

The 3rd International Workshop on Uncertainty in Greenhouse Gas Inventories,
Lviv Ukraine, 22-24 September 2010

Improving resolution of spatial inventory with a statistical inference approach

Joanna Horabik
Zbigniew Nahorski



Systems Research Institute of Polish Academy of Sciences
Laboratory of Computer Modelling



Motivation

Development of spatially distributed GHG inventories crucially depends on availability of highly spatially resolved activity data. For instance, in Poland activity data relevant to GHG emissions at present is available for 17 country regions, with no more accurate spatial resolution. Information of higher spatial resolution can be obtained for some land use and line emission sources which are related to GHG emissions. These are, however, only *proxy* data about activities.

We propose to apply methods of spatial statistics to produce higher resolution activity data, taking advantage of more detailed land use information.

Outline

1. Statistical framework for disaggregation procedure:

- Modelling spatial dependence with a conditionally autoregressive structure
- A link between the coarse and fine grids
- Model estimation and prediction in a fine grid

2. Data example: a quadruple disaggregation of ammonia inventory emissions in Pomorskie province (Poland) based on the CORINE land use map.



Model specification in a *fine* grid

- Y_i - a random variable associated with a (missing) value of interest y_i in a *fine* grid

$$Y_i | \mu_i \sim \text{Gau}(\mu_i, \sigma_Y^2), \quad i = 1, \dots, n$$

- Modelling μ_i - available covariates explain part of the spatial pattern in observations, and the remaining part is captured through **the conditional autoregressive CAR** structure.

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

w_{ij} - the adjacency weights: $w_{ij}=1$ if j is a neighbour of i and 0 otherwise; $w_{ii}=0$

$w_{i+} = \sum_j w_{ij}$ - the number of neighbours of area i

\mathbf{x}_i - covariates of area i

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$ - regression coefficients

τ^2 - a variance parameter.

- The joint probability distribution of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$

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

\mathbf{X} - the (design) matrix with vectors \mathbf{x}_i

\mathbf{D} - a diagonal matrix with $[\mathbf{D}]_{ii} = w_{i+}$

$[\mathbf{W}]_{ij} = w_{ij}$ - a matrix with adjacency weights



Model specification in a coarse grid

- Equivalently, we can write (*) as

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \text{Gau}_n(\mathbf{0}, \mathbf{N}) \quad (**)$$

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

- The model for a coarse grid (aggregated) data is obtained by multiplication of (**) with an **aggregation matrix** \mathbf{C} consisting of 0's and 1's, indicating which cells have to be aggregated together

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

where N is a number of observations in a coarse grid.

- We treat the random variable $\boldsymbol{\lambda} = \mathbf{C}\boldsymbol{\mu}$ as 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

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

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



Estimation

- The joint unconditional distribution of \mathbf{Z}

$$\mathbf{Z} \sim \text{Gau}_N(\mathbf{CX}\boldsymbol{\beta}, \mathbf{M} + \mathbf{CNC}^T)$$

where $\mathbf{M} = \sigma_z^2 \mathbf{I}_N$.

- The log likelihood associated with \mathbf{Z}

$$L(\boldsymbol{\beta}, \sigma_z^2, \tau^2, \rho) = -\frac{1}{2} \log |\mathbf{M} + \mathbf{CNC}^T| - \frac{N}{2} \log(2\pi) - \frac{1}{2} (\mathbf{z} - \mathbf{CX}\boldsymbol{\beta})^T (\mathbf{M} + \mathbf{CNC}^T)^{-1} (\mathbf{z} - \mathbf{CX}\boldsymbol{\beta})$$

- With fixed σ_z^2, τ^2 and ρ , the log likelihood is maximised for

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

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

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

Further maximisation of $L(\sigma_z^2, \tau^2, \rho)$ is performed numerically.

