42/2007

Raport Badawczy

Research Report

RB/51/2007

Bayesian spatial model for analysis of emission inventory data

J. Horabik, Z. Nahorski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Pracowni zgłaszający pracę:
Prof. dr hab. inż. Zbigniew Nahorski

Warszawa 2007

Bayesian spatial model for analysis of emission inventory data

Joanna Horabik, Zbigniew Nahorski
Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
{Joanna.Horabik, Zbigniew.Nahorski}@ibspan.waw.pl

Abstract

This paper aims to contribute to comparing gridded inventories of atmospheric emissions with different resolution. We propose a hierarchical Bayesian model with high resolution emission assessments treated as dependent variable, and spatially explicit activity data treated as covariates. The results of our example suggest excluding from further analysis two initially considered covariates, and indicate existence of another spatially correlated factor. The point of the contribution is that including spacial scale helps to improve emission inventories.

Keywords: conditionally autoregressive model; emission data; hierarchical Bayes; spatial emissions

1 Introduction

The contribution is focused on a spatial aspect of inventories for atmospheric pollutants. This perspective is motivated with situations when two kinds of inventories for the same area and for the same pollutant are available: based on a bottom-up and top-down procedures. Although all inventories can have features of both bottom-up, and of top-down type; the main difference is the following. The bottom-up procedure of inventory provides detailed (high resolution) information on source types, locations and emissions. On the other hand, a top-down inventory procedure generally provides low spatial resolution. When activity data (e.g. land use, vehicle or other) are available, a top-down inventory is spatially distributed, using these statistics and appropriate emission factors. The idea is then to compare this map with a reference inventory based on a bottom-up procedure, and try to conclude on the relevance of activity data used for disaggregation.

This kind of analysis has been already performed in some studies. Specifically, we were motivated with the paper of Winiwarter *et al.*, 2003. In this

paper two sets of data on NOx (Nitrogen oxides) emissions over the same spatial grid for the Greater Athens, Greece were compared. While the authors formulate their conclusions mainly based on a visual comparison of maps, we would like to provide a quantitative approach.

When performing statistical inference of spatial inventory data we account for the fact that values at proximate locations tend to be more alike, which motivate use of spatial statistics. Secondly, since for each grid cell we have information on aggregated emission values, these are areal data. A popular tool for incorporating this kind of spatial information is conditionally autoregressive (CAR) model developed by Besag, 1974.

The aim of the present paper is to explore the usefulness of CAR model to analyse data from spatially explicit emission inventory. The outline of the study is the following. Section 2 presents our illustrative data set. Bayesian model for emission data is described in Section 3. Results of the analysis are contained in Section 4, and Section 5 concludes.

2 Initial exploration of the data

Our illustration is provided by data on CO_2 emissions (in tones) reported in municipalities of southern Norway (see Figure 1). The data come from StatBank in Statistics Norway (available at `http://www.ssb.no`). The map comprises 259 municipalities. We use a log transformation on the emission data to ensure a constant variance. For each municipality three kinds of covariate information are available (Figure 2). Covariates are also log transformed for further analysis. Let us then denote: y_i - (log) CO_2 emissions (in tones), $x_{i,1}$ - (log) total area (in km^2), $x_{i,2}$ - (log) population, $x_{i,3}$ - (log) area covered by roads (in km^2), $i = 1, \ldots, 259$ are numbered spatial cells. An initial linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i \quad i = 1, \ldots, 259$$

was considered. It showed that each covariate is significant (for each parameter coefficient p-value was lower than 2E(-10)), and coefficient of determination was $R^2 = 0.87$.

