Raport Badawczy

Research Report

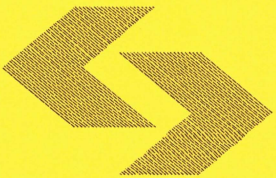
RB/8/2014

Methodology for spatial scaling of GHG activity data

Z. Nahorski, J. Horabik

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 Zakładu zgłaszający pracę:
Prof. dr hab. inż. Zbigniew Nahorski

Warszawa 2014

D 3.2

Version	1
Date	20.06.2014
Author	SRI
Dissemination level	PP
Document reference	D 3.2

GESAPU

Geoinformation technologies, spatio-temporal approaches, and full carbon account for improving accuracy of GHG inventories

Deliverable 3.2. Methodology for spatial scaling of GHG activity data

Zbigniew Nahorski, Joanna Horabik
Systems Research Institute, Polish Academy of Sciences, Poland

Delivery Date: M36

Project Duration	24 June 2010 – 23 June 2014 (48 Months)
Coordinator	Systems Research Institute of the Polish Academy of Sciences (SRI)
Work package leader	Systems Research Institute of the Polish Academy of Sciences (SRI)

Disclaimer

All rights reserved

This project is supported by funding by the European Commission: FP7-PEOPLE-2009-IRSES, Project n° 247645.

Project: #247645. Call: FP7-PEOPLE-2009-IRSES, Marie Curie Actions—International Research Staff Exchange Scheme (IRSES).

Work package 3. Improving accuracy of inventories by means of spatio-temporal statistical methods

Deliverable 3.2. Methodology for spatial scaling of GHG activity data

Content

List of figures

List of tables

1 Introduction

Greenhouse gas (GHG) emission inventories serve as a basic tool for verification of international treaties aimed at constraing global warming. Despite all their drawbacks and limitations [19], national GHG inventories provide invaluable information on anthropogenic emission sources, and, indirectly, on effectiveness of undertaken emission abatement measures. Constant efforts of IPCC community seek to improve the inventory procedure and to limit underlying uncertainties and imprecision [15].

Although the greenhouse gases directly are not harmful for human health, their spatial distribution is of great importance. For instance, a network of ecosystem long-term observation sites is launched across Europe to understand behavior of the global carbon cycle and greenhouse gas emissions. The activities are conducted within the Integrated Carbon Observation System infrastructure. Another approach is to develop a spatially resolved GHG inventory. All of these efforts open new opportunities for improvement of emission reduction activities, including among others attribution of sources and sinks.

The present study was conducted as a part of the 7FP Marie Curie Actions project *Geoinformation technologies, spatio-temporal approaches, and full carbon account for improving accuracy of GHG inventories*. One of the main aims of the project is to develop a spatial inventory of GHG for Poland. The task comprises estimation of GHG related activity data, which need to be spatially resolved in this case, and their corresponding emission factors. In terms of considered sectors, subsectors and separate emission source groups, the IPCC guidelines [13] provide relevant methodology, and it is followed throughout the project. The main GHG emission sectors include energy (fossil fuel burning from stationary and mobile sources), industry and agriculture.

Development of spatial GHG inventory crucially depends on availability of low resolution activity data. In Poland, relevant information needs to be acquired from national/regional totals. A procedure of allocation into smaller spatial units (like districts, municipalities and finally 2x2km grid) differs among various emission sectors. Basically, all the emission sources are categorised as line, area or large point emission sources; further steps differ significantly for each group. For large point sources, such as power/heat stations or refinery plants, corresponding emissions are associated directly with a particular object located in space. Line sources, like roads, railways or pipelines, are usually analyzed by cutting line objects into sections using respective grids. Area sources comprise e.g. agricultural fields, urban areas as well as highly dense urban transportation network. In this case, a procedure of spatial allocation depends on methods and technologies of fossil fuel combustion in a considered sector [2]. A common approach though is a spatial allocation made in a proportion to some related indicators, i.e. proxy data, which are available in a finer grid. This solution to a large extent relies on subjective assumptions, and usually there is no mean for verification of the results obtained.

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 [9, 8]. 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 [16, 17]

Within the Work Package 3 (Deliverable 3.2) of GESAPU project, the statistical scaling methods have been developed in order to support the procedure of compiling high resolution activity data. In this report we describe the original method for allocating GHG activity data to finer spatial scales conditional on covariate information, such as land use, observable in a fine grid. The proposition is suitable for spatially correlated, area emission sources.

The approach resembles to some extent the method of Chow and Lin (1971) [4], 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 area data, i.e. the conditional autoregressive (CAR) model. Although the CAR specification is typically used in epidemiology [1], it was also successfully applied for modelling air pollution over space [14], [20]. Compare also [11] 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. In addition, the standard errors of estimated parameters are provided, based on the Fisher information matrix. We demonstrate a usefulness of the disaggregation method for spatially correlated area sources, in particular for agricultural sector.

