Spatial Analysis of Geographic Data

Julian Bernauer

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

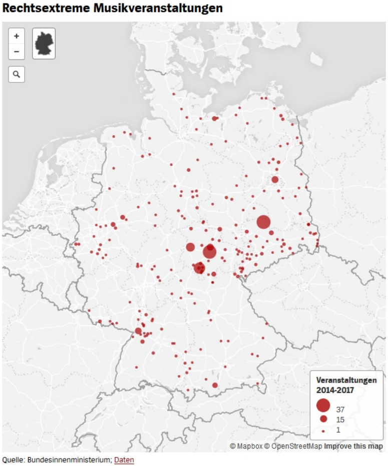
6 March 2018

Implementing Spatial Analysis in R

Overview

1. Starting your own project: initial ideas/problems?
2. Getting fancy with geodata in R: more packages and maps; tidyverse, ggplot2 and plotly
3. Prelude Bayesian analysis for varied outcomes: categorical and count

Running example: a proxy for ...?



`http://www.spiegel.de/politik/deutschland/`

`rechtsrock-bundesweit-starke-zunahme-von-neonazi-konzerten-daten-a-1195798.html`

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

Spatial autoregressive model - SAR

Selb (2006)

$$y_i = \rho W y_i + X_i \beta + \epsilon_i \tag{1}$$
$$\epsilon_i = \lambda W \epsilon_i + \mu_i \tag{2}$$

Still linear...

Non-linear dependent variables

Non-linear dependent variables

- Binary categorical data: logit link
- Count data: Poisson distribution
- Variations for count data given the running example: negative binomial/zero-inflation
- Not covered: ordinal, nominal, or conditional logit, survival models...

Logistic regression

King (1989: 98-101, "Unifying Political Methodology", CUP)

$$Y_i = f_{bern}(y_i|\pi_i) \tag{3}$$

$$\pi_i = g(x_i, \beta) \tag{4}$$

$$\pi_i = \frac{1}{1 + exp(-x_i\beta)} \tag{5}$$

→ Maps binary 0/1 (e.g. attack or not) on explanatory variables via latent π generated by the logistic link function
→ Interpretation β: effect of x on logit log odds of y (compute quantities of interest/predicted probabilities)

Count: Poisson

King (1989: 48-51, 121-124)

$$f(y_i|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \text{for} \quad \lambda > 0 \quad \text{and} \quad y_i = 0, 1, \ldots, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$E(Y_i) = \lambda = exp(x_i\beta) \tag{7}$$

→ Upper equation: for single observation; lower equation: expected value across observations
→ Count with no upper bound; constant and independent event probabilities: negative binomial relaxes this assumption
→ Zero-inflated versions: add an equation for modelling zeros if some observations are very likely to bear them

Bayesian approach to spatial data

Why - pragmatic reasons

- As mentioned, advanced estimation procedures are needed
- Maximum likelihood methods are also cool and do a good job (see King 1989)
- For complex models (e.g. a zero-inflated spatial multilevel count model) out-of-the-box models are hard to find
- Bayesian statistics eases estimation as it bypasses (approximate) numerical solutions by simulations
- Interpretation is more intuitive (parameters as distributions)

Bayesian approach to spatial data

Why - theoretical background

Model (θ) and data (y)

$$Pr(\theta|y) = \frac{Pr(\theta, y)}{Pr(y)} \tag{8}$$

$$= \frac{Pr(\theta)Pr(y|\theta)}{Pr(y)} \tag{9}$$

- We know $Pr(y)/Pr(y|\theta))$ and want $Pr(\theta|y)$: Probability of model given data
- Bayes works via $Pr(\theta)$: Prior for model
- Maximum likelihood avoids $Pr(\theta)$ and looks for relative likelihood of models

Conclusions

Take-away messages

- R offers virtually unlimited possibilities to obtain and visualize spatial data
- Obviously, outcome do not need to be linear in spatial models
  $\rightarrow$ I prefer moving to a Bayesian modelling framework

Thank you for your attention!