Prediction



- Y_0 - random variable associated with missing emission values in a **fine** grid, and having the mean μ_0

- Assume

$$Y_0 | \mu_0 \sim \text{Gau}(\mu_0, \sigma_Y^2)$$
$$\mu_0 | \mu_{j, j \neq 0} \sim \text{Gau}\left(\mathbf{x}_0^T \boldsymbol{\beta} + \rho \sum_{j \neq 0} \frac{w_{0j}}{w_{0+}} (\mu_j - \mathbf{x}_j^T \boldsymbol{\beta}), \frac{\tau^2}{w_{0+}}\right), \quad j = 1, \dots, n$$

- The predictor of Y_0 , that is optimal in terms of the minimum mean squared error, is given by $E(Y_0 | \mathbf{z})$.
- First, we calculate the distribution of $\boldsymbol{\mu} / \mathbf{z}$ based on the distributions of $\boldsymbol{\mu}, \mathbf{Z} | \mathbf{C}\boldsymbol{\mu}, \mathbf{Z}$
$$\boldsymbol{\mu} / \mathbf{z} \sim \text{Gau}_n(\mathbf{W}\mathbf{V}, \mathbf{W})$$

$$\mathbf{W} = (\mathbf{C}^T \mathbf{M}^{-1} \mathbf{C} + \mathbf{N}^{-1})^{-1}$$

$$\mathbf{V} = \mathbf{C}^T \mathbf{M}^{-1} \mathbf{z} + \mathbf{N}^{-1} \mathbf{X}\boldsymbol{\beta}$$



- Next, we develop the predictor $E(Y_0 | \mathbf{z})$

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

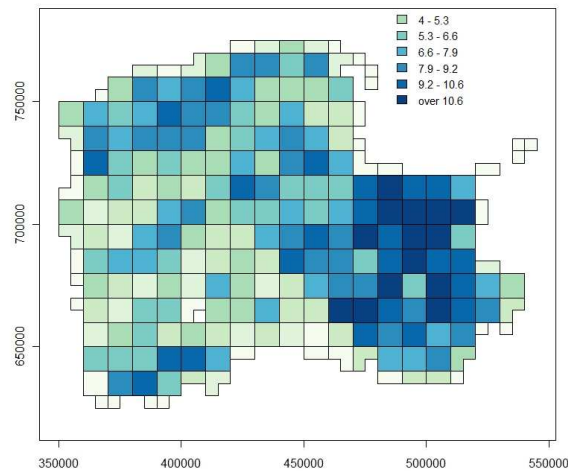
- To calculate the rightmost expectation, we recall that $E(\boldsymbol{\mu} | \mathbf{z}) = \mathbf{WV}$. Denoting the j th element of the vector \mathbf{WV} with l_j , we get the predictor in the form

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



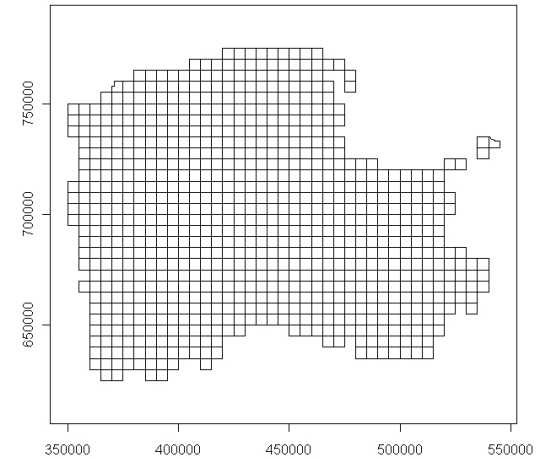
Case study

Inventory of NH₃ emissions from fertilization (in tonnes per year) reported for the Pomorskie voivodship and available in a 10km×10km grid



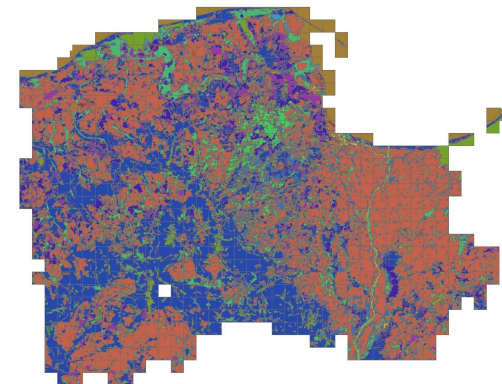
To be disaggregated into

5km×5km grid



Making use of

CORINE land cover map





... But this is just an exercise to *examine performance of the disaggregation procedure*, so we actually know the true emissions in a 5km×5km grid. Thus,

1. we fit the model to a *coarse* data and predict NH₃ emissions for a *fine* grid;
2. we check these results with true inventory emissions of a 5km×5km grid.

