A spatio-temporal anisotropic model applied to ozone data

Dani Gamerman, UFRJ, Brazil
Eliane R. Rodrigues – IMATE-UNAM, Mexico
Mario H. Tarumoto, UNESP, Brazil
Guadalupe Tzintzun, SEMARNAT, Mexico

16 June 2014

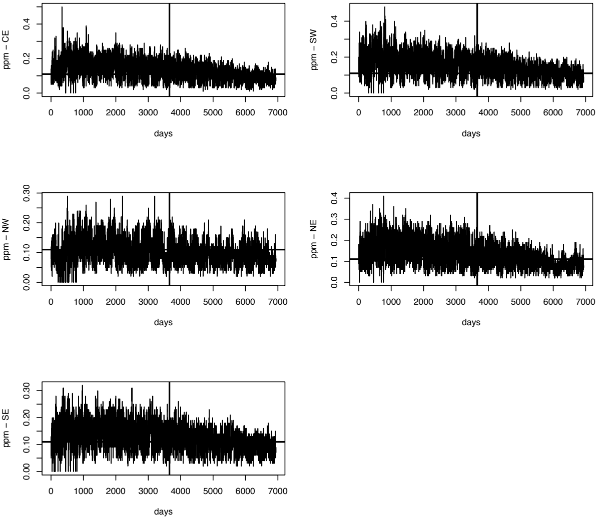
Introduction

- High levels of pollution in large cities throughout the world

- Deterioration in health if sensitive population (newborn, elderly, ill) are exposed to a concentration of 0.11 parts per million (0.11ppm) for a long time Bell *et al.*, 2004, 2005, 2007)

- In Mexico City levels are lower then in the 90's but still high

Daily maximum ozone 1990-2008



Aim

- describe the behaviour of a given air pollutant in terms of the (mean) number of times that a threshold is surpassed in a time interval of interest
- predict the behaviour of the data at places where measurements may not be taken
- non-homogeneous Poisson model with spatial component in the parameters of the rate function
- parameters are estimated using a Bayesian point of view and Markov chain Monte Carlo (MCMC) algorithms: programmed in R
- models are applied to ozone data from the monitoring network of Mexico City

Mexico City case

- Mexico: levels higher than 0.11ppm for hour or more (NOM, 2002)
- Mexico City: levels higher than 0.2ppm $\longrightarrow$ environmental alert
- Environmental alerts: region instead of in the entire city
- threshold considered here: 0.11ppm

Mathematical Model

- $N_O \geq 1$ and $N_U \geq 0$: number of sites where data can and cannot be collected, respectively
- $K_i \geq 0$: number of days in which the threshold 0.11ppm is surpassed in the time interval $[0, T_i)$ ($T_i > 0$) at site i
- $\mathbf{D}_i = \{d_{1,i}, d_{2,i}, \ldots, d_{K_i,i}\}$: set of days in which the exceedances occurred (observed data) at site i
- $\mathbf{D}^{(O)} = \{\mathbf{D}_1^{(O)}, \mathbf{D}_2^{(O)}, \ldots, \mathbf{D}_{N_O}^{(O)}\}$: set of observed data
- $\mathbf{D}^{(U)} = \{\mathbf{D}_1^{(U)}, \mathbf{D}_2^{(U)}, \ldots, \mathbf{D}_{N_U}^{(U)}\}$: set of exceedance days at locations where measurements may not be taken

Mathematical Model

- $M_t^{(i)} \geq 0$: records the number of times that the ozone concentration surpasses the threshold of interest at site i during the time interval $[0, t)$, $t \geq 0$
- $M^{(i)} = \{M_t^{(i)} : t \geq 0\}$: non-homogeneous Poisson process
- where for $t, \alpha, \beta \geq 0$
  - $\lambda^{(i)}(t) = (\alpha_i/\beta_i)(t/\beta_i)^{\alpha_i - 1}$: rate function
  - $m^{(i)}(t) = (t/\beta_i)^{\alpha_i}$: mean function

Mathematical Model

- $\boldsymbol{\theta}^{(O)} = ((\alpha_i^{(O)}, \beta_i^{(O)}),\ i = 1, 2, \ldots, N_O)$: parameters of the Poisson model related to the sites where measurements may be directly taken
- $\boldsymbol{\theta}^{(U)} = ((\alpha_j^{(U)}, \beta_j^{(U)}),\ j = 1, 2, \ldots, N_U)$: corresponding parameters related to the sites where measurements may not be taken directly

