International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Joanna Horabik-Pyzel, Nadia Charkovska, Olha Danylo, Zbigniew Nahorski, Rostyslav Bun

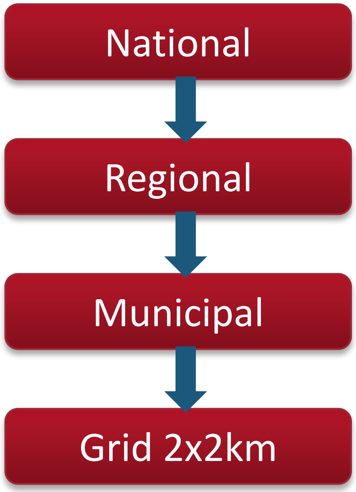
Conditionally autoregressive model for spatial disaggregation of activity data in GHG inventory: Application for agriculture sector in Poland

4th International Workshop on Uncertainties in Atmospheric Emissions
Kraków, 7-9 October 2015

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.

Goal

Application of **statistical spatial scaling methods** to produce higher resolution activity data



Disaggregation framework

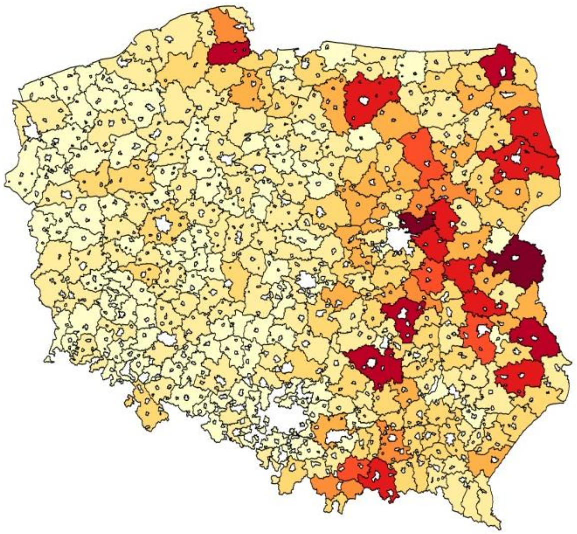
Classification of inventory sectors

Energy (fossil fuel burning)
- power/heat production
- residential
- transport
- industry and construction
- others

Industry (chemical processes)
Agriculture

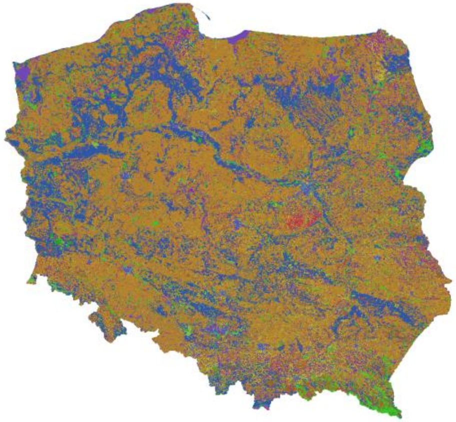
Task

Livestock data available
in ***disticts***



to be disaggregated
into ***municipalities,***

making use of detailed
land cover map.



CAR model for areal data

- **Conditionally autoregressive (CAR)** formulation of process $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$

$$\theta_i \mid \theta_{j, j \neq i} \sim Gau\left(\rho \sum_{j \neq i} \frac{w_{ij}}{w_{i+}} \theta_j, \frac{\tau^2}{w_{i+}}\right), \qquad i, j = 1, \ldots, n$$

w_{ij} - neighbour weights: *1* for neighbour, *0* otherwise

$w_{i+} = \sum_j w_{ij}$ - number of neighbours of cell *i*

τ^2- variance parameter

- Joint probability distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$

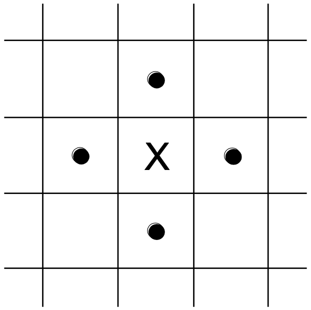
$$\boldsymbol{\theta} \sim Gau_n\left(\mathbf{0}, \tau^2 (\mathbf{D} - \rho \mathbf{W})^{-1}\right)$$