The proposed methodology, in its basic version, is described in section 3.1; this part has been already presented in [12], which is attached to this report as an Appendix II. In addition, the basic model is extended for the case of various regression models in each region (here, voivodeship); see section 3.2. The performance of the method for livestock data in agricultural sector of GHG inventory is presented and discussed in section 4. The considered case study shows an allocation from districts to municipalities (i.e. an irregular grid); the proposed approach is compared with the results of naive disaggregation in proportion to a single covariate (here, population density). For an application of the technique to a regular grid (fourfold and ninefold disaggregation), see Appendix II. It also presents the performance of the method evaluated with respect to explanatory power of covariates available in a fine grid, i.e. it is shown to which extent inclusion of a spatial correlation structure can compensate for less adequate covariate information.

2 The data set

Considered is a livestock dataset (cattle, pigs, horses, poultry, etc.) based on agricultural census 2010, and available from the Central Statistical Office of Poland - Local Data Bank [10]. The goal is to allocate relevant livestock amounts from district (powiaty) into municipality (gminy) levels.

In particular, for horses the data are available also in municipalities, and this fact enables verification of the proposed disaggregation method. Therefore, in what follows we consider the task of disaggregation of number of horses reported for 314 districts into 2171 municipalities, taking advantage of covariate information observable for municipalities,

compare Figure 1. Only rural municipalities are considered in the study, see Figure 2.

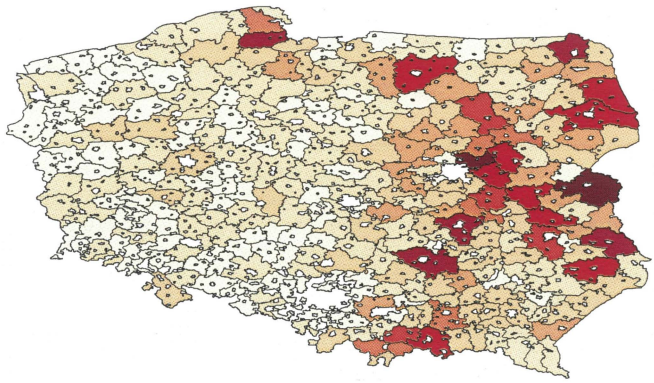


Figure 1: Livestock data (horses) available for districts

As explanatory variables we use population density (denoted $\boldsymbol{x}_1$) and land use information. For the latter, the CORINE Land Cover map, available from the European Environment Agency [7], was employed. For each rural municipality we calculate the area of agricultural classes, which may be related to livestock farming, see Figure 3. Three CORINE classes were considered (the CORINE class numbers are given in brackets):

- Arable land (2.1); denoted $\boldsymbol{x}_2$
- Pastures (2.3); denoted $\boldsymbol{x}_3$
- Heterogeneous agricultural areas (2.4), which include subclasses Complex cultivation patterns (2.4.2) and Land principally occupied by agriculture, with significant areas of natural vegetation (2.4.3); denoted $\boldsymbol{x}_4$.

The results of the disaggregation with the proposed procedure are further compared with the results of allocation proportional to population of municipalities. Here, we stress once more that only rural municipalities are considered in the study. Otherwise, allocation of number of horses in a proportion to population would be meaningless. This naive approach, however, gave rise for a modification of the basic version of the method. Namely, we account for the fact that a relationship of farmed livestock with available covariates is diversified across the country - we allow for various regression models for regions. In this particular case study, we treat 16 voivodeships (województwa) as regions. This extention of the model is described in section 3.2.

Figure 2: The net of rural municipalities

3 The disaggregation framework

3.1 The basic model

3.1.1 Model formulation

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 $i = 1, ..., n$ of a fine grid (n denotes the overall number of cells in a fine grid). Assume that each random variable Y_i follows Gaussian distribution with the mean μ_i and variance σ_Y^2

$$Y_i|\mu_i \sim \mathcal{N}\left(\mu_i, \sigma_Y^2\right). \tag{1}$$

Given the values μ_i and σ_Y^2, the random variables Y_i are assumed independent. The values $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^n$ represent the true process underlying distribution of activity data in our case study, and the (missing) observations are related to this process through a measurement error of variance σ_Y^2. The model for the underlying process $\boldsymbol{\mu}$ is formulated as a sum of regression component with available covariates, and a spatially varying random effect.

