Spatial disaggregation of pollutant concentration data [1]

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

Abstract: The purpose of this study is to develop a method for allocating pollutant concentrations to finer spatial scales conditional on covariate information observable in a fine grid. Spatial dependence is modeled with the conditional autoregressive structure. The maximum likelihood approach to inference is employed, and the optimal predictors are developed to assess missing concentrations in a fine grid. The method is developed for a practical application of an output from the dispersion model CALPUFF run for Warsaw agglomeration.

Keywords: Air pollutant concentration, conditional autoregressive structure, spatial disaggregation

1 Introduction

Atmospheric dispersion models constitute a basic tool for air quality control. Further usage of output from dispersion models include, among others, health impact assessments. For improved risk assessments, it is often required to develop air quality data in a resolution higher than the one readily available from dispersion models.

Making inference on variables at points or grid cells different from those of the data is referred to as the change of support problem. Several approaches have been proposed to address the problem. The geostatistical solution for realignement from point to areal data is provided by block kriging (Gotway & Young 2002, Gelfand 2010). In the case that data are observed at areal units and inference is sought at a new level of spatial aggregation, areal weighting offers a straightforward approach. Some improved approaches with better covariate modeling were also proposed e.g. in Mugglin & Carlin 1998, and Mugglin *et al.* 2000.

In the following we present an approach for areal to areal data realignement, which accounts for a tendency toward spatial clustering, and is focused on application to air quality. The idea stems from the method proposed in Chow & Lin (1971) for time series, see also Polasek *et al.* (2010). Regarding an assumption on residual covariance structure, we apply the conditional autoregressive (CAR) specification. While the CAR structure is extensively used in epidemiology, it can be also applied for modeling air pollution over space (Kaiser *et al.* 2002, McMillan *et al.* 2010).

[1]The research of Joanna Horabik was supported by Ministry of Science and Higher Education under the Iuventus Plus project No. 0128/H03/2010/70.

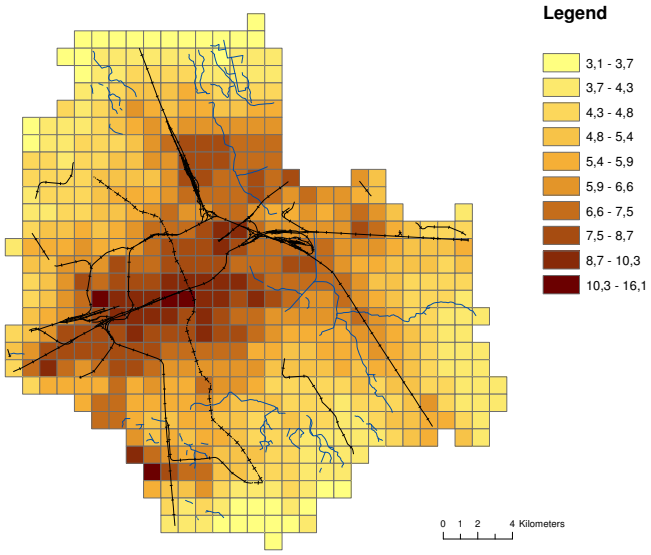

Figure 1: SO_2 concentration ($\mu g/m^3$) in a 1 km grid

2 Motivating data set

The study concerns air pollution concentrations (PM_{10}, NO_x and SO_2 among others) obtained from the dispersion model CALPUFF. A 1 km grid for Warsaw area comprises 563 grid cells. Health risk studies, conducted in parallel, motivated our search for the air pollution map in a 0.5 km resolution. The dispersion model output represents an average pollutant concentration over each 1 km grid cell. This value, multiplied by a cell area, reflects a pollutant level in a grid cell, and it constitutes the value to be disaggregated.

In addition, available covariate information characterizes transportation, area and point emission sources of the city in a 0.5 km grid.

3 The disaggregation framework

We begin with the model specification in a fine 0.5 km grid. Let Y_i denote a random variable associated with a missing value of pollutant, say SO_2, level y_i defined at each cell i, $i = 1, ..., n$ of a fine grid. Assume that random variables Y_i follow a Gaussian distribution with the mean μ_i and variance σ_Y^2, and given these values Y_i are independent. The values $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^n$ represent the true process underlying SO_2 level, and the (missing) observations are related to this process through a measurement error of variance σ_Y^2. The model for the underlying SO_2 process is formulated as a sum of regression component with available covariates, and a spatially varying random effect. The applied CAR structure follows an assumption of similar random

effects in adjacent cells, and it is given through the specification of full conditional distribution functions

$$\mu_i|\mu_{j,j\neq i} \sim \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \rho\sum_{j\neq i}\frac{w_{ij}}{w_{i+}}\left(\mu_j - \boldsymbol{x}_j^T\boldsymbol{\beta}\right), \frac{\tau^2}{w_{i+}}\right),\ i,j = 1,...,n \tag{1}$$

where w_{ij} are the adjacency weights; w_{i+} is the number of neighbours of area i; $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is a regression component with explanatory covariates for area i and a respective vector of regression coefficients, and τ^2 is a variance parameter. The joint distribution of the process $\boldsymbol{\mu}$ is (Cressie, 1993)

$$\boldsymbol{\mu} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}\right), \tag{2}$$

where $\boldsymbol{X}$ is a design matrix with vectors $\boldsymbol{x}_i$; $\boldsymbol{D}$ is an $n\times n$ diagonal matrix with w_{i+} on the diagonal; and $\boldsymbol{W}$ is an $n\times n$ matrix with adjacency weights w_{ij}. Equivalently, we can write (2) as $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{N})$, with $\boldsymbol{N} = \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}$.

