Preferred citation style for this presentation

Axhausen, K.W. (2007) Problems with errors: A brief introduction, Englishseminar 2007, Schliersee, February 2007.

Problems with errors: A brief introduction

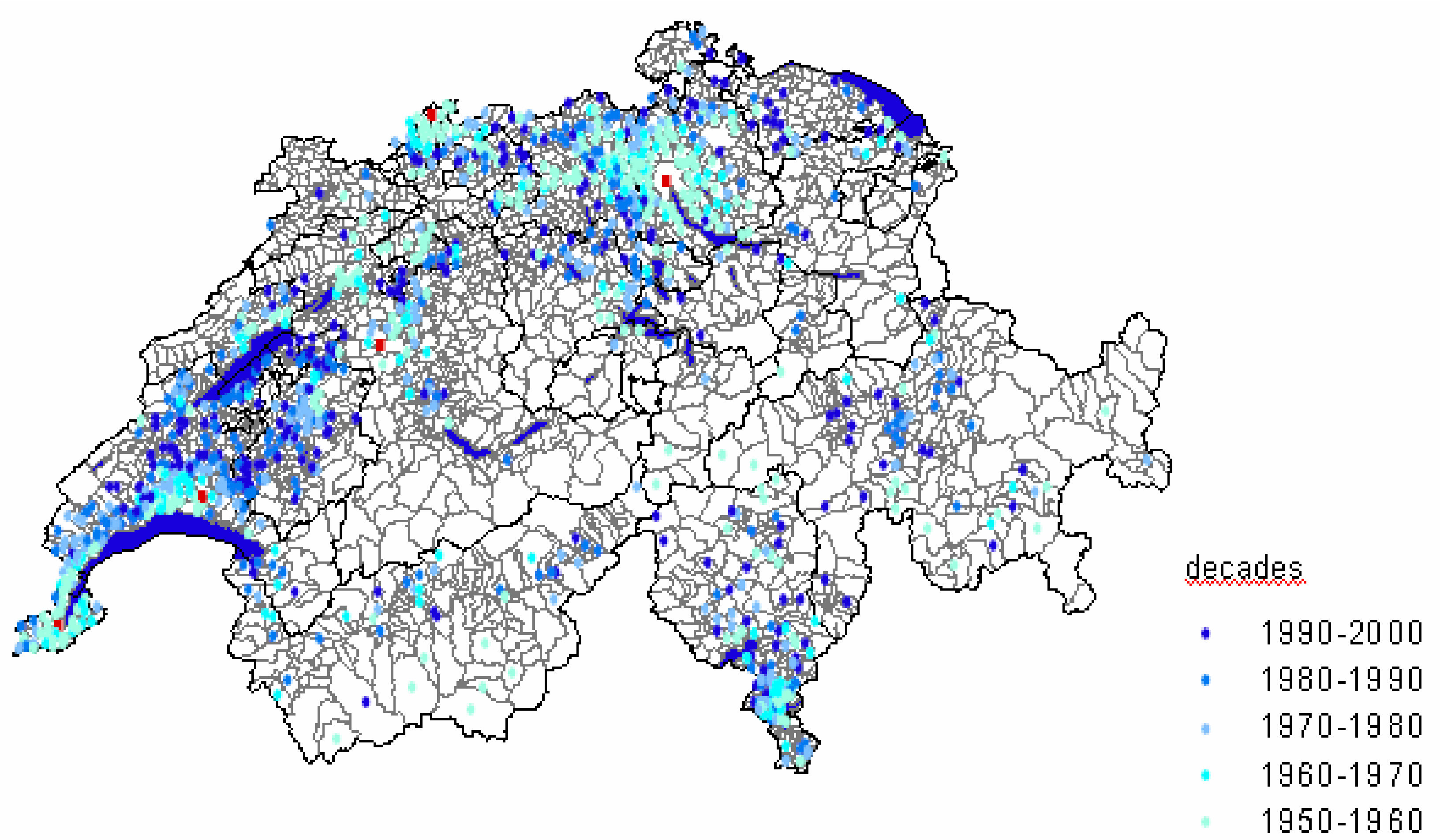
KW Axhausen

IVT
ETH
Zürich

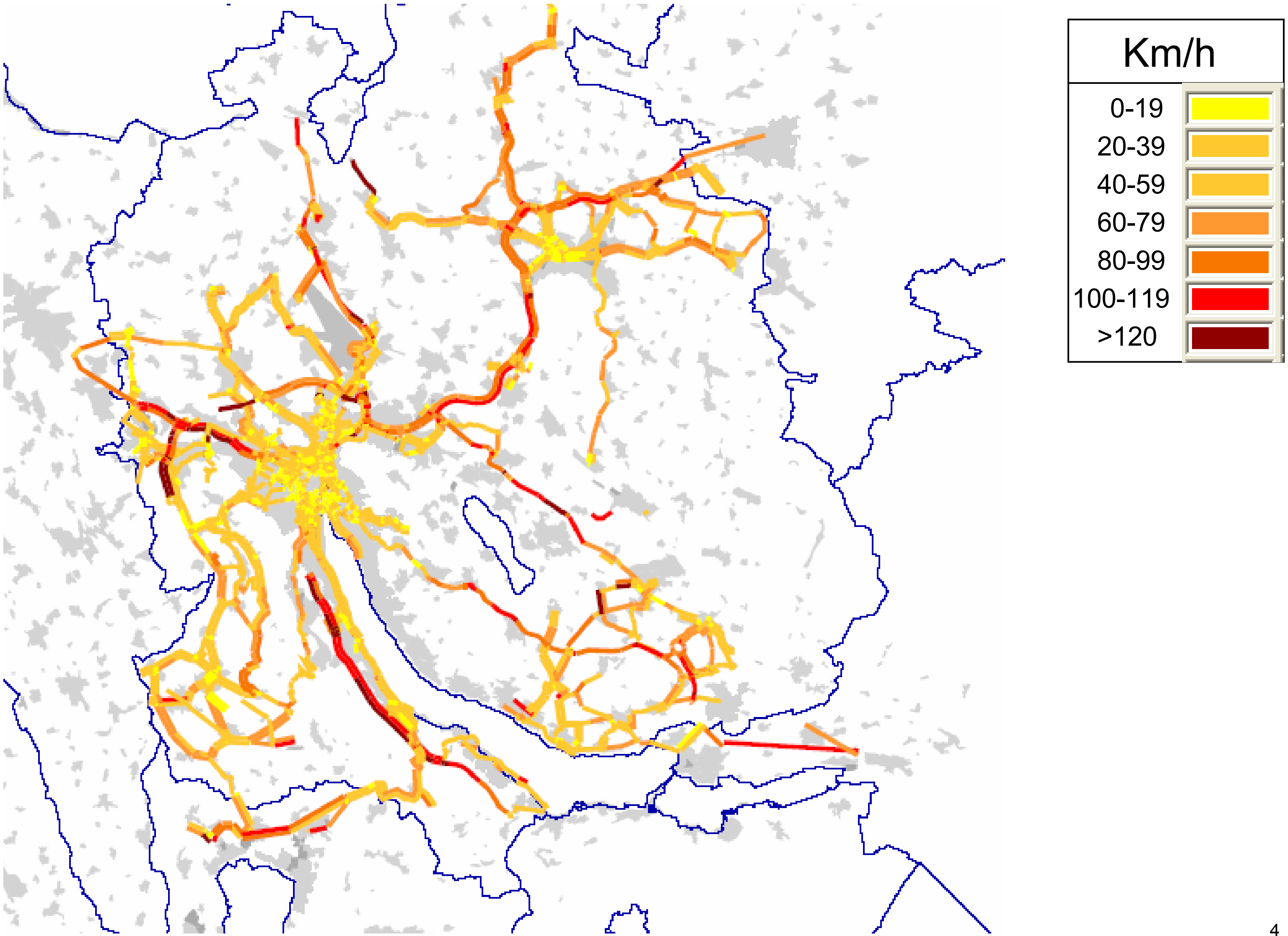
February 2007

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Example: Population growth of Swiss municipalities



Example: Link speeds (Kanton Zürich)



Software

Hierarchial (multilevel) models:

- MLWin
- Any software, which estimates “mixed models” (SAS, GLIM, LIMDEP, etc.)

Spatial error and lag models:

- S or R (integrated into ArcGIS)
- GeoDA
- LeSage’s Econometric Toolbox for MatLab

Starting point

OLS assumes:

$$y = X\beta + \varepsilon$$
$$\varepsilon \sim iid\ N(0,\sigma)$$

y	Dependent variable
β	Vector of parameters
X	Matrix of independent variables
ε	Error
σ	Variance of the error

What can go wrong ?

Heteroscedacity 1

$$\varepsilon \sim \hat{y}$$

Heteroscedacity 2

$$\varepsilon \sim x$$

Collinearity

$$\operatorname{cov}(x_i, x_j) \neq \begin{Bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{Bmatrix}$$

What else can go wrong ?

Spatial or temporal vicinity

$$\mathrm{cov}(\varepsilon_n, \varepsilon_m) \neq \begin{Bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{Bmatrix}$$

Hierarchical regression (Simplest 2-level model)

$$y_{ij} = \beta_{0ij} x_0 + \beta_{1ij} x_{1ij}$$

with:

$$\beta_{0ij} = \underbrace{\beta_0}_{\text{fixed part}} + \underbrace{u_{0j} + \varepsilon_{0ij}}_{\text{random part}}$$

