

35/2009

**Raport Badawczy**  
**Research Report**

**RB/3/2009**

**A statistical model for spatial  
inventory data:  
a case study of N<sub>2</sub>O emissions  
in municipalities of southern  
Norway**

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 2009

A statistical model for spatial inventory data:  
a case study of N<sub>2</sub>O emissions in municipalities  
of southern Norway<sup>†</sup>

Joanna Horabik, Zbigniew Nahorski

Systems Research Institute, Polish Academy of Sciences

Newelska 6, 01-447 Warsaw, Poland

Tel: 4822 3810 100, Fax: 4822 3810 105

{Joanna.Horabik, Zbigniew.Nahorski}@ibspan.waw.pl

**Abstract.** In this paper we apply a linear regression with spatial random effect to model geographically distributed emission inventory data. The study presented is on N<sub>2</sub>O emission assessments for municipalities of southern Norway and on activities related to emissions (proxy data). Taking advantage of spatial dimension of the emission process, the method proposed is intended to improve inventory extension beyond its earlier coverage. For this, the proxy data are used. The conditional autoregressive model is used to account for spatial correlation between municipalities. Parameter estimation is based on the maximum likelihood method and the optimal predictor is developed. The results indicate that inclusion of a spatial dependence component lead to improvement in both representation of the observed data set and prediction.

**Keywords:** geographically distributed emission inventories, conditionally autoregressive models, spatial prediction

## 1. Introduction

This study focus on a spatial aspect of inventories for atmospheric pollutants. The study is motivated with situations when emission inventory is to be expanded beyond its present coverage, where relevant activity data are missing. In absence of measured data (activities) contributing to emissions, proxy data 'about' activities can be used. The aim is to provide a tool to improve inventory, developed with proxy data, by taking advantage of spatial correlation of an emission process.

In case of greenhouse gases, spatial resolution is usually not crucial for the emission effect as such. However, there are several situations when spatial dimension is needed. In elaborated models of climate change, for instance model HadAM3 of the British Meteorological Office (Pope et al., 2000), transport of greenhouse gases is modelled similarly to other pollutants, e.g. SO<sub>x</sub> and NO<sub>x</sub>. With growing resolution, for

<sup>†</sup> Praca zaakceptowana do druku w czasopiśmie *Climatic Change*.



instance in national models of this kind, a need for finer grid of inventory data becomes important. The proposed method can be used for this purpose. Other examples include validations of regional inventories by field measurements or by inverse modelling in top down approaches (Ciais et al., 2009; Rivier et al., 2009).

The topic of spatial heterogeneity of greenhouse gas emissions and sequestration can be addressed in various ways. For instance, the spatial distribution of greenhouse gas emissions for Ukraine has been presented in (Bun et al., 2009). (Theloke et al., 2007) develop a methodology for spatial (and temporal) disaggregation of greenhouse gas annual country totals. (Oijen and Thomson, 2009) used a process-based forest model which accounts for spatial distribution of climate and soil; a Bayesian calibration was employed to quantify uncertainties.

When performing a statistical inference of spatial inventory data, we account for the fact that values at proximate locations tend to be more alike. This can be modelled by using spatial statistics. Moreover, since for each grid cell we have information on aggregated emission values, these are called areal data (also known as lattice data). A popular tool for incorporating this kind of spatial information is the conditional autoregressive (CAR) model proposed by (Besag, 1974). As opposed to the geostatistical models with spatially continuous data, the CAR models have been developed to account for a situation where the set of all possible spatial locations is countable. The idea is to define a model in terms of the conditional distribution of the observation at one location given its values at other neighbouring locations. Applications of the CAR model include, among others, mapping diseases in counties and modelling particulate matter air pollution in space and time (Kaiser et al., 2002).

The aim of the present paper is to demonstrate usefulness of the CAR model to analyse data from spatially distributed emission inventory. With available proxy data related to emissions and an independent set of (modelled or measured) emission assessments, one may build a suitable regression model. Inclusion of the spatial component is intended to improve estimation results, compensating for weaker explanatory power of proxy information. Based on the model, we develop the optimal predictor to extend the inventory.

The outline of the study is the following. Section 2 presents an illustrative data set, including an initial non-spatial model. As a next step, the model is enriched with a spatial random effect. We use the conditional autoregressive structure to account for a spatial correlation between neighbouring areas (municipalities, in this case). The model is characterized in Section 3. It comprises model formulation, estimation and prediction. Results are presented in Section 4 - we fit

the spatial model and assess its predictive performance by means of a cross-validation procedure. Section 5 contains final remarks.

