# 4th International Workshop
## on Uncertainty in Atmospheric Emissions
**7-9 October 2015, Krakow, Poland**

# PROCEEDINGS

# 4th International Workshop on Uncertainty in Atmospheric Emissions

## 7–9 October 2015, Kraków, Poland

# PROCEEDINGS

4<sup>th</sup> International Workshop on Uncertainty in Atmospheric Emissions
7- 9 October 2015, Cracow, Poland

Printed from the material submitted by the authors.

# About the Workshop

The assessment of greenhouse gases and air pollutants (indirect GHGs) emitted to and removed from the atmosphere is high on the political and scientific agendas. Building on the UN climate process, the international community strives to address the long-term challenge of climate change collectively and comprehensively, and to take concrete and timely action that proves sustainable and robust in the future. Under the umbrella of the UN Framework Convention on Climate Change, mainly developed country parties to the Convention have, since the mid-1990s, published annual or periodic inventories of emissions and removals, and continued to do so after the Kyoto Protocol to the Convention ceased in 2012. Policymakers use these inventories to develop strategies and policies for emission reductions and to track the progress of those strategies and policies. Where formal commitments to limit emissions exist, regulatory agencies and corporations rely on emission inventories to establish compliance records.

However, as increasing international concern and cooperation aim at policy-oriented solutions to the climate change problem, a number of issues circulating around uncertainty have come to the fore, which were undervalued or left unmentioned at the time of the Kyoto Protocol but require adequate recognition under a workable and legislated successor agreement. Accounting and verification of emissions in space and time, compliance with emission reduction commitments, risk of exceeding future temperature targets, evaluating effects of mitigation versus adaptation versus intensity of induced impacts at home and elsewhere, and accounting of traded emission permits are to name but a few.

The *4th International Workshop on Uncertainty in Atmospheric Emissions* is jointly organized by the *Systems Research Institute of the Polish Academy of Sciences*, the Austrian-based *International Institute for Applied Systems Analysis*, and the *Lviv Polytechnic National University*. The 4th Uncertainty Workshop follows up and expands on the scope of the earlier Uncertainty Workshops – the *1st Workshop* in 2004 in Warsaw, Poland; the *2nd Workshop* in 2007 in Laxenburg, Austria; and the *3rd Workshop* in 2010 in Lviv, Ukraine.

# Steering Committee

Rostyslav BUN (Lviv Polytechnic National University, UA)
Matthias JONAS (International Institute for Applied Systems Analysis, AT)
Zbigniew NAHORSKI (Polish Academy of Sciences, PL) – Chair

# Scientific Committee

Evgueni GORDOV (Siberian Center for Environmental Research & Training, RU)
Piotr HOLNICKI-SZULC (Polish Academy of Sciences, PL)
Joanna HORABIK-PYZEL (Polish Academy of Sciences, PL)
Olgierd HRYNIEWICZ (Polish Academy of Sciences, PL)
Katarzyna JUDA-REZLER (Warsaw University of Technology, PL)
Petro LAKYDA (National University of Life and Environmental Sciences of Ukraine, UA)
Myroslava LESIV (Lviv Polytechnic National University, UA)
Gregg MARLAND (Appalachian State University, USA)
Sten NILSSON (Forest Sector Insights AB, SE)
Tom ODA (Univ. Space Research Association, NASA Goddard Space Flight Center, USA)
Stefan PICKL (Universität der Bundeswehr München, Germany)
Elena ROVENSKAYA (International Institute for Applied Systems Analysis, AT)
Kazimierz RÓŻAŃSKI (AGH University of Science and Technology in Cracow, PL)
Dmitry SCHEPASCHENKO (International Institute for Applied Systems Analysis, AT)
Anatoly SHVIDENKO (International Institute for Applied Systems Analysis, AT)
Jacek SKOŚKIEWICZ (National Centre for Emissions Management, PL)
Philippe THUNIS (EC Joint Research Centre Ispra, EU)
Marialuisa VOLTA (University of Brescia, IT)

