and statistical efficiency of the MCMC method.

Once the statistical identification is carried out and validated (e.g., by simulations), it is important to seek a meaningful interpretation of the results, bearing in mind updates in the environmental policies by local and central authorities as one possible source of the emerging change-points. Certain care is needed in practical applications of the Poisson model, since “false” change-points may be induced by meteorological conditions, seasonality, missing values and cluster correlations in threshold exceedances. We have addressed these issues by the data preprocessing (including deseasonalisation and declusterisation), supported by a graphical and statistical assessment of the results.

The application of non-homogeneous Poisson processes with multiple change-points may provide a suitable modelling framework in the air quality management, which can be used to analyse pattern changes in the violation of air quality standards by various pollutants. In particular, these methods may be instrumental in obtaining important feedback about environmental actions and assessing their impact on the dynamical patterns of potentially hazardous pollutants.

Acknowledgements

J. Gyarmati-Szabó was supported by an EPSRC Doctoral Training Grant, by the Strategic Fund of the Institute for Transport Studies and by a Postgraduate Research

Scholarship of the School of Mathematics (University of Leeds). L.V. Bogachev was partially supported by a Leverhulme Research Fellowship.

References

Achcar, J.A., Rodrigues, E.R., Tzintzun, G., 2009. Using non-homogeneous Poisson models with multiple change-points to estimate the number of ozone exceedances in Mexico City. Environmetrics, published online, doi: 10.1002/env.1029 (accessed 30 March 2010).

Akman, V.E., Raftery, A.E., 1986. Asymptotic inference for a change-point Poisson process. Annals of Statistics 14, 1583–1590.

AQA, 2010. The UK Air Quality Archive, http://www.airquality.co.uk/data_and_statistics.php (accessed 6 April 2010).

AQAP, 2004. Air Quality Action Plan, Leeds City Council, http://www.leeds.gov.uk/files/Internet2007/2010/15/doc%2012.pdf (accessed 16 October 2010).

AQMA, 1997. Air Quality Management Areas. Defra http://aqma.defra.gov.uk/. Information about Leeds: http://aqma.defra.gov.uk/maps.php?map_name=fulluk&la_id=143 (accessed 26 July 2010).

AQS, 2007. The Air Quality Strategy for England, Scotland, Wales and Northern Ireland, Vols. 1, 2. Cm 7169 NIA 61/06-07, Defra, http://www.defra.gov.uk/environment/quality/air/airquality/strategy/ (accessed 5 May 2008).

AURN, 2010. UK Automatic Urban and Rural Network, Defra, Leeds Centre (station #54), http://aurn.defra.gov.uk/stations/viewStation.php?id=54 (accessed 21 October 2010).

Ayres, J.G., 2002. Chronic effects of air pollution. Occupational and Environmental Medicine 59, 147–148.

Brimblecombe, P., 1998. History of urban air pollution. In: Fenger, J. et al. (Eds.), Urban Air Pollution — European Aspects, Kluwer, Dordrecht, pp. 7–20.

CCME, 2002. The Ongoing Challenge of Managing Carbon Monoxide Pollution in Fairbanks, Alaska. National Academy Press, Washington; http://www.nap.edu/catalog/10378.html (accessed 6 October 2010).

Clapp, L.J., Jenkin, M.E., 2001. Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NO_x in the UK. Atmospheric Environment 35, 6391–6405.

Cox, D.R., 1981. Statistical analysis of time series: Some recent developments. Scandinavian Journal of Statistics 8, 93–108.

Cox, D., Lewis, P.A.W., 1966. The Statistical Analysis of Series of Events. Methuen, London.

ELLR, 2010. East Leeds Link Road. Leeds City Council, www.leeds.gov.uk/ellr (accessed 1 August 2010).

EUR, 2003. Trends in Vehicle and Fuel Technologie. Report EUR 20746 EN, EC Joint Research Centre, http://ftp.jrc.es/EURdoc/eur20746en.pdf (accessed 16 October 2010).

Faulkner, M., Russell, P., 2010. Review of Local Air Quality Management. Report, IHPC, http://www.defra.gov.uk/environment/quality/air/airquality/local/documents/laqm-report.pdf (accessed 16 October 2010).

Gibbons, J.D., Chakraborti, S., 2003. Nonparametric Statistical Inference, 4th edition. Marcel Dekker, New York.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Introducing Markov chain Monte Carlo. In: Gilks, W.R. et al. (Eds.), Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC, London, pp. 1–19.

Glaister, S., 2002. UK transport policy 1997–2001. Oxford Review of Economic Policy 18, 154–186.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Hathaway, P., 2001. The effect of the fuel 'protest' on road traffic. In: Transport Trends: 2001 Edition. TSO, London; http://web.archive.org/web/20040224182437/http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dft_transstats_505952.pdf (accessed 1 August 2010).

HOVL, 2004. Case study: High occupancy vehicle lanes, A647 Stanningley Road, Leeds. In: Bus Priority: The Way Ahead, 2nd edition, http://www.dft.gov.uk/pgr/regional/buses/bpf/busprioritythewayahead12/ (accessed 8 October 2010).

LCCA, 2010. 9th Leeds City Centre Audit 2010. Leeds City Council, http://www.leedsinitiative.org/WorkArea/DownloadAsset.aspx?id=18786 (accessed 10 October 2010).

Leadbetter, M.R., Lindgren, G., Rootzén, H., 1983. Extremes and Related Properties of Random Sequences and Processes. Springer, New York.

Lewis, P.A.W., Shedler, G.S., 1976. Statistical analysis of non-stationary series of events in a data base system. IBM Journal of Research and Development 20, 465–482.

MET-O, 2010. Met Office UK. Climate: North East England, http://www.metoffice.gov.uk/climate/uk/ne/ (accessed 8 October 2010).