## 2. Preliminary explorations

Our illustration is provided with the data set on  $N_2O$  emissions reported in 2006 for municipalities of southern Norway. In 2006, the main contributors to the country total  $N_2O$  emissions were as follows (National Inventory Report, 2008). 47% of emissions were attributed to agriculture, with agricultural soil as the most important source. The second most important source was production of nitric acid in two plants, which accounted for 37%. Nitric acid is used in the production of fertilizer. Emissions from road traffic amounted to 4%. The remaining 12% included emissions from e.g. manure management and waste-water handling.

The considered map of southern Norway covers 259 municipalities out of 431 in the whole of Norway. The data come from the StatBank (available at <http://www.ssb.no>) in Statistics Norway. According to the StatBank identification system, the area of our interest covers the municipalities of numbers from 0101 to 1449. One of the mentioned nitric acid plants is operating in Porsgrunn municipality, which is a relatively small municipality located near the southern coast of the considered area, see also (Perez-Ramirez, 2007). Emissions from this kind of point sources are usually reported and there is no need to model them. In our analysis we do not consider emissions from this source.

The municipalities have been chosen by the StatBank as the smallest unit for geographical distribution of emissions. Details on the Norwegian emission model can be found in (Sandmo, 2009).

Out of the statistics available in the StatBank at the municipal level, we consider the following variables that might explain spatial distribution of  $N_2O$  emissions. Figures on livestock and detailed statistics on agricultural usage are the ones that are the most relevant to the  $N_2O$  emissions. However, these data sets contained a large number of missing values, and as such were of poor quality. Emissions from agriculture can be generally characterized with data on agricultural area in use as well as with data on persons employed in agriculture. Regarding emissions from stationary and mobile sources, data on population can be of use. Besides the Porsgrunn plant, emissions from fertilizer production occurs in a small number of municipalities. There is a lack of statistics on relevant production, financial data or employment at the municipal level (Flugsrud et al., 2009).

Independence of the above-mentioned variables from the emission data was consulted with the inventory preparers from Statistics Norway (Flugsrud et al., 2009). We found out that for the municipal emission assessments they used figures from the agricultural statistics that are both more detailed and more comprehensive than those described above. In addition, a model that estimates emissions of ammonia from agriculture were used, as were figures on energy use.

Let us denote<sup>1</sup>

- $y_i$  - N<sub>2</sub>O emissions (tonnes) (Table 03535),  $y = (y_1, \dots, y_n)^T$
- $x_{i,1}$  - agricultural area in use (decare) (Table 06462),  
 $x_1 = (x_{1,1}, \dots, x_{n,1})^T$
- $x_{i,2}$  - persons employed in agriculture (Table 03324),  
 $x_2 = (x_{1,2}, \dots, x_{n,2})^T$
- $x_{i,3}$  - population (Table 05231),  $x_3 = (x_{1,3}, \dots, x_{n,3})^T$ .

Figure 1 presents a scatterplot matrix for these data. We notice that the relationship between  $y$  and  $x_1$  is more pronounced than between  $y$  and  $x_2$ . Our aim is to explore opportunities for improvements of inventory prepared in the absence of information on agricultural area ( $x_1$ ) activity, but using data on persons employed in agriculture ( $x_2$ ) as its proxy. Therefore we define a multiple regression model

$$Y_i = \beta_0 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \quad (1)$$

where  $\varepsilon_i$  are independent random variables following normal distribution with mean equal 0 and variance  $\sigma^2$  and  $i = 1, \dots, 259$  indexes municipalities. In the sequel we compare results of the above model to the one with variable  $x_1$  instead of  $x_2$ . We distinguish between an observation ( $y_i$ ) and a random variable ( $Y_i$ ) generating this observation. In the model (1) regression coefficients of the covariates  $x_2$  and  $x_3$  have p-values equal 2e-16 and 2.07e-09, respectively. The model explains 79% of variability in N<sub>2</sub>O emissions - coefficient of determination is  $R^2 = 0.79$ .

Residuals of the model, that is responses minus fitted values, are presented in Figure 2: a residual plot (a) and a map (b). From a residual map we can identify the cluster of municipalities with underestimated emissions (yielding positive residuals) in the eastern part; moreover municipalities with highly overestimated emissions (yielding negative residuals) are located in the western region. In Figure 2(a) residuals are plotted against municipality numbers. Since municipalities are not randomly numbered and neighbouring areas usually have

<sup>1</sup> In brackets we report a number of the table containing the data set available from the StatBank web site as of October 2009.