Mathematical Model

- **Likelihood function**

$$
\begin{aligned}
L(\mathbf{D}^{(O)} \mid \boldsymbol{\theta}^{(O)}, \boldsymbol{\theta}^{(U)}) &= L(\mathbf{D}^{(O)} \mid \boldsymbol{\theta}^{(O)}) \\
&= \prod_{i=1}^{N_O} L(\mathbf{D}_i^{(O)} \mid \boldsymbol{\theta}^{(O)}) \\
&= \prod_{i=1}^{N_O} L(\mathbf{D}_i^{(O)} \mid \alpha_i^{(O)}, \beta_i^{(O)}),
\end{aligned}
$$

Mathematical Model

- **Likelihood function**

$$L(\mathbf{D}^{(O)} \mid \boldsymbol{\theta}^{(O)}, \boldsymbol{\theta}^{(U)}) = \prod_{i=1}^{N_O} \left[\left(\frac{\alpha_i^{(O)}}{\beta_i^{(O)}} \right)^{K_i} e^{-\left(T_i/\beta_i^{(O)}\right)^{\alpha_i^{(O)}}} \left(\prod_{k=1}^{K_i} \left[\frac{d_{k,i}}{\beta_i^{(O)}} \right]^{\alpha_i^{(O)}-1} \right) \right].$$

Mathematical Model

- Independent Gaussian processes are assumed to rule the spatial variation of the parameters
  - $\boldsymbol{\alpha} = (\alpha_i^{(O)}, \alpha_j^{(U)};\ i = 1, 2, \ldots, N_O,\ j = 1, 2, \ldots, N_U)$
  - $\boldsymbol{\beta} = (\beta_i^{(O)}, \beta_j^{(U)};\ i = 1, 2, \ldots, N_O,\ j = 1, 2, \ldots, N_U)$

  in their logarithmic form

Mathematical Model

- assume
  - $\log \boldsymbol{\alpha} = (\log \alpha_i^{(O)}, \log \alpha_j^{(U)};\ i = 1, 2, \ldots, N_O,\ j = 1, 2, \ldots, N_U)$
    $\log \boldsymbol{\beta} = (\log \beta_i^{(O)}, \log \beta_j^{(U)};\ i = 1, 2, \ldots, N_O,\ j = 1, 2, \ldots, N_U)$

    (multivariate Normal distributions)
    - with mean vectors

      $$\boldsymbol{\mu}^{\alpha} = (\mu^{\alpha_i^{(O)}}, \mu^{\alpha_j^{(U)}};\ i = 1, 2, \ldots, N_O,\ j = 1, 2, \ldots, N_U)$$

      $$\boldsymbol{\mu}^{\beta} = (\mu^{\beta_i^{(O)}}, \mu^{\beta_j^{(U)}};\ i = 1, 2, \ldots, N_O,\ j = 1, 2, \ldots, N_U),$$
    - variance-covariance matrices Σ^{α} and Σ^{β}
  - prior independence between $\log \boldsymbol{\alpha}$ and $\log \boldsymbol{\beta}$

Mathematical Model

- variance-covariance matrices:
  - $\Sigma^\alpha = \left(\nu_{ij}^\alpha\right)_{i,j=1,2,\ldots,N_O+N_U}$
  - $\Sigma^\beta = \left(\nu_{ij}^\beta\right)_{i,j=1,2,\ldots,N_O+N_U}$
- where we take

$$\nu_{ij}^\alpha = \sigma_{\alpha_i}\,\sigma_{\alpha_j}\,\exp\left(-\phi_\alpha\,\|d(s_i)-d(s_j)\|\right)$$

$$\nu_{ij}^\beta = \sigma_{\beta_i}\,\sigma_{\beta_j}\,\exp\left(-\phi_\beta\,\|d(s_i)-d(s_j)\|\right)$$

Mathematical Model

- with
  - ϕ_α and ϕ_β: parameters that need to be estimated
  - σ_{α_k} and σ_{β_k}: the standard deviations of $\log \alpha_k$ and $\log \beta_k$
  - $\boldsymbol{\sigma}_\alpha = (\sigma_{\alpha_i};\ i = 1, 2, \ldots, N_O + N_U)$

