Spatial Analysis of Geographic Data

Julian Bernauer

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

20 February 2018

Concepts of Spatial Analysis

Overview

1. Round of introduction
2. Goals of the course
3. Some concepts of spatial analysis: W, mainly
4. Introducing the running example: Possible link between AfD vote and attacks on refugees
5. Exercise based on German federal election data

Goals of the course

Get to know concepts

Geodata, interdependence and W

Learn about models

Spatial correlation and regression varieties

Easy implementation

R and open-ended Bayesian modeling

Ready for application

Policy diffusion, voting, conflict...

A few basics

Geographic or Spatial Analysis?

Geography has to do with the surface of the earth, spatial is more general.
Or: "Space is More than Geography" (Franzese and Hays 2008: 745, citing Beck et al. 2006).

First law of geography

"Everything is related to everything else, but near things are more related than distant things" (Tobler 1970: 236).

But careful!

"[T]he substantive content of ... proximity ... extend[s] well beyond simple physical distance..." (Franzese and Hays 2008: 745).

Discuss: Neighbours

Who (which countries) are "neighbours" - think in terms of interdependence

- Tax policy
- Migration
- Constitutions (character of political institutions)

→ Contiguity vs. connectivity

Discuss: Neighbours

Who (which countries) are "neighbours" - think in terms of interdependence

- Tax policy: alternative locations, e.g. Ireland and Germany (see Franzese and Hays 2008)
- Migration: attraction and access, e.g. Syria: Germany, but also Turkey
- Constitutions: former colonies, e.g. UK and Barbados

→ Just to be sure: North and South Korea might be geographic neighbours, but Germany and Japan are much more interdependent in terms of trade

Interdependence

Franzese and Hays 2008, CPS

1. Equate spatial interdependence and diffusion (743)
2. Mechanisms: Coercion, competition, learning, emulation and migration
3. Reactions can depend on units properties such as veto players
4. Their example: Tax policy

A first spatial model

$$y_{it} = \rho \sum_{j \neq i} w_{ij} y_{jt} + \beta_d^{'} d_{it} + \beta_s^{'} s_t + \beta_{sd}^{'} (d_{it} \otimes s_t) + \epsilon_{it} \qquad (1)$$

Spatial lag model with interactions by Franzese and Hays (2008: 754)

Issues with the spatial model

Franzese and Hays 2008, CPS

1. Galton's problem: Ignoring interdependence leads to bias in favour of non-spatial factors
2. Originally study of social complexity and marriage institutions
3. Autocorrelation due to "borrowing" of marriage institutions
4. Reduces effective sample size
5. But also chance to analyze interdependence -> Rather "Galton's Opportunity"?

Issues with the spatial model

Franzese and Hays 2008, CPS

1. Endogeneity of **W**
2. Lagged dependent variable puts "some observation's left-hand sides on others' right-hand sides" (Franzese and Hays 2008: 756) and produces bias in favour of interdependence
3. France explains Germany, and Germany explains France...

Issues with the spatial model

Modifiable areal unit problem

1. Type of aggregation affects results
2. Zoning effect: different boundaries
3. Scale effect: different number of areal units
4. Possible solution: simulation

(Fischer and Wang 2011: 10)

Types of geographic data

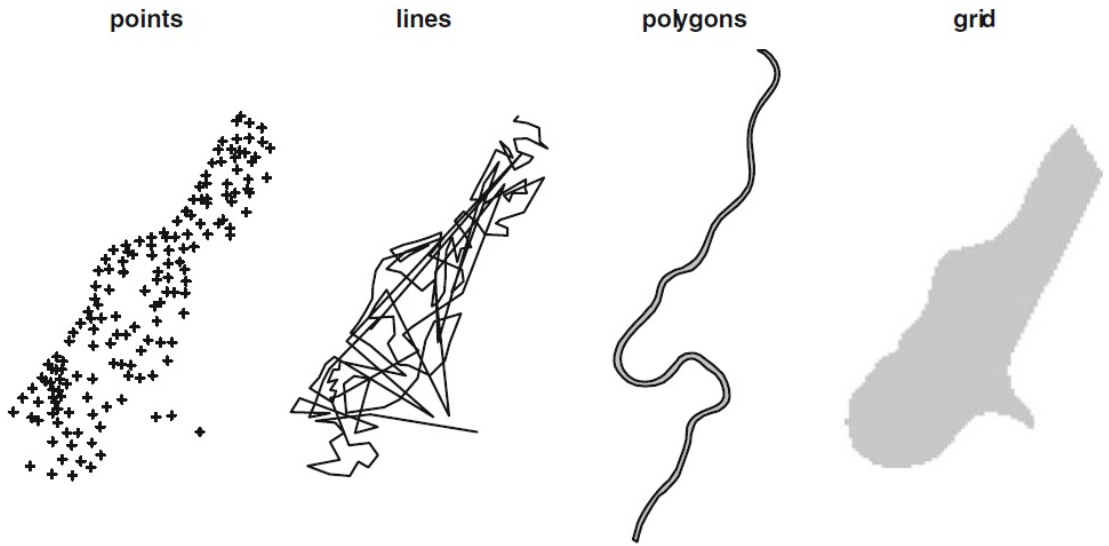


Fig. 3.1 The meuse data set: sample points, the sample path (*line*), the Meuse river (*ring*) and the *gridded* study area