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

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

τ^2 - a variance parameter

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



X - Cell *i*

● - Neighbours of cell *i* ($w_{ij}=1$)
So, $w_{i+}=4$

Disaggregation model → Specification in a *fine* grid

Y_i - random variable associated with emissions in a ***fine*** grid

$$Y_i \mid \mu_i \sim Gau\left(\mu_i, \sigma_Y^2\right), \quad i = 1,\ldots,n$$

- Modelling $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^T$

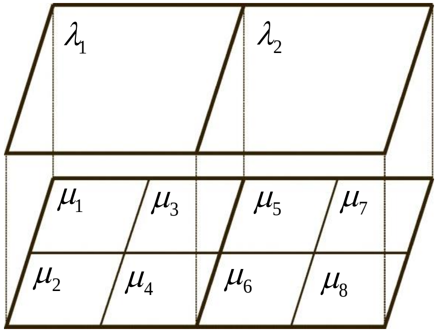
$$\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{\theta}, \qquad \boldsymbol{\theta} \sim Gau_n(\mathbf{0}, \boldsymbol{N}) \qquad (*)$$

available covariates of cell *i* ← $\boldsymbol{X\beta}$; $\boldsymbol{\theta}$ → **spatial CAR structure**

$\boldsymbol{X}$ – (design) matrix with covariates

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

→ Specification in a *coarse* grid



- The model for a ***coarse*** grid: multiplication of (*) with **aggregation matrix** $\boldsymbol{C}_{N\times n}$ indicating which cells are aligned together

$$\boldsymbol{C\mu} = \boldsymbol{CX\beta} + \boldsymbol{C\theta} \qquad \boldsymbol{C\theta} \sim Gau_N\left(\mathbf{0}, \boldsymbol{CNC}^T\right)$$

N - # of observations in a *coarse* grid

n – # of observations in a *fine* grid

- $\boldsymbol{\lambda} = \boldsymbol{C\mu}$ - the mean process for random variables $\boldsymbol{Z} = (Z_1,\ldots,Z_N)^T$ of the ***coarse*** grid

$$\boldsymbol{Z} \mid \boldsymbol{\lambda} \sim Gau_N\left(\boldsymbol{\lambda}, \sigma_Z^2 \boldsymbol{I}_N\right)$$

Estimation

- The joint unconditional distribution of $\boldsymbol{Z}$

$$\boldsymbol{Z} \sim Gau_N\left(\boldsymbol{CX\beta}, \boldsymbol{M} + \boldsymbol{CNC}^T\right)$$

where $\boldsymbol{M} = \sigma_Z^2 \boldsymbol{I}_N$, $\boldsymbol{N} = \tau^2\left(\boldsymbol{D} - \rho\boldsymbol{W}\right)^{-1}$

- Maximum likelihood estimation based on the joint distribution of $\boldsymbol{Z}$.
- Analytical solution for $\boldsymbol{\beta}$, further maximisation performed numerically.
- Expected Fisher information matrix used to get standard errors of parameters.

Prediction in a *fine* grid

- The process $\boldsymbol{\mu}$ underlying emission inventory is of our primary interest, with the optimal predictor given by $E(\boldsymbol{\mu}|\boldsymbol{z})$.
- The joint distribution of $(\boldsymbol{\mu}, \boldsymbol{Z})$

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

which yield both the predictor $\hat{\boldsymbol{\mu}} = E\left(\boldsymbol{\mu} \mid \boldsymbol{z}\right)$ and its error $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\mu}}^2 = Var\left(\boldsymbol{\mu} \mid \boldsymbol{z}\right)$

- The joint distribution of $(\boldsymbol{\mu}, \boldsymbol{Z})$ allows us to see the approach in analogy to *block kriging*.

