

Improving resolution of a spatial air pollution inventory with a statistical inference approach

Joanna Horabik · Zbigniew Nahorski

Received: 2 January 2013 / Accepted: 1 December 2013 / Published online: 28 January 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract This paper presents a novel approach to allocation of spatially correlated data, such as emission inventories, to finer spatial scales, conditional on covariate information observable in a fine grid. Spatial dependence is modelled with the conditional autoregressive structure introduced into a linear model as a random effect. The maximum likelihood approach to inference is employed, and the optimal predictors are developed to assess missing values in a fine grid. An example of ammonia emission inventory is used to illustrate the potential usefulness of the proposed technique. The results indicate that inclusion of a spatial dependence structure can compensate for less adequate covariate information. For the considered ammonia inventory, the fourfold allocation benefited greatly from incorporation of the spatial component, while for the ninefold allocation this advantage was limited, but still evident. In addition, the proposed method allows correction of the prediction bias encountered for the upper range emissions in the linear regression models.

1 Introduction

The development of high-resolution emission inventories is essential for designing suitable abatement measures. Spatial distributions of emissions can serve as an input for atmospheric dispersion models, which in turn may produce concentration maps of pollutants contributing to the adverse health effects, like ammonia emissions. For other air pollutants, such as greenhouse gases (GHG), spatial patterns become helpful in improving identification of distributed emission sources.

Numerous issues underlying preparation of spatially resolved GHG inventory were discussed e.g. in Boychuk and Bun ([this issue](#)), Bun et al. 2010 or Thiruchittampalam et al. 2010. In general, the task crucially depends on availability of spatially distributed activity data. For instance, at present in Poland the activity data relevant to GHG emissions can be obtained at a level of country regions (voivodships). Information of higher spatial resolution can be often obtained only for some proxy data related to GHG emissions, such as land use and linear emission sources. Recently, also nighttime lights observed by a satellite have been used for

This article is part of a Special Issue on “Third International Workshop on Uncertainty in Greenhouse Gas Inventories” edited by Jean Ometto and Rostyslav Bun.

J. Horabik (✉) · Z. Nahorski

Polish Academy of Sciences, Systems Research Institute, ul. Newelska 6, 01-447 Warsaw, Poland
e-mail: Joanna.Horabik@ibspan.waw.pl

more accurate estimation of spatial distribution of CO₂ emissions (Ghosh et al. 2010; Oda and Maksyutov 2011).

Typically, the regression models have been applied for spatial allocation of emission data (Dragosits et al. 1998; Oda and Maksyutov 2011). However, emissions in general tend to be spatially correlated, which provides opportunity for potential improvements. This idea motivated us to develop a more advanced approach for accurate disaggregation of air pollution data.

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

In this study we propose to apply methods of spatial statistics to produce higher resolution emission inventory data, taking advantage of more detailed land use information. The approach resembles to some extent the method of Chow and Lin (1971), originally proposed for disaggregation of time series based on related, higher frequency series. Here, a similar methodology is employed to disaggregate spatially correlated data.

Regarding an assumption on residual covariance, we apply the structure suitable for areal data, i.e. the conditional autoregressive (CAR) model. Although the CAR specification is typically used in epidemiology (Banerjee et al. 2004), it was also successfully applied for modelling air pollution over space (Kaiser et al. 2002; McMillan et al. 2010). Compare also Horabik and Nahorski (2010) for another application of the CAR structure to model spatial inventory of GHG emissions. The maximum likelihood approach to inference is employed, and the optimal predictors are developed to assess missing concentrations in a fine grid.

The application part of the study concerns an ammonia (NH₃) emission inventory in a region of Poland. Ammonia is emitted mainly by agricultural sources such as livestock production and fertilized fields. Its high concentrations can lead to acidification of soils, forest decline, and eutrophication of waterways. Ammonia emissions are also recognized for their importance in contributing to fine particulate matter; hence its spatial distribution is of great importance. However, agricultural emission sources cannot be measured directly, and spatial emission patterns need to be assessed otherwise. This issue was addressed, among others, by Dragosits et al. 1998, where agricultural and land cover data were used to disaggregate the national NH₃ emission totals across Great Britain. We demonstrate that the straightforward approaches based on linear dependences might be improved by introducing a spatial random effect.

Nevertheless, the proposed approach is of wider applicability, and can be used in numerous situations where higher resolution of spatial data is needed. In the context of the greenhouse gasses, the method might be particularly adequate to improve resolution of these activity data which tend to be spatially correlated. The plausible sectors include agriculture, transportation and forestry. Improved resolution may in turn contribute to reduction in uncertainties underlying GHG inventories.

2 Disaggregation framework

This section presents the statistical approach to the issue of spatial disaggregation. We have available data on a spatially distributed variable (inventory of emissions) integrated in a coarse grid. The aim is to estimate the distribution of this variable in a fine grid, conditional on some

explanatory variables observable in this grid. It is assumed that the variable of interest is spatially correlated. Its residual covariance structure is set and the conditional autoregressive model is applied. An additional important assumption of the method is that the covariance structure of the variable in a coarse grid is the same as that in a fine grid.

Below we specify the model and provide details on its estimation in the coarse grid as well as on prediction in the fine grid.

2.1 Model

Fine grid We begin with the model specification in a fine grid. Let Y_i denote a random variable associated with a missing value of interest y_i defined at each cell i for $i=1, \dots, n$ of a fine grid (n denotes the overall number of cells in a fine grid). Assume that each random variable Y_i follows the Gaussian distribution with the mean μ_i and variance σ_Y^2

$$Y_i | \mu_i \sim \text{Gau}(\mu_i, \sigma_Y^2). \tag{1}$$

Given the values μ_i and σ_Y^2 , the random variables Y_i are assumed independent, thus the joint distribution of $\mathbf{Y}=(Y_1, \dots, Y_n)^T$ conditional on the mean process $\boldsymbol{\mu}=(\mu_1, \dots, \mu_n)^T$ is the Gaussian

$$\mathbf{Y} | \boldsymbol{\mu} \sim \text{Gau}_n(\boldsymbol{\mu}, \sigma_Y^2 \mathbf{I}_n), \tag{2}$$

where \mathbf{I}_n is the $n \times n$ identity matrix; the superscript T stands for the transpose.