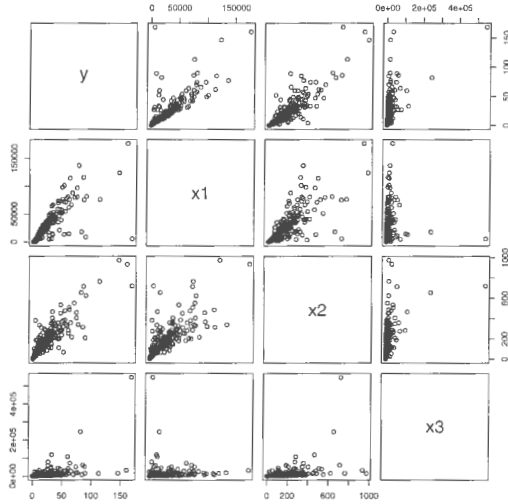


Figure 1. Scatterplot matrix showing plausible relations between data on: N<sub>2</sub>O emissions ( $y$ ), agricultural area ( $x_1$ ), persons employed in agriculture ( $x_2$ ) and population ( $x_3$ ) in municipalities.

close identification numbers, we again note that there exist regions with underestimated and overestimated emissions.

We check spatial correlation in the residuals using the Moran's  $I$  statistic

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (\varepsilon_i - \bar{\varepsilon})(\varepsilon_j - \bar{\varepsilon})}{\sum_i (\varepsilon_i - \bar{\varepsilon})^2},$$

where  $\varepsilon_i$  - a residual of linear regression in the area  $i$ ,  $\bar{\varepsilon}$  - the mean of residuals,  $w_{ij}$  - the adjacency weights ( $w_{ij} = 1$  if  $j$  is a neighbour of  $i$  and 0 otherwise, also  $w_{ii} = 0$ ). We consider two municipalities as neighbours if they share common border. Moran's  $I$  can be recognized as a modification of the correlation coefficient. It accounts for correlation between residuals in area  $i$  and nearby locations and takes values approximately on the interval  $[-1, 1]$ . Higher (positive) values of  $I$  suggest stronger positive spatial association. Under the null hypothesis, where  $\varepsilon_i$  are independent and identically distributed,

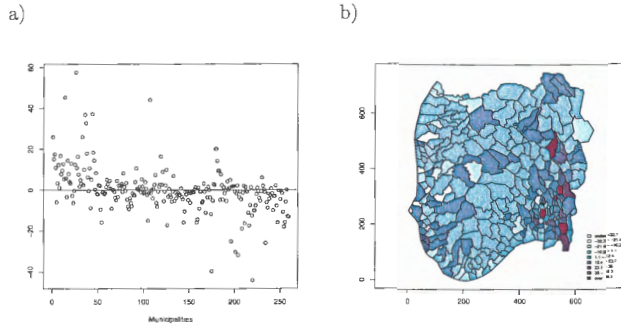


Figure 2. Residuals from the linear model with covariates on persons employed in agriculture ( $x_2$ ) and population ( $x_3$ ).

$I$  is asymptotically normally distributed, with the mean and variance known, see e.g. (Banerjee et al., 2004).

In the case of the residuals from model (1) with covariates on  $x_2$  and  $x_3$  Moran's  $I$  is equal to 0.2466. The corresponding test statistic  $z$  (Moran's  $I$  standardized with the asymptotic mean and variance) is equal to  $z = 6.6953$  while  $z_{cr} = 2.33$  at the significance level  $\alpha = 0.01$ . Thus we reject the null hypothesis of no spatial correlation of errors. Moran's  $I$  is, however, recommended as an exploratory information on spatial association, rather than a measure of spatial significance (Banerjee et al., 2004).

### 3. Modelling spatial correlation

In this section we develop a model to characterize the spatial distribution of  $N_2O$  emissions in municipalities. Further, we provide details on the model estimation, prediction and an applied cross-validation procedure. The calculations were accomplished using the statistical software R (R Development Core Team, 2008).

#### 3.1. THE MODEL

Let  $Y_i$  denote a stochastic variable associated with the observed emission ( $y_i$ ) defined at each spatial location  $i$  for  $i = 1, \dots, n$ . It is assumed that the random variables  $Y_i$  for  $i = 1, \dots, n$  follow normal distribution



with the mean  $\mu_i$  and common variance  $\sigma^2$

$$Y_i | \mu_i, \sigma^2 \sim \mathcal{N}(\mu_i, \sigma^2). \quad (2)$$

The collection of all  $Y_i$  is denoted as  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ . Given the values of  $\mu_i$  and  $\sigma^2$ , stochastic variables  $Y_i$  are assumed independent, thus the joint distribution of the process  $\mathbf{Y}$  conditional on the mean process  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$  is

$$\mathbf{Y} | \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n), \quad (3)$$

where  $\mathbf{I}_n$  is an identity  $n \times n$  matrix.

