Raport Badawczy
Research Report

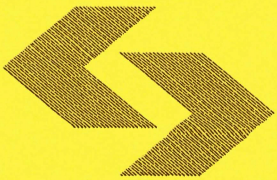
RB/9/2014

Statistical methodology for verification of GHG inventory maps

J. Verstraete, Z. Nahorski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

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

Warszawa 2014

D 3.4

Version	1
Date	20.06.2014
Author	SRI
Dissemination level	PP
Document reference	D 3.4

GESAPU

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

Deliverable 3.4. Statistical methodology for verification of GHG inventory maps

Jörg Verstraete, Zbigniew Nahorski
Systems Research Institute, Polish Academy of Sciences, Poland

Delivery Date: M42

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 3.4. Statistical methodology for verification of GHG inventory maps

Content

Introduction: statistical assimilation of independent knowledge
References

Solving the map overlay problem with a fuzzy approach

1. **Introduction**
2. **The map overlay problem**
   - **2.1. Problem description**
   - **2.2. Current solution methods**
     - **2.2.1. Areal weighting**
     - **2.2.2. Spatial smoothing**
     - **2.2.3. Regression methods**
   - **2.3. Using additional knowledge**
     - **2.3.1. Data fusion**
     - **2.3.2. Intuitive approach to grid remapping**
3. **Using intelligent techniques**
   - **3.1. Introduction to fuzzy sets and fuzzy inference**
     - **3.1.1. Fuzzy sets**
     - **3.1.2. Fuzzy inference system**
   - **3.2. Defining the inference system**
4. **Experiments**
   - **4.1. Description of the results**
   - **4.2. Observations of the methodology**
   - **4.3. Future developments**
5. **Conclusion**

References

Appendix 1: Solving the map overlay problem with a fuzzy approach, supplement

S1. Defining the inference system
- **S1.1 Concept**
- **S1.2 Defining the parameters**
- **S1.3 Defining the rules**
- **S1.4 Defining the fuzzy sets**

Appendix 2: Parameters to use a fuzzy rulebase approach to remap gridded spatial data

1. **Introduction**
   - **1.1. Problem description**

1.2. Current solution methods
2. Rule base approach
2.1. Reasoning with added knowledge
2.2. Fuzzy Inference System
3. Discussion of parameters
3.1. Current
3.2. Data study
4. Conclusion
References

Appendix 3: Automatically identifying suitable rulebase parameters in the context of solving the map overlay problem

1. Introduction
2. Problem description
2.1. Map overlay problem
2.2. Current approaches
2.3. Additional data approach
3. AI Algorithm
3.1. Prerequisites
3.2. Translating the problem to fit a rulebase
4. Parameters
4.1. What are the *best* parameters?
4.2. Relating to the output
4.3. Proper range
4.4. Example
5. Parameter selection
5.1. Data generation
5.2. Calculation
6. Experiments
6.1. Prerequisites
6.2. Case 1
6.3. Case 2
7. Conclusion and future work
References

Appendix 4: A fuzzy rulebase approach to remap gridded spatial data: initial observations

1. Introduction
1.1. Problem description
1.2. Current solution methods
2. Rulebase approach
2.1. A different look at the problem
2.2. Emulating the intelligent reasoning
2.3. Parameters and range
2.4. Rulebase construction
3. Experiments
4. Conclusion
References

List of figures

Figure 1. Different data distributions within a grid cell that result in the same value for the grid cell are shown in (a). The examples are: a single point source of value 100, two point sources of value 50, a line source of value 100 and an area source of value 100. Each of these are such that they are in one gridcell, which then has the value 100. When viewing the gridcell, it is not known what the underlying distribution is. Different incompatible grids are shown in (b)-(e): a relative shift (b), a different gridsize (c), a different orientation (d) and a combination (e).

Figure 2. Examples to explain the problem: (a) Problem illustration: remapping grid *A* onto grid *B*, (b) Areal weighting: the value of each output cell is determined by the amount of overlap, (c) Areal smoothing: the value of each output cell is determined resampling a smooth surface that is *fitted* over the input data, (d) Intelligent reasoning using additional data: grid *C* supplies information on the distribution, which can be used to determine values in the output grid.

Figure 3. Illustrations for the different cases from Table 1: *A* is the input grid, *B* is the output grid and *C* the auxiliary grid. The gridcells are drawn above each other for visibility purposes, but should cover each other as shown on Figure 2(d). The size of the circles reflects the relative value of the associated cell (a small circle is shown for 0 values, for illustration purposes).