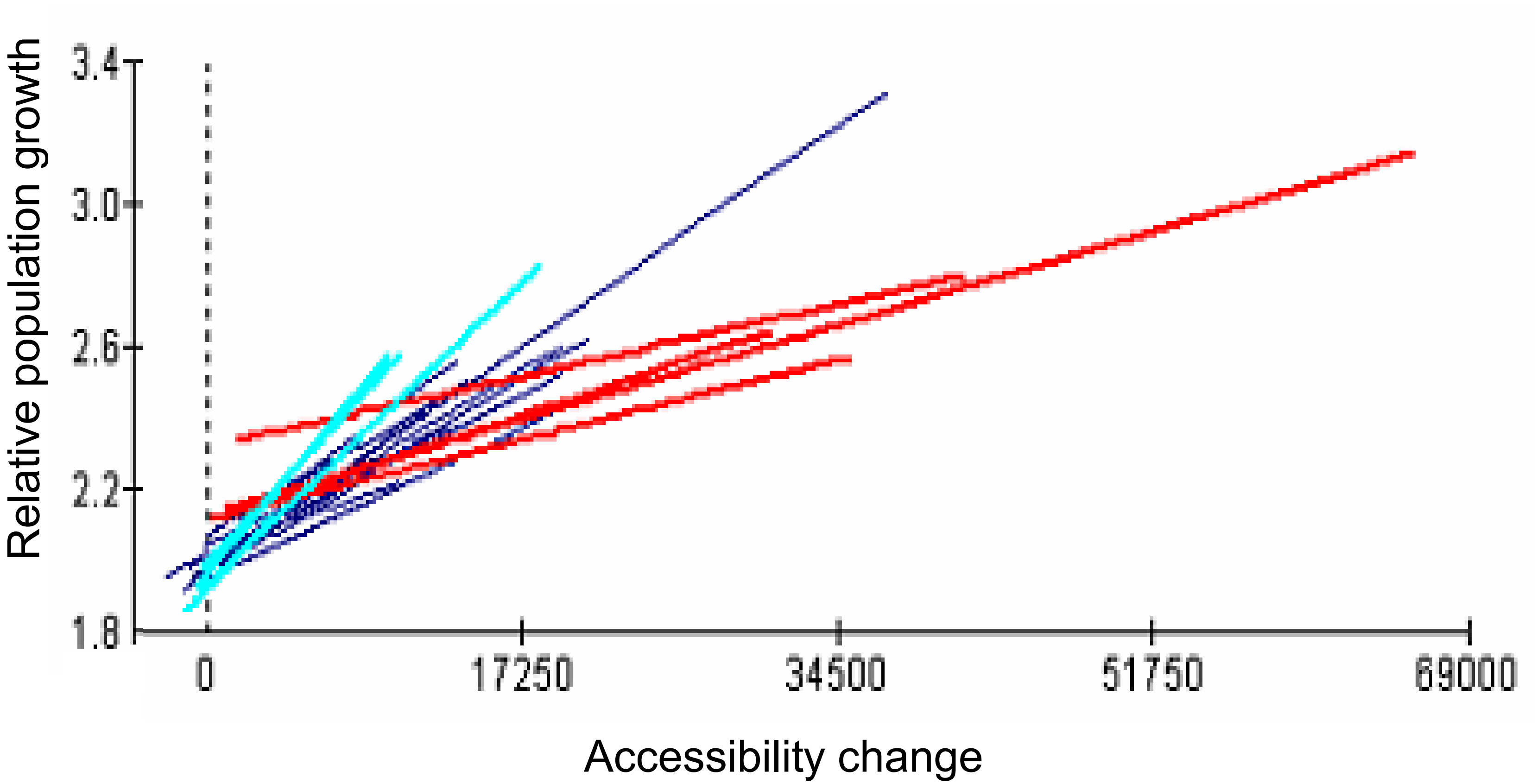
and:

$$\beta_{1ij} = \underbrace{\beta_1}_{\text{fixed part}} + \underbrace{u_{1j} + \varepsilon_{1ij}}_{\text{random part}}$$

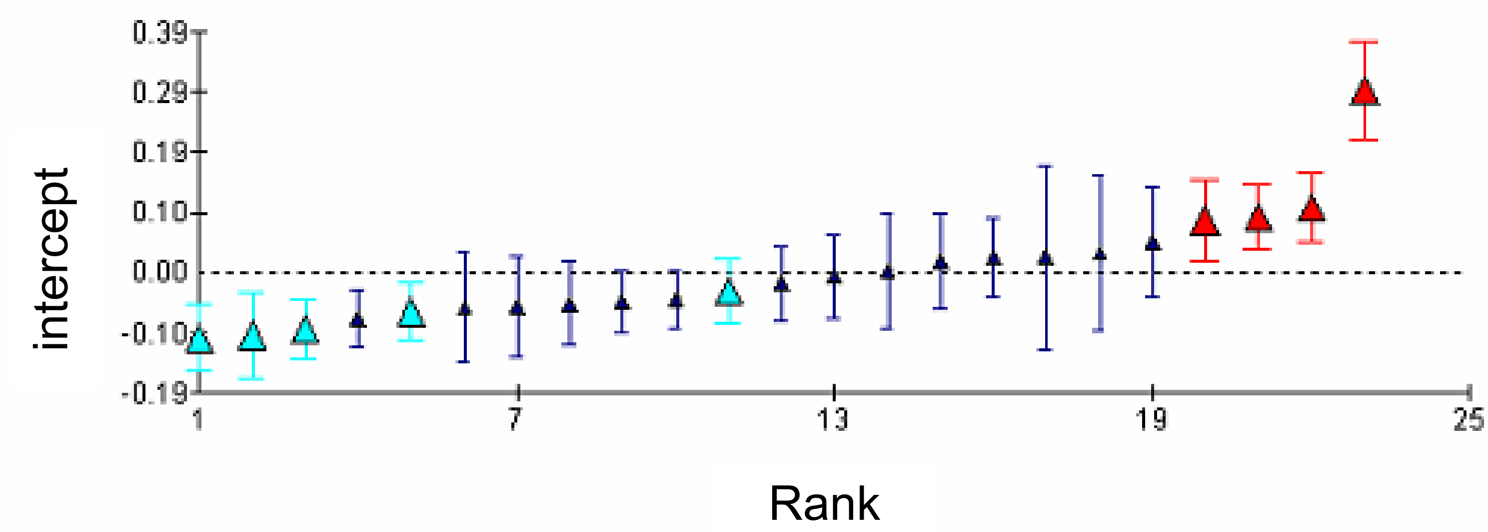
Example:

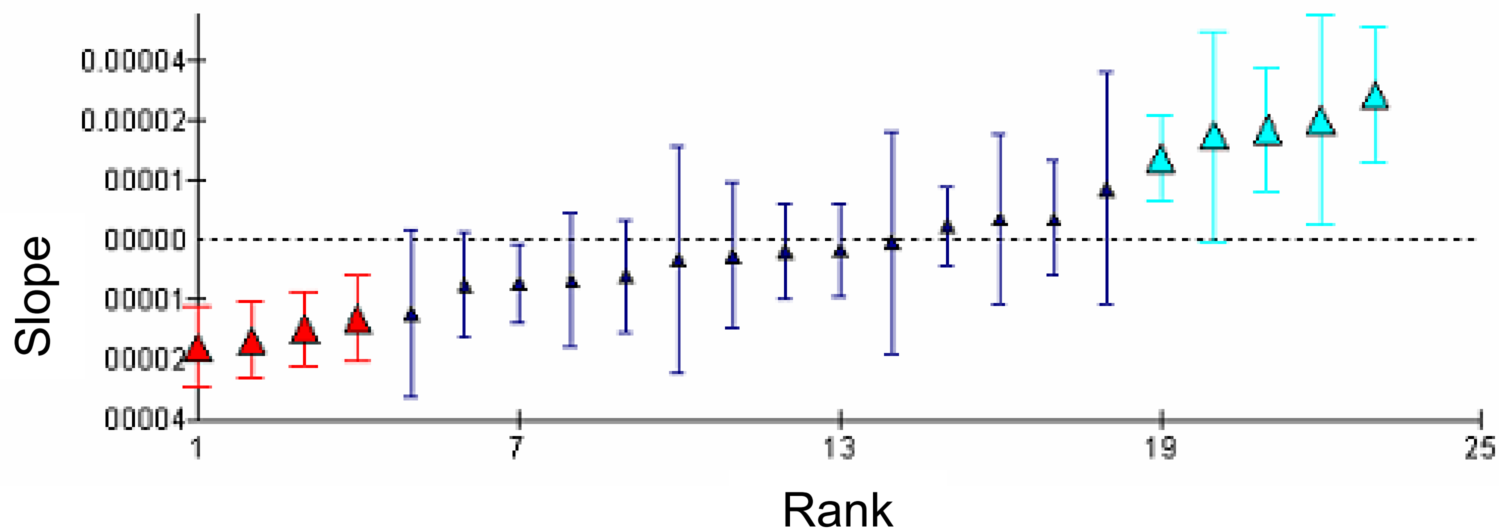
y	Relative population growth
$\beta_{0,1}$	Parameter
x_0	Constant
x_1	Change in accessibility
u	Systematic error (departure of the *j*-th Cantons intercept (slope) from the overall value)
ε	Error (departure of the *i*-th municipality's actual score from the predicted score) $\varepsilon \sim iid\ N(0, \sigma)$
i	Level 1 (Municipality)
j	Level 2 (Kanton)