    $\boldsymbol{\sigma}_\beta = (\sigma_{\beta_i};\ i = 1, 2, \ldots, N_O + N_U)$

Mathematical Model

- where
  - s_i indicates the UTM coordinates of site i
  - we take

$$R = \begin{pmatrix} \cos\psi_a & -\sin\psi_a \\ \sin\psi_a & \cos\psi_a \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & 0 \\ 0 & \psi_r^{-1} \end{pmatrix},$$

    - ψ_a: the anisotropy angle
    - $\psi_r > 1$: anisotropy ratio

    (Cressie 1991; Diggle and Ribeiro 2007; Schmidt and Rodríguez 2010)

  - $A = R \times X$
  - $d(s_k) = A \times s_k$

Mathematical Model

- posterior distribution of the vector of parameters

$$P(\bar{\boldsymbol{\theta}} \,|\, \mathbf{D}^{(O)}) \propto L(\,\mathbf{D}^{(O)} \,|\, \boldsymbol{\theta}^{(O)})\, P(\boldsymbol{\alpha}^{(O)} \,|\, \boldsymbol{\mu}^{\alpha^{(O)}}, \phi_\alpha, \psi_a, \psi_r, \boldsymbol{\sigma}_\alpha)$$

$$P(\boldsymbol{\alpha}^{(U)} \,|\, \boldsymbol{\alpha}^{(O)}, \boldsymbol{\mu}^{\alpha}, \phi_\alpha, \psi_a, \psi_r, \boldsymbol{\sigma}_\alpha)$$

$$P(\boldsymbol{\beta}^{(O)} \,|\, \boldsymbol{\mu}^{\beta^{(O)}}, \phi_\beta, \psi_a, \psi_r, \boldsymbol{\sigma}_\beta)$$

$$P(\boldsymbol{\beta}^{(U)} \,|\, \boldsymbol{\beta}^{(O)}, \boldsymbol{\mu}^{\beta}, \phi_\beta, \psi_a, \psi_r, \boldsymbol{\sigma}_\beta)$$

$$P(\boldsymbol{\mu}^{\alpha})\, P(\boldsymbol{\mu}^{\beta})\, P(\phi_\alpha)\, P(\phi_\beta)$$

$$P(\psi_a)\, P(\psi_r)\, P(\boldsymbol{\sigma}_\alpha)\, P(\boldsymbol{\sigma}_\beta)$$

$$P(\mathbf{D}^{(U)} \,|\boldsymbol{\alpha}^{(U)}, \boldsymbol{\beta}^{(U)})$$

Mathematical Model

- $\log(\boldsymbol{\alpha}^{(U)})$: multivariate Normal distribution
  - mean

$$\overline{\mu}^{\boldsymbol{\alpha}^{(U)}} = \boldsymbol{\mu}^{\boldsymbol{\alpha}^{(U)}} + \Sigma^{\alpha^{(U)},\alpha^{(O)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha)$$

$$\left[\Sigma^{\alpha^{(O)},\alpha^{(O)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha)\right]^{-1} \left(\log \boldsymbol{\alpha}^{(O)} - \boldsymbol{\mu}^{\alpha^{(O)}}\right)$$

  - variance-covariance matrix

$$\overline{\Sigma}^{\alpha^{(U)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha^{(U)}) = \Sigma^{\alpha^{(U)},\alpha^{(U)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha)$$

$$-\Sigma^{\alpha^{(U)},\alpha^{(O)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha)$$

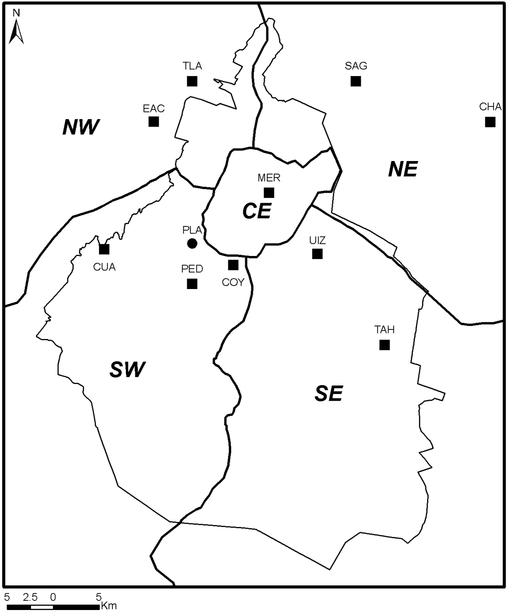
$$\left[\Sigma^{\alpha^{(O)},\alpha^{(O)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha)\right]^{-1} \Sigma^{\alpha^{(O)},\alpha^{(U)}}(\psi_a, \psi_r, \phi_\alpha, \boldsymbol{\sigma}_\alpha)$$