The mean $\boldsymbol{\mu}$ represents the true process underlying emissions, and the (missing) observations are related to this process through a measurement error with the variance σ_Y^2 . The model for the mean process is formulated as a sum of the regression component with available covariates, and a spatially varying random effect. For this, the conditional autoregressive model is used. The CAR model is given through the specification of the full conditional distribution functions of μ_i for $i=1, \dots, n$ (Cressie 1993; Banerjee et al. 2004)

$$\mu_i | \boldsymbol{\mu}_{-i} \sim \text{Gau} \left(\mathbf{x}_i^T \boldsymbol{\beta} + \rho \sum_{\substack{j=1 \\ j \neq i}}^n \frac{w_{ij}}{w_{i+}} (\mu_j - \mathbf{x}_j^T \boldsymbol{\beta}), \frac{\tau^2}{w_{i+}} \right), \tag{3}$$

where $\boldsymbol{\mu}_{-i}$ denotes all elements in $\boldsymbol{\mu}$ but μ_i , w_{ij} are the adjacency weights ($w_{ij}=1$ if j is a neighbour of i and 0 otherwise, also $w_{ii}=0$); $w_{i+}=\sum_j w_{ij}$ is the number of neighbours of an area i ; \mathbf{x}_i is a vector containing 1 as its first element (for the intercept β_0) and k explanatory covariates of an area i as the next elements; $\boldsymbol{\beta}=(\beta_0, \beta_1, \dots, \beta_k)^T$ is a vector of regression coefficients. For calculation of the adjacency weights we use the Queen Method, i.e. two cells are considered neighbours if they share a side or a vertex. The CAR structure follows an assumption of similar random effects in adjacent cells; this is reflected in the second summand of the conditional expected value of μ_i , which is proportional to the average values of remainders $\mu_j - \mathbf{x}_j^T \boldsymbol{\beta}$ for neighbouring sites (i.e. when $w_{ij}=1$). This proportion is calibrated with the parameter ρ . Thus ρ reflects the strength of spatial association. The variance of the full conditional distribution of μ_i is inversely proportional to the number of neighbours w_{i+} , and τ^2 is a variance parameter.

Given (3), the joint probability distribution of the process μ is as follows, see e.g. Banerjee et al. (2004)

$$\mu \sim \text{Gau}_n(\mathbf{X}\beta, \tau^2(\mathbf{D} - \rho\mathbf{W})^{-1}), \tag{4}$$

where \mathbf{X} is the matrix whose rows are the vectors \mathbf{x}_i^T

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix};$$

\mathbf{D} is an $n \times n$ diagonal matrix with w_{i+} on the diagonal; and \mathbf{W} is an $n \times n$ matrix with adjacency weights w_{ij} . Equivalently we can write (4) as

$$\mu = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \text{Gau}_n(\mathbf{0}, \Omega), \tag{5}$$

where $\Omega = \tau^2(\mathbf{D} - \rho\mathbf{W})^{-1}$.

Coarse grid The model for a coarse grid (aggregated) observed data is obtained by multiplication of (5) with the $N \times n$ aggregation matrix \mathbf{C} consisting of 0's and 1's, indicating which cells have to be aggregated together

$$\mathbf{C}\mu = \mathbf{C}\mathbf{X}\beta + \mathbf{C}\varepsilon \quad \mathbf{C}\varepsilon \sim \text{Gau}_N(\mathbf{0}, \mathbf{C}\Omega\mathbf{C}^T) \tag{6}$$

where N is the number of observations in a coarse grid. Now, suppose that the random variable $\lambda = \mathbf{C}\mu$ is the mean process for random variables $\mathbf{Z} = (Z_1, \dots, Z_N)^T$ associated with observations $\mathbf{z} = (z_1, \dots, z_N)^T$ of the aggregated model

$$\mathbf{Z} | \lambda \sim \text{Gau}_N(\lambda, \sigma_Z^2 \mathbf{I}_N). \tag{7}$$

Thus, random variables $Z_i, i=1, \dots, N$ are conditionally independent

$$Z_i | \lambda_i \sim \text{Gau}(\lambda_i, \sigma_Z^2) \tag{8}$$

where λ_i is the i -th element of the vector λ .

2.2 Estimation and prediction

Having available observations of Z_i in the coarse grid, we can estimate parameters $\beta, \sigma_Z^2, \tau^2$ and ρ with the maximum likelihood (ML) method. First, from (6) and (7) the joint unconditional distribution of \mathbf{Z} is derived

$$\mathbf{Z} \sim \text{Gau}_N(\mathbf{C}\mathbf{X}\beta, \mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T), \tag{9}$$

where $\mathbf{M} = \sigma_Z^2 \mathbf{I}_N, \mathbf{I}_N$ is the $N \times N$ identity matrix; see e.g. Lindley and Smith (1972). Next, the log likelihood function associated with (9) is formulated

$$L(\beta, \sigma_Z^2, \tau^2, \rho) = -\frac{1}{2} \log |\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T| - \frac{N}{2} \log(2\pi) - \frac{1}{2} (\mathbf{z} - \mathbf{C}\mathbf{X}\beta)^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} (\mathbf{z} - \mathbf{C}\mathbf{X}\beta),$$

where $|\cdot|$ denotes the determinant. With fixed σ_z^2 , τ^2 and ρ , the above log likelihood is maximised for

$$\beta(\sigma_z^2, \tau^2, \rho) = \left[(\mathbf{CX})^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \mathbf{CX} \right]^{-1} (\mathbf{CX})^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \mathbf{z},$$

which substituted back into the function $L(\beta, \sigma_z^2, \tau^2, \rho)$ provides the profile log likelihood

$$\begin{aligned} L(\sigma_z^2, \tau^2, \rho) &= -\frac{1}{2} \log |\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T| - \frac{N}{2} \log(2\pi) \\ &\quad - \frac{1}{2} \left[\mathbf{z} - \mathbf{CX} \left[(\mathbf{CX})^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \mathbf{CX} \right]^{-1} (\mathbf{CX})^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \mathbf{z} \right]^T \\ &\quad \times (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \times \left[\mathbf{z} - \mathbf{CX} \left[(\mathbf{CX})^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \mathbf{CX} \right]^{-1} (\mathbf{CX})^T (\mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T)^{-1} \mathbf{z} \right]. \end{aligned}$$