Figure: Bivand et al. 2013: 60

"A Polygon is a two-dimensional surface stored as a sequence of points defining an exterior bounding ring and zero or more interior rings." (https://docs.microsoft.com/en-us/sql/relational-databases/spatial/polygon)

Types of dependence and chess analogy: of rooks and queens

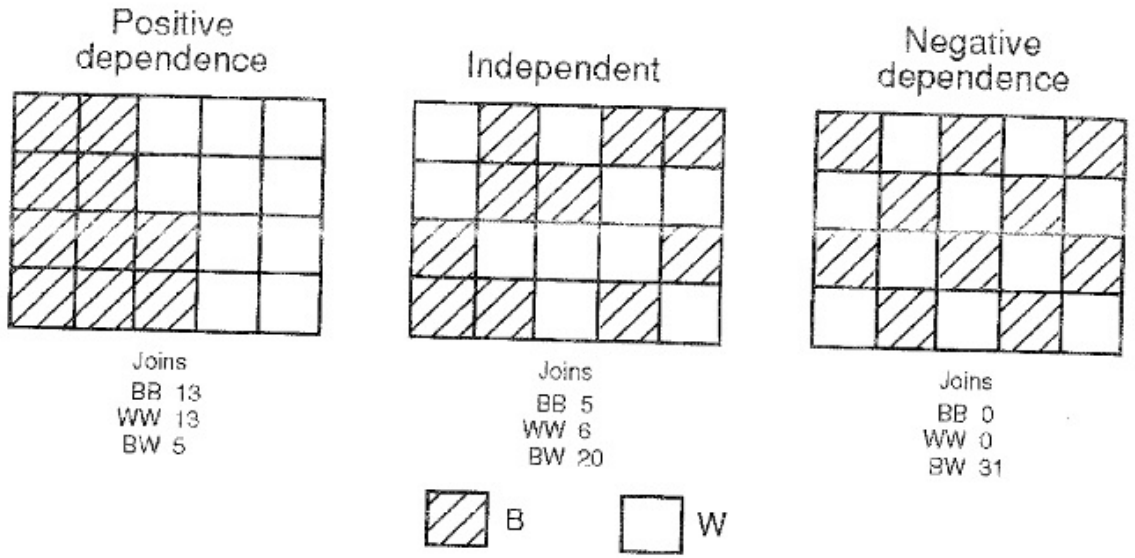


Figure 2.9 Join-counts for different map patterns

Figure: Haining 2003: 80

Bishop rule, anyone?

Discuss: negative spatial correlation?

Can you think of negative correlations?

- ...

Discuss: negative spatial correlation?

Can you think of negative correlations?

- Plants compete for soil nutrients
- Fast-food chains tend to scatter across space
- Allies of hegemonic countries tend to reduce their military budget if the hegemon spends more (Neumayer and Plümper 2016)
- Labor market policies encourage free-riding on behalf of neighbours (Franzese and Hays 2006)

→ You would not plant all your potatoes in one place, would you?

Aside: Mapping what type of information?