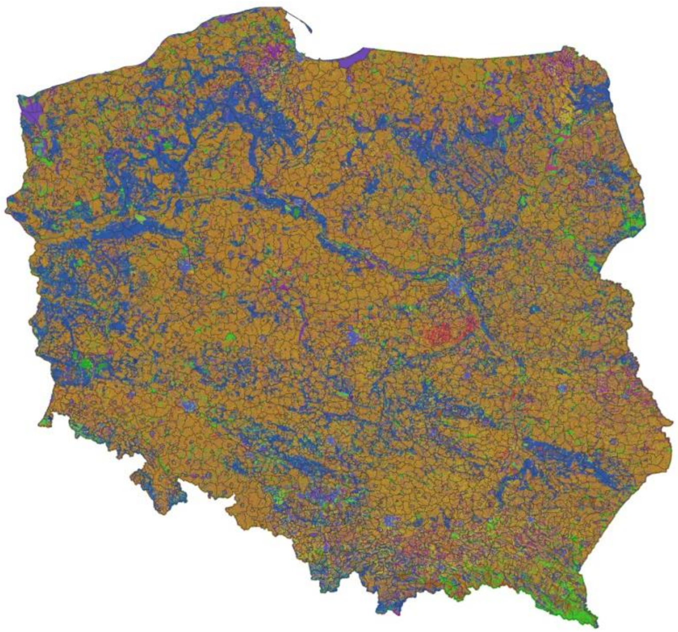
Case study

- **Livestock** (cattle, pigs, horses etc.) data based on agricultural census 2010
- Disaggregation from 314 districts (*powiaty*) into 2171 municipalities *(gminy*) needed
- Only **rural municipalities** considered

→ For **horses**, data are available also in municipalities, which enabled verification of the method.

Explanatory variables

- **Population**
- **Considered CORINE classes**
  - Pastures (231)
  - Complex cultivation patterns (242)
  - Principally agriculture, with natural vegetation (243)

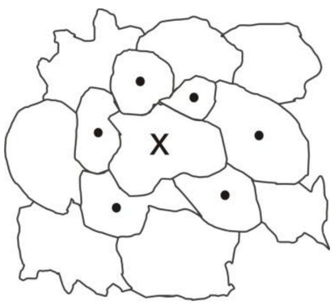


For each municipality, we calculate the area of these land use classes, that can be related to livestock farming.

Livestock data available in *districts*

Fit model **CAR**
Prediction in *municipalities*

Regression & spatial effect



Fit model **LM**
Prediction in *municipalities*

Regression based on population & land use
No spatial effect

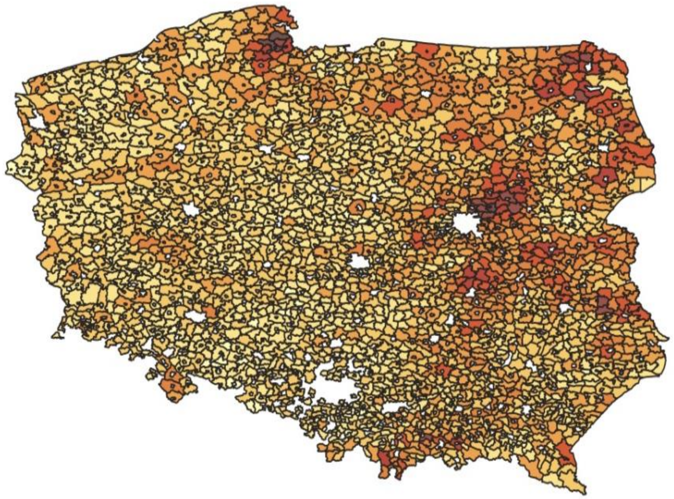
Fit model **NAIVE**
Prediction in *municipalities*

