Statistical spatial modeling of gridded air pollution data

Joanna Horabik, Zbigniew Nahorski

Systems Research Institute of Polish Academy of Sciences

Workshop on Uncertainty in GHG Inventories,
IIASA, 27-28 September 2007

Motivation

- Focus on a *spatial* aspect of emission inventories.
- This perspective is motivated with situations when two independent inventories are available (Winiwarter et.al., 2003):
  - **bottom-up** inventory which was constructed from a detailed knowledge of source types, locations and their emissions
  - **top-down** inventory - with low spatial resolution - which can be distributed into grid cells using activity data and appropriate weighting factors

We apply statistical spatial model to compare bottom-up inventory with spatially explicit activity data, which we treat as covariate information.

Outline

1. Statistical framework:
   - **Conditionally Autoregressive model** - based on Markov property extended to space
2. Illustrative data set and results
3. Extensions
   - space-varying regression models
   - space-time settings

Model

- $\boldsymbol{Y}' = (Y_1, \ldots, Y_n)$ - bottom-up emissions

$$Y_i \sim N(\mu_i, \sigma^2) \qquad i = 1, \ldots, n$$

- **Conditionally autoregressive (CAR)** formulation of a process μ_i: covariate information + spatially correlated residuals

$$\mu_i | \mu_j, i \neq j \sim N\left(\boldsymbol{x}_i' \boldsymbol{\beta} + \frac{1}{w_{i+}} \sum_{j \in \boldsymbol{N_i}} (\mu_j - \boldsymbol{x}_j' \boldsymbol{\beta}), \frac{\tau^2}{w_{i+}} \right)$$

$\boldsymbol{x}_i'$ - explanatory spatial covariates
$\boldsymbol{\beta}'$ - parameter coefficients
N_i - set of neighbors of area i
w_{i+} - number of neighbors
τ^2 - variance parameter

- Joint distribution of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is improper:

$$\boldsymbol{\mu} \sim N\left(\begin{bmatrix} \boldsymbol{x}_1'\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n'\boldsymbol{\beta} \end{bmatrix}, \tau^2 \begin{bmatrix} w_{1+} & & -w_{ij} \\ & \ddots & \\ -w_{ij} & & w_{n+} \end{bmatrix}^{-1} \right)$$

$$w_{i+} = \sum_{j \in N_i} w_{ij}$$

w_{ij} - neighbor weights: 1 for neighbors, 0 otherwise

- Model parameters are estimated with the **Bayes theorem**:

$$p(\boldsymbol{\beta}, \sigma^2, \tau^2 | \boldsymbol{Y}, \boldsymbol{X}) \propto L(\boldsymbol{Y}|\boldsymbol{\mu}, \sigma^2) p(\boldsymbol{\mu}|\boldsymbol{\beta}, \boldsymbol{X}, \tau^2) p(\boldsymbol{\beta}) p(\tau^2) p(\sigma^2)$$

  - The likelihood function $L(\boldsymbol{Y}|\boldsymbol{\mu}, \sigma^2)$ is based on the assumption

$$Y_i \sim N(\mu_i, \sigma^2) \qquad i = 1, \ldots, n$$

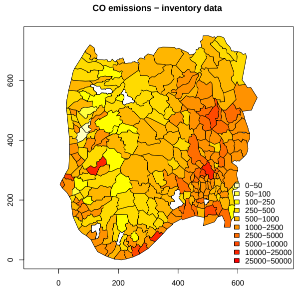
  - CAR distribution for $p(\boldsymbol{\mu}|\boldsymbol{\beta}, \boldsymbol{X}, \tau^2)$
  - Remaining vague priors for:

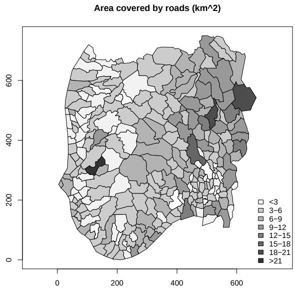
$$p(\boldsymbol{\beta}),\ p(\tau^2),\ p(\sigma^2)$$

- Posterior distributions of parameters are obtained using MCMC - Gibbs sampler algorithm

Data set

- CO emissions reported in municipalities of southern Norway (y_i)
- 259 municipalities
- Covariates for each municipality:
  - total area (x_1)
  - population (x_2)
  - area covered by roads (x_3)





- Initial linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

  showed that each covariate is significant, also $R^2 = 0.87$
- ...but the residuals are spatially correlated.