Mathematical Model

- prior distributions
  - $\boldsymbol{\mu}^{\alpha}$ and $\boldsymbol{\mu}^{\beta}$: product of appropriate Normal distributions
  - ϕ: inverse Gamma function $\mathrm{IG}(a, b)$, $a > 2$
  - $P(\psi_a)$: uniform distribution on $(0, \pi)$
  - $P(\psi_r)$: Pareto distribution with hyperparameters (c, d)
- additionally, we consider (see Schmidt and Rodríguez, 2010)

$$-\log(0.05) = \frac{\phi \max \|d(s) - d(s')\|}{2},$$

Monitoring stations considered



Application: Mexico City ozone data

- 01 January 2005 to 31 December 2009
- data

	SAG	CHA	EAC	TLA	MER	UIZ	TAH	PED	CUA	COY
mean	0.072	0.075	0.082	0.077	0.084	0.09	0.083	0.099	0.089	0.096
sd	0.026	0.024	0.031	0.029	0.032	0.031	0.029	0.037	0.033	0.034
K_i	129	116	319	223	365	492	346	708	457	621

- $T_i = T = 1826$
- $N_O = 10, \ N_U = 1$

Application: Mexico City ozone data

- Gibbs sampling algorithm, Metropolis-Hasting algorithm and Kriging: programmed in R
- Parameters estimation
  - sample of size 5000
  - five chains
  - burn-in period: 30000 steps

Application: Mexico City ozone data

Station		Mean	SD	95% Credible Interval
EAC	α	0.81	0.04	(0.73, 0.88)
	β	1.45	0.48	(0.66, 2.58)
TLA	α	0.84	0.05	(0.76, 0.94)
	β	2.98	1.11	(1.4, 5.7)
SAG	α	0.77	0.05	(0.68, 0.89)
	β	3.51	1.57	(1.35, 7.3)
CHA	α	0.69	0.05	(0.6, 0.78)
	β	1.73	0.83	(0.59, 3.72)
MER	α	0.76	0.04	(0.68, 0.85)
	β	0.82	0.35	(0.29, 1.66)
UIZ	α	0.84	0.03	(0.78, 0.9)
	β	1.07	0.29	(0.58, 1.77)
TAH	α	0.8	0.04	(0.73, 0.88)
	β	1.18	0.44	(0.55, 2.27)
PED	α	0.82	0.03	(0.77, 0.9)
	β	0.63	0.22	(0.36, 1.24)
CUA	α	0.8	0.04	(0.73, 0.88)
	β	0.86	0.32	(0.4, 1.68)
COY	α	0.8	0.03	(0.74, 0.86)
	β	0.57	0.19	(0.3, 0.99)
PLA	α	0.87	0.054	(0.2, 2.25)
	β	1.43	0.57	(0.52, 2.73)