Solely in proportion to population

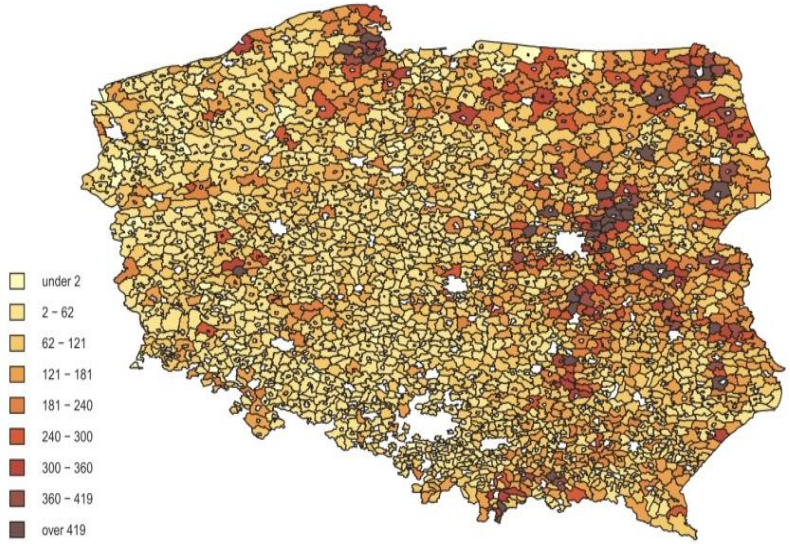
VERIFICATION: Data on horses available in *municipalities*