# Local Organizing Committee

Joanna HORABIK-PYZEL
Jolanta JARNICKA - Chair
Weronika RADZISZEWSKA
Jörg VERSTRAETE

# Conditionally autoregressive model for spatial disaggregation of activity data in GHG inventory: Application for agriculture sector in Poland

Joanna Horabik-Pyzel[1], Nadiia Charkovska[2], Olha Danylo[2,3], Zbigniew Nahorski[1], and Rostyslav Bun[2,4]

[1] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
*Joanna.Horabik@ibspan.waw.pl*
[2]Lviv Polytechnic National University, Lviv, Ukraine
[3]International Institute for Applied Systems Analysis, Laxenburg, Austria
[4]Academy of Business in Dąbrowa Górnicza, Poland

**Abstract**

This report presents a novel approach for allocation of spatially correlated data, such as emission inventories, into 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. The usefulness of the proposed technique is shown for agricultural sector of GHG inventory in Poland. An example of allocation of livestock data (a number of horses) from district to municipality level is analysed. The results indicate that the proposed method outperforms a naive and commonly used approach of proportional distribution.

**Keywords:** GHG inventory, agricultural sector, spatial correlation, disaggregation, conditional autoregressive model

## 1. Introduction

Spatially resolved inventories of greenhouse gases (GHG) contribute valuable information for an assessment of carbon sources and sinks. Various authors point out that a regional or local formulation improves accuracy of the assessment. Quality of these inventories is subject to various conditions; particularly, it depends on availability of high resolution activity data.

In case of national GHG inventories, relevant information about low resolution activity data needs to be acquired from national/regional totals. A procedure of allocation into smaller spatial units (like districts, municipalities, and finally 2x2km grid cells) 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. 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 [1]. A common approach though, is a spatial allocation made in a proportion to some related indicators that are available in a finer grid.

In this study, the statistical scaling method is developed in order to support the procedure of compiling high resolution activity data. We propose the 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.

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 [2], it was also successfully applied for modelling air pollution over space [3]. We demonstrate usefulness of the proposed technique for the agricultural sector of GHG national inventory in Poland. The example considers an allocation of livestock data (a number of horses) from district to municipality level.

A part of the methododology described in section 3.1 was already presented in [4]. This contribution extends the basic model for the case of various regression models in each region (here voivodeship); see section 3.2. Performance of the method for livestock data in agricultural sector of GHG inventory is presented in section 4.

## 2. Inventory livestock dataset

Considered is a livestock dataset (cattle, pigs, horses, poultry, etc.) for the territory of Poland, based on the agricultural census 2010, and available from the Central Statistical Office of Poland - Local Data Bank [5]. The goal is to allocate relevant livestock amounts from districts (*powiaty*) to municipalities (*gminy*).

In particular, for horses the data are available also in municipalities, and this fact enables us to verify 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. Only rural municipalities are considered in the study.

As explanatory variables we use population density (denoted $x_1$) and land use information. For the latter, the CORINE Land Cover map, available from the European Environment Agency [6], was employed. For each rural municipality we calculate the area of agricultural classes, which may be related to livestock farming. Three CORINE classes were considered (the CORINE class numbers are given in brackets):
-   Arable land (2.1); denoted $x_2$
-   Pastures (2.3); denoted $x_3$
-   Heterogeneous agricultural areas (2.4); denoted $x_4$.

The results of the disaggregation with the proposed procedure are further compared with the results of allocation proportional to population of municipalities. 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 case study, we treat 16 voivodeships (*województwa*) as regions.