Appendix 1.

Figure S1. Simplified example using two grids: the output grid *B* and the auxiliary grid *C* use the same raster. The cover the same region of interest the input grid *A*, but the gridcells do not overlap nicely: grid *A* has two gridcells, whereas grids *B* and *C* have three gridcells.

Figure S2. Example of the fuzzy sets used to define low, medium and high values for both input and auxiliary grid.

Figure S3. The sets to define the labels of the output values of the inference system.

Appendix 2

Figure 1. Example showing the different sources that yield a similar grid cell: single point source, two point sources, line source and area source.

Fig. 2. Example cases for the case study. Cases (a) and (b) are used to show how auxiliary cells should influence the output, cases (c) and (d) are used to show how the input cells that neighbour the overlapping input cell influence the output.

Fig. 3. Example cases for the case study, (a) and (b) are used to show the influence of auxiliary cells that overlap the neighbouring input cell.

Fig. 4. Example to illustrate possible definitions for the limits of the fuzzy sets.

Appendix 3

Fig. 1. The three geometries used to test the algorithms and their approximation as input grids. Geometry (a) contains two line sources with a constant value, geometry (b) contains 3 area sources with a constant value and geometry (c) contains different line patters with varying associated values. Greyscales are used to illustrated the values: higher values are shaded darker.

Fig. 2. Test case 1. Target grid (top) and resulting segment grid (bottom) (a). The auxiliary grid (top) and segment grid (bottom), for the first geometry (b), second geometry (c), and reference geometry (d).

Fig. 3. Test case 2. Target grid (top) and resulting segment grid (bottom) (a). The auxiliary grid (top) and segment grid (bottom), for the first geometry (b), second geometry (c), and reference geometry (d).

Appendix 4

Fig. 1. Example of an input grid (2x2, in bold) that needs to be remapped onto a target grid (3x3, dotted line). Different additional data are represented by the thick lines in (a) and (b).

Fig. 2. (a) generated input data with grid, (b) ideal solution for target 1, (c) areal weighting for target 1, (d) ideal solution for target 2, (e) areal weighting solution for target 2. Darker shades represent higher associated values; but the scale between different grids does not match. For each grid, black indicates the highest occurring colour in that grid; the lighter the colour, the lower the associated value.

Fig. 3. Case 1: low resolution auxiliary data: (a) auxiliary data, (b) result, (c) detail of the remapping of the input data and case 2: High resolution auxiliary data: (d) auxiliary data, (e) result, (f) detail of the remapping of the input data.

Fig. 4. Case 3: low resolution rotated auxiliary data: (a) auxiliary data, (b) result, (c) detail of the remapping of the input data and Case 4: low resolution rotated auxiliary data and rotated target: (a) auxiliary data, (b) result, (c) detail of the remapping of the input data.

List of tables

Table 1. Overview of the cases used in the simulations. The grid layout is illustrated on Figure 2(d), these results are graphically illustrated on Figure 3.

Appendix 1

Table S1. Example values for the conceptual example to reason about the gridcells.

Appendix 2

Table 1. Example and expected values for the examples in figure 2a.

Table 2. Example and expected values for the examples in figure 2b.

Table 3. Example and expected values for the examples in figure 2c and 2d.

Table 4. Example and expected values for the examples in figure 3.

Table 5. Different range definitions over the input cells.

Table 6. Different range definitions over the output cells.

Appendix 3

Table 1. The correlations for the parameters in the different datasets in the two considered testcases, in decreasing order of correlation.

Appendix 4

Table 1. Properties of the results of the 4 examples.

Introduction: statistical assimilation of independent knowledge

It is difficult to talk about verification of inventory maps in a strict sense, since the true values of emissions are not known. With this respect we can talk rather about testing the results or validation of the models and methods used to obtain the emission results.

There are several methods providing independent observations, and they were discussed in Deliverable 3.3. Namely, the following four methods have been presented:

- using nighttime lights,
- estimation of fossil fuel emissions from tracer measurements, mainly $^{14}CO_2$,
- atmospheric inversion to estimate fluxes,
- flux tower observations.