Example: Swiss population growth

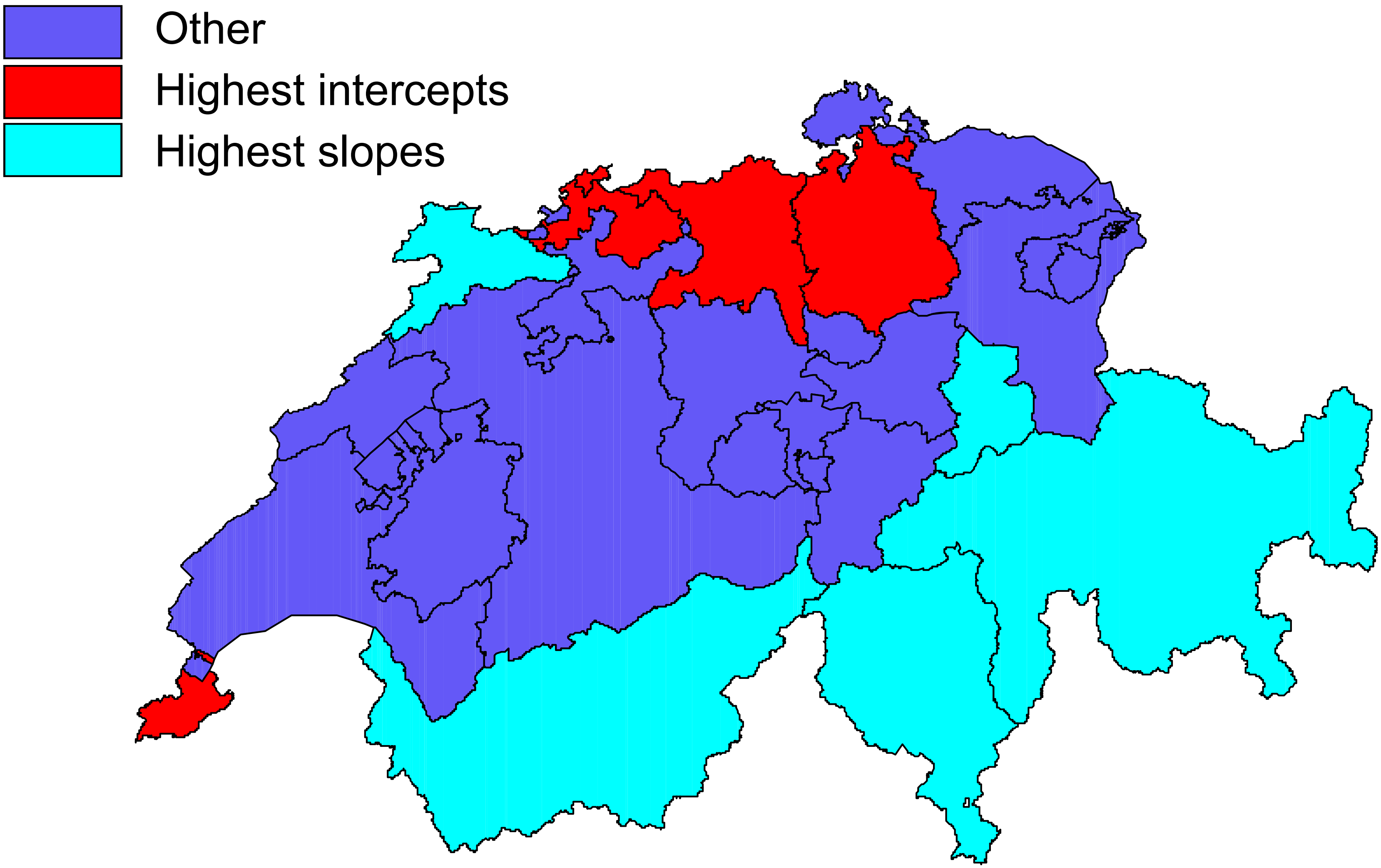


Example: “Systematic errors”





Example: Neighbourhoods in Swiss population growth



Spatial regression models

Spatial autoregressive model (SAR):

$$y = \rho W_A y + X\beta + \varepsilon \qquad \varepsilon \sim iid\ N(0,\sigma)$$

Spatial error model (SEM)

$$y = X\beta + u \qquad u = \lambda W_E u + \varepsilon$$

Spatial autoregressive and spatial error model combined (SAC):

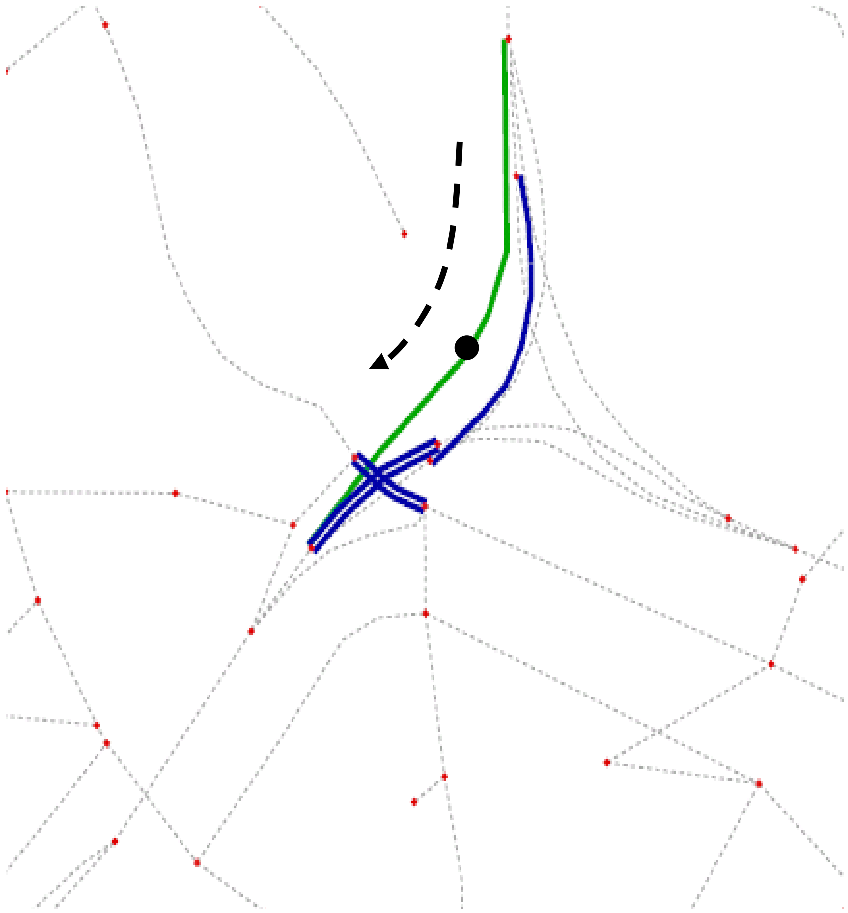
$$y = \rho W_A y + X\beta + u \qquad u = \lambda W_E u + \varepsilon$$