CORINE land use information



For each grid cell we calculate area of these land use classes, which can be related to NH_3 emissions. Considered CORINE classes :

- Non-irrigated arable land (211)
- Fruit tree and berry plantations (222)
- Pastures (231)
- Complex cultivation patterns (242)
- Principally agriculture, with natural vegetation (243)

We will examine models with all the above classes (set 1), and compare the results with models including only Non-irrigated arable land and Complex cultivation patterns (set 2).

Secondly, we compare a linear regression with independent (iid) errors vs. spatially correlated errors modelled by the CAR process.

Results



Table 1. Maximum likelihood estimates

	CAR1		LM1		CAR2		LM2	
	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.	Est.	Std.Err.
β_0	-	-	-	-	0.376	9.27e-02	0.452	5.45e-02
β_1	1.13e-07	3.26e-09	1.09e-07	2.46e-09	1.08e-07	5.27e-09	9.58e-08	4.43e-09
β_2	2.55e-07	1.94e-07	4.48e-07	1.97e-07	-	-	-	-
β_3	9.77e-08	1.19e-08	1.08e-07	1.07e-08	-	-	-	-
β_4	1.17e-07	2.13e-08	1.21e-07	1.76e-08	1.22e-07	2.89e-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	0.112	1.95	0.184
τ^2	0.536	0.082	-	-	0.807	0.122	-	-
ρ	0.948	9.98e-04	-	-	0.972	1.74e-04	-	-



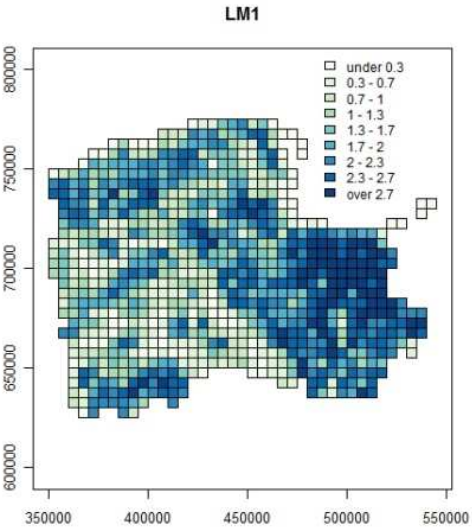
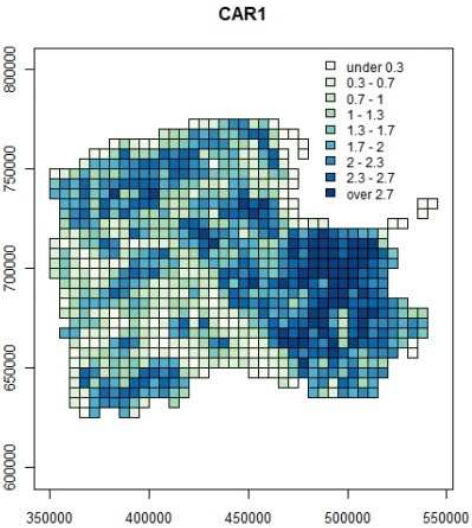
- The models are compared using the Akaike Information Criterion (AIC)
 $AIC = -2L + 2p$
with p denoting a number of parameters.

Lower the AIC, the better a model is.

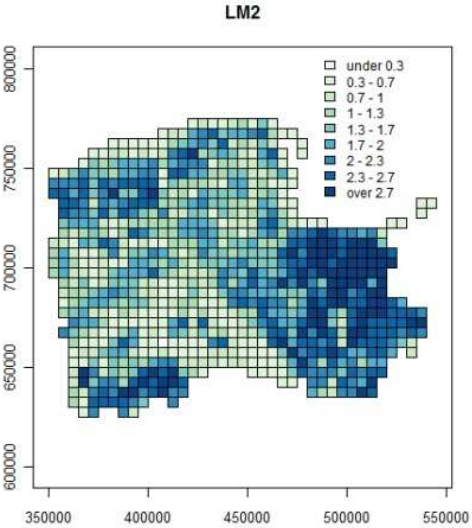
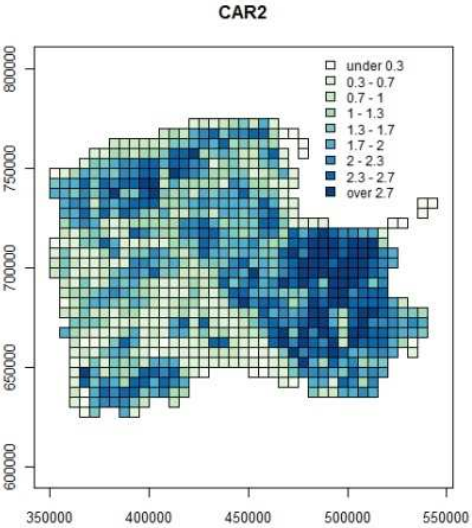
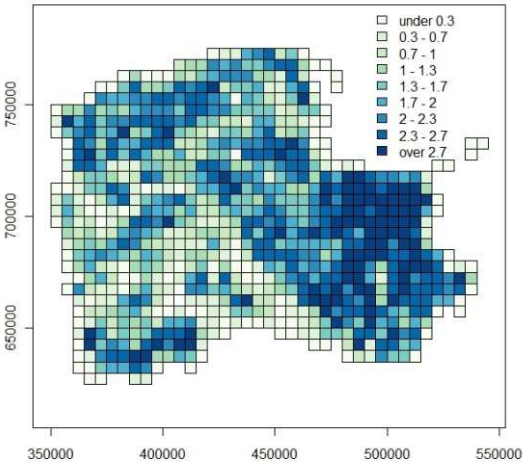
Model	$-L$	AIC
CAR1	312.3	640.7
LM1	336.5	685.1
CAR2	365.4	742.8
LM2	394.8	797.6