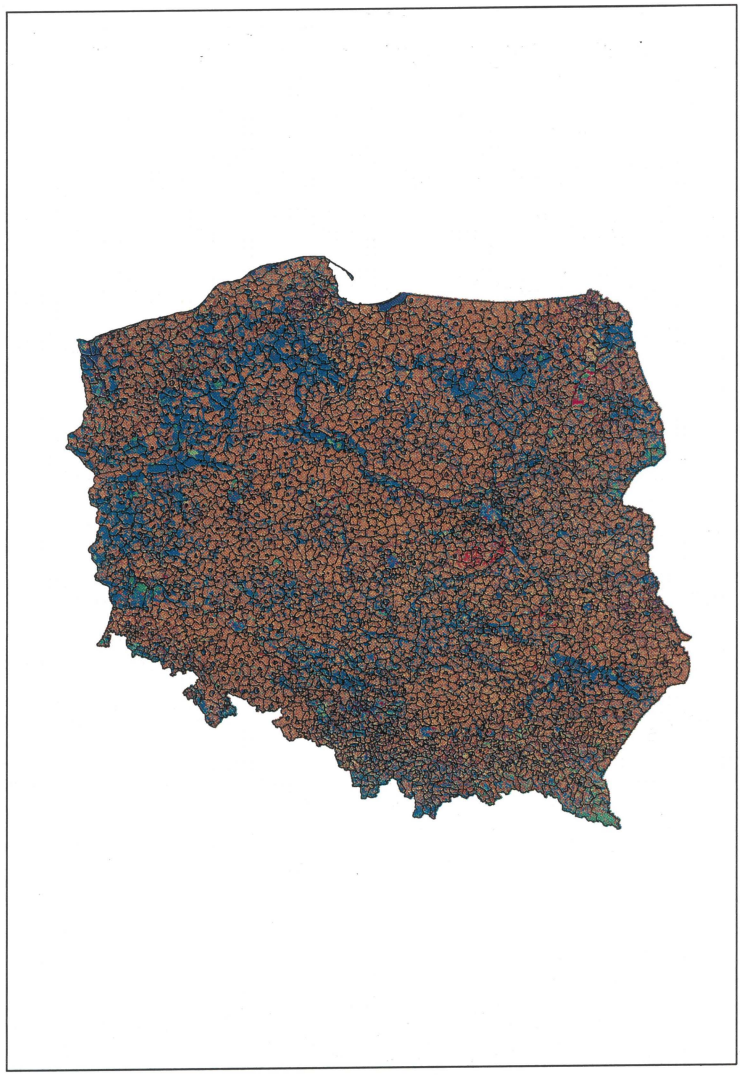


Figure 3: CORINE land use map of Poland, with the net of rural municipalities

Spatial correlation is modelled with the conditional autoregressive structure CAR. Following an assumption of similar random effects in adjacent cells, it is given through the specification of full conditional distribution functions [5], [8]

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

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 [5], [8]

$$\boldsymbol{\mu} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}\right), \tag{3}$$

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 (3) as $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{N})$, with $\boldsymbol{N} = \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}$.

Coarse grid. The model for the data observed at the district level is obtained by the multiplication of $\boldsymbol{\mu}$ with an $N \times n$ *aggregation matrix* $\boldsymbol{C}$, where N is a number of observations on the district level

$$\boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\epsilon},\quad \boldsymbol{C}\boldsymbol{\epsilon} \sim \mathcal{N}_N\left(\boldsymbol{0}, \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right). \tag{4}$$

The matrix $\boldsymbol{C}$ 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

$$\boldsymbol{Z}|\boldsymbol{\lambda} \sim \mathcal{N}_N\left(\boldsymbol{\lambda}, \sigma_Z^2\boldsymbol{I}_N\right). \tag{5}$$

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

3.1.2 Maximum likelihood estimation

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 $\boldsymbol{M} = \sigma_Z^2\boldsymbol{I}_N$.

Next, the log likelihood function associated with (5) is formulated

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma_Z^2, \tau^2, \rho\right) &= -\frac{1}{2}\log\left|\boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right| - \frac{N}{2}\log\left(2\pi\right) \\ &\quad -\frac{1}{2}\left(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T\right)^{-1}\left(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}\right), \end{aligned} \tag{6}$$

where $|\cdot|$ denotes the determinant. The analytical derivation is limited to the regression coefficients $\boldsymbol{\beta}$, and further maximisation of the profile log likelihood is performed numerically.

3.1.3 Evaluating the Fisher information matrix

The standard errors of parameter estimators are calculated with the Fisher information matrix. Let us denote with $\hat{\theta}$ a vector of maximum likelihood estimates for a model parametrized by θ with associated log likelihood $L(\theta)$. The Fisher information matrix is defined as

$$\boldsymbol{J}(\theta) = E\left[-\boldsymbol{L}^{(2)}(\theta)\right],$$

where E denotes the expected value, and $\boldsymbol{L}^{(2)}$ is the Hessian matrix of second order partial derivatives of the log likelihood function. Inverting this matrix gives the asymptotic covariance matrix of the maximum likelihood estimators, i.e. the Cramér-Rao (lower) bound.