Further maximisation of $L(\sigma_z^2, \tau^2, \rho)$ is performed numerically, including checks on ρ to ensure that the matrix $\mathbf{D} - \rho\mathbf{W}$ is non-singular, see Banerjee et al. (2004).

To obtain the standard errors of the estimated parameters, one needs to derive the Fisher information matrix. The asymptotic variance-covariance matrix of the ML estimators is obtained by inverting the expectation of the negative of the second derivatives (the Hessian) of the log likelihood function, and the expectation is evaluated at the ML estimates. In other words, the expected Fisher information matrix is used to obtain the standard errors of parameters. Calculation of the Hessian with respect to the regression coefficients is relatively straightforward, but it becomes more burdensome for the covariance parameters. A detailed derivation of the explicit formulas for the expected Fisher information matrix will be provided elsewhere; here we report the standard errors of the parameter estimators obtained in the case study.

To estimate the required values in a fine grid, the following prediction procedure is applied. Note that our primary interest is the underlying emission inventory process μ . The predictors optimal in the minimum mean squared error sense are given by $E(\mu|z)$. The joint distribution of (μ, \mathbf{z}) is given by

$$\begin{bmatrix} \mu \\ \mathbf{z} \end{bmatrix} \sim \text{Gau}_{N+n} \left(\begin{bmatrix} \mathbf{X}\beta \\ \mathbf{CX}\beta \end{bmatrix}, \begin{bmatrix} \Omega & \Omega \mathbf{C}^T \\ \mathbf{C}\Omega & \mathbf{M} + \mathbf{C}\Omega\mathbf{C}^T \end{bmatrix} \right). \tag{10}$$

The distribution (10) allows for full inference, yielding both the predictor $\hat{\mu} = \hat{E}(\mu|z)$ and its error $\hat{\sigma}_\mu^2 = \text{Var}(\mu|z)$

$$\hat{\mu} = \mathbf{X}\hat{\beta} + \hat{\Omega}\mathbf{C}^T (\hat{\mathbf{M}} + \mathbf{C}\hat{\Omega}\mathbf{C}^T)^{-1} [\mathbf{z} - \mathbf{CX}\hat{\beta}] \tag{11}$$

$$\hat{\sigma}_\mu^2 = \hat{\Omega} - \hat{\Omega}\mathbf{C}^T (\hat{\mathbf{M}} + \mathbf{C}\hat{\Omega}\mathbf{C}^T)^{-1} \mathbf{C}\hat{\Omega}, \tag{12}$$

where $\hat{\cdot}$ denotes the estimated values.

3 Case study

3.1 Data

The proposed procedure is illustrated using a real dataset of gridded inventory of NH_3 (ammonia) emissions from fertilization (in tonnes per year) reported in the northern region of Poland (the Pomorskie voivodship). The inventory grid cells are of a regular size $5 \text{ km} \times$

5 km, and the whole of cadastral survey compiles $n=800$ cells, denoted $y=(y_1, \dots, y_n)^T$, see Fig. 1. For explanatory information we use the CORINE Land Cover Map for this region, available at the European Environment (2010). Specifically, for each grid cell we calculate the area of these land use classes, which are related to ammonia emissions. The following CORINE classes were considered (the CORINE class numbers are given in brackets):

- Non-irrigated arable land (211), denoted $x_1=(x_{1,1}, \dots, x_{n,1})^T$;
- Fruit tree and berry plantations (222), denoted $x_2=(x_{1,2}, \dots, x_{n,2})^T$;
- Pastures (231), denoted $x_3=(x_{1,3}, \dots, x_{n,3})^T$;
- Complex cultivation patterns (242), denoted $x_4=(x_{1,4}, \dots, x_{n,4})^T$;
- Principally agriculture, with natural vegetation (243), denoted $x_5=(x_{1,5}, \dots, x_{n,5})^T$.

Performance of the proposed disaggregation framework depends on a few factors. Perhaps the most crucial ones are the following two: (i) explanatory power of covariates available in the fine grid, and (ii) an extent of disaggregation, which is connected with preservation of the spatial correlation. The impact of both these features will be evaluated in our case study.

Regarding the first factor, we will examine models with all the above land use classes (set 1), and compare the results with models including only two of them: non-irrigated arable land and complex cultivation patterns (set 2). This subset of land use classes was chosen on the basis of its explanatory power. When limiting the number of explanatory variables, these two covariates provided the best results. Secondly, we compare linear regression with independent (iid) errors versus spatially correlated errors modelled by the CAR process. We consider the following models:

- Model CAR1: - CAR errors, set 1 of covariates;
- Model LM1: - iid errors, set 1 of covariates;
- Model CAR2: - CAR errors, set 2 of covariates;
- Model LM2: - iid errors, set 2 of covariates.

This setting of four models is intended to enable the analysis of extent to which a limited number of explanatory information can be compensated by spatial modelling.