Results: Prediction for *municipalities*

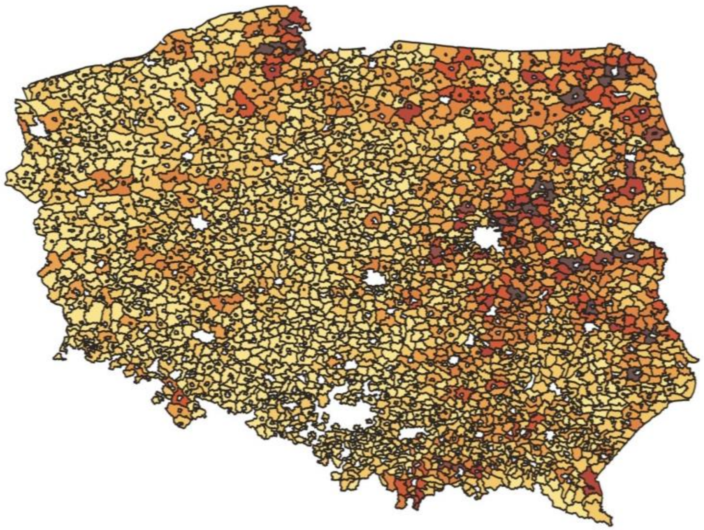
Model CAR



DATA

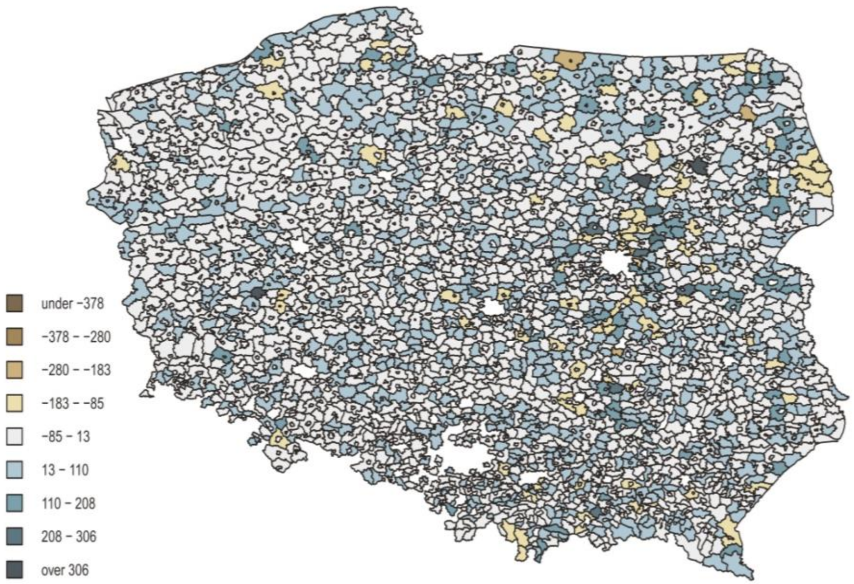


NAIVE prediction – proportional to population

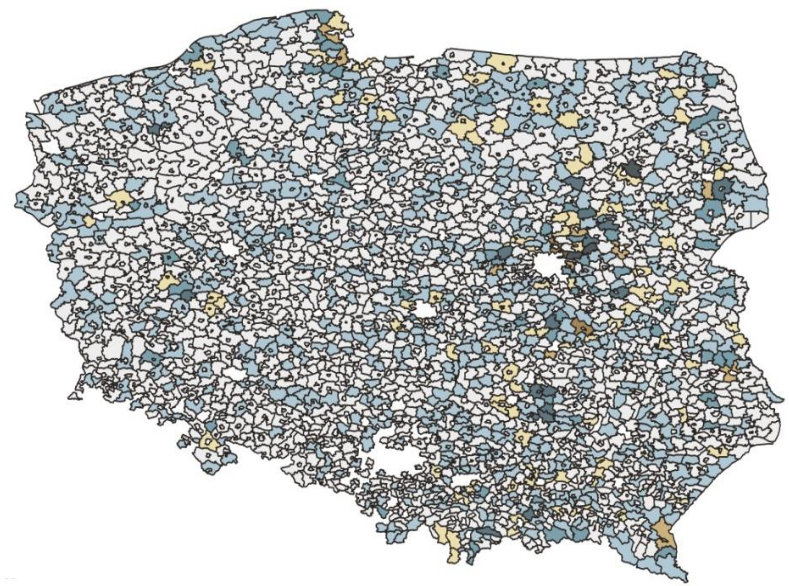


Residuals from predicted values

Model CAR



NAIVE prediction

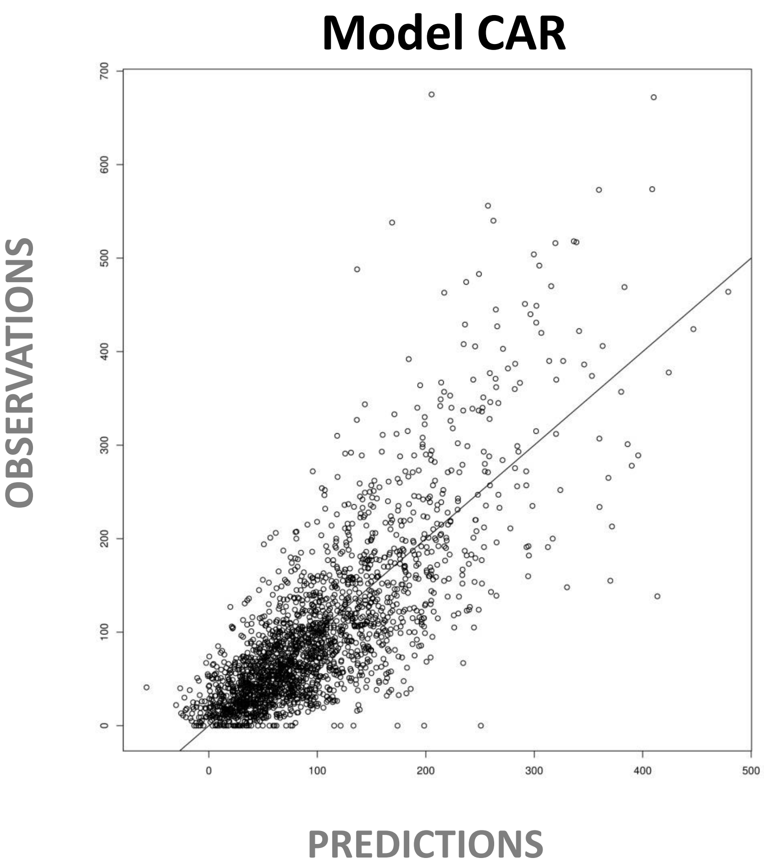


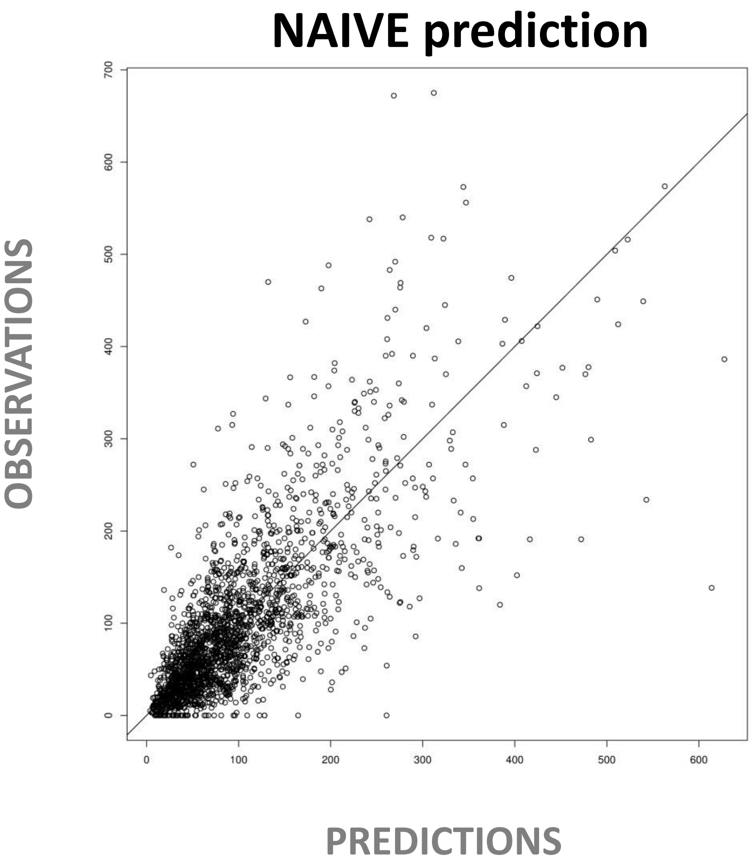
$d_i = y_i - y_i^*$

y_i - data
y_i^* - prediction

	MSE	Avg$\|d_i\|$	Min(d_i)	Max(d_i)	r
CAR	3069.4	38.37	-275	469	0.784
Naive	3374.4	38.17	-475	403	0.766
LM	5641.2	51.28	-357	522	0.555

Scatterplots of data vs. predictions





Modification #1: Various regression models in regions

$l=1,\ldots,L$ index regions, e.g. voivodships, $n = \sum_{l=1}^{L} n_l$

X* - block diagonal matrix of covariates

Separate sets of regression coefficients and variance $\sigma^2_{Y,l}$ for each voivodship

$$
X^* = \left[\begin{array}{cccc|c|cccc}
1 & x_{11}^1 & \cdots & x_{1k}^1 & & & & & \\
\vdots & & \ddots & \vdots & & & & & \\
1 & x_{n_1 1}^1 & & x_{n_1 k}^1 & & & & & \\
\hline
 & & & & \ddots & & & & \\
\hline
 & & & & & 1 & x_{11}^L & \cdots & x_{1k}^L \\
 & & & & & & & \ddots & \vdots \\
 & & & & & 1 & x_{n_L 1}^L & & x_{n_L k}^1
\end{array}\right]
\qquad
\beta^* = \left[\begin{array}{c}
\beta_0^1 \\ \vdots \\ \beta_k^1 \\ \hline \vdots \\ \hline \beta_0^L \\ \vdots \\ \beta_k^L
\end{array}\right]
$$