To estimate $\boldsymbol{J}(\theta)$, either the *expected* or the *observed* Fisher information matrices can be used [3, 18]. The expected information matrix is defined as

$$\boldsymbol{\mathcal{J}}(\hat{\theta}) = \left[E\left(-\boldsymbol{L}^{(2)}(\theta)\right)\right]\Big|_{\theta=\hat{\theta}}$$

The observed information matrix is defined as minus the second derivative of the log likelihood function at $\hat{\theta}$ given data:

$$\boldsymbol{\mathcal{I}}(\hat{\theta}) = \left[-\boldsymbol{L}^{(2)}(\theta)\right]\Big|_{\theta=\hat{\theta}}$$

It should be noted that $\boldsymbol{\mathcal{J}}(\hat{\theta})$, unlike $\boldsymbol{\mathcal{I}}(\hat{\theta})$, is a maximum likelihood estimator of $\boldsymbol{J}(\theta)$. $\boldsymbol{\mathcal{I}}(\hat{\theta})$ is actually only an approximation of $\boldsymbol{J}(\theta)$ that may be, however, easier to compute for complicated models, where the theoretical Fisher information matrix may be difficult to determine. For instance, for state-space models the expected information matrix was shown to estimate more accurately the true Fisher information [3]. On the other hand, in [6] the authors argue in favour of the observed information matrix over the expected one.

In what follows, we derive the expected and observed Fisher information matrices for the considered disaggregation model. For notational convenience, we shall henceforth denote $\boldsymbol{V} = \boldsymbol{M} + \boldsymbol{C}\boldsymbol{N}\boldsymbol{C}^T$, $\boldsymbol{U} = \boldsymbol{z} - \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}$, and $\boldsymbol{P} = (\boldsymbol{D} - \rho\boldsymbol{W})^{-1}$.

Let us denote the derivative vector of the log likelihood function (6) as

$$L^{(1)} = \left(L_{\beta}^T, L_{\sigma_Z^2}, L_{\tau^2}, L_{\rho}\right)^T.$$

It comprises the following elements:

$$L_{\beta} = \boldsymbol{z}^T\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{X} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{C}^T\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{X}$$

$$L_{\sigma_Z^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{V}^{-1}\right) + \frac{1}{2}\boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{U}$$

$$L_{\tau^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T\right) + \frac{1}{2}\boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T\boldsymbol{V}^{-1}\boldsymbol{U}$$

$$L_{\rho} = -\frac{1}{2}\mathrm{tr}\left(\tau^2\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{W}\boldsymbol{P}\boldsymbol{C}^T\right) + \frac{1}{2}\tau^2\boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{W}\boldsymbol{P}\boldsymbol{C}^T\boldsymbol{V}^{-1}\boldsymbol{U},$$

where tr$(\cdot)$ denotes the trace of a matrix. The respective diagonal elements of the second derivative matrix[1]

$$\text{diag}(\boldsymbol{L}^{(2)}) = \left[\text{diag}\left(\boldsymbol{L}_{\beta\beta}\right), L_{\sigma_Z^2\sigma_Z^2}, L_{\tau^2\tau^2}, L_{\rho\rho}\right]$$

are as follows:

$$\boldsymbol{L}_{\beta\beta} = -(\boldsymbol{CX})^T\boldsymbol{V}^{-1}\boldsymbol{CX} \tag{7}$$

$$L_{\sigma_Z^2\sigma_Z^2} = \frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\right) - \boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{U} \tag{8}$$

$$L_{\tau^2\tau^2} = \frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{CPC}^T\boldsymbol{V}^{-1}\boldsymbol{CPC}^T\right) - \boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{CPC}^T\boldsymbol{V}^{-1}\boldsymbol{CPC}^T\boldsymbol{V}^{-1}\boldsymbol{U} \tag{9}$$

$$\begin{aligned} L_{\rho\rho} \;=\; & \frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPC}^T\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPC}^T\right. \\ & \left.-2\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPWPC}^T\right) \\ & -\boldsymbol{U}^T\left(\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPC}^T\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPC}^T\right. \\ & \left.-\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPWPC}^T\right)\boldsymbol{V}^{-1}\boldsymbol{U} \end{aligned} \tag{10}$$

The Fisher information matrix becomes

$$\boldsymbol{J} = -E\left[\boldsymbol{L}^{(2)}\right] = \text{diag}\left(\boldsymbol{J}_\beta, J_{\sigma_Z^2}, J_{\tau^2}, J_\rho\right).$$

Consequently, we obtain

$$\boldsymbol{J}_\beta = (\boldsymbol{CX})^T\boldsymbol{V}^{-1}\boldsymbol{CX} \tag{11}$$

$$J_{\sigma_Z^2} = \frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\right) \tag{12}$$

$$J_{\tau^2} = \frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{CPC}^T\boldsymbol{V}^{-1}\boldsymbol{CPC}^T\right) \tag{13}$$

$$J_\rho = \frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPC}^T\boldsymbol{V}^{-1}\tau^2\boldsymbol{CPWPC}^T\right). \tag{14}$$

Evaluating the above expressions (11)-(14) at the values of ML parameter estimators $\hat{\theta}$ yields the expected information matrix $\boldsymbol{\mathcal{J}}(\hat{\theta})$. Evaluating the negative of expressions (7)-(10) at $\hat{\theta}$ yields the observed information matrix $\boldsymbol{\mathcal{I}}(\hat{\theta})$. The Cramér-Rao lower bound of the estimators' variances is estimated with a reciprocal of these values.