Residuals of the linear regression were checked for spatial correlation using Moran's I statistic:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})}{\sum_i (\epsilon_i - \bar{\epsilon})^2}$$

where ϵ_i - a residual of linear regression in area i, $\bar{\epsilon}$ - mean of residuals, w_{ij} - adjacency weights ($w_{ij} = 1$ if j is a neighbour of i, and 0 otherwise, also $w_{ii} = 0$). Under a null hypothesis where ϵ_i are independent and identically distributed, I is asymptotically normally distributed, with the mean and

variance defined, see e.g. Banerjee et.al. 2004, Kopczewska, 2006. In our case the test statistic (standardized Moran's I) is equal $z = 4.65$ ($z_{cr} = 2.33$ at the significance level $\alpha = 0.01$), which suggests evidence against a null hypothesis of no spatial correlation of errors. Moran's I is, however, recommended just as an exploratory information on spatial association, rather than a measure of spatial significance.

3 Modelling spatial correlation

In this section, we develop a Bayesian model to characterize the spatial distribution of CO_2 emissions in municipalities. We first model the data. Let Y_i denote stochastic variable associated with the response of interest (inventory data) defined with high resolution at each spatial location i for $i = 1, \ldots, n$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. It is assumed that the random variables Y_i follow normal distribution with mean μ_i and common variance σ^2. Let the mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ be such that $\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{\theta}$. Then

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{\theta}, \Sigma) \tag{1}$$

where Σ is a diagonal $n \times n$ matrix with elements σ^2. $\boldsymbol{X}$ is the $n \times (k+1)$ matrix containing explanatory covariates and a vector of 1s in the first column for the intercept. $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of coefficients. $\boldsymbol{\theta}$ is a vector of correlated random variables. Thus, conditionally on the parameters $\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2$, stochastic variables Y_i are independent.

Next we describe the random component $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. Correlation of variables θ_i allows us to model spatial dependence between the variables Y_i. Since the data from inventories are aggregated for each grid cell and available on a discrete space, we make use of a conditionally autoregressive (CAR) model, which is based on a (space) Markov property. The CAR model is given through specification of full conditional distribution functions:

$$\theta_i | \theta_j, i \neq j \sim \mathcal{N}\left(\frac{1}{w_{i+}} \sum_{j \in N_i} \theta_j, \frac{\tau^2}{w_{i+}}\right) \tag{2}$$

with N_i being a set of neighbours of area i, w_{i+} being a number of neighbours of area i and τ^2 is a variance parameter. Conditional expected value of θ_i is the average with fixed values of those variables θ_j which are neighbours of site i. Conditional variance is inversely proportional to the number of neighbours w_{i+}.

Given (2), the joint probability distribution of $\boldsymbol{\theta}$ is the following (Banerjee *et al.*, 2004; Cressie, 1993)

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{0}, \tau^2 (\boldsymbol{D} - \boldsymbol{W})^{-1}\right) \tag{3}$$

where $\boldsymbol{D}$ is $n \times n$ diagonal matrix with w_{i+} elements on the diagonal and $\boldsymbol{W}$ is $n \times n$ matrix with adjacency weights w_{ij}. Equivalently, (3) may be rewritten as

$$p(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\tau^2}\sum_{i\neq j} w_{ij}(\theta_i - \theta_j)^2\right] \tag{4}$$

Estimation of unknown parameters $\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2$ is done with the Bayesian approach. The joint posterior distribution of these parameters is proportional to the product of the likelihood function associated with Equation (1); the distribution for spatial random component $\boldsymbol{\theta}$ in Equation (3); and specified prior distributions for the remaining parameters. Improper uniform distributions are employed for each of the β parameters. The inverse variance parameters $1/\sigma^2$ and $1/\tau^2$ are assumed $Gamma(0.01, 0.01)$ distribution, where $Gamma(a, b)$ distribution is parametrized with mean equal to a/b.

Combination of all model assumptions allows to derive full conditionals for all of the parameters in a closed-form. The full conditional distributions for our model can be found in the Appendix. Gibbs sampling is used to update all parameters. Calculations were accomplished both using the `WinBUGS` package (Lunn *et al.*, 2000) and writing our own function in `R` (`www.r-project.org`).