Our approach to modelling the mean  $\mu_i$  expresses the observation that available covariates explain part of the spatial pattern in observations, and the remaining part is captured through a regional clustering. To this extent we make use of the conditional autoregressive model. The CAR structure is given through specification of the full conditional distribution functions for  $i = 1, \dots, n$

$$\mu_i | \mu_{j, j \neq i} \sim \mathcal{N} \left( \mathbf{x}_i^T \boldsymbol{\beta} + \rho \sum_{j \neq i} \frac{w_{ij}}{w_{i+}} (\mu_j - \mathbf{x}_j^T \boldsymbol{\beta}), \frac{\tau^2}{w_{i+}} \right) \quad (4)$$

with  $w_{i+} = \sum_j w_{ij}$  being the number of neighbours of area  $i$ ;  $\mathbf{x}_i$  is a vector containing 1 for the intercept  $\beta_0$  and  $k$  explanatory covariates of area  $i$  e.g. population;  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$  is a vector of regression coefficients and  $\tau^2$  is a variance parameter. An additional parameter  $\rho$  is introduced into (4) in order to remedy singularity of the covariance function in the joint distribution of  $\boldsymbol{\mu}$ , for more details see e.g. (Banerjee et al., 2004).

The second summand of the conditional expected value of  $\mu_i$  (a remainder) is proportional to the average value of remainders  $\mu_j - \mathbf{x}_j^T \boldsymbol{\beta}$  for those areas  $j$  which are the neighbours of the site  $i$  (i.e.  $w_{ij} = 1$ ). The proportion is calibrated with parameter  $\rho$ . The conditional variance is inversely proportional to the number of neighbours  $w_{i+}$ .

Given (4), the joint probability distribution of the process  $\boldsymbol{\mu}$  is the following (Banerjee et al., 2004; Cressie, 1993)

$$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \tau^2 (\mathbf{D} - \rho\mathbf{W})^{-1}), \quad (5)$$

where  $\mathbf{X}$  is the (design) matrix containing transposed vectors  $\mathbf{x}_i$ ,  $i = 1, \dots, n$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix};$$

$D$  is an  $n \times n$  diagonal matrix with  $w_{i+}$  on the diagonal; and  $W$  is an  $n \times n$  matrix with adjacency weights  $w_{ij}$ .

### 3.2. ESTIMATION

Estimation of unknown parameters  $\beta, \rho, \sigma^2$  and  $\tau^2$  is based on the maximum likelihood approach. From (3) and (5) we obtain the joint unconditional distribution of  $Y$

$$Y \sim \mathcal{N}(X\beta, M + N), \quad (6)$$

where for notational simplicity  $M = \sigma^2 I_n$  and  $N = \tau^2 (D - \rho W)^{-1}$  were introduced. To see this let us write (3) as  $Y = \mu + v$ , where  $v \sim \mathcal{N}(0, M)$  and (5) in the form of  $\mu = X\beta + \nu$ , where  $\nu \sim \mathcal{N}(0, N)$ . Together we obtain  $Y = X\beta + \nu + v$ , which is a sum of a constant and two independent normal random variables with the distribution  $\nu + v \sim \mathcal{N}(0, M + N)$ . Compare also the lemma of Lindley and Smith (Lindley and Smith, 1972).

The log likelihood associated with (6) is, see e.g. (Papoulis and Pillai, 2002)

$$\begin{aligned} \mathcal{L}(\beta, \rho, \sigma^2, \tau^2) = & -\frac{1}{2} \log(|M + N|) - \frac{n}{2} \log(2\pi) \\ & - \frac{1}{2} (\mathbf{y} - X\beta)^T (M + N)^{-1} (\mathbf{y} - X\beta), \end{aligned} \quad (7)$$

where  $|\cdot|$  denotes a determinant and  $\mathbf{y}$  is a vector containing the observations  $y_i$ ,  $i = 1, \dots, n$ . With fixed  $\rho, \sigma^2$  and  $\tau^2$ , the log likelihood (7) is maximized for

$$\hat{\beta}(\rho, \sigma^2, \tau^2) = (X^T (M + N) X)^{-1} X^T (M + N) \mathbf{y}, \quad (8)$$

which substituted back into (7) provides the profile log likelihood

$$\begin{aligned} \mathcal{L}(\rho, \sigma^2, \tau^2) = & -\frac{1}{2} \log(|M + N|) - \frac{n}{2} \log(2\pi) \\ & - \frac{1}{2} \left( \mathbf{y} - X (X^T (M + N) X)^{-1} X^T (M + N) \mathbf{y} \right)^T \\ & \times (M + N)^{-1} \\ & \times \left( \mathbf{y} - X (X^T (M + N) X)^{-1} X^T (M + N) \mathbf{y} \right). \end{aligned} \quad (9)$$