3.1.4 Prediction in a fine grid

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

$$\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{Z} \end{bmatrix} \sim \mathcal{N}_{n+N}\left(\begin{bmatrix} \boldsymbol{X\beta} \\ \boldsymbol{CX\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{N} & \boldsymbol{NC}^T \\ \boldsymbol{CN} & \boldsymbol{M}+\boldsymbol{CNC}^T \end{bmatrix}\right). \tag{15}$$

[1]Since the off-diagonal elements of the information matrix equal zero, those derivatives are not calculated here.

The distribution (15) allows for full inference, yielding both the predictor and its error

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

3.2 A modification: Various regression models in regions

Next, we adjust the model to reflect possibly diversified regression component across regions. In the considered study of national GHG inventory, we will analyse various regression models for 16 voivodeships indexed with $l = 1, \ldots, L$. Then, all n municipalities are associated with their corresponding voivodeship l, and let n_l denote a number of municipalities in a region l

$$n = \sum_{l=1}^{L} n_l.$$

To accommodate the modification, consider a block diagonal matrix of covariates $\boldsymbol{X}^*$, where each block corresponds to a region $l = 1, \ldots, L$ and contains covariates only for municipalities of this region

$$\boldsymbol{X}^* = \left[\begin{array}{cccc|c|cccc} 1 & x^1_{11} & \cdots & x^1_{1k} & & & & & \\ \vdots & & \ddots & \vdots & & & & & \\ 1 & x^1_{n_1 1} & & x^1_{n_1 k} & & & & & \\ \hline & & & & \ddots & & & & \\ \hline & & & & & 1 & x^L_{11} & \cdots & x^L_{1k} \\ & & & & & 1 & \ddots & \vdots & \\ & & & & & 1 & x^L_{n_L 1} & & x^1_{n_1 k} \end{array}\right]$$

Also a vector of regression coefficients needs to be modified into $\boldsymbol{\beta}^*$, comprising separate sets of regression coefficients for each region

$$\boldsymbol{\beta}^* = \left[\begin{array}{c} \beta^1_0 \\ \vdots \\ \beta^1_k \\ \hline \vdots \\ \hline \beta^L_0 \\ \vdots \\ \beta^L_k \end{array}\right]$$

and the process $\boldsymbol{\mu}$ is redefined as

$$\boldsymbol{\mu} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim Gau_n(\boldsymbol{0}, \Omega).$$

To complete the setting, variance parameters $(\sigma_{Y,l})^2$ and $(\sigma_{Z,l})^2$ are introduced for each region $l = 1, \ldots, L$.

4 Results

First, Table 1 presents estimation results (parameters with their standard errors) for models with and without a spatial component, denoted CAR and LM respectively. Note that β_2 - land use class Arable land turned to be statistically insignificant in this setting. Introducing spatial CAR structure increases standard error of estimated parameters, as compared with LM model. However, for assessment of goodness of fit for these models Table 2 should be referred to.

Table 1: Maximum likelihood estimates

	CAR		LM	
	Est.	Std.Err.	Est.	Std.Err.
β_0	8.525	0.1605	-6.981	0.0389
β_1	3.517	0.0148	1.932	0.0042
β_2	-	-	-	-
β_3	0.916	0.0034	1.786	0.0010
β_4	3.912	0.0055	5.032	0.0013
σ_Z^2	0.961	0.4052	1.506	0.1202
τ^2	1.683	0.1569	-	-
ρ	0.9889	2.62e-06	-	-

Table 2 contains the analysis of residuals ($d_i = y_i - y_i^*$, where y_i^* - predicted values) for considered models. We report the mean squared error mse, the minimum and maximum values of d_i as well as the sample correlation coefficient r between the predicted and observed values. From here it is obvious that the spatial CAR structure considerably improve the results obtained with the model of independent errors LM. For comparison, we also include the results obtained with an allocation done proportionally to population in municipalities; this approach is called NAIVE. It is a straightforward, commonly used approach in this area of application. Here we note that the NAIVE approach provides reasonable results, but CAR model outperforms it in terms of all the reported criteria. The decrease of the mean squared error is from 3374.4 for NAIVE to 3069.4 for CAR, which gives 9% improvement.