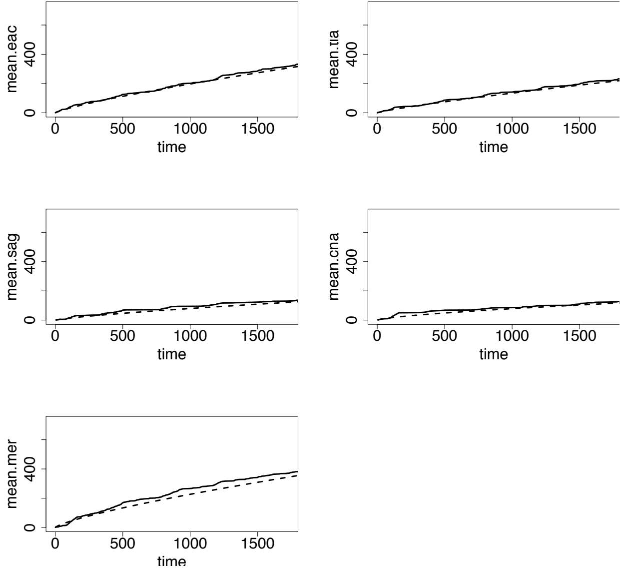
Application: Mexico City ozone data

Station		Mean	SD	95% Credible Interval
EAC	μ^{α}	- 0.61	0.47	(-1.51, -0.3)
	μ^{β}	0.33	0.44	(-0.56, 1.18)
TLA	μ^{α}	- 0.48	0.48	(-1.4, -0.45)
	μ^{β}	1.04	0.4	(0.27, 1.84)
SAG	μ^{α}	- 0.45	0.48	(-1.4, -0.51)
	μ^{β}	1.22	0.5	(0.25, 2.21)
CHA	μ^{α}	- 0.47	0.49	(-1.43, -0.48)
	μ_{β}	0.57	0.53	(-0.48, 1.59)
MER	μ^{α}	- 0.65	0.47	(-1.57, -0.27)
	μ^{β}	-0.22	0.5	(-1.22, 0.76)
UIZ	μ^{α}	- 0.49	0.47	(-1.4, -0.46)
	μ^{β}	0.11	0.34	(-0.59, 0.75)
TAH	μ^{α}	- 0.44	0.5	(-1.4, -0.55)
	μ^{β}	0.18	0.4	(-0.6, 0.99)
PED	μ^{α}	- 0.97	0.49	(-1.9, -0.03)
	μ^{β}	-0.77	0.45	(-1.64, 0.15)
CUA	μ^{α}	- 0.67	0.49	(-1.63, -0.32)
	μ^{β}	-0.2	0.46	(-1.11, 0.72)
COY	μ^{α}	- 0.59	0.08	(-1.8, -0.06)
	μ^{β}	-0.75	0.44	(-1.57, 0.15)
PLA	μ^{α}	-1.96	0.43	(-2.85, -1.22)
	μ^{β}	-0.81	0.19	(-1.21, -0.46)

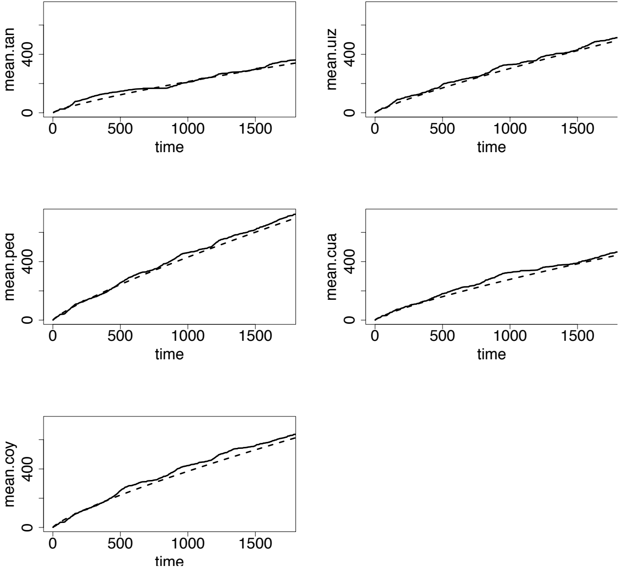
Application: Mexico City ozone data

	Mean	SD	95% Credible Interval
ψ_a	$1.58 \approx \pi/2$	0.9	(0.09, 3.07)
ψ_r^{-1}	0.84	0.13	(0.51, 0.995)
ϕ_α	1.07E-04	9.23E-05	(3.49E-05, 2.89E-04)
ϕ_β	1.92E-04	1.13E-04	(8.11E-05, 4.115E-04)

Observed and estimated accumulated means



Observed and estimated accumulatedmeans

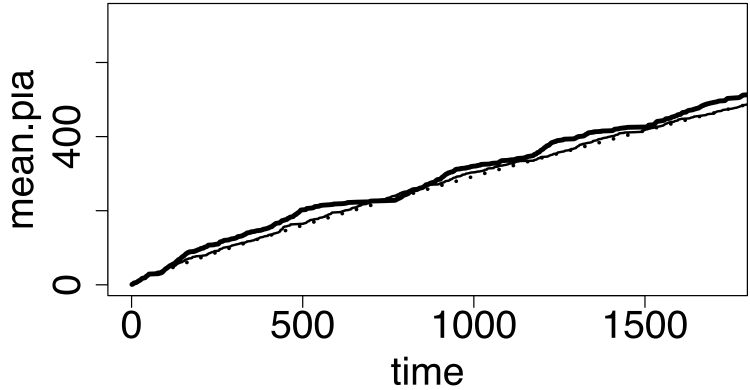


Prediction: station Plateros (PLA)