## 3. The disaggregation framework

### 3.1 The basic model

First, the model is specified on a level of *fine* grid. Let $Y_i$ denote a random variable associated with an unknown 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). The random variables $Y_i$ are assumed to follow the Gaussian distribution with the mean $\mu_i$ and variance $\sigma_Y^2$

$$Y_i|\mu_i \sim Gau(\mu_i, \sigma_Y^2)$$

Given the values $\mu_i$ and $\sigma_Y^2$, the random variables $Y_i$ are assumed independent. The mean $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^n$ represents the true process underlying emissions, and the (unknown)

observations are related to this process through a measurement error with the variance $\sigma_Y^2$. The approach to modeling $\mu_i$ expresses an assumption that available covariates explain part of the spatial pattern, and the remaining part is captured through a spatial dependence. The CAR scheme follows an assumption of similar random effects in adjacent cells, and it is given through the specification of full conditional distribution functions of $\mu_i$ for $i = 1,...,n$

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

where $\boldsymbol{\mu}_{-i}$ denotes all elements in $\boldsymbol{\mu}$ but $\mu_i$, $w_{ij}$ 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 $i$; $\boldsymbol{x}_i^T \boldsymbol{\beta}$ is a regression component with proxy information available for area $i$ and a respective vector of regression coefficients; $\tau^2$ is a variance parameter. Thus, the mean of the conditional distribution $\mu_i | \boldsymbol{\mu}_{-i}$ consists of the regression part and the second summand, which is proportional to the average values of remainders $\mu_j - \boldsymbol{x}_j^T \boldsymbol{\beta}$ for neighbouring sites (i.e. when $w_{ij} = 1$). The proportion is calibrated with the parameter $\rho$, reflecting strength of a spatial association. Furthermore, the variance of the conditional distribution $\mu_i | \boldsymbol{\mu}_{-i}$ is inversely proportional to a number of neighbours $w_{i+}$.

The joint distribution of the process $\boldsymbol{\mu}$ is the following (for the derivation see [2])
$$\boldsymbol{\mu} \sim Gau_n(\boldsymbol{X}\boldsymbol{\beta}, \tau^2(\boldsymbol{D} - \rho \boldsymbol{W})^{-1}) \tag{1}$$
where $\boldsymbol{D}$ 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 (1) as
$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim Gau_n(\boldsymbol{0}, \boldsymbol{\Omega}) \tag{2}$$
with $\boldsymbol{\Omega} = \tau^2(\boldsymbol{D} - \rho \boldsymbol{W})^{-1}$.

The model for a *coarse* grid of (aggregated) observed data is obtained by multiplication of (2) with the $N \times n$ *aggregation matrix* $\boldsymbol{C}$, where $N$ is a number of observations in a coarse grid
$$\boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon}, \quad \boldsymbol{C}\boldsymbol{\varepsilon} \sim Gau_n(\boldsymbol{0}, \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}^T) \tag{3}$$
The aggregation 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 (in a coarse grid)
$$\boldsymbol{Z} | \boldsymbol{\lambda} \sim Gau_N(\boldsymbol{\lambda}, \sigma_Z^2 \boldsymbol{I}_N)$$

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

Model parameters $\boldsymbol{\beta}, \sigma_Z^2, \tau^2$ and $\rho$ are estimated with the maximum likelihood method based on the joint unconditional distribution of observed random variables $\boldsymbol{Z}$
$$\boldsymbol{Z} \sim Gau_N(\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}, \ \sigma_Z^2 \boldsymbol{I}_N + \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}^T) \tag{4}$$
The log likelihood function associated with (4) is formulated, and the analytical derivation is limited to the regression coefficients $\boldsymbol{\beta}$; further maximization of the profile log likelihood is performed numerically.