From the maps of predicted values for the models CAR and NAIVE (Figure 4) it is difficult to spot a meaningful difference. The map of residuals (Figure 5) and scatterplot (Figure 6) are slightly more informative.

Next, we considered the models with various regression coefficients in voivodeships but having the same same set of covariates (β_0,β_1, β_3 and β_4); the models are denoted CAR* and LM*, respectively. Note that the model CAR* gives much worse results than the models CAR and NAIVE.

Further, considered were the models with varying across regions both the coefficients and sets of covariates. Only statistically significant covariates were chosen. Table 3 includes regression coefficients along with their standard errors for all the considered regions (voivodeships), indexed with l. A reference list with the voivodship names is included in the Appendix I.

Table 2: Analysis of residuals ($d_i = y_i - y_i^*$)

	mse	$\min(d_i)$	$\max(d_i)$	r
CAR	3069.4	-275	469	0.784
LM	5641.2	-357	522	0.555
CAR*	3437.0	-258	512	0.763
LM*	4876.1	-374	546	0.651
CAR**	3124.9	-256	446	0.783
LM**	4427.6	-352	472	0.674
NAIVE	3374.4	-475	403	0.766

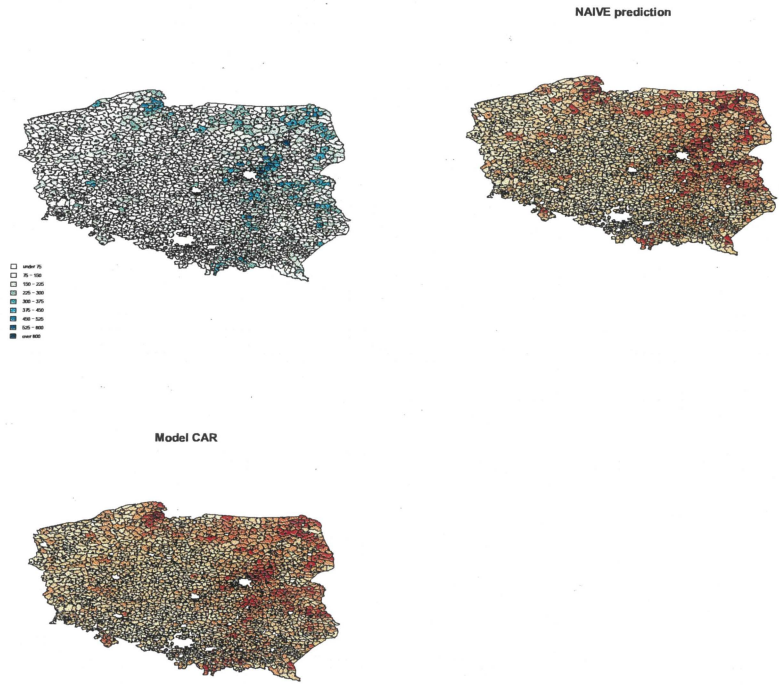


Figure 4: Original data in municipalities and predicted values for the models NAIVE and CAR

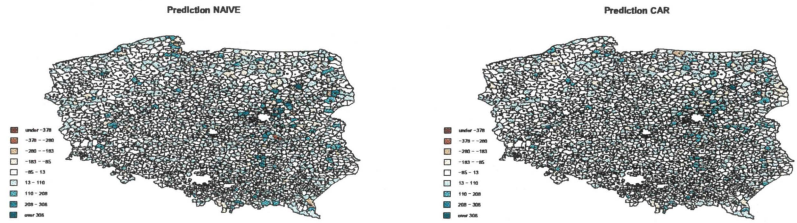


Figure 5: Residuals from predicted values for the models NAIVE and CAR

Table 3: Maximum likelihood estimates of the models CAR** and LM**

	CAR**		LM**		CAR**		LM**	
	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.
	l=1				l=2			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	3.514	0.0528	1.289	0.0098	5.227	0.0592	3.431	0.0099
β_2^l	-	-	-	-	-	-	-	-
β_3^l	1.593	0.0221	2.063	0.0060	0.588	0.0194	1.032	0.0044
β_4^l	1.344	0.0322	3.049	0.0052	4.759	0.0288	2.909	0.0048
$(\sigma_Z^l)^2$	1.281	1.1759	0.559	0.1552	1.0905	1.6542	0.368	0.1194
	l=3				l=4			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	23.849	0.0966	24.729	0.0331	-3.349	0.0967	-2.611	0.0301
β_2^l	-1.546	0.0085	-1.679	0.0033	-	-	-	-
β_3^l	4.632	0.0196	4.308	0.0043	3.056	0.0164	2.447	0.0043
β_4^l	1.622	0.0187	2.119	0.0051	6.271	0.0512	5.129	0.0150
$(\sigma_Z^l)^2$	0.974	2.2569	2.616	0.8273	0.852	1.7905	0.614	0.2509
	l=5				l=6			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	6.392	0.0678	6.409	0.0272	0.729	0.0407	-2.221	0.0122
β_2^l	-	-	-	-	-	-	-	-
β_3^l	-	-	-	-	1.662	0.0205	4.276	0.0066
β_4^l	1.726	0.0253	2.122	0.0117	4.080	0.0199	5.117	0.0062
$(\sigma_Z^l)^2$	0.938	1.6488	2.0944	0.6463	1.382	2.7181	2.723	0.8835
	l=7				l=8			