Regarding the second factor, we test the disaggregation from 10×10 km to 15×15 km (coarse) grids into a $5 \text{ km} \times 5 \text{ km}$ (fine) grid. To examine performance of the disaggregation procedure, first the original fine grid emissions are aggregated into respective coarse grid cells. Next, the proposed model is fitted and ammonia emissions are predicted for a $5 \text{ km} \times 5 \text{ km}$ (fine) grid. Finally, the obtained results are checked with the original inventory emissions of a $5 \text{ km} \times 5 \text{ km}$ (fine) grid. Thus, our simulation study tests the cases of a fourfold and ninefold

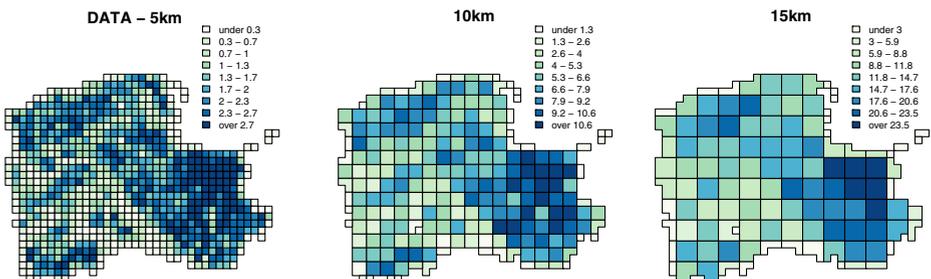


Fig. 1 Ammonia emissions: inventory data in 5 km grid, and aggregated values in 10 km and 15 km grids

disaggregation. The aggregated values of the two coarse grids as well as the actual inventory data in the fine grid are shown in Fig. 1.

3.2 Results of disaggregation from the 10 km grid

This subsection presents the model testing results for disaggregation from the 10 km grid. Table 1 (the upper part) displays the maximum likelihood estimates (denoted by Est.) and standard errors (denoted by Std.Err.) of the parameter estimators for each model. Note that in the models with set 1 of covariates (CAR1, LM1) the regression coefficient β_0 was dropped as it was statistically insignificant. In the table, we can observe that the ML estimates of the regression coefficients are similar for all the models. From the ratio of regression coefficients and its respective standard errors (i.e. the t -test statistic), we can roughly conclude that all the considered land use classes are statistically significant; in fact, in each case respective p -values proved to be less than 0.05 (not shown). Next, let us turn our attention to the error part of the models. Significantly lower values of σ_Z^2 estimates under both the CAR models, as compared with their linear regression counterparts, indicate that greater variability is explained by the models with spatially correlated errors than by the corresponding models with independent errors. As expected, among the spatially correlated models, both variance parameters σ_Z^2 and τ^2 are higher for CAR2 than for CAR1 model with five land use classes as explanatory variables. Furthermore, the parameter ρ reflects strength of the spatial correlation. Note that $\rho=0$ corresponds to a model with independent errors, see also Banerjee et al. (2004) for more details. A value of parameter ρ is higher for CAR2 model, which illustrates that in the models of limited explanatory power, the importance of spatial correlation becomes more pronounced.

The results of the four models are also summarized using the Akaike criterion (AIC). The idea of AIC is to favour a model with a good fit and to penalize it for a number of parameters; models with smaller AIC are preferred to models with larger AIC. Table 2 (the upper part) displays AIC for each model, and additionally it reports the negative log likelihood (-L). Naturally, the models with set 1 of covariates provide much better results than the models with another set. Among these respective sets, the models with the spatial structure considerably improve results obtained with the models of independent errors. Note, that this improvement is higher for the models with set 2 of covariates ($797.6-742.8=54.8$) than for the models with set 1 of covariates ($685.1-640.7=44.4$).

The values of ammonia emissions predicted in a $5\text{ km} \times 5\text{ km}$ grid (y_i^*) are featured in Fig. 2. Differences between the four models are negligible, although a visual comparison with the original emissions in Fig. 1 (the left-hand-side plot) suggests that the both models based on set 1 of covariates (CAR1, LM1) provide slightly better results. Since the mapped emission values are classified into just 9 bins, therefore some features might not be easily distinguishable on the maps in Fig. 2. To remedy this, Fig. 3 presents the model residuals ($d_i=y_i-y_i^*$). Now the difference in prediction results among the models is evident—the best results are obtained for CAR1 model and the worst for LM2 model.

At this point it must be stressed that the values predicted in a fine grid (y_i^*) are calculated with the formula (11) based on the aggregated values of 10 km grid; the calculations are made as if the true emissions were unknown. On the other hand, recall that these true emissions in the fine grid (y_i) are available; see the left-hand-side map in Fig. 1. From now on, our analysis is based on a comparison between the prediction results obtained with the proposed technique and the original fine grid ammonia emissions (observations).

Figure 4 presents, for each model, a scatterplot of predicted values y_i^* versus observations y_i . The straight line has slope 1, thus if the predicted values are close to the original data, points

Table 1 Maximum likelihood estimates

	CAR1		LM1		CAR2		LM2	
	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.
10 km grid								
β_0	–	–	–	–	0.386	9.29e-02	0.452	5.45e-02
β_1	1.13e-07	3.26e-09	1.09e-07	2.46e-09	1.06e-07	5.03e-09	9.58e-08	4.43e-09
β_2	2.56e-07	1.94e-07	4.48e-07	1.97e-07	–	–	–	–
β_3	9.77e-08	1.19e-08	1.08e-07	1.08e-08	–	–	–	–
β_4	1.18e-07	2.13e-08	1.21e-07	1.76e-08	1.27e-07	2.72e-08	1.60e-07	2.22e-08
β_5	1.27e-07	1.32e-08	1.35e-07	1.11e-08	–	–	–	–
σ_Z^2	0.334	0.073	1.165	0.109	0.522	0.111	1.95	0.184
τ^2	0.536	0.082	–	–	0.807	0.124	–	–
ρ	0.948	9.98e-04	–	–	0.972	9.98e-04	–	–
15 km grid								
β_0	–	–	–	–	0.424	1.04e-01	0.476	6.82e-02
β_1	1.12e-07	3.95e-09	1.09e-07	3.42e-09	1.00e-07	7.01e-09	9.35e-08	5.79e-09
β_2	–	–	–	–	–	–	–	–
β_3	1.07e-07	1.84e-08	1.16e-07	1.55e-08	–	–	–	–
β_4	1.24e-07	2.77e-08	1.29e-07	2.34e-08	1.56e-07	3.65e-08	1.75e-07	2.79e-08
β_5	1.27e-07	1.65e-08	1.33e-07	1.49e-08	–	–	–	–
σ_Z^2	2.339	0.424	3.50	0.474	2.681	0.548	5.55	0.753
τ^2	0.214	0.088	–	–	0.414	0.088	–	–
ρ	0.966	4.91e-04	–	–	0.982	5.55e-05	–	–