4 Results

Spatial CAR models have been applied to the Norway emission data. Using DIC statistics (Spiegelhalter *et al.*, 2002) we compare various combinations of covariate data between the spatial and linear regression models (see Table 1). Fit measures $\bar{D}$ and effective number of parameters p_D (a measure of complexity) are also displayed in Table 1.

[Table 1]

We note that the best result (DIC=108) is obtained for two spatial models: the one with two covariate information CAR(x_1, x_3); and the one only with information on the area covered by roads CAR(x_3). This means that covariate data on area x_1 does not add any meaningful information, and so CAR(x_1, x_3) model should be chosen for further analysis. This model outperforms among others the CAR model with all the covariates. In case of a simple linear regression the situation was the opposite - we got better results including all the three covariates. We conclude that there exists a missing, spatially correlated variable contributing to overall emissions much better than the initial variables x_1, x_2. Table 1 shows also results for other combinations of covariate data. For instance for CAR (x_1, x_2) model, we have less parameters compared with the case of three covariates and thus

lower complexity ($p_D = 65$), on the other hand the model fit is much worse ($\bar{D} = 782$).

Parameter estimates for three chosen models are shown in Table 2. Comparing results for model CAR (x_1, x_2, x_3) with results for linear regression, we see that although the 95% confidence (bayesian) intervals for β_1 and β_2 does not include zero, their values moved towards zero considerably. On the other hand, estimate of β_3 remained almost the same. It generally confirms our previous conclusion.

[Table 2]

Maps of posterior mean for two CAR models are shown on Figure 3. It can be noticed that model CAR (x_3) maps the original data (Figure 1) slightly better than the model CAR (x_1, x_2, x_3).

[Figure 3]

5 Concluding remarks

We have shown the application of spatial conditionally autoregressive model to examine influence of activity data towards independent, bottom-up inventory. Our results suggest excluding from further analysis two initially considered covariates, and indicate existence of another, spatially correlated factor. Generally, such situation - that we get better results just for a subset of covariates plus a spatial component - is not unusual. The point of this contribution was to make use of this approach for comparison of inventory data.

It should be noted that our exercise is to some extend illustrative and in a more realistic application more informative results could be obtained. For example, a potentially problematic part of inventory are emission point sources (plants), which are correctly reported in a bottom-up approach but are missing in datasets with activity information (Winiwarter, 2007). The proposed method seems to be capable to identify such cases.

Trying to extend this model little further, one may think of the case where CAR prior is used for parameter coefficients $\boldsymbol{\beta}$. This approach might be useful when considering space-varying emission factors. These models are mentioned for instance in Gamerman and Lopes, 2006, while Gamerman *et al.*, 2003 provide computational details for sampling schemes in a relevant MCMC algorithm.

Appendix: MCMC Algorithm

Below we present full conditional distributions for the model. The conditional distribution for some parameter vector $\boldsymbol{Z}$ given all other random

quantities is denoted $[\boldsymbol{Z}|\cdot]$.

$$
\begin{aligned}
&[\boldsymbol{\beta}|\cdot] \sim N\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{\theta}), \left(\boldsymbol{X}'\tfrac{1}{\sigma^2}\boldsymbol{I}\boldsymbol{X}\right)^{-1}\right) \\
&[\boldsymbol{\theta}|\cdot] \sim N\left(\left(\boldsymbol{I}+\tfrac{\sigma^2}{\tau^2}\boldsymbol{K}\right)^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}), \left(\tfrac{1}{\sigma^2}\boldsymbol{I}+\tfrac{1}{\tau^2}\boldsymbol{K}\right)^{-1}\right) \\
&\text{where } \boldsymbol{K}=\boldsymbol{D}-\boldsymbol{W} \\
&[\sigma^2|\cdot] \sim \mathrm{IG}\left(\alpha+\tfrac{n}{2}, \gamma+\tfrac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\theta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\theta})\right) \\
&[\tau^2|\cdot] \sim \mathrm{IG}\left(\alpha+\tfrac{n}{2}, \gamma+\tfrac{1}{2}\textstyle\sum_{j\neq i} w_{ij}(\theta_j-\theta_i)^2\right)
\end{aligned}
$$

References

[1] Banerjee S, Carlin BP, Gelfand AE. 2004. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall: London.