Results

- Model comparison using DIC statistics (lower the DIC the better a model)

$$\bar{D} + p_D = DIC$$

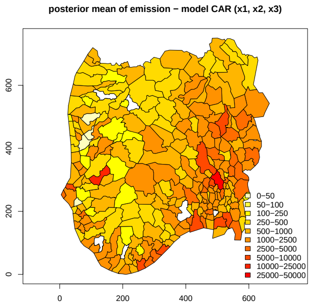
$\bar{D}$ - posterior deviance (a measure of fit)
p_D - effective number of parameters (a measure of complexity)

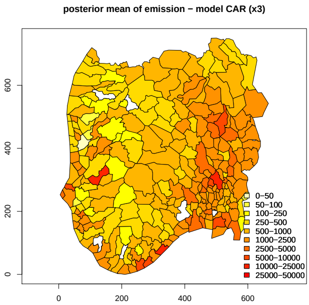
Model	$\bar{D}$	p_D	DIC
CAR (x_1, x_2, x_3)	217	108	325
CAR (x_1, x_2)	790	60	850
CAR (x_3)	-377	317	-60
linear regression (x_1, x_2, x_3)	415	5	420
linear regression (x_3)	588	3	591

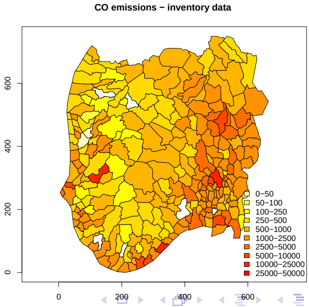
- **Conclusion:** missing, spatially correlated variable is contributing to overall emissions much better than the initial variables x_1, x_2.

Table: Parameter estimates

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)


posterior mean of emission − model CAR (x1, x2, x3)
0
200
400
600
0−50
50−100
100−250
250−500
500−1000
1000−2500
2500−5000
5000−10000
10000−25000
25000−50000


posterior mean of emission − model CAR (x3)
0
200
400
600
0−50
50−100
100−250
250−500
500−1000
1000−2500
2500−5000
5000−10000
10000−25000
25000−50000


CO emissions − inventory data
0
200
400
600
0−50
50−100
100−250
250−500
500−1000
1000−2500
2500−5000
5000−10000
10000−25000
25000−50000

Extension I: Space-varying regression model

- CAR prior for parameter coefficients $\boldsymbol{\beta}$

$$Y_i \sim N(x_i' \beta_i, \sigma^2) \qquad i = 1, \ldots, n$$

$$p(\beta_1, \ldots, \beta_n) \propto \exp\left[-\frac{1}{2\tau^2} \sum_{i \neq j} w_{ij} (\beta_i - \beta_j)^2 \right]$$

- The setting could be of potential use when considering *spatially varying emission factors*.

Extension II: Space-time model

Accounting for seasonal variations and regional structure:

$$Y(s,t) \sim N\left(\mu(s) + M(t, \beta(s)) + X(s,t), \sigma_Y^2(s)\right)$$

- site-specific mean - CAR model

$$\mu(s)$$

- seasonal component with spatially varying amplitudes

$$M = f(s)\sin(\omega t) + g(s)\cos(\omega t)$$

- space-time, non seasonal process:

$$X(t) = HX(t-1) + \eta(t)$$

(Wikle, Berliner, Cressie, 1998)

To sum up

Application of CAR structure to examine influence of activity data towards independent, bottom-up inventory

- **'Basic' CAR model**: capable to identify cases where some factors (e.g. emission point sources) are correctly reported in a bottom-up approach but are missing in activity data
- **CAR prior for parameter coefficients** β: can be helpful when spatially varying emission factors are considered
- **Space-time setting**: to account for regional structure and different dynamics of activity data

References

Banerjee S. *et.al.*(2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC Press.

Cressie, N. (1993) *Statistics for spatial data*, Revised edition, Wiley.

Gamerman, D. and Lopes, H.F. (2006) *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference*, 2nd edition, Chapman and Hall/CRC Press.

Winiwarter, W. *et.al.* (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.