are close to the straight line. This setting, once again, illustrates much better explanatory power of models based on all the land use classes (set 1 of covariates). It also illustrates the importance of the spatial structure component. In the case of models CAR2 and LM2, the introduction of spatial dependence has evidently improved the accuracy of prediction. Whereas in the case of models CAR1 and LM1, the applied spatial structure considerably limited a number of highly overestimated predictions (points below the straight line).

Table 2 Model comparison and analysis of residuals ($d_i = y_i - y_i^*$)

Model	-L	AIC	MSE	min(d_i)	max(d_i)	r
10 km grid						
CAR1	312.3	640.7	0.064	-1.717	1.104	0.961
LM1	336.5	685.1	0.186	-2.544	0.268	0.882
CAR2	365.4	742.8	0.158	-1.917	1.362	0.901
LM2	394.8	797.6	0.291	-2.498	1.765	0.808
15 km grid						
CAR1	220.6	455.3	0.136	-2.428	0.646	0.915
LM1	222.9	455.9	0.189	-2.600	0.516	0.880
CAR2	240.4	492.8	0.190	-2.132	1.446	0.880
LM2	248.1	504.4	0.295	-2.511	1.746	0.807

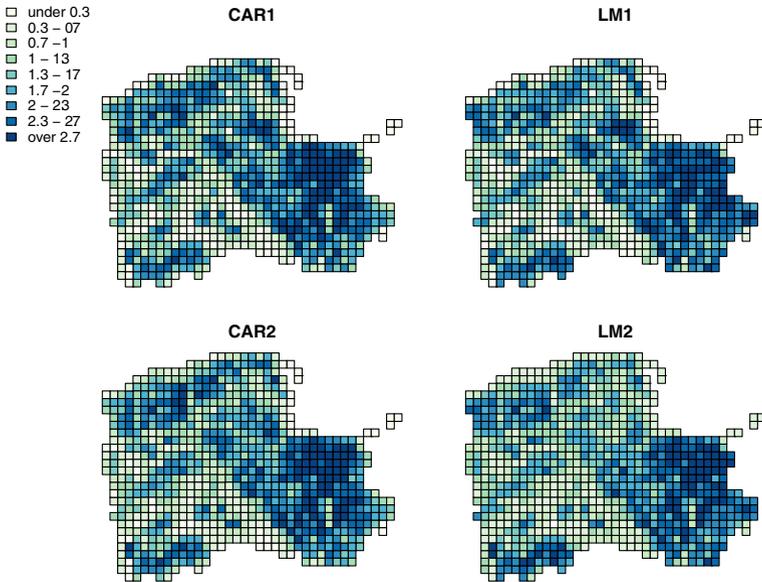


Fig. 2 Ammonia emissions predicted in a fine grid—disaggregation from 10 km grid

Furthermore, we note that for a prevailing number of cases in the high emission range (emissions over 1.5 tonnes) the linear regression LM1 provides biased (underestimated) predictions, while CAR1 model allows overcoming this deficiency. This is due to the fact that the analysed emissions are spatially correlated, that is, cells located nearby tend to have

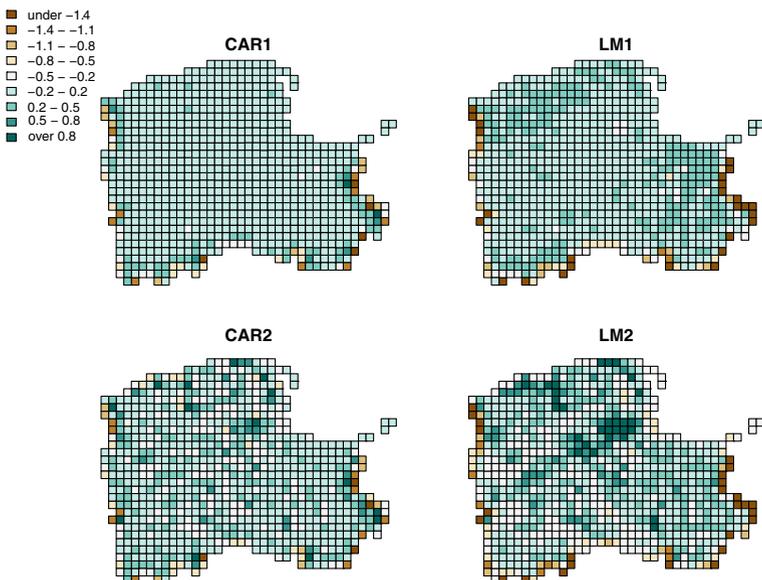


Fig. 3 Residuals from predicted values—disaggregation from 10 km grid

similar values. In particular, the majority of high emission values are located in the eastern part of the voivodship as well as in the north-west stripe along the coastline (compare Fig. 1). The covariates applied in the linear regression LM1 explain emission variability only to some extent, and the point is that the unexplained variability remains spatially correlated. This can be noticed on the map in Fig. 3 for LM1 model, where clusters of residual values (0.2–0.5) in the mentioned areas indicate underestimated predictions. The autocorrelation term in the model CAR1 allows for this feature. In Fig. 4 it can be seen as a slope of a dotted line, which is visibly higher than 1 for LM1 model, while for CAR1 it lines up with the one of slope 1.

