Spatial Analysis of Geographic Data

Julian Bernauer

Data and Methods Unit
MZES
julian.bernauer@mzes.uni-mannheim.de

27 February 2018

Spatial Models

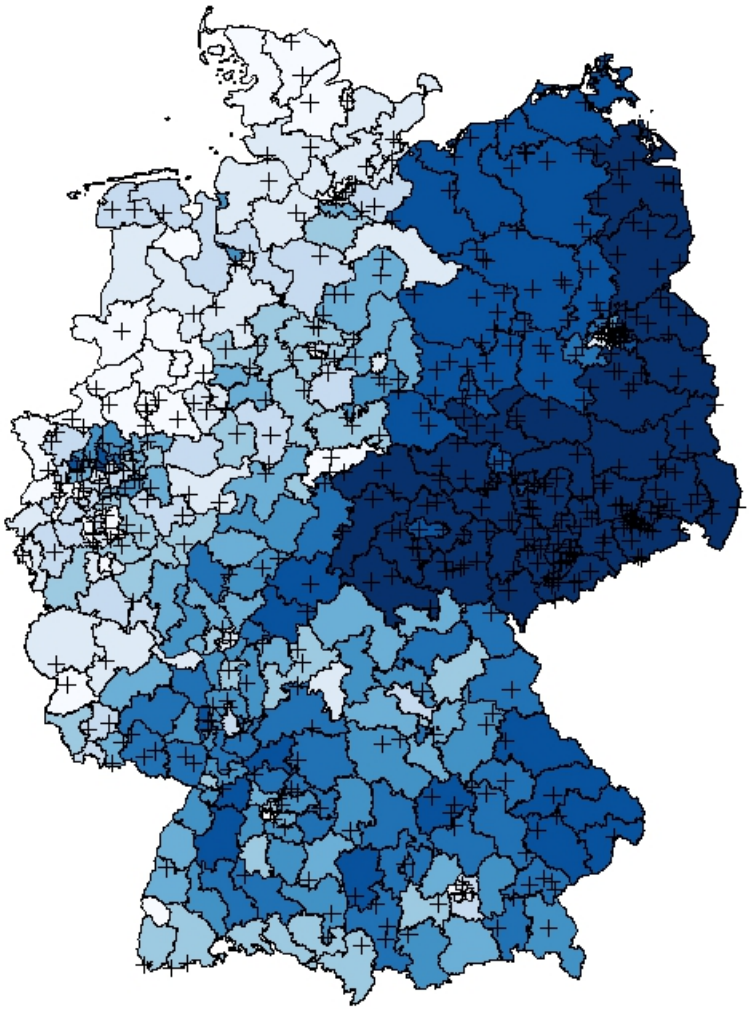
Overview

1. Recap last week
2. Starting from the running example - AfD vote and attacks
3. Spatial regression: meeting SEM, SLM, SAR, SLX and Durbin
4. Spatial correlation: global and local
5. In R: exercise on running example
6. Your project: look for data and explore spatial mechanisms
7. Time and space, varied outcomes (count...): next weeks

Recap

- Spatial analyis = dependency = geography and more
- Connectivity matrix **W** → specification, standardization, transformation, dimensionality, directionality (Neumayer and Plümper 2016)
- Example of AfD vote in the 2017 BTW election and attacks on refugees = spatial problem, causally challenged

AfD vote and attacks on refugees – arson and assaults



Back to the running example

Attacks on refugees in Germany

- Naive expectation: AfD vote shares and the level of attacks on refugees are positively related
- Generating a "climate" which makes attacks more likely?
- We will discuss:
  → Which model(s) do you think is (are) appropriate here?
  → Do you see mechanisms justifying spatial errors or lags (dependent/independent variables)?
  → Advanced: Multiple W? Any W not based on physical proximity?

It's a violation!