Prediction in a *fine* grid



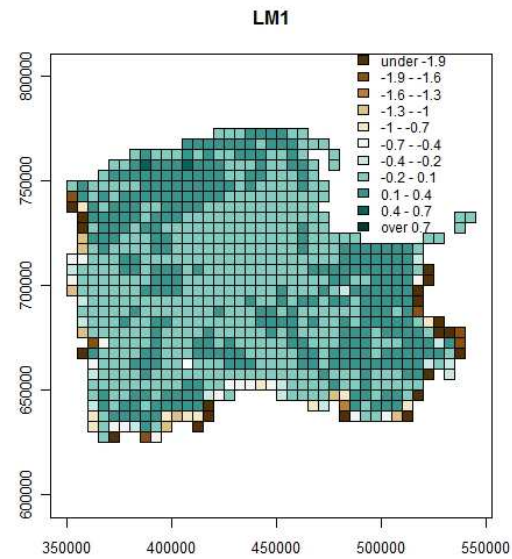
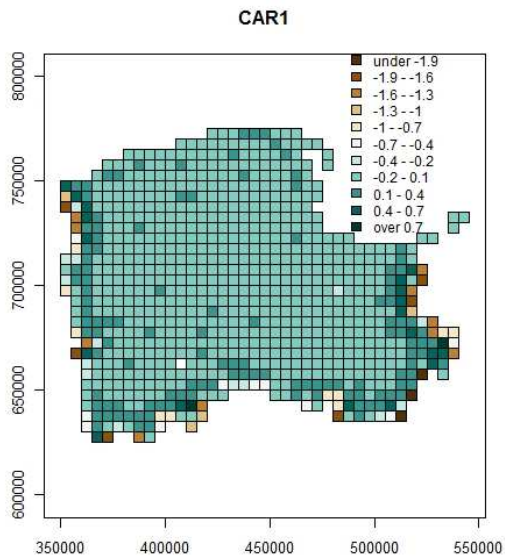
True NH₃ emission



The difference is indistinguishable.

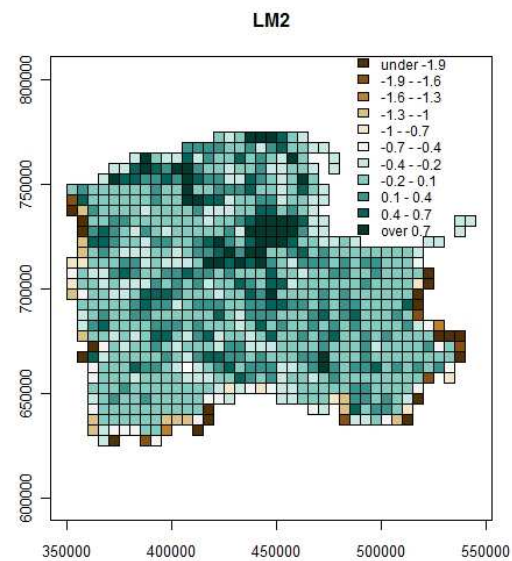
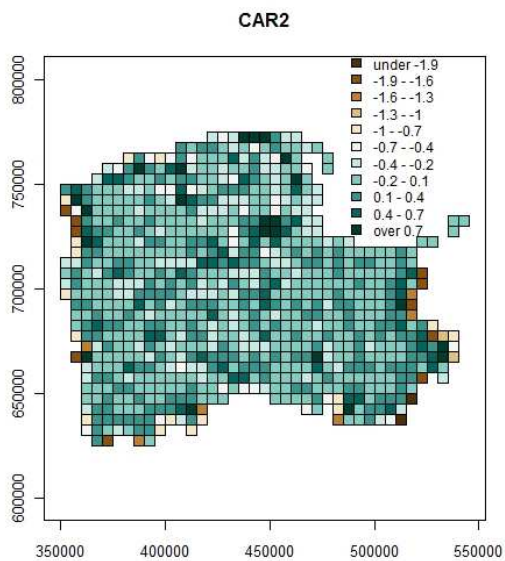