NAEI, 2010. UK Emissions of Air Pollutants 1970 to 2008. Report, National Atmospheric Emission Inventory (NAEI), http://www.airquality.co.uk/reports/cat07/1009030925_2008_Report_final270805.pdf (accessed 16 October 2010).

Raftery, A.E., 1989. Comment: Are ozone exceedance rates decreasing?.Statistical Science 4, 378–381.

Raftery, A.E., 1996. Hypothesis testing and model selection. In: Gilks, W.R. et al. (Eds.), Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC, London, pp. 163–187.

Shively, T.S., 1991. An analysis of the trend in ground-level ozone using non-homogeneous Poisson processes. Atmospheric Environment 25, 387–395.

Smith, R.L., 1989. Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. Statistical Science 4, 367–377.

Smith, R.L., 2004. Statistics of extremes, with applications in environment, insurance, and finance. In: Finkenstädt, B., Rootzén, H. (Eds.), Extreme Values in Finance, Telecommunications, and the Environment, Chapman & Hall/CRC, Boca Raton, pp. 1–78.

Smith, R.L., Shively, T.S., 1995. Point process approach to modeling trends in tropospheric ozone based on exceedances of a high threshold. Atmospheric Environment 29, 3489–3499.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment 35, 617–630.

TMA, 2004. Traffic Management Act 2004 (c. 18). TSO, London; http://www.statutelaw.gov.uk/content.aspx?activeTextDocId=1606563 (accessed 26 July 2010).

Venables, W., 2002. Modern Applied Statistics with S, 4th edition. Springer, New York.

1
2
3

Figure captions

Fig. 1. Cumulative count plots of threshold exceedances for the observed raw data (*black*) and for the data with the missing values estimated via the moving-average with a two-sided window of size ± 65 days (*grey*). Black ticks on the time axis indicate the occurrence of missing values; red points on the grey curve show the imputed exceedance values.

Fig. 2. *Top:* observed daily maximum concentrations of NO_2, NO and CO (*black*) and the estimated yearly periodic trend for the log-transformed data (*red*). *Bottom:* deseasonalised data, with the yearly periodic trend removed.

Fig. 3. Estimated MCMC posterior distribution of the number k of change-points in the step rate function $\lambda(t)$, obtained for the observed threshold exceedance data (*red*) and for data simulated via non-homogeneous Poisson process with the estimated step function $\hat{\lambda}(t)$ as the rate (*blue*).

Fig. 4. Estimated MCMC posterior density functions of the marginal distributions for the locations s_j of the identified change-points (cf. Table 2), obtained for the observed threshold exceedance data (*red*) and for data simulated via non-homogeneous Poisson process with the estimated step function $\hat{\lambda}(t)$ as the rate (*blue*). The estimates were computed using a Gaussian kernel with bandwidth 95 days.

Fig. 5. Estimated MCMC posterior density functions of the univariate marginal distributions for the heights h_j of the estimated change-points (cf. Table 3), obtained for the observed threshold exceedance data (*red*) and for data simulated via the estimated step rate function $\hat{\lambda}(t)$ (*blue*). The estimates were computed using a Gaussian kernel with bandwidth 0.003 days^{-1}.

Fig. 6. MCMC results for each pollutant: estimated rate functions (*left axes*) and the corresponding cumulative plots for data sets simulated via non-homogeneous Poisson process (*right axes*). Colour-coded plots correspond to the posterior mean rate $\bar{\lambda}(t)$ (*red*) and a step rate function $\hat{\lambda}(t)$ estimated from the posterior distribution (*blue*). *Black* dotted graphs show the cumulative plots of observed threshold exceedances. Dashed vertical lines indicate the locations of estimated change-points.

Fig. 7. Validation of the MCMC performance in the step rate Poisson model estimation. The rate graphs (*left axes*) show the step rate function $\hat{\lambda}(t)$ (*blue*) estimated for the observed threshold exceedances from the MCMC posterior distribution, and the posterior mean rate $\bar{\lambda}(t)$ (*red*) obtained for a data set simulated via a non-homogeneous Poisson process with rate $\hat{\lambda}(t)$. Cumulative plots (*right axes*) correspond to data sets simulated via non-homogeneous Poisson processes with rate $\hat{\lambda}(t)$ (*blue*) or $\bar{\lambda}(t)$ (*red*).

Fig. 8. Bayesian validation of the MCMC results using replicated counting processes. The graphs show cumulative plots for simulated data, non-conditional (*left*) and conditional (*right*) on the total number of observed threshold exeedances. *Black*: cumulative plots for the original exceedance data. *Grey*: 1000 replicated trajectories simulated each with a rate function independently sampled from the MCMC posterior distribution. *Red*: pointwise average of these trajectories. *Blue*: pointwise average of another 1000 simulated trajectories of a Poisson counting process with the step rate function $\hat{\lambda}(t)$ estimated from the posterior distribution.

Fig. 9. Combined log-linear/periodic trend of the form (17) (*red*) fitted to the the log-transformed raw data (*black*).

Table captions

Table 1

Comparison of the MCMC estimation results for NO_2 before and after imputation of missing values. Here n = number of exceedances in $[0,T]$; k = number of change-points; s_1 = location of the (first) change-point; h_0, h_1 = step heights of the rate function on the intervals $(0,s_1)$, (s_1,T), respectively.

Table 2

Posterior modes of the estimated change-points s_j and the corresponding lower and upper quartiles, expressed as day numbers (ranging from 1 to 6206) as well as calendar dates (dd-mm-yy).

Table 3

Posterior modes of the estimated heights h_j of the unknown step rate function $\lambda(t)$, with the corresponding lower and upper quartiles.