Further maximization of  $\mathcal{L}(\rho, \sigma^2, \tau^2)$  is performed numerically. One also needs to ensure that the matrix  $D - \rho W$  is nonsingular. This

is guaranteed if  $\lambda_1^{-1} < \rho < \lambda_n^{-1}$ , where  $\lambda_1 < \dots < \lambda_n, \lambda_i \neq 0, i = 1, \dots, n$  are the eigenvalues of  $D^{-1/2}WD^{-1/2}$ , see (Banerjee et al., 2004; Cressie, 1993). Our optimization procedure takes this constraint into account.

### 3.3. PREDICTION

Consider a random variable  $Y_0$  associated with emissions at an unobserved location and let  $\mu_0$  denote the mean corresponding to this variable. We assume that the distribution of  $Y_0|\mu_0$  is of the form (2) and the distribution of  $\mu_0|\boldsymbol{\mu}$  is of the form (4). The predictor of the observation  $Y_0$ , that is optimal in terms of minimum mean squared error, is given by  $E(Y_0|\mathbf{y})$ . It should be stressed that knowledge on covariates  $\mathbf{x}_0$  is required to calculate the predictor in the location considered.

To begin with, we derive the conditional distribution of  $\mu|\mathbf{y}$  based on (3), (5) and (6)

$$\mu|\mathbf{y} \sim \mathcal{N}(BC, B) \quad (10)$$

with  $B = (M^{-1} + N^{-1})^{-1}$  and  $C = M^{-1}\mathbf{y} + N^{-1}\mathbf{X}\beta$ .

Next we develop the predictor  $E(Y_0|\mathbf{y})$ , see also (Kaiser et al., 2002). In deriving the formula we will make use of the following property of the conditional expected value:  $Y_0 = E(Y_0|\mu_0)$  and analogously  $\mu_0 = E(\mu_0|\boldsymbol{\mu})$ . We have

$$\begin{aligned} E(Y_0|\mathbf{y}) &= E[E(Y_0|\mu_0)|\mathbf{y}] = E[\mu_0|\mathbf{y}] = E[E(\mu_0|\boldsymbol{\mu})|\mathbf{y}] \\ &= E\left[\mathbf{x}_0^T\beta + \rho \sum_j \frac{w_{0j}}{w_{0+}} (\mu_j - \mathbf{x}_j^T\beta) \mid \mathbf{y}\right] \\ &= \mathbf{x}_0^T\beta - \rho \sum_j \frac{w_{0j}}{w_{0+}} \mathbf{x}_j^T\beta + E\left[\rho \sum_j \frac{w_{0j}}{w_{0+}} \mu_j \mid \mathbf{y}\right]. \quad (11) \end{aligned}$$

We use the expression (10) to calculate the rightmost expectation in the last equality of (11) and denoting the  $j$ th element of the vector  $BC$  with  $l_j$ , we get the predictor

$$E(Y_0|\mathbf{y}) = \mathbf{x}_0^T\beta + \rho \sum_j \frac{w_{0j}}{w_{0+}} (l_j - \mathbf{x}_j^T\beta). \quad (12)$$

In order to assess quality of the prediction we perform a leave-one-out cross-validation procedure. The idea is to fit a model to a data set from which a single observation was dropped. This observation is considered as unobserved and its value is calculated using the predictor (12). The operation is repeated for each observation ( $n$  times). The

difference between the observation  $y_i$  and the prediction  $y_i^*$ ,  $d_i = y_i - y_i^*$ , constitutes a base to quantify prediction error. We summarize it forming the mean squared error

$$\text{mse} = \frac{1}{n} \sum_i (y_i - y_i^*)^2, \quad (13)$$

which should be as low as possible, indicating how well a model predicts data. We report also the minimum and maximum value of  $d_i$ , average values of the absolute differences  $|d_i|$ , and the sample correlation coefficient  $r$  between the predicted and observed values.

#### 4. Results

The spatial CAR model has been applied to the emission data. In addition, we estimate the linear regression (1) denoted  $\text{LM}(x_2, x_3)$ , as well as the model  $\text{LM}(x_1, x_3)$  with the variable on agricultural area ( $x_1$ ) instead of the number of people employed in agriculture ( $x_2$ ). The linear regression models are also estimated with the maximum likelihood approach.