Interdependence

- Basic regression assumptions: errors independent and identically distributed
- Violated by definition in spatially structured data
- Thrust: Turning this from a problem into a substantively interesting enterprise

Spatial regression

Variants: substance vs. nuisance

- Spatial lag model (ρ): Assumes that levels of dependent variable y in connected units affects y in unit of interest
  → Matter of substantial influence, e.g. diffusion
  → Issue of endogeneity
- Spatial error model (λ): Assumes that spatially clustered latent variables spreading across units cause dependence
  → Matter of nuisance/control
  → Less problematic properties

Spatial lag model - SLM

Notation following Selb (2006), replacing row-standardized W^* by raw W

Assumption: $\lambda = 0$

$$y_i = \rho W y_i + X_i \beta + \mu_i \tag{1}$$

Running example

Attacks on refugees

Do you see a spatial lag and of what kind?

Running example: spatial lag

- Theory: Contagion between attacks
- Implementation: Just the direct neighbours? Also accumulation or differentiation between smaller and larger attacks?
- → Requires strategy to deal with endogeneity

Spatial error model - SEM

Selb (2006)

Assumption: $\rho = 0$

$$y_i = X_i\beta + \epsilon_i \tag{2}$$

$$\epsilon_i = \lambda W\epsilon_i + \mu_i \tag{3}$$

Running example

Attacks on refugees

Do you see a spatial error and why?

Running example: spatial error

- Theory: Regionally clustered latent variables
- Example: Area of influence of neo-Nazi groups
- Other ideas for regionally clustered latent variables?
- East Germany as an example? → Can be explained by structural factors? (Jäckle and König 2017, WEP)
- Other factors? Consequences of economic problems after reunification? (Lack of) social capital?
- → In combination with spatial lag?!

Spatial autoregressive model - SAR

Selb (2006)

$$y_i = \rho W y_i + X_i \beta + \epsilon_i \tag{4}$$

$$\epsilon_i = \lambda W \epsilon_i + \mu_i \tag{5}$$

By the way: linear regression model

Selb (2006)

Assumption: $\lambda = 0$
Assumption: $\rho = 0$

$$y_i = X_i\beta + \mu_i \tag{6}$$

Durbin model

Fischer and Wang (2011: 37)

Spatially lagged dependent and independent variable

$$y_i = \rho W y_i + X_i W \beta + \mu_i \tag{7}$$

Running example

Attacks on refugees

Do you see a spatially lagged independent variable?

Running example: spatial independent variable

- Halo effect: Threat without contact
- Share of foreigners in surrounding districts relative to in the district
- Special case of lagged independent variable
- See Martig and Bernauer (2016, SPSR)
- Lagged x does not pose much problem for estimation

Dealing with endogeneity

Franzese and Hays (2006, CPS)

- Endogeneity in the spatial lag model estimated with least squares: y on the right-hand side
- Bias in favour of interdependence
- Instrumentation or spatial maximum likelihood or S-ML
  → Tendency towards S-ML, which manages to move all y and W_y to the left-hand side
  → Instruments affecting regressor but not outcome are hard to find, even though spatial lags of non-spatial regressors offer themselves
- Ignoring dependency often much worse than bias from lagged y (Franzese and Hays 2006: 757)

Example model specification and choice

Selb (2006: 306-7)

- Example: spatial patterns in the NSDAP vote
- Both a spatial lag and a spatial error specification would pass a singular test → no way to tell right one
- Lagrange Multiplier Tests based on OLS residuals and W (Selb 2006: 303) for model chocie: rather spatial error model
- Considering spatial error removes effect of unemployment present in OLS model

Interpretation

Fischer and Wang (2011: 41-44)

- Simultaneous feedback in lag models complicates interpretation
- Change in an explanatory variable (lagged y) changes dependent variable somewhere else
- Suggestion: "scalar summary measures" averaging effects/simulation