As to the prediction of missing values in a fine grid, the underlying mean process $\boldsymbol{\mu}$ is of our primary interest. The predictors optimal in terms of the mean squared error are given by the conditional expected value $E(\boldsymbol{\mu}|\boldsymbol{z})$. The joint distribution of $(\boldsymbol{\mu}, \boldsymbol{Z})$ is
$$\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{Z} \end{bmatrix} \sim Gau_{n+N}\left( \begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\Omega}\boldsymbol{C}^T \\ \boldsymbol{C}\boldsymbol{\Omega} & \sigma_Z^2 \boldsymbol{I}_N + \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}^T \end{bmatrix} \right) \tag{5}$$
The distribution (5) yields both the predictor $\widehat{E(\boldsymbol{\mu}|\boldsymbol{z})}$ and its error $\widehat{Var(\boldsymbol{\mu}|\boldsymbol{z})}$

$$E\widehat{(\mu|z)} = X\widehat{\beta} + \widehat{\Omega}C^T\left(\widehat{\sigma_Z^2}I_N + C\widehat{\Omega}C^T\right)^{-1}\left[z - CX\widehat{\beta}\right]$$

$$Var\widehat{(\mu|z)} = \widehat{\Omega} - \widehat{\Omega}C^T\left(\widehat{\sigma_Z^2}I_N + C\widehat{\Omega}C^T\right)^{-1}C\widehat{\Omega}$$

The standard errors of parameter estimators are calculated with the Fisher information matrix based on the log likelihood function, see [7].

### 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 $l = 1,...,L$. Then, all $n$ municipalities are associated with their corresponding voivodeship $l$, and let $n_l$ denote a number of municipalities in a region $l$.

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

$$X^* = \begin{bmatrix} 1 & x_{11}^1 & \cdots & x_{1k}^1 & & & & \\ \vdots & & \ddots & \vdots & & & & \\ 1 & x_{n_1 1}^1 & & x_{n_1 k}^1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & x_{11}^L & \cdots & x_{1k}^L \\ & & & & & & & \ddots & \vdots \\ & & & & & 1 & x_{n_L 1}^L & & x_{n_L k}^L \end{bmatrix} \qquad \beta^* = \begin{bmatrix} \beta_0^1 \\ \vdots \\ \beta_k^1 \\ \vdots \\ \beta_0^L \\ \vdots \\ \beta_k^L \end{bmatrix}$$

Also a vector of regression coefficients has to be modified into $\beta^*$, comprising separate sets of regression coefficients for each region (see above), and the process $\mu$ is redefined as $\mu = X^*\beta^* + \varepsilon$, $\varepsilon \sim Gau_n(0, \Omega)$. To complete the setting, variance parameters $\sigma_{Y,l}^2$ and $\sigma_{Z,l}^2$ are introduced for each region $l=1,...,L$.

### 4. Results

First, Table 1 presents the estimation results (parameters with their standard errors) for the models with and without a spatial component, denoted CAR and LM respectively. Note that in this setting the variable $\beta_2$ (land use class *Arable land*) turned out to be statistically insignificant. Introduction of the spatial CAR structure increased the standard error of estimated parameters, as compared with LM model. However, for an assessment of goodness of fit for these models Table 2 should be referred to.

**Table 1**. Maximum likelihood estimates.

| | CAR | | LM | |
|---|---|---|---|---|
| | Estimate | Std. Error | Estimate | Std. Error |
| $\beta_0$ | 8.525 | 0.1605 | -6.981 | 0.0389 |
| $\beta_1$ | 3.517 | 0.0148 | 1.932 | 0.0042 |
| $\beta_2$ | - | - | - | - |
| $\beta_3$ | 0.916 | 0.0034 | 1.786 | 0.0010 |
| $\beta_4$ | 3.912 | 0.0055 | 5.032 | 0.0013 |
| $\sigma_z^2$ | 0.961 | 0.4052 | 1.506 | 0.1202 |
| $\tau^2$ | 1.683 | 0.1569 | - | - |
| $\rho$ | 0.9889 | 2.62e-06 | - | - |

Table 2 contains the analysis of residuals ($d_i = y_i - y_i^*$, where $y_i^*$ - predicted values) for the 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 improves the results obtained with the model of independent errors LM. For comparison, we also include the results obtained with the allocation proportional to population in municipalities; this setting is called NAIVE. It is a straightforward and commonly used approach in this area of application. Here we note that the NAIVE approach provides reasonable results, but the 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 1), it is difficult to spot a meaningful difference.