The results are compared using the Akaike Information Criterion (AIC), which is a suitable tool for comparison of models estimated with the maximum likelihood method. The AIC is calculated as a sum of twice the negative log likelihood  $\mathcal{L}(\theta)$  and twice the number of parameters  $p$ :

$$\text{AIC} = -2\mathcal{L}(\theta) + 2p.$$

The term  $-2\mathcal{L}(\theta)$  measures how well the model fits the data; the larger this value, the worse the fit. Model complexity is summarized by the number of parameters  $p$ . The idea of the AIC is to favour a model with a good fit and to penalize for the number of parameters. Thus models with smaller AIC are preferred to models with larger AIC.

Table I. Model comparison for the linear regressions (LM) and the spatial model (CAR)

Model	$-\mathcal{L}$	AIC
$\text{LM}(x_2, x_3)$	1622.27	3252.55
$\text{CAR}(x_2, x_3)$	1552.32	3116.65
$\text{LM}(x_1, x_3)$	1281.98	2573.97

For the estimated models both the negative log likelihood and the AIC are displayed in Table I. The applied spatial structure improved the results considerably. The negative log likelihood  $-\mathcal{L}$  decreased from 1622 for the linear regression  $\text{LM}(x_2, x_3)$  to 1552 for the spatial model with the same set of covariates  $\text{CAR}(x_2, x_3)$ . The spatial model includes only two parameters ( $\rho$  and  $\tau^2$ ) more than its linear regression counterpart. In terms of the AIC criterion the spatially enriched model is preferred (has lower the AIC), since the decrease in the negative log likelihood overwhelms increased model complexity.

To put this improvement in a perspective, we present results for the non-spatial model  $\text{LM}(x_1, x_3)$  with the variable on agricultural area. Spatially explicit model  $\text{CAR}(x_2, x_3)$  with the proxy is still much worse than the model  $\text{LM}(x_1, x_3)$ . The latter has  $-\mathcal{L} = 1282$  and  $\text{AIC} = 2574$ . In terms of the negative log likelihood  $-\mathcal{L}$ , the gain achieved by taking into account a spatial correlation can be summarized as a 20.5% improvement over the initial model. Parameter estimates for the models are reported in Table II.

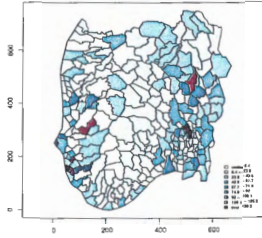
Table II. Estimated parameter values

Model	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\sigma^2$	$\rho$	$\tau^2$
$\text{LM}(x_2, x_3)$	-1.882	-	0.129	0.00012	15.494	-	-
$\text{CAR}(x_2, x_3)$	-1.965	-	0.128	0.00013	15.127	0.9984	0.6186
$\text{LM}(x_1, x_3)$	0.177	0.0007	-	0.00031	15.494	-	-

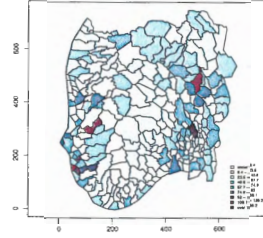
We regard the method as a tool, which can help to extend spatial coverage of inventories in a situation, when the inventories are based on proxy data. The motivation behind is that proxy data is more frequently available than measured data. This task calls for prediction. To evaluate predictive performance of the method, we employ a cross-validation technique. The procedure was applied to the spatial model and its non-spatial counterpart with the same set of proxy variables, see Table III. We note again that observation  $y_i$  is not accounted for in the construction of the predictor  $y_i^*$ , thus a model is reestimated for each observation separately. In a case of the spatial model, it is a time consuming procedure.

Cross-validation results are also displayed in Figure 3 as predicted values for the respective models, along with the observations. It can be noticed that the spatial model predicts the original data slightly better. However, we suspect that some of the differences might have been masked because the mapped values are binned into 9 classes. Therefore, in Figure 4 we present the model residuals  $d_i$ . Here we can clearly see

a)



b)



c)

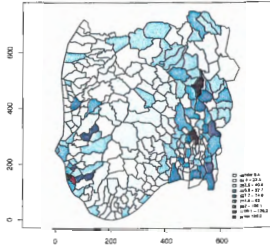


Figure 3. Predicted values in the model  $CAR(x_2, x_3)$  (a); predicted values in the linear regression  $LM(x_2, x_3)$  (b); observed emission (c).

that for the linear regression in the eastern part there is a cluster of municipalities with highly underestimated values (positive residuals). Application of the spatial random effect to some extent remedied this deficiency.