Solely one of the above approaches may be presently considered for comparison with the very high resolution inventory obtained by modeling and disaggregation of rough resolution data, like the 2 km × 2 km inventory for Poland, developed in the GESAPU. This is by means of nighttime lights. For example, ODIAC inventory provides the data with the resolution 1 km × 1 km. There are, however, some limitations. Firstly, nighttime lights allow only for estimation of energy emissions, and secondly, the accuracy of the nighttime lights inventory seems to be distinctly smaller than the accuracy of inventory by modeling. The latter is quite accurate, with uncertainty even of the order 2-5%, especially for the energy sector. Nevertheless, the comparison of GESAPU and ODIAC inventories showed that the results were close, except for about 10% which could be attributed to high point emission sources and misalign of the grids in both maps. More precisely, some high point emission sources in the ODIAC map were misplaced, and those with high values that were counted in a grid cell which did not match precisely a grid in the other map, also contributed to high differences. Some quantitative results of comparison have been presented in the Deliverable 3.3, also a proposition how to combine both maps is described there. To apply it, estimates of uncertainties are needed, and no such estimate is given for ODIAC inventory in Oda & Maksyutov (2011). It is also possible to use statistical tests for comparison of both data sets, like the *t*-test for two dependent samples when the distributions are Gaussian, or the Wilcoxon matched-pairs signed-ranks test when they are not, see e.g. Sheskin (1997). Basically, both tests check whether the differences of the corresponding values from two samples has zero mean within the statistical rigors, that is with a prespecified level of significance. The difference between these test is that in the former one, the parametric test statistic is obtained assuming the Gaussian distribution, while in the latter a nonparametric rank test is used. In this sense, as of now, the nighttime lights inventories provide the best possibility of comparison and improvement of inventories obtained from modeling and disaggregation. One of the problems to be solved here is a mismatch of the map overlay.

Atmospheric tracer measurements have been used for estimation of fossil fuel emissions. However, this way only the dilution of $^{14}CO_2$ in the atmosphere due to fossil fuel combustion is measured. This can be used for detection of changes in fossil fuel emission, as it was outlined in Levin & Rödenbeck (2008). They used the *t*-test statistic for comparison of 5-year mean $^{14}CO_2$ concentrations in the starting and last years of the Kyoto Protocol agreement. An interesting conclusion was, that they were unable to detect a significant difference in the mean using two high precision measurements in the Schauinsland and Heidelberg stations. Whether this was caused by a lack of real significant reduction of the fossil fuel emissions across Europe, or by low sensitivity of the method, is a matter of further studies. Although this method allows for constraining emissions, particularly from bigger areas and/or bigger sources, a transport models have to be used in order to get some reduction in emission uncertainty (see e.g. Rayner et al., 2010). They used an area grid with 0.25^0 resolution, and certainly applying their method to a very high resolution would reveal problems both in bigger computation expenses and much smaller improvement. Here, one of the difficulties is a in proper definition of an area from which the fossil emission come; it is another manifestation of a map overlay problem. It is worth to add that the estimates of fluxes obtained from the atmospheric tracer measurements are much more uncertain than those obtained from inventories. However, the former may be regarded as more *objective*, while the latter as more *subjective*.

Atmospheric inversion for estimating fluxes is another method that can help in comparison of inventory results with independent information. This method has been used mainly for the estimation of CO_2 fluxes from biospheric sources and ocean, assuming that the fossil fuel emissions are exact (i.e. much less uncertain). Knowing rather low uncertainty of biospheric inventory estimates, which are of main interest in our studies, the inversion methods may introduce useful additional information. They can help both in improving the inventory values and in reduction of their uncertainty. As before, the problems stem from scarce measurement sites, and from uncertainty as to the area from which emissions come. That is why these methods were applied to the problems of rough spatial resolution. However, Lauvaux et al. (2008) have been able to get results with the 8 km × 8 km resolution, although in a relatively small region (300 km × 300 km), and with intensive measurement efforts, including aircraft measurements. The inverse methods are quite general and can be used for other greenhouse gases. For example, Thompson et al. (2011) used them in estimating N_2O fluxes. Also other gridded information can be incorporated into inventory estimates. Needless to say, the atmospheric inverse results can be also compared with inventory data using e.g. the earlier mentioned *t*-test or Wilcoxon test. For this, the very high resolution inventory data would have to be aggregated, which makes this idea rather futile.

Probably, the most promising results for comparison and assessment of inventory emissions for the biosphere fluxes can be obtained from flux tower observations. At present however, the flux towers are extremely rare; at the territory of Poland and Ukraine there are no such installations yet. The measurements from a flux tower can be representative at most for a one very high resolution cell. With this respect, this option seems to be more useful for calibration of biosphere emission models used for preparation of inventory, than for significant tests of annual inventories given in high resolution gridded form.