The residuals d_i are further analysed in Table 2 (the upper part). Namely, the mean squared error (MSE) is calculated

$$MSE = \frac{1}{n} \sum_i (y_i - y_i^*)^2,$$

and it should be as low as possible. The mean squared error reflects how well a model predicts data. In Table 2 we report also the minimum and maximum values of d_i , and the sample correlation coefficient r between the predicted y_i^* and observed y_i values. In terms of both the mean squared error and the coefficient r , the best model is CAR1 and the poorest one is LM2, following the previous assessments. Interestingly, the remaining two models changed their ranks compared with the AIC criterion. That is, CAR2 model has lower $MSE=0.158$ and higher coefficient $r=0.901$ than the linear model based on set 1 of covariates (LM1 model with $MSE=0.186$ and $r=0.882$). This proves that the model with a limited number of covariates but having a spatial component (CAR2) can provide better disaggregation results than the model based solely on linear regression, even though its covariate information is richer (LM1). Note that the analysis based on residuals is more robust than the AIC rating, which basically tests a model fit to the aggregated data.

Following the formula (12), we also calculate the prediction error. Since in the present case study the correct values of predicted emissions are known, we are in a position to compare the

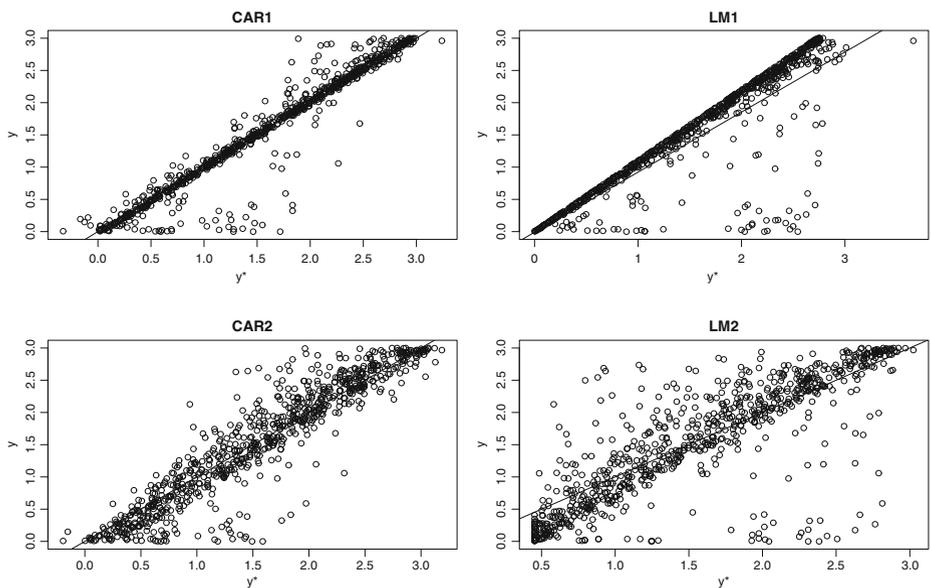


Fig. 4 Predicted (y^*) versus observed (y) values—disaggregation from 10 km grid

prediction error with the actual residuals (more precisely, with its absolute values). In Fig. 5 these values are presented for CAR2 model. It is noticeable that the prediction error is significantly underestimated, and moreover, it does not reflect the diversification of the actual residuals properly. Note that in the both maps the highest errors are reported on the border of the domain; this fact is known in spatial modelling as the edge effect.

3.3 Results of disaggregation from the 15 km grid

Next, we present the results of disaggregation from the 15 km grid. The conducted analysis is similar to the one of the 10 km grid and, where appropriate, both settings are compared.

The lower part of Table 1 contains the maximum likelihood estimates for the 15 km grid data. In the models with set 1 of covariates, the regression coefficient β_0 was again dropped. Moreover, in all the models at this level of aggregation the land use class “Fruit tree and berry plantations” (β_2) was statistically insignificant, and thus it was also dropped. The remaining land use classes were informative, with respective p -values lower than 0.05.

As regards the error part, all the comments reported for 10 km disaggregation remain valid also here, although their degree is significantly lower. Both CAR models provide lower values of σ_2^2 than their linear regression counterparts. However, the reduction of unexplained variability between the models, for instance, LM1 and CAR1 is only 1.5 (3.5/2.339), while it was over 3 (1.165/0.334) for respective models of 10 km disaggregation. This suggests that the spatial correlation strength of the 15 km grid model is smaller than the 10 km grid one. Thus, here the CAR models are less competitive than the LM models, as compared to the former grid.

The values of AIC criterion and of the negative log likelihood (-L) are reported in the lower part of Table 2. Similarly as for the disaggregation from a 10 km grid, also in this case the models based on set 1 of covariates provide better results. The CAR structure improves obtained linear regression results of both respective covariate sets. Note, however, that in the setting of 15 km disaggregation, the impact of the spatial component is not that substantial anymore as it was previously. Again, a bigger improvement is noted for the models with a limited number of covariates (504.4–492.8=11.6 in terms of the AIC criterion), and the gain from incorporation of the spatial component is only marginal for the models with set 1 of covariates (455.9–455.3=0.6).