Values for analysis of prediction error for the two models are given in Table III. The spatial model showed noticeable improvement over the linear regression. In particular, the mean squared error was reduced by about 15% from 135 to 115. A smaller reduction is noted for the average in absolute residuals. With inclusion of spatial dependence we obtain higher minimum and lower maximum residuals, which yields a reduction of over 14% in a spread of differences  $d_i$ .

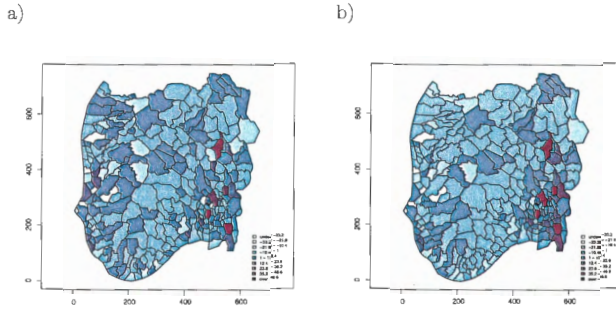


Figure 4. Residuals from cross-validated values for the model  $CAR(x_2, x_3)$ (a); and for the model  $LM(x_2, x_3)$  (b)

Table III. Cross-validation results

Model	mse	avg( d )	min(d)	max(d)	$\tau$
$LM(x_2, x_3)$	134.67	7.06	-44.63	58.03	0.877
$CAR(x_2, x_3)$	115.38	6.87	-41.57	46.60	0.896

In Figure 5 predicted values  $y_i^*$  are plotted against the observations  $y_i$  for the two models. An overall impression is that the spatial model provides better predictions. This is confirmed by a higher value of sample correlation coefficient  $\tau$ , see Table III. It should be noted however, that small value observations (i.e. below ca 10 tonnes) are predicted more accurately with a linear regression approach. This observation is related to a general feature of the conditional autoregressive models, which tend to over-smooth data.

## 5. Concluding remarks

The goal of this study was to demonstrate that emission inventories may be improved by making efficient use of spatial information. We consider a case study with a geographically distributed inventory for  $N_2O$ . Suppose we wish to spatially expand the inventory beyond the present coverage. We have some proxy data available both for the

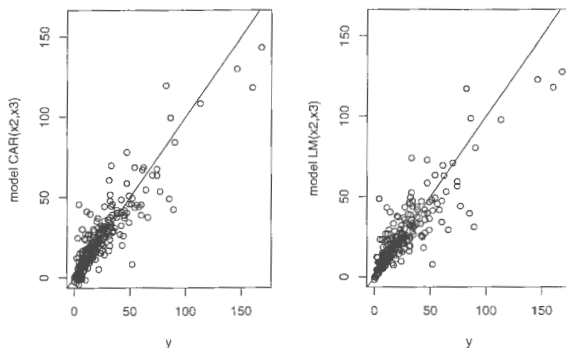


Figure 5. Predicted values vs. observed values for the model  $CAR(x_2, x_3)$  (a); and for the model  $LM(x_2, x_3)$  (b)

present inventory area and in a predictive capacity. The proxy data is, however, of limited adequacy.

The idea is to take advantage of potentially existing spatial correlation to improve the outcome. First, the task includes model estimation based on available measured/modelled inventory as a first step. Second, an appropriately constructed predictor is employed to produce emission assessment out of the proxy information. To model spatial dependencies we make use of the conditional autoregressive structure. It was introduced into a linear regression as a random effect.

The results indicate that inclusion of a spatial dependence component lead to improvement in both representation of the observed data set and prediction. Specifically, introduction of spatial random effect into a model with less adequate covariate (on number of people employed in agriculture) improved estimation results by over 20% of what would be obtained using more relevant activity data (on agricultural area). In terms of prediction, a 15% reduction in the mean squared error was achieved.

The presented application of the method seems to be particularly suitable to  $N_2O$  emissions, since  $N_2O$  emission pathways include among others agriculture and soil emissions. These factors tend to be spatially correlated and have quite often been modelled with spatial tools,



e.g (Sigua and Hudnall, 2008). Based on a study of 15 national greenhouse gas inventories, (Leip, 2009) note that  $N_2O$  emissions from agricultural soils are dominating the uncertainty of not only the agricultural sector, but also the overall greenhouse gas inventory for many countries.