The model for the CALPUFF output data observed in a 1 km grid is obtained by multiplication of $\boldsymbol{\mu}$ with an $N \times n$ *aggregation matrix* C, where N is a number of observations in a 1 km grid

$$\boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\epsilon},\quad \boldsymbol{C}\boldsymbol{\epsilon} \sim \mathcal{N}_N\left(\boldsymbol{0}, \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right). \tag{3}$$

The matrix C consists of 0's and 1's, indicating which cells have to be aligned together. The random variable $\boldsymbol{\lambda} = \boldsymbol{C}\boldsymbol{\mu}$ is treated as the mean process for variables $\boldsymbol{Z} = \{Z_i\}_{i=1}^N$ associated with observations $\boldsymbol{z} = \{z_i\}_{i=1}^N$ of the aggregated model

$$\boldsymbol{Z}|\boldsymbol{\lambda} \sim \mathcal{N}_N\left(\boldsymbol{\lambda}, \sigma_Z^2\boldsymbol{I}_N\right). \tag{4}$$

Also at this level, the underlying process $\boldsymbol{\lambda}$ is related to $\boldsymbol{Z}$ through a measurement error with variance σ_Z^2.

The parameters $\boldsymbol{\beta}, \sigma_Z^2, \tau^2$ and ρ are estimated with the maximum likelihood method based on the joint unconditional distribution

$$\boldsymbol{Z} \sim \mathcal{N}_N\left(\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right),$$

where $\boldsymbol{M} = \sigma_Z^2\boldsymbol{I}_N$. The analytical derivation is limited to the regression coefficients $\boldsymbol{\beta}$, and further maximisation of the profile log likelihood is performed numerically. The standard errors of estimators are calculated with the expected Fisher information matrix.

Regarding the missing values in a fine 0.5 km grid, the underlying SO_2 process is of our primary interest. The predictors optimal in terms of the minimum mean squared error are given by $E(\boldsymbol{\mu}|\boldsymbol{z})$. The joint distribution of $(\boldsymbol{\mu}, \boldsymbol{Z})$ is

$$\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{Z} \end{bmatrix} \sim \mathcal{N}_{n+N}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{N} & \boldsymbol{N}\boldsymbol{C}^T \\ \boldsymbol{C}\boldsymbol{N} & \boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T \end{bmatrix}\right). \tag{5}$$

The distribution (5) allows for full inference, yielding both the predictor and its error

$$\begin{aligned} \widehat{E(\boldsymbol{\mu}|\boldsymbol{z})} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{N}}\boldsymbol{C}^T\left(\widehat{\boldsymbol{M}} + \boldsymbol{C}\widehat{\boldsymbol{N}}\boldsymbol{C}^T\right)^{-1}\left[\boldsymbol{z} - \boldsymbol{C}\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right] \\ \widehat{Var(\boldsymbol{\mu}|\boldsymbol{z})} &= \widehat{\boldsymbol{N}} - \widehat{\boldsymbol{N}}\boldsymbol{C}^T\left(\widehat{\boldsymbol{M}} + \boldsymbol{C}\widehat{\boldsymbol{N}}\boldsymbol{C}^T\right)^{-1}\boldsymbol{C}\widehat{\boldsymbol{N}}. \end{aligned}$$

Note that in the predictor $\widehat{E(\boldsymbol{\mu}|\boldsymbol{z})}$, a naive regression forecast is corrected with a residual on the aggregated level distributed over respective grid cells.

4 Concluding remarks

To conclude, the change of support problem in our study is addressed by defining the underlying air pollution process to be an aggregation for respective grid cells. The joint distribution (5) allows to view the approach in analogy to block kriging (Gelfand 2010, p.524).

The application part of the study is under development.

References

Chow G. C., Lin A. (1971) Best linear unbiased interpolation, distribution, and extrapolation of time series by related series, *The Review of Economics and Statistics*, 53, 372-375.

Cressie N.A.C. (1993) *Statistics for Spatial Data*, Wiley, New York.

Gelfand A.E. (2010) Misaligned Spatial Data: The Change of Support Problem, in: *Handbook of Spatial Statistics*, Gelfand A. E., Diggle P. J., Fuentes M., Guttorp P. (Eds.), Chapman & Hall/CRC, 517-539.

Gotway C.A., Young L.J. (2002) Combining incompatible spatial data, *Journal of the American Statistical Association*, 97, 632-648.

Kaiser M. S., Daniels M. J., Furakawa K., Dixon P. (2002) Analysis of particulate matter air pollution using Markov random field models of spatial dependence, *Environmetrics*, 13, 615-628.

McMillan N.J., Holland D.M., Morara M., Feng J. (2010) Combining numerical model output and particulate data using Bayesian space-time modeling, *Environmetrics*, 21, 48-65.

Mugglin A.S., Carlin B.P. (1998) Hierarchical modeling in geographical information systems: Population interpolation over incompatible zones, *Journal of Agricultural, Biological and Environmental Statistics*, 3,111-130.

Mugglin A.S., Carlin B.P., Gelfand A.E. (2000) Fully model-based approaches for spatially misaligned data, *Journal of the American Statistical Association*, 95, 877-887.

Polasek W., Llano C., Sellner R. (2010) Bayesian methods for completing data in spatial models, *Review of Economic Analysis*, 2, 194-214.