with W: neighborhood matrix (contiguity matrix) with row sum=1

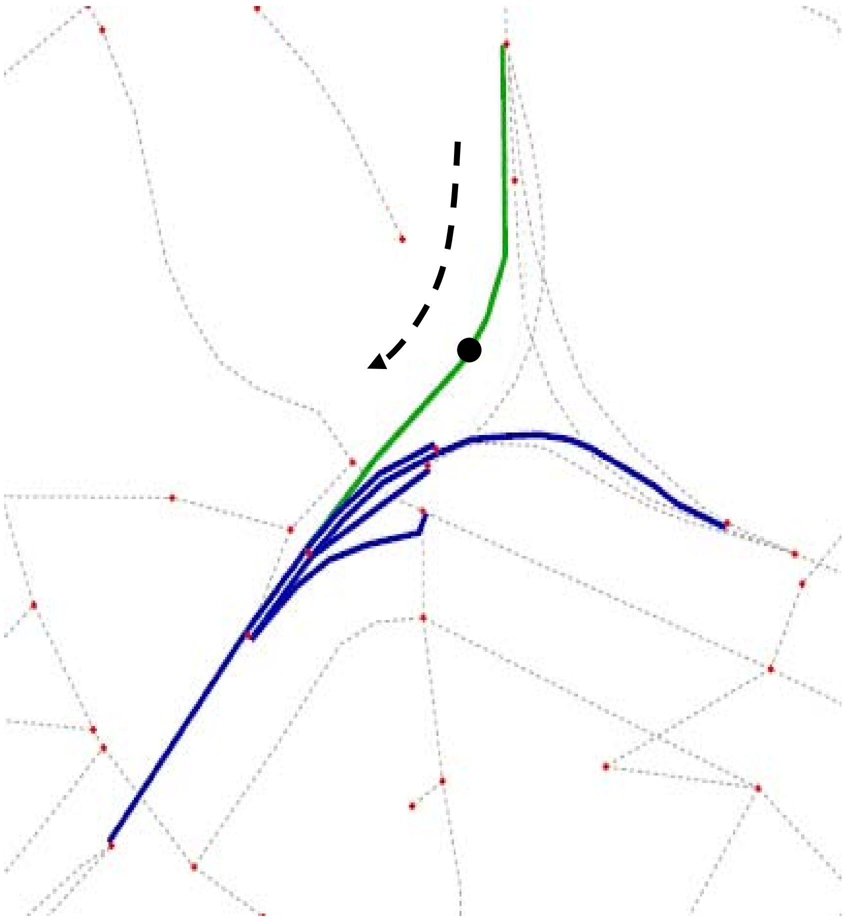
ρ: influence factor of spatial autoregressive dependence