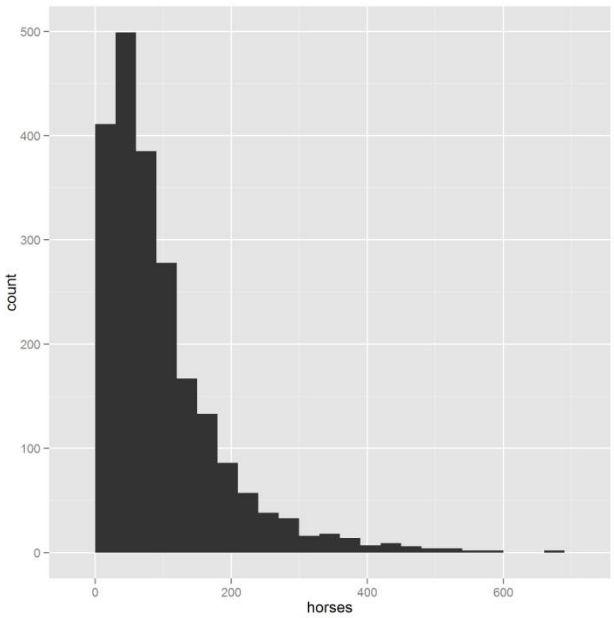
$$\boldsymbol{\mu} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

$$\boldsymbol{\epsilon} \sim Gau_n(\boldsymbol{0}, \boldsymbol{N})$$

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

	MSE	Avg$|d_i|$	Min(d_i)	Max(d_i)	r
CAR	3069.4	38.37	-275	469	0.784
CAR*	3124.9	38.99	-256	446	0.783
Naive	3374.4	38.17	-475	403	0.766
LM	5641.2	51.28	-357	522	0.555

Modification #2: Accounting for data skeweness



Distributions of activity data (here: horses) are highly right skewed → the assumption of normality should be revised.

Potential approaches

- log transformation
- truncated normal distribution
- trans-Gaussian kriging
- ...

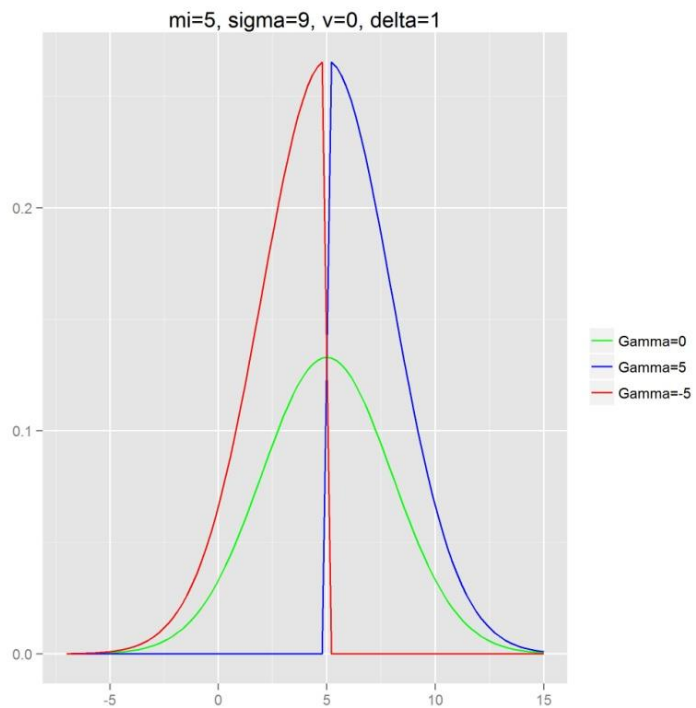
However, none of the listed options support summation of random variables
→ this is required in the developed model of spatial scaling (aggregation matrix $\boldsymbol{C}_{N \times n}$)