For the four considered models, the maps of ammonia emissions disaggregated from the 15 km grid and predicted in the fine grid provided visually similar results (not shown). The residual maps proved to be more informative, see Fig. 6. While for the 10 km disaggregation

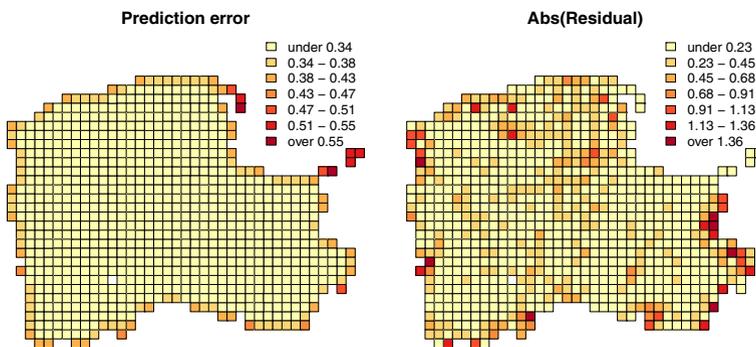


Fig. 5 Prediction error and absolute values of residuals for CAR2 model. Note that the maps are drawn in different scales

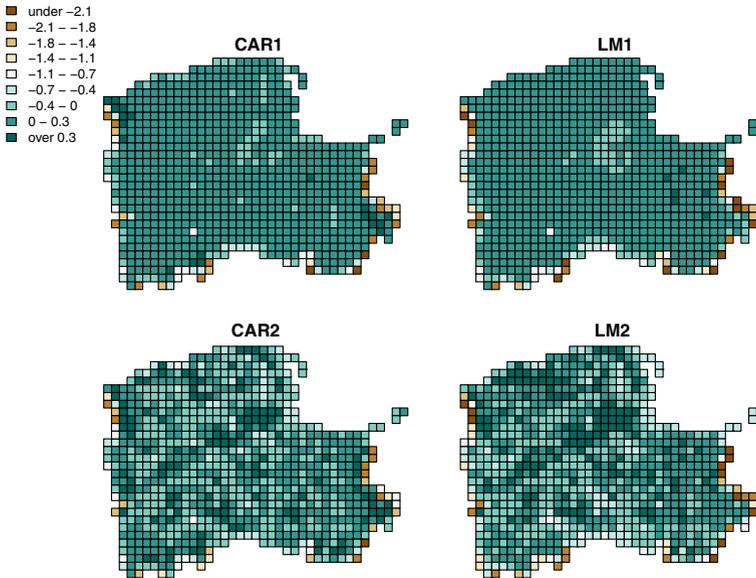


Fig. 6 Residuals from predicted values—disaggregation from 15 km grid

the residual maps clearly indicated discrepancies among the models, here it is not easily visible. The models based on set 1 of covariates (CAR1, LM1) provide smaller residuals. However, the differences between the spatial models and their linear regression counterparts seem to be negligible.

Again, Table 2 (the lower part) provides further analysis of residuals. The mean squared error MSE and the correlation coefficient r yield a consistent ranking of the models. Obviously the best model is CAR1 with $r=0.915$ and $MSE=0.136$, while the poorest one is LM2 with $r=0.807$ and $MSE=0.295$. When it comes to the remaining two models, LM1 slightly outperforms CAR2 (in terms of the mean squared error). Note that this order is reversed when compared with the results of the 10 km grid disaggregation (the upper part of the table). Therefore, when disaggregating from the 10 km grid, the spatial structure is more informative than some of the covariates, but this is not true anymore when disaggregating from the 15 km grid. From this we conclude that in this particular case study, the proposed framework offers an efficient tool for a quadruple and nine-times disaggregation, but it may become less adequate for higher order allocations.

The actual interplay among the four models is illustrated on the scatterplots in Fig. 7. In general, the 15 km disaggregation preserves the features reported previously—the performance of respective models is analogous as for the 10 km disaggregation. It means that for the models based on set 2 of covariates, the spatial correlation significantly improves prediction quality. Also for the other two models, the introduction of spatial structure is still beneficial as it allows correction of the prediction bias and a slight reduction in the number of overestimates. We highlight the difference between the models CAR2 and LM1 that yield almost the same MSE and coefficient r , but provide completely distinct plots, see Fig. 7. The residuals of CAR2 model are more dispersed owing to a limited set of explanatory covariates. On the other hand, improved covariate modelling of LM1 leads to the residuals gathered close to the diagonal, but a lack of spatial averaging results in larger amount of overestimated values. Altogether, the assessment of residuals for both models becomes the same.

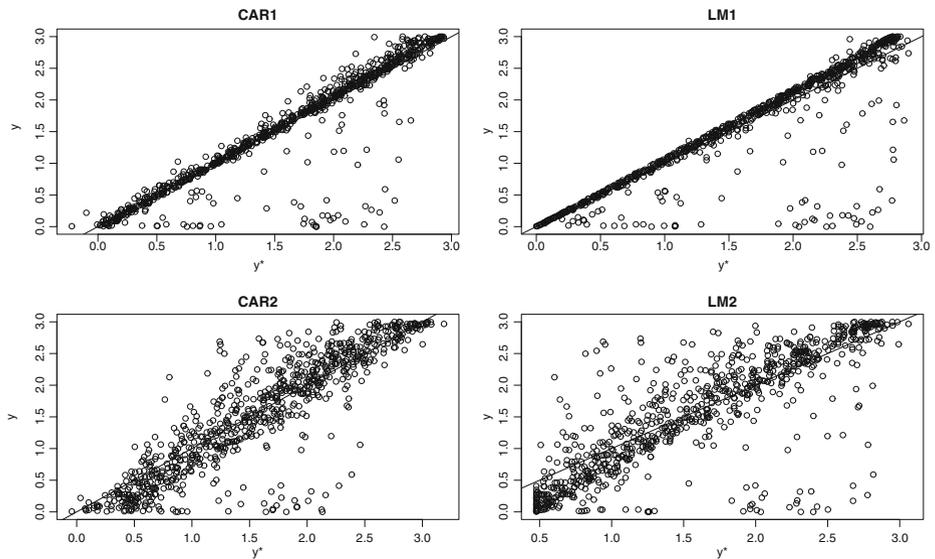


Fig. 7 Predicted (\hat{y}) versus observed (y) values—disaggregation from 15 km grid

4 Discussion and conclusions

The major objective of this study was to demonstrate how a variable of interest (here, emissions) available in a coarse grid plus information on covariates available in a finer grid can be combined together to provide the variable of interest in a finer grid, and therefore to improve its spatial resolution. We proposed a relevant disaggregation model and illustrated the approach using a real dataset of ammonia emission inventory. The idea is conceptually similar to the method of Chow and Lin (1971), originally designed for time series data; see also Polasek et al. (2010). It was applied to the spatially correlated data, and spatial dependence was modelled with the conditional autoregressive structure introduced into a linear model as a random effect.