[2] Besag J. 1974. Spatial interactions and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36**: 192-236.

[3] Cressie N. 1993. *Statistics for Spatial Data.* Revised edition, Wiley.

[4] Gamerman D, Moreira ARB, Rue H. 2003. Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis* **42**: 513-533.

[5] Gamerman D, Lopes HF. 2006. *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference.* 2nd edition. Chapman and Hall: London.

[6] Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325-337.

[7] Kopczewska K. 2006. *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*, CeDeWu, Warszawa (in Polish).

[8] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* **64**: 583-639.

[9] Winiwarter W, Dore Ch, Hayman G, Vlachogiannis D, Gounaris N, Bartzis J, Ekstrand J, Tamponi M, Maffeis G. 2003. Methods for comparing gridded inventories of atmospheric emissions - application for Milan province, Italy and the Greater Athens Area, Greece. *The Science of the Total Environment* **303**: 231-243.

[10] Winiwarter W. 2007. Personal communication.

Table 1: Model comparison using DIC statistics

Model	$\bar{D}$	p_D	DIC
CAR (x_1, x_2, x_3)	224	106	330
CAR (x_1, x_2)	782	65	847
CAR (x_1, x_3)	-177	285	108
CAR (x_2, x_3)	-147	276	129
CAR (x_3)	-173	281	108
linear regression (x_1, x_2, x_3)	415	5	420
linear regression (x_1, x_2)	898	4	902
linear regression (x_1, x_3)	560	4	564
linear regression (x_2, x_3)	554	4	558
linear regression (x_3)	588	3	591

Table 2: Parameter estimates (95% credible intervals are given in brackets)

Param.	Linear regression	model CAR (x_1, x_2, x_3)	model CAR (x_3)
β_0	4.027	4.169 (3.91, 4.46)	4.794 (4.72, 4.87)
β_1	-0.308	-0.198 (-0.26, -0.13)	-
β_2	0.266	0.182 (0.13, 0.23)	-
β_3	1.497	1.462 (1.38, 1.53)	1.322 (1.27, 1.38)

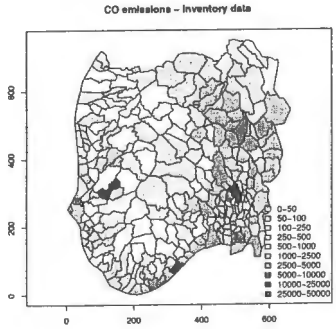

Figure 1: CO_2 emission data in tones

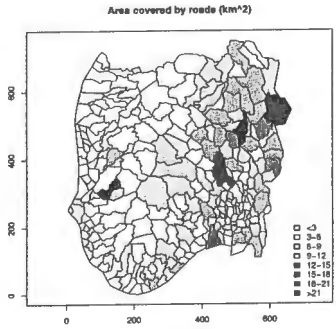

Figure 2: Area covered by roads in km^2

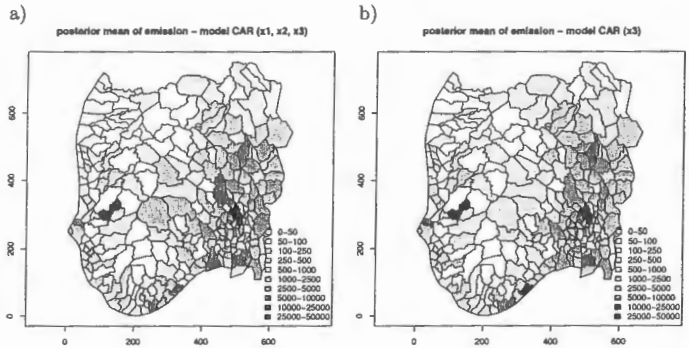

Figure 3: Posterior mean of emission in tones for model CAR (x_1, x_2, x_3) (a) and model CAR (x_3) (b)