Table 3: (continued)

	CAR**		LM**		CAR**		LM**	
	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.
β_0^l	-	-	-	-	-	-	-	-
β_1^l	2.332	0.0348	4.452	0.0250	3.739	0.0648	3.491	0.0145
β_2^l	-	-	-	-	-	-	-	-
β_3^l	-	-	-	-	0.731	0.0438	0.489	0.0122
β_4^l	7.698	0.0148	8.459	0.0111	-	-	-	-
$(\sigma_Z^l)^2$	1.127	1.4045	7.5264	1.749	0.955	2.134	0.640	0.2731
	l=9				l=10			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	-	-	-	-	-	-	-	-
β_2^l	0.652	0.0078	0.686	0.0021	0.956	0.0038	0.897	0.0013
β_3^l	2.543	0.0166	1.865	0.0056	-	-	-	-
β_4^l	3.660	0.0157	3.135	0.0039	2.857	0.0101	4.322	0.0035
$(\sigma_Z^l)^2$	1.227	1.7052	0.998	0.3080	0.809	2.1353	2.145	0.8106
	l=11				l=12			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	11.063	0.0655	14.421	0.0200	2.562	0.0543	1.170	0.0097
β_2^l	-0.456	0.0045	-0.625	0.0013	0.1315	0.0097	0.523	0.0013
β_3^l	-	-	-	-	-	-	-	-
β_4^l	5.397	0.0163	4.034	0.0053	2.595	0.0390	2.142	0.0069
$(\sigma_Z^l)^2$	1.139	1.8027	1.301	0.4602	1.016	2.6822	0.636	0.2182
	l=13				l=14			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	-	-	-	-	16.235	0.0585	14.090	0.0318
β_2^l	-0.114	0.0056	-0.073	0.0021	-	-	-	-
β_3^l	-	-	-	-	-	-	-	-
β_4^l	7.445	0.0229	7.368	0.0070	1.569	0.0147	3.273	0.0107
$(\sigma_Z^l)^2$	0.515	1.7805	1.735	0.6805	0.858	1.1953	3.189	1.0349
	l=15				l=16			
β_0^l	-	-	-	-	-	-	-	-
β_1^l	2.367	0.0312	2.001	0.0100	13.159	0.0630	10.993	0.0189
β_2^l	0.615	0.0031	0.458	0.0012	-	-	-	-
β_3^l	1.652	0.0095	1.793	0.0038	-	-	-	-
β_4^l	-	-	-	-	0.379	0.0237	-0.160	0.0089
$(\sigma_Z^l)^2$	0.627	0.993	1.303	0.3311	0.634	1.4092	1.018	0.339
τ^2	1.647	0.1536	-	-				
ρ	0.9913	1.59e-06	-	-				

The reported values of estimated parameters for CAR** and LM** show considerable differences across the voivodeships, not only in terms of estimated values of regression coefficients, but also in terms of their significance. Moreover, from Table 2 we note that this setting (CAR**) provides comparable results to CAR.

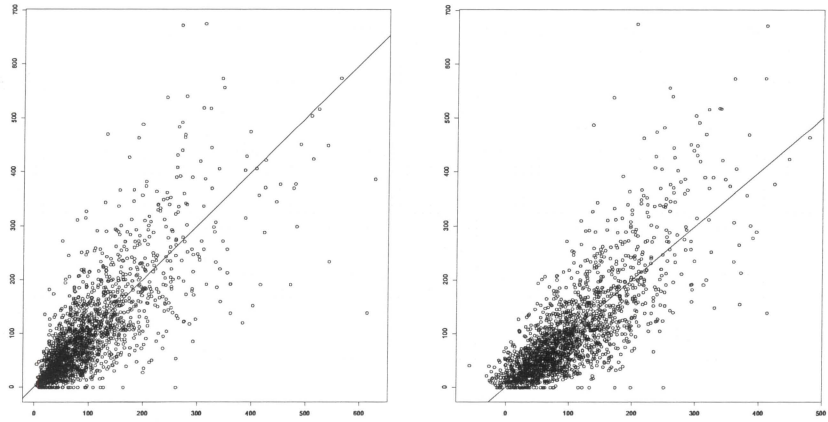


Figure 6: Scatterplot of predictions (y_i^*) against observations (y_i) for the models NAIVE (left) and CAR (right)

5 Concluding remarks and discussion

The study presents the first attempt to apply the spatial scaling model for the GHG inventory in Poland. The task was to allocate spatially correlated data to finer spatial scales, conditional on covariate information observable in a fine grid. Spatial dependence is set and it is assummed not to change with the change of grid. It 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. The usefulness of the proposed technique is shown on an example of allocation of livestock data (a number of horses) from district to municipality level.