The model allows for this part of a spatial variation which has not been explained by available covariates. Thus, if the covariate information does not correctly reflect a spatial distribution of a variable of interest, there is potential for improving the approach with a relevant model of a spatial correlation. The underlying assumption of the method is that the covariance structures of the variable in a coarse grid and in a fine grid are the same. In the present study of ammonia emissions examined in 5 km, 10 km, and 15 km grids, this assumption proved to be reasonable.

Performance of the proposed framework was evaluated with respect to the following two factors: explanatory power of covariates available in a fine grid, and the extent of disaggregation. The results indicate that inclusion of a spatial dependence structure can compensate for less adequate covariate information. For the considered ammonia inventory, the fourfold allocation benefited greatly from the incorporation of the spatial component, while for the ninefold allocation this advantage was limited, but still evident. In addition, the proposed method allowed to correct the prediction bias encountered for upper range emissions in the linear regression models.

We note that in this case study we used the original data in a fine grid to assess the quality of resulting predictions. For the purpose of potential applications, we developed also a relevant measure of prediction error (the formula 12). Although not entirely faultless, it is the first attempt to quantify the prediction error in situations, where original emissions in a fine grid are not known.

Other approaches, such as a geostatistical model, might be potentially used in the case of spatial allocation. Application of the geostatistical approach brings us to the concept of block kriging (Gelfand 2010). However, it should be stressed that geostatistics is more appropriate for point referenced data, while our proposition is dedicated to the case of emission inventories which involve a real data. Thus, the choice between these two options should be considered on a case by case basis.

Another possibility to deal with the issue of spatial disaggregation could be to use some expert knowledge and logical inference; compare Verstraete (this issue) for a fuzzy inference system to the map overlay problem.

The described method opens the way to uncertainty reduction of spatially explicit emission inventories, hence the future work will also include testing the proposed disaggregation framework for inventories of greenhouse gasses.

Acknowledgments The study was conducted within the 7FP Marie Curie Actions IRSES project No. 247645. J. Horabik acknowledges support from the Polish Ministry of Science and Higher Education within the funds for statutory works of young scientists. This contribution is also supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing; project financed from The European Union within the Innovative Economy Operational Programme 2007–2013 and European Regional Development Fund. Z. Nahorski was financially supported by the statutory funds of the Systems Research Institute of Polish Academy of Sciences.

The authors gratefully acknowledge the provision of data for the case study from Ekometria – Biuro Studiów i Pomiarów Proekologicznych in Gdańsk, Poland.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, Boca Raton
- Boychuk K, Bun R (this issue) Regional spatial cadastres of GHG emissions in energy sector: accounting for uncertainty
- Bun R, Hamal K, Gusti M, Bun A (2010) Spatial GHG inventory at the regional level: accounting for uncertainty. *Clim Chang* 103(1–2):227–244
- Chow GC, Lin A (1971) Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *Rev Econ Stat* 53(4):372–375
- Cressie NAC (1993) Statistics for spatial data. John Wiley & Sons, New York
- Dragosits U, Sutton MA, Place CJ, Bayley AA (1998) Modelling the spatial distribution of agricultural ammonia emissions in the UK. *Environ Pollut* 102(S1):195–203
- European Environment Agency (2010) Corine land cover 2000. <http://www.eea.europa.eu/data-and-maps/data>. Cited August 2010
- Gelfand AE (2010) Misaligned spatial data: the change of support problem. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P (eds) Handbook of spatial statistics. Chapman & Hall/CRC, Boca Raton
- Ghosh T, Elvidge CD, Sutton PC et al (2010) Creating a global grid of distributed fossil fuel CO₂ emissions from nighttime satellite imagery. *Energies* 3:1895–1913
- Gotway CA, Young LJ (2002) Combining incompatible spatial data. *J Am Stat Assoc* 97:632–648
- Horabik J, Nahorski Z (2010) A statistical model for spatial inventory data: a case study of N₂O emissions in municipalities of southern Norway. *Clim Chang* 103(1–2):263–276
- 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
- Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. *J Roy Stat Soc B* 34:1–41
- McMillan AS, Holland DM, Morara M, Fend J (2010) Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* 21:48–65

- Mugglin AS, Carlin BP (1998) Hierarchical modeling in geographical information systems: population interpolation over incompatible zones. *J Agric Biol Environ Stat* 3:111–130
- Mugglin AS, Carlin BP, Gelfand AE (2000) Fully model-based approaches for spatially misaligned data. *J Am Stat Assoc* 95:877–887
- Oda T, Maksyutov S (2011) A very high-resolution (1 km×1km) global fossil fuel CO₂ emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmos Chem Phys* 11: 543–556
- Polasek W, Llano C, Sellner R (2010) Bayesian methods for completing data in spatial models. *Rev Econ Anal* 2: 194–214
- Thiruchittampalam B, Theloke J, Uzbisich M et al. (2010) Analysis and comparison of uncertainty assessment methodologies for high resolution Greenhouse Gas emission models. In: *Proceedings of the 3rd International Workshop on Uncertainty in Greenhouse Gas Inventories*, Lviv Polytechnic National University, Ukraine, 22–24 Sept 2010
- Verstraete J (this issue) Solving the map overlay problem with a fuzzy approach