Accounting for spatial scale of inventories may have one more aspect. One may compare estimation results for alternative proxy data used and try to conclude on their relevance. Similar analysis has been already performed in some studies, e.g. see (Winiwarter et al., 2003). In that study two sets of data on  $NO_x$  (Nitrogen oxides) emissions over the same spatial grid for the Greater Athens, Greece were compared. The authors examine significance of area, line and point emission sources on the basis of statistical exploratory tools and a visual comparison of maps. In the case study presented here, we believe the problem is more of data availability rather than lack of knowledge on the relevant covariates. Therefore our focus remains on prediction.

The applied spatial model proved to be especially successful when dealing with underestimated emission assessments. Further developments of the method would be required to deal with the problem of over-smoothed values for low emission observations.

### Acknowledgements

We are grateful to Ketil Flugsrud, Trond Sandmo and Kathrine Loe Hansen from Statistics Norway, Division of Environmental Statistics for comprehensive information on the inventory data. In addition, we are thankful to two undisclosed reviewers for their valuable comments that considerably helped us to shape the final version of the manuscript.

### References

- Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman and Hall: London
- 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
- Bun R, Hamal K, Gusti M, Bun A (2009) Spatial GHG inventory on regional level: Accounting for uncertainty. *Climatic Change*. This issue
- Ciais P, Rayner P, Chevallier F, Bousquet P, Logan M, Peylin P, Ramonet M (2009) Combining top-down and bottom-up information for estimating  $CO_2$  fluxes: Methods and perspectives. *Climatic Change*. This issue
- Cressie N (1993) *Statistics for spatial data*. Revised edition, Wiley
- Flugsrud K, Sandmo T, Hansen KL (2009) Personal communication

- Kaiser MS, Daniels MJ, Furakawa K, Dixon P (2002) Analysis of particulate matter air pollution using Markov random field models of spatial dependence. *Environmetrics* 13:615-628
- Leip A (2009) The uncertainty of the uncertainty ... On the example of the quality assessment of the greenhouse gas inventory for agriculture in Europe. *Climatic Change*. This issue
- Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society* 34(B):1-41
- National Inventory Report - Norway (2008), Greenhouse Gas Emissions 1990-2006 Reported According to the UNFCCC Reporting Guidelines. Available at <http://www.sft.no/publikasjoner/2416/ta2416.pdf>
- Oijen M van, Thomson A (2009) Towards Bayesian uncertainty quantification for forestry models used in the U.K. GHG inventory for LULUCF. *Climatic Change*. This issue
- Papoulis A, Pillai SU (2002) *Probability, random variables and stochastic processes*. Fourth edition, McGraw Hill
- Perez-Ramirez J (2006) Prospects of N<sub>2</sub>O emission regulations in the European fertilizer industry. *Applied Catalysis B: Environmental* 70:31-35
- Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dynamics* 16:123-146
- R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at [www.r-project.org](http://www.r-project.org) ISBN 3-900051-07-0
- Rivier L, Peylin Ph, Ciais Ph, Gloor M, Rdenbeck C, Geels C, Karstens U, Bousquet Ph, Brandt J, Heimann M (2009) European CO<sub>2</sub> fluxes from atmospheric inversions using regional and global transport models. *Climatic Change*. This issue
- Sandmo T (ed.) (2009) *The Norwegian emission inventory 2009. Documentation of methodologies for estimating emissions of greenhouse gases and long-range transboundary air pollutants*. Statistics Norway, Report 2009/10
- Sigua GC, Hudnall WH (2008) Kriging analysis of soil properties. *Journal of Soils Sediments* 8:192-202
- Theloke J, Pfeiffer H, Pregger T, Scholz Y, Koble R, Kummer U, Nicklass D, Thiruchittampalam B, Friedrich R (2007) Development of a methodology for temporal and spatial resolution of greenhouse gas emission inventories for validation. In: *Proceedings of the 2nd International Workshop on Uncertainty in Greenhouse Gas Inventories*, IIASA 27-28.09.2007. Available at <http://www.ibspan.waw.pl/ghg2007/GHG-total.pdf>
- Winiwarter W, Dore Ch, Hayman G 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

the 1990s, the number of people in the world who are undernourished has increased from 600 million to 800 million.

There are a number of reasons for this increase. One of the main reasons is that the world population has increased from 5 billion in 1987 to 6 billion in 2000, and is projected to reach 9 billion by 2050.

Another reason is that the world's food production has not kept pace with the increase in population. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural land has decreased. In 1987, there were 1.4 billion hectares of agricultural land, but in 2000, there were only 1.3 billion hectares.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

There are a number of reasons for this. One of the main reasons is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

Another reason is that the world's agricultural production has become more inefficient. In 1987, the world produced 2.1 billion tonnes of food, but in 2000, it only produced 2.4 billion tonnes.