The results of the disaggregation with the proposed procedure were compared with the allocation proportional to population of municipalities. An improvement over the naive, proportional approach of 9% in terms of the mean squared error was reported. In addition, we extended the model to allow for various regression models in regions (here voivodeships). Numerous features of the proposed method require further investigation.

The proposed method provided good results for livestock activity data of agricultural sector. Apart from the reported above study, the approach was also applied in a residential sector for disaggregation of natural gas consumption in households. In that case, with disaggregation featured from voivodeships into municipalities, the results turned to be quite modest. This was partly due to limited spatial correlation of the analysed process and too large extent of disaggregation. The method is feasible for disaggregation from districts into municipalities, but not from voivodeships into municipalities.

It should be stressed that the primary asset of the proposed approach is the possibility to asses significance of considered regression coefficients. The widely used proportional distribution of activity data can be based only on expert judgements, providing no means for outcome verification.

Acknowledgement

The study was conducted within the 7FP Marie Curie Actions IRSES project No. 247645 *Geoinformation technologies, spatio-temporal approaches, and full carbon account for improving accuracy of GHG inventories.* The support from the Polish Ministry of Science and Higher Education within the funds for statutory works of young scientists is gratefully acknowledged.

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.

References

[1] Banerjee S., Carlin B.P., Gelfand A.E. (2004) *Hierarchical modeling and analysis for spatial data.* Chapman & Hall/CRC.

[2] Boychuk K., Bun R., Regional spatial cadastres of GHG emissions in Energy sector: Accounting for uncertainty, *Climatic Change*, under revision.

[3] Cavanaugh J.E., Shumway R.H. (1996) *On computing the expected Fisher information matrix for state-space model parameters, Stat. Probabil. Lett.*, 26:347-355.

[4] 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(4):372-375.

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

[6] Efron B., Hinkley D.V. (1978) *Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, Biometrika* 65(3):457-87.

[7] European Environment Agency (2000) Corine Land Cover 2000. `http://www.eea.europa.eu/data-and-maps/data` Accessed November 2012.

[8] Gelfand A.E., Diggle P.J., Fuentes M., Guttorp P., Eds. (2010) *Handbook of Spatial Statistics*, Chapman & Hall/CRC.

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

[10] Główny Urząd Statystyczny (2012) Bank Danych Lokalnych. `http://www.stat.gov.pl/bdlen/app/strona.html?p_name=indeks` Accessed November 2012.

[11] Horabik J., Nahorski Z. (2010) A statistical model for spatial inventory data: a case study of N_2O emissions in municipalities of southern Norway, *Climatic Change* 103(1-2):263-276.

[12] Horabik J., Nahorski Z. (2014) Improving resolution of a spatial air pollution inventory with a statistical inference approach, *Climatic Change* 124(3):575-589.

[13] IPCC (1996) IPCC Guidelines for National Greenhouse Gas Inventories. Volume 1, 2, and 3. Intergovernmental Panel on Climate Change, London.

[14] 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.

[15] Lim B., Boileau P., Bonduki Y. et al. (1999) Improving the quality of national greenhouse gas inventories, *Environmental Science & Policy* 2:335-346.

[16] 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.

[17] 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.

[18] Monahan J.F. (2001) *Numerical methods of statistics.* Cambridge University Press.

[19] Rypdal K., Winiwarter W. (2001) Uncertainties in greenhouse gas emission inventories - evaluation, comparability and implications, *Environmental Science & Policy* 4:107-116.

[20] McMillan A.S., Holland D.M., Morara M., Fend J. (2010) Combining numerical model output and particulate data using Bayesian space-time modeling, *Environmetrics* 21:48-65.

Appendix I

Table 4: List of voivodships

l	Voivodship
1	Dolnośląskie
2	Kujawsko-Pomorskie
3	Lubelskie
4	Lubuskie
5	Łódzkie
6	Małopolskie
7	Mazowieckie
8	Opolskie
9	Podkarpackie
10	Podlaskie
11	Pomorskie
12	Śląskie
13	Swiętokrzyskie
14	Warmińsko-Mazurskie
15	Wielkopolskie
16	Zachodniopomorskie

Appendix II

Reprint of the article
Horabik J., Nahorski Z. (2014) Improving resolution of a spatial air pollution inventory with a statistical inference approach, *Climatic Change* 124(3):575-589