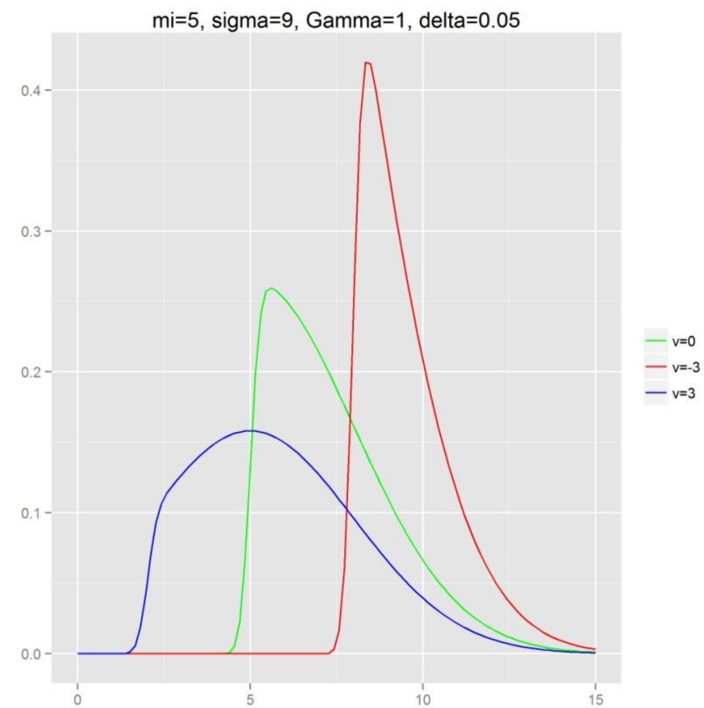
Closed skew normal (CSN) distribution

- Introducing skewness to the normal distribution, while the distribution is closed under marginalisation and conditioning (Dominiguez-Molina et al. 2003)
- Density of multivariate CSN distribution: $\boldsymbol{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, \Gamma, \nu, \Delta)$

$$f_{p,q}(\boldsymbol{y} \mid \boldsymbol{\mu}, \Sigma, \Gamma, \nu, \Delta) = K \phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) \Phi_q(\Gamma(\boldsymbol{y} - \boldsymbol{\mu}); \nu, \Delta)$$

where $K = \Phi_q^{-1}(0; \Delta + D\Sigma D^T)$, $\phi_p(\cdot)$ - standard normal pdf, $\Phi_q(\cdot)$ - standard normal cdf





Properties of closed skew normal distribution

- Closed under *linear transformation* (Gonzalez-Farias et al. 2004)
  Let $\boldsymbol{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, \Gamma, \nu, \Delta)$ and $A_{n \times p}$. Then

$$A\boldsymbol{Y} \sim CSN_{n,q}(\boldsymbol{\mu}_A, \Sigma_A, \Gamma_A, \nu, \Delta_A)$$

  → important for the disaggregation method

- *Conditioning property* → important for estimation (Gibbs sampler within MCMC)

Concluding remarks

We propose a novel approach for allocation of activity data to finer grids, conditional on available covariate information.

Concluding remarks

We propose a novel approach for allocation of activity data to finer grids, conditional on available covariate information.

- Good results for livestock activity data of **agricultural sector**, but very limited e.g. in residential sector (natural gas consumption in households)

The approach is applicable for **AREA emission sources** which are **spatially correlated**.

Concluding remarks

We propose a novel approach for allocation of activity data to finer grids, conditional on available covariate information.

- Good results for livestock activity data of **agricultural sector**, but very limited e.g. in residential sector (natural gas consumption in households)

The approach is applicable for **AREA emission sources** which are **spatially correlated**.

- The method is feasible for disaggregation from *districts* to *municipalities*, but not from *voivodeships* to *municipalities*.

Concluding remarks

We propose a novel approach for allocation of activity data to finer grids, conditional on available covariate information.

- Good results for livestock activity data of **agricultural sector**, but very limited e.g. in residential sector (natural gas consumption in households)

The approach is applicable for **AREA emission sources** which are **spatially correlated**.

- The method is feasible for disaggregation from *districts* to *municipalities*, but not from *voivodeships* to *municipalities*.

- Comparison with the *naive (proportional) disaggregation*:
  In the case study, 9% improvement in terms of the *mean squared error*.

In general, it provides the assessment of the **significance of regression coefficients** and **uncertainty of calculated values**.

TAKE HOME MESSAGE:

A ***structure of dataset*** can give us an opportunity to develop an improved / alternative modeling approach, and thus to provide a better insight.

International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

TAKE HOME MESSAGE:

A ***structure of dataset*** can give us an opportunity to develop an improved / alternative modeling approach, and thus to provide a better insight.

Thank you for your attention!

International Institute for Applied Systems Analysis
www.iiasa.ac.at