It is evident from the above considerations that some possibilities exist for testing the inventory results, validation of models and methods used, and even for incorporation of additional knowledge to improve the results and their uncertainty. They are mainly in the form of independent estimates of emission fluxes in different spatial scales and different scarcity. There may exist estimates in isolated cells in the fine scale, group of cells, estimates in rough scale or even bigger regions. All this information can be incorporated in the Bayes approach described in the Deliverable D3.3, at available scales. This means that some information is used directly in the fine scale, while other may possibly help in improving rough scale estimates, which are then disaggregated to fine scale, but with improved accuracy.

For completeness of the reasoning, the Bayes procedure is summarized below. Let $\boldsymbol{y}_{\text{obs}}$ be a n-vector of the independent estimates in the considered scale cells, and $\boldsymbol{x}$ is a n-vector of real unknown fluxes (emissions) from these cells. Then we have

$$\boldsymbol{y}_{\text{obs}} = \boldsymbol{x} + \boldsymbol{\psi} \tag{1}$$

where $\boldsymbol{\psi}$ is a n-vector of errors, which is modeled as a random variable with the Gaussian distribution of a covariance matrix $\mathbf{C}_y$

$$p(\boldsymbol{\psi}) = \left[(2\pi)^m \det \mathbf{C}_y\right]^{-1} \exp\left\{-\frac{1}{2}\boldsymbol{\psi}^T \mathbf{C}_y^{-1}\boldsymbol{\psi}\right\} \tag{2}$$

On the other hand, it is assumed that uncertain information $\boldsymbol{x}_{\text{prior}}$ on fluxes is also available, so that

$$\boldsymbol{x} = \boldsymbol{x}_{\text{prior}} + \boldsymbol{\vartheta} \tag{3}$$

where again, uncertainty is modeled as a random vector with the Gaussian distribution independent of $p(\boldsymbol{\psi})$ and having covariance matrix $\mathbf{C}_x$

$$p(\boldsymbol{\vartheta}) = \left[(2\pi)^m \det \mathbf{C}_x\right]^{-1} \exp\left\{-\frac{1}{2}\boldsymbol{\vartheta}^T \mathbf{C}_x^{-1}\boldsymbol{\vartheta}\right\} \tag{4}$$

Now we are looking for the conditional probability $p(\boldsymbol{x}|\boldsymbol{y}_{\text{obs}})$. From the Bayes theory we have

$$p(\boldsymbol{x}|\boldsymbol{y}_{\text{obs}}) = \frac{p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y}_{\text{obs}})} \tag{5}$$

As the Jacobians in the transformations (1) and (3) is equal 1, then

$$p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x}) = \left[(2\pi)^m \det \mathbf{C}_y\right]^{-1} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_{\text{obs}} - \boldsymbol{x})^T \mathbf{C}_y^{-1}(\boldsymbol{y}_{\text{obs}} - \boldsymbol{x})\right\} \tag{6}$$

$$p(\boldsymbol{x}) = [(2\pi)^m \det \mathbf{C}_x]^{-1} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_{\text{prior}})^T \mathbf{C}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_{\text{prior}})\right\} \tag{7}$$

Thus, the conditional probability $p(\boldsymbol{x}|\boldsymbol{y}_{\text{obs}})$ is proportional to

$$p(\boldsymbol{x}|\boldsymbol{y}_{\text{obs}}) \backsim \exp\left\{-\tfrac{1}{2}\left[(\boldsymbol{y}_{\text{obs}}-\boldsymbol{x})^T \mathbf{C}_y^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{x}) + (\boldsymbol{x}-\boldsymbol{x}_{\text{prior}})^T \mathbf{C}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_{\text{prior}})\right]\right\} \tag{8}$$

Now, assuming that it is unique, the value $\hat{\boldsymbol{x}}$ which maximizes the above conditional probability is taken as the estimator. Namely, it is the value which minimizes the following cost function

$$J = (\boldsymbol{y}_{\text{obs}}-\boldsymbol{x})^T \mathbf{C}_y^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{x}) + (\boldsymbol{x}-\boldsymbol{x}_{\text{prior}})^T \mathbf{C}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_{\text{prior}}) \tag{9}$$

The solution can be found analytically. As the matrices $\mathbf{C}_y$ and $\mathbf{C}_x$ are symmetric, the derivative of J with respect to $\boldsymbol{x}$ is

$$\tfrac{1}{2}\frac{dJ}{d\boldsymbol{x}} = -\mathbf{C}_y^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{x}) + \mathbf{C}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_{\text{prior}}) \tag{10}$$