Spatial correlation/spatial statistics

Variants

- Moran's I
- Geary's c
- Local variants
- Moran scatterplot

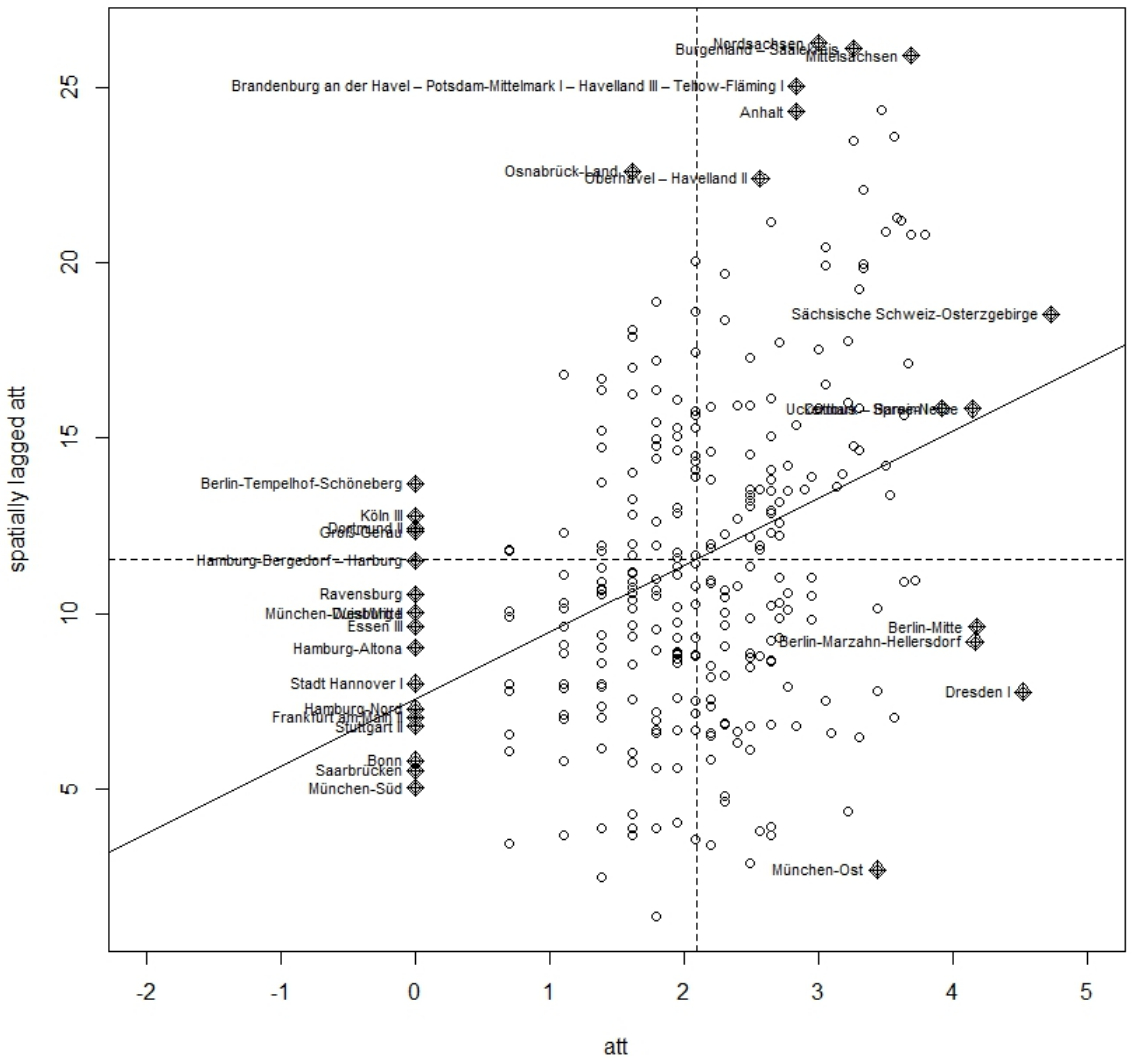
Moran's I and Geary's c

Moran's I for global spatial association based on cross-products (Fischer and Wang 2011: 23)

$$I = \frac{n}{W_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n} (z_i - \overline{z})^2} \tag{8}$$

→ Geary's c is similar but works with squared differences; both statistics do not range between −1 and 1
→ Local variants find hot or cold spots of spatial association

Moran scatterplot



Now...

...application time

Spatial regression and correlation

Your own project: getting started

- Think about your own spatial analysis project
- Look for map and substantial data
- Think about spatial mechanisms - geographic proximity or not, lag or error, further complications such as multilevel structures...
- We'll have a discussion to solve a few early problems

Conclusions

Take-away messages

- Depending on the type of spatial connectivity, use spatial lags, errors, or both
- Lagged dependent variables are endogeneous
- Use visualization and tests of spatial dependency for model choice

Thank you for your attention!