λ: influence factor of spatial dependence of error

Neighbors: Euclidean distance vs. network distance



Five spatially symmetric nearest neighbors by Euclidean distance

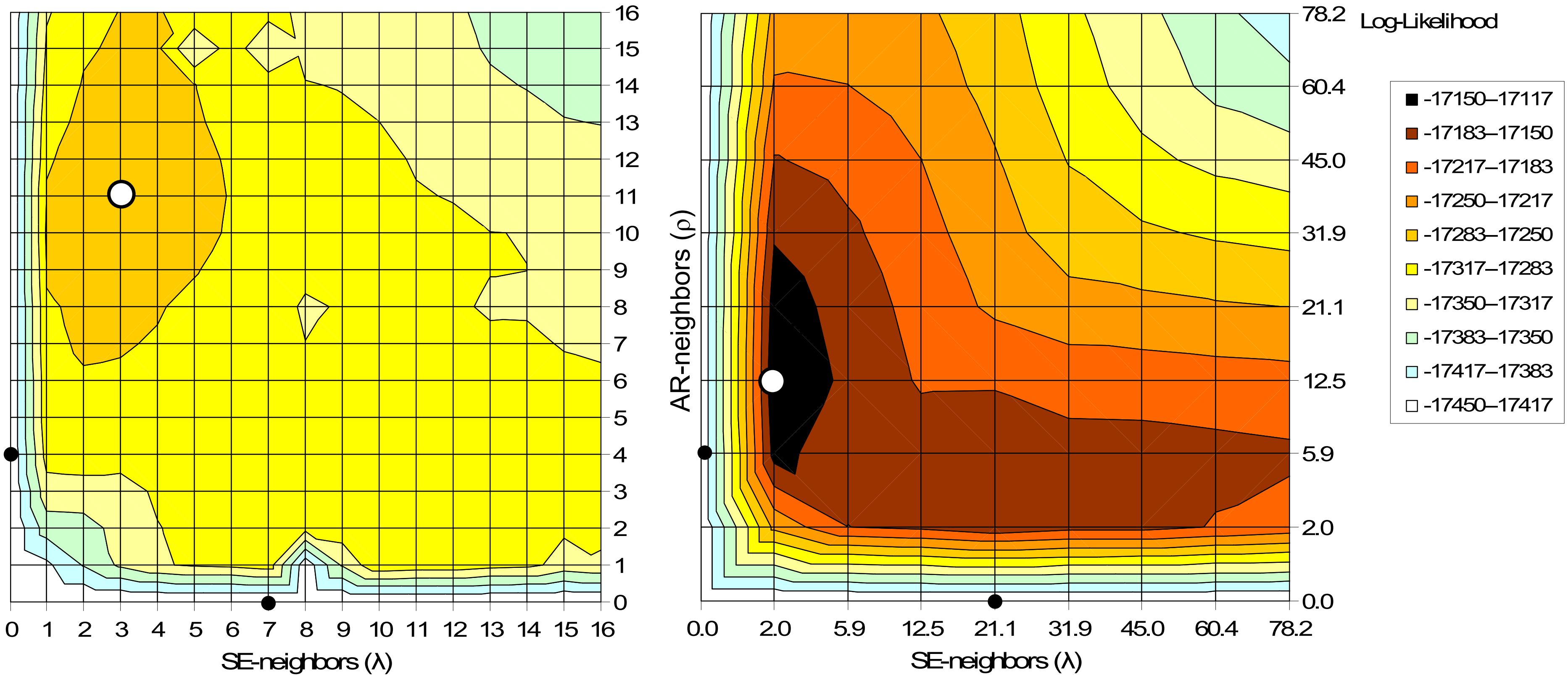


Five neighbors within a network distance of up to two intersections

Log-Likelihood Measure: Choosing Best Model (W-Matrix)

Nearest neighbors by Euclidean distance

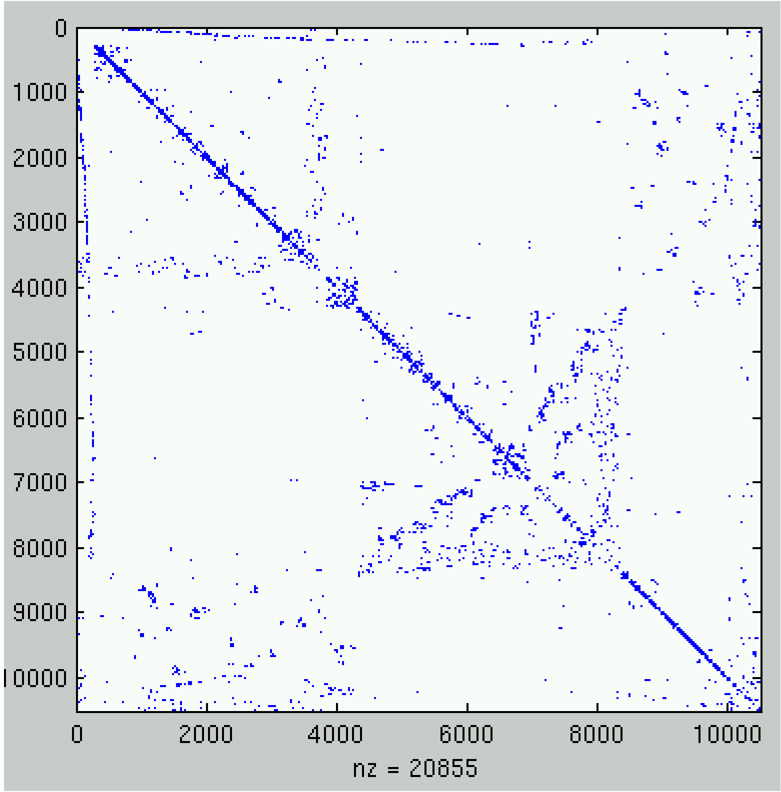
Nearest neighbors by network distance of intersections



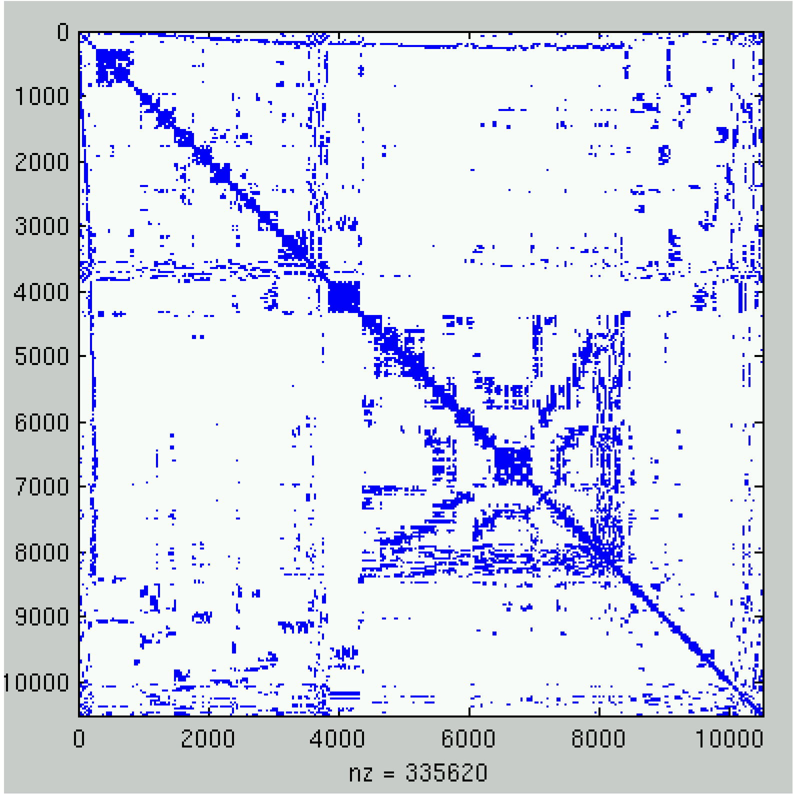
SAR:	$\rho_4=0.249$	-
SEM:	-	$\lambda_7=0.374$
SAC:	$\rho_{11}=0.242$	$\lambda_3=0.173$