Residuals from predicted values



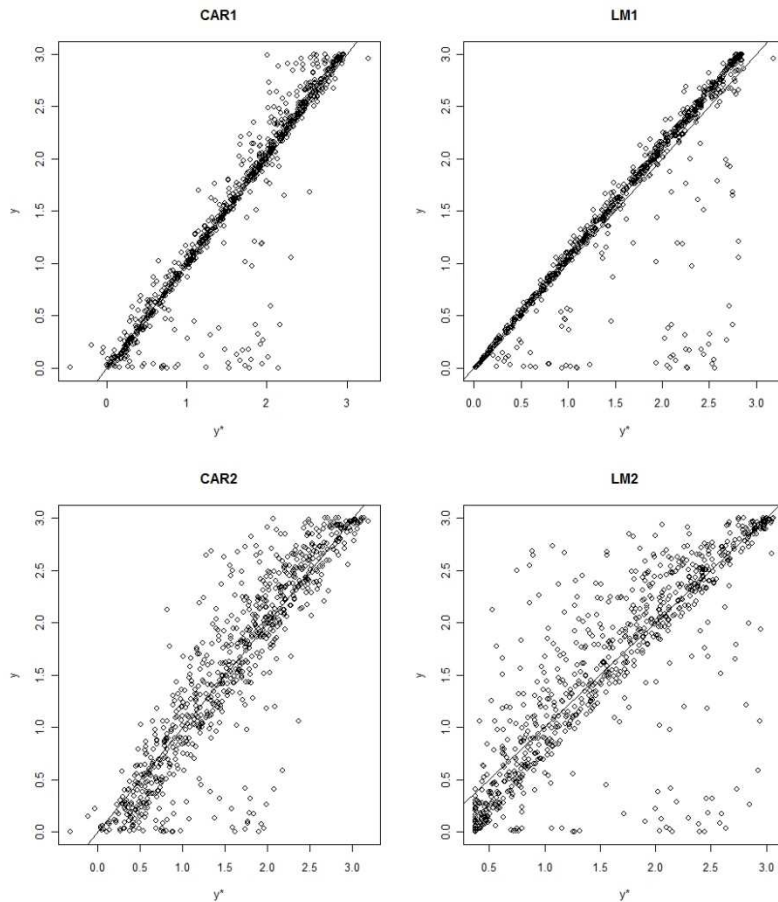
$$d_i = y_i - y_i^*$$

y_i - observation
 y_i^* - prediction





Further analysis of residuals



The mean squared error

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

Model	<i>mse</i>	$\min(d_i)$	$\max(d_i)$	<i>r</i>
CAR1	0.102	-2.144	0.989	0.937
LM1	0.188	-2.562	0.433	0.882
CAR2	0.158	-1.917	1.362	0.901
LM2	0.291	-2.498	1.765	0.808

$$(d_i = y_i - y_i^*)$$

Concluding remarks



- The objective of this study was to demonstrate how a variable of interest (available in a coarse grid) plus information on some related covariates (available in a finer grid) can be combined together to provide the *variable of interest in a finer grid*.
- Performance of the proposed framework depends on
 - explanatory power of covariates available in a fine grid
 - strength of a spatial dependence within the predicted value
 - extent of disaggregation.
- In our simulation study we used original data in a fine grid to assess quality of resulting predictions. For the purpose of future applications, the proposed disaggregation framework should be completed with a *measure of prediction error*.
- Future work devoted to a case of greenhouse gas inventories...