Supposing that the inverted below matrix is nonsingular, the derivative is zero for

$$\hat{\boldsymbol{x}} = \left(\mathbf{C}_y^{-1}+\mathbf{C}_x^{-1}\right)^{-1}\left(\mathbf{C}_y^{-1}\boldsymbol{y}_{\text{obs}} + \mathbf{C}_x^{-1}\boldsymbol{x}_{\text{prior}}\right) =$$

$$= \left(\mathbf{C}_y^{-1}+\mathbf{C}_x^{-1}\right)^{-1}\left[\mathbf{C}_y^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{x}_{\text{prior}}) + \left(\mathbf{C}_y^{-1}+\mathbf{C}_x^{-1}\right)\boldsymbol{x}_{\text{prior}}\right]$$

which finally gives the Bayes estimator of the fluxes

$$\hat{\boldsymbol{x}} = \boldsymbol{x}_{\text{prior}} + \left(\mathbf{C}_y^{-1}+\mathbf{C}_x^{-1}\right)^{-1}\mathbf{C}_y^{-1}(\boldsymbol{y}_{\text{obs}}-\boldsymbol{x}_{\text{prior}}) \tag{11}$$

If the matrix $\mathbf{C}_y^{-1}+\mathbf{C}_x^{-1}$ is singular, then a singular value decomposition (SVD) can be used.

It can be demonstrated that having inserted $\hat{\boldsymbol{x}}$ to (8), one gets a Gaussian distribution. Then, as the expression under the exponent in a Gaussian distribution is quadratic, we can find the inverse of the covariance matrix $\hat{\mathbf{C}}_x$ of the estimator from the second derivative of J. Differentiating (10) gives

$$\frac{1}{2}\frac{d^2J}{d\boldsymbol{x}^2} = \mathbf{C}_y^{-1} + \mathbf{C}_x^{-1}$$

thus

$$\hat{\mathbf{C}}_x = \left(\mathbf{C}_y^{-1} + \mathbf{C}_x^{-1}\right)^{-1} = \mathbf{C}_x - \mathbf{C}_x\left(\mathbf{C}_x + \mathbf{C}_y\right)^{-1}\mathbf{C}_x \tag{12}$$

This matrix allows us to estimate statistical uncertainty of the Bayesian estimator. The most right hand expression in (12) is more convenient for numerical calculations, since only one matrix has to be inverted, while three matrices has to be inverted in the middle expression. In case of big number of observations n, this may give quite a saving in computation time.

An evident difficulty in using independent information is in practically unavoidable problems of incompatible grids, when two maps overlay. This incompatibility introduces errors which can substantially contribute to uncertainty of results obtained from common

processing of such incompatible maps. As of now, the solutions to this problem are under development. The typical approach is to partition the emissions proportionally to the area or, optionally, to some other proxy variables, like population. Usually, a lot of additional knowledge exist, which can be used for more advanced and, at the same time, more accurate allocation of emissions. Sometimes it is available even in a non-numerical form. This type of additional knowledge is difficult for processing, and often it may be even impossible using the probability terms and statistical approach in a strict sense. Consequently, this problem is considered in the sequel by means of the intelligent computation and fuzzy logic approach.

References

Lauvaux T., Uliasz M., Sarrat C., Chevallier F., Bousquet P., Lac C., Davis K.J., Ciais P., Denning A.S., Rayner P.J. (2008) Mesoscale inversion: first results from the CERES campaign with synthetic data. *Atmospheric Chemistry and Physics*, 8:3459-3471.

Levin I., Rödenbeck C. (2008) Can the envisaged reductions of fossil fuel CO_2 emissions be detected by atmospheric observations? *Naturwissenshaften*, 95:203-208, DOI: 10.1007/s00114-007-0313-4.

Oda T., Maksyutov S. (2011) A very high-resolution (1 km × 1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics*, 11:543-556, DOI: 10.5194/acp-11-543-2011.

Rayner P.J., Raupach M.R., Paget M., Peylin P., Koffi E. (2010) A new global gridded dataset if CO_2 emissions from fossil fuel combustion: Methodology and evaluation. *Journal of Geophysical Research*, 115, D19306, DOI: 10.1029/2009JD013439.

Sheskin D.J. (1997) *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, Boca Raton, Florida, USA.

Thompson R.L., Gerbig C., Rödenbeck C. (2011) A Bayesian inversion estimate of N_2O emissions for western and central Europe and the assessment of aggregation error. *Atmospheric Chemistry and Physics*, 11:3443-3458, DOI: 10.5194/acp-11-3443-2011.