- substitute the respective estimated α and β in $\lambda_{PLA}(t)$
- sample of size 5000 to estimate the value of K_{PLA}
- estimated value: 504 (22.79) – actual value: 518
- (Cox and Lewis, 1966): given that K events occurred in $[0, T]$
  - $t_1, t_2, \ldots, t_K$: order statistics of a sample with distribution

$$F(t) = m(t)/m(T), \quad t \in [0, T]$$

Observed and estimated means: PLA



Comments

- associate: $1 = \text{EAC}$, $2 = \text{TLA}$, $3 = \text{SAG}$, $4 = \text{CHA}$, $5 = \text{MER}$, $6 = \text{UIZ}$, $7 = \text{TAH}$, $8 = \text{PED}$, $9 = \text{CUA}$, $10 = \text{COY}$, $11 = \text{PLA}$

- $\rho_{\boldsymbol{\alpha}}$ and $\rho_{\boldsymbol{\beta}}$ the correlation matrices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, obtained using the generated values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ provided by the MCMC algorithm

Comments

$$\rho_{\boldsymbol{\alpha}} = \begin{pmatrix} 1.0 & 0.32 & 0.02 & -0.014 & 0.13 & 0.045 & 0.009 & 0.14 & 0.19 & 0.104 & -0.013 \\ & 1.0 & 0.092 & 0.005 & 0.13 & 0.057 & 0.017 & 0.074 & 0.077 & 0.058 & -0.007 \\ & & 1.0 & 0.13 & 0.13 & 0.091 & 0.036 & 0.029 & 0.015 & 0.068 & -0.015 \\ & & & 1.0 & 0.034 & 0.066 & 0.041 & 0.012 & -0.002 & 0.007 & 0.0008 \\ & & & & 1.0 & 0.25 & 0.042 & 0.13 & 0.088 & 0.27 & -0.026 \\ & & & & & 1.0 & 0.15 & 0.112 & 0.037 & 0.207 & 0.012 \\ & & & & & & 1.0 & 0.078 & 0.006 & 0.089 & -0.007 \\ & & & & & & & 1.0 & 0.24 & 0.397 & -0.04 \\ & & & & & & & & 1.0 & 0.15 & -0.022 \\ & & & & & & & & & 1.0 & -0.023 \\ & & & & & & & & & & 1.0 \end{pmatrix}$$

Comments

$$
\rho_{\boldsymbol{\beta}} = \begin{pmatrix}
1.0 & 0.15 & -0.036 & -0.035 & 0.062 & -0.081 & 0.049 & 0.15 & 0.103 & 0.064 & -0.0076 \\
 & 1.0 & 0.083 & -0.046 & -0.043 & 0.086 & 0.0061 & 0.157 & 0.023 & -0.055 & -0.071 \\
 & & 1.0 & -0.0003 & 0.061 & 0.136 & 0.047 & -0.0013 & 0.049 & 0.013 & 0.024 \\
 & & & 1.0 & -0.13 & 0.032 & -0.028 & -0.04 & -0.049 & 0.011 & 0.0595 \\
 & & & & 1.0 & -0.071 & 0.096 & 0.11 & 0.038 & 0.155 & -0.076 \\
 & & & & & 1.0 & 0.007 & -0.011 & 0.138 & 0.099 & -0.0082 \\
 & & & & & & 1.0 & -0.038 & -0.026 & 0.103 & -0.078 \\
 & & & & & & & 1.0 & 0.224 & 0.209 & -0.149 \\
 & & & & & & & & 1.0 & 0.133 & -0.067 \\
 & & & & & & & & & 1.0 & -0.036 \\
 & & & & & & & & & & 1.0
\end{pmatrix}
$$

Comments

- impose a spatial relationship on the parameters of the non-homogeneous Poisson model
- infer the behaviour of the process in sites where measurements could not be taken directly
- mechanism for estimating the number of times that an environmental threshold would be surpassed in a site where measurements cannot be taken directly
- predicts well the behaviour of ozone in sites where measurements are not available
- describe well that behaviour at sites where measurements are available