SAR:	$\rho_{5.9}=0.298$	-
SEM:	-	$\lambda_{21.1}=0.632$
SAC:	$\rho_{12.5}=0.299$	$\lambda_{2.0}=0.142$

The CPU memory problem of the W-Matrices



1 intersection (~2 neighbors)



5 intersections (~32 neighbors)

Predictions

Ordinary least squares and weighted least squares (OLS, WLS):

$$\hat{y} = X\beta$$

Spatial autoregressive model (SAR):

$$\hat{y} = (I - \rho W_A)^{-1} X\beta$$

Spatial error model (SEM):

$$\hat{y} = X\beta$$

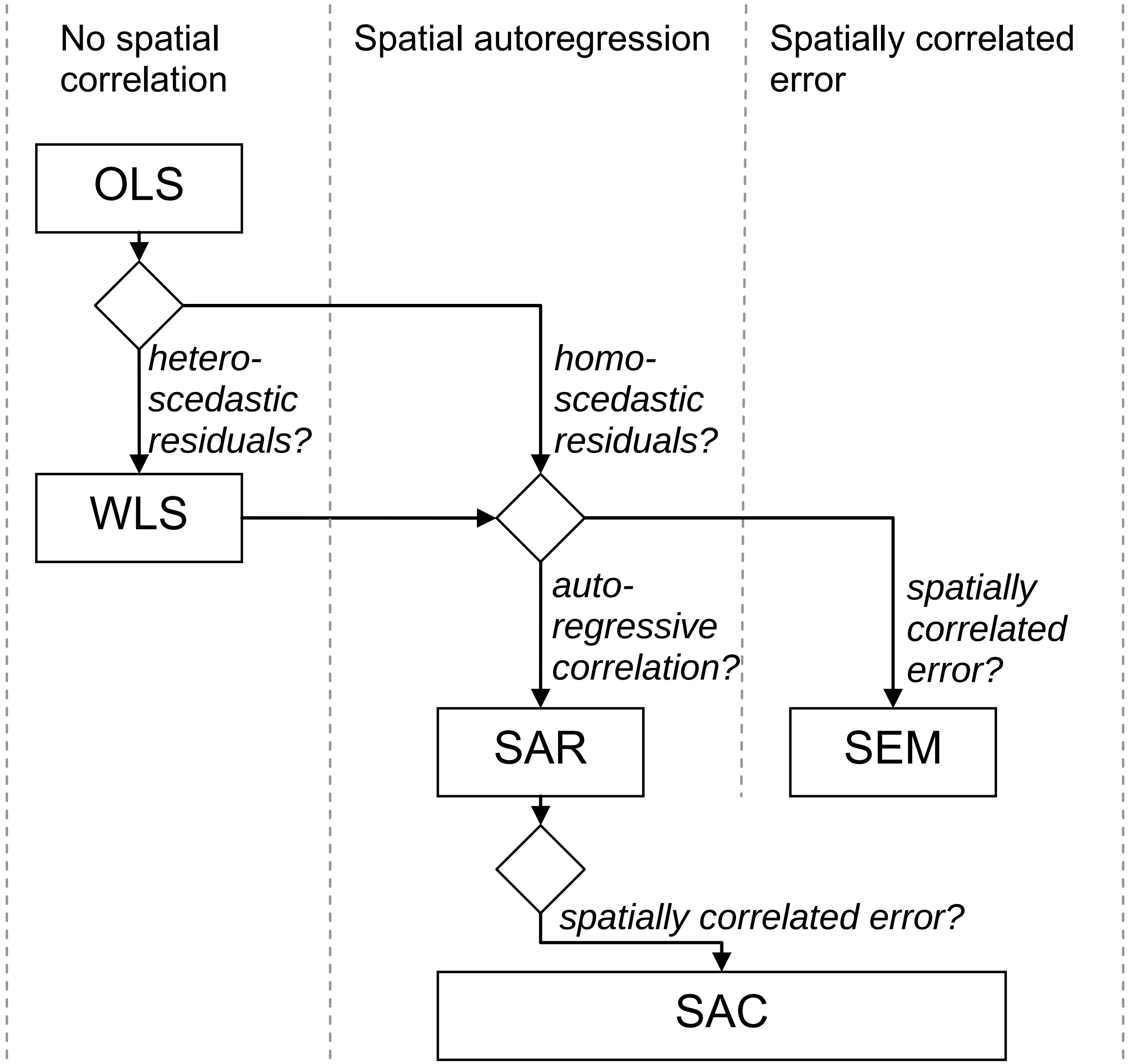
Spatial autoregressive and spatial error model combined (SAC):

$$\hat{y} = (I - \rho W_A)^{-1} X\beta$$

Do you always check ?

- Collinearity (Covariance matrix of the X) ?
- Correlations between y and ε ?
- Correlations between X and ε ?
- Presence of structural units (e.g. organisational membership, social networks and groups, cohorts, life style groups, residential clustering) ?
- Temporal correlations (DW – Test) ?
- Spatial correlations (Moran's I) ?

Choosing appropriate regression model



Thanks to

Martin Tschopp (Hierarchical models)

Michael Bernard and Jeremy Hackney (Spatial lag and error models)