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

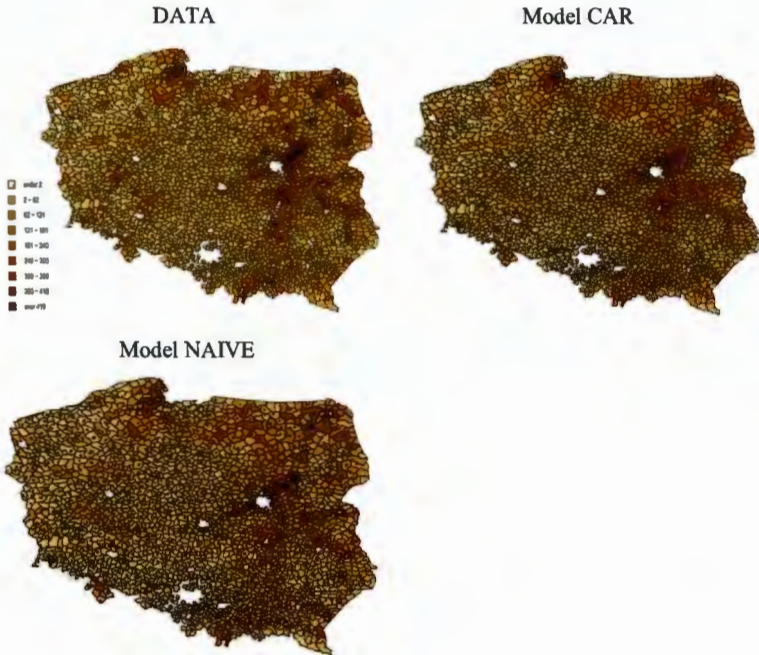|       | *mse*  | min($d_i$) | max($d_i$) | r     |
|-------|--------|------------|------------|-------|
| CAR   | 3069.4 | -275       | 469        | 0.784 |
| LM    | 5641.2 | -357       | 522        | 0.555 |
| NAIVE | 3374.4 | -475       | 403        | 0.766 |
| CAR*  | 3437.0 | -258       | 512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

DATA                                    Model CAR



Model NAIVE



**Figure 1.** Original data in municipalities as well as predicted values for the models NAIVE and CAR.

Next, we considered the models with various regression coefficients in regions (voivodeships) but having the same set of covariates ($\beta_0$, $\beta_1$, $\beta_3$, $\beta_4$); the models are denoted CAR* and LM*, respectively, for the spatial and non-spatial approaches. Note that the model CAR* gives much worse results than the models CAR and NAIVE.

Further, considered were the models (CAR** and LM**) where, both, the coefficients as well as sets of covariates vary across the regions. Only the statistically significant covariates were chosen. Due to a lack of space, we do not provide here the table with the regression coefficients and their standard errors for all the considered regions. We only report that the values of estimated parameters for CAR** and LM** showed considerable differences across the voivodeships, not only in terms of the estimated values, but also in terms of their significance. From Table 2 we note that this setting (CAR**) provides the results comparable to that of CAR.

## 5. Concluding remarks

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. The results of the disaggregation with the proposed procedure were compared with the allocation proportional to population; an improvement of 9% in terms of the mean squared error was reported. The model was extended to allow for various regression covariates in regions (here voivodeships). Numerous features of the method require further investigation.

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

### Acknowledgements

## References

[1] Boychuk, K. and R. Bun (2014) Regional spatial inventories (cadastres) of GHG emissions in the Energy sector: Accounting for uncertainty, Climatic Change 124:561-574.

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

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

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

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

[6] European Environment Agency (2000) Corine Land Cover 2000.
http://www.eea.europa.eu/data-and-maps/data

[7] Horabik, J. and Z. Nahorski (2014) The Cramér-Rao lower bound for the estimated parameters in a spatial disaggregation model for areal data. In: D. Filev et al.: Intelligent Systems 2014, Springer International Publishing, pp. 661-668.