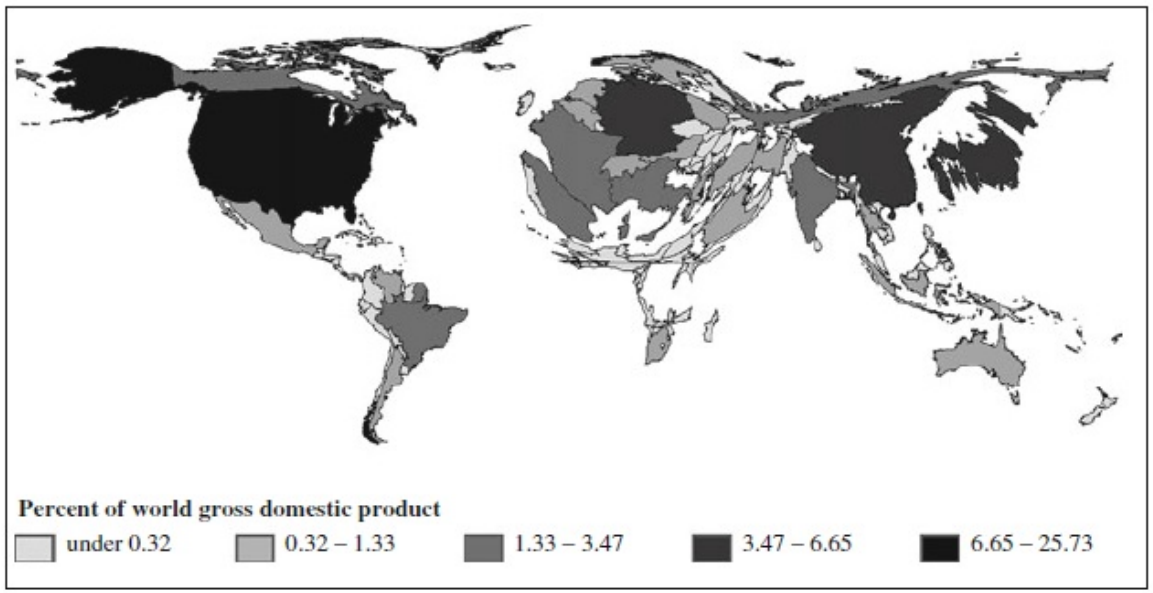


Figure: Fischer and Wang 2011: 19

Meeting W

(W)hat it is: A definition

"W, the connectivity matrix that links observations with each other, by definition determines which observations spatially depend on each other - and to what degree they do so."
(Neumayer and Plümper 2015: 175)

→ W defines spatial dependence
→ W is not about the actual dependent variable assumed to be spatially structured, but the spatial structure itself!
→ For instance, refugee flows are spatially structured, where the spatial structure is (also) defined by the level of border openness

Map to graph

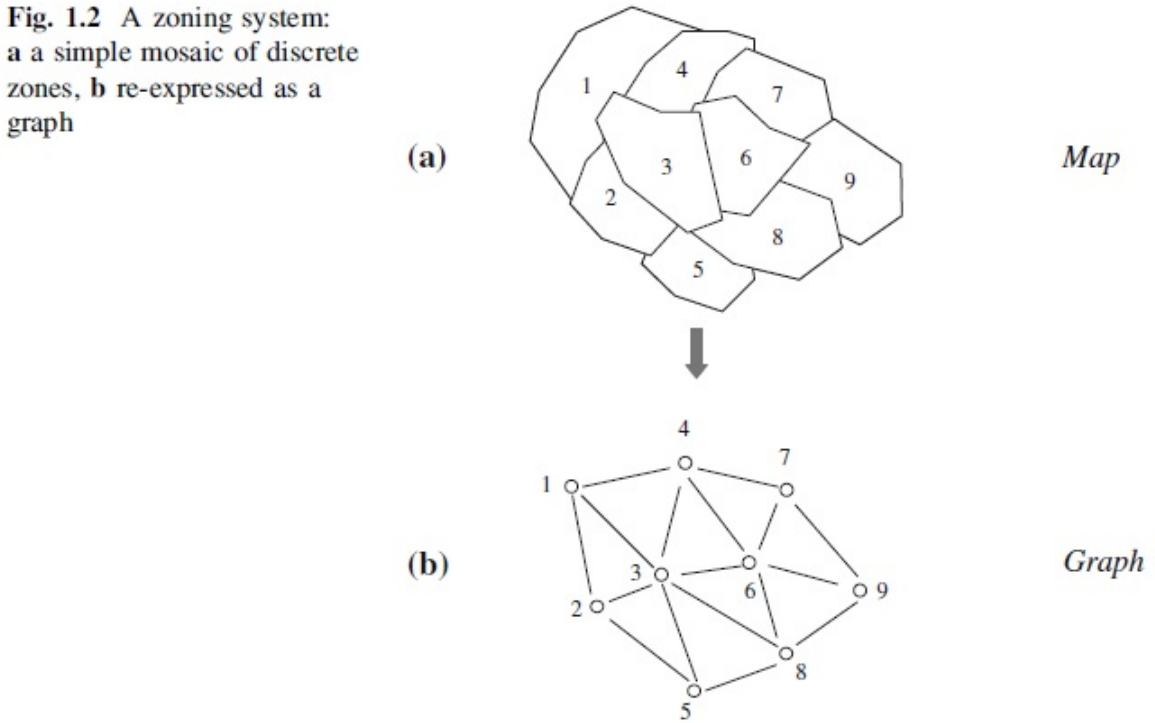


Figure: Fischer and Wang 2011: 9

→ Example of W based on geography/contiguity/common borders
→ Alternatives for spatial interdependence: distance (bands) or not based on geography...

W - in its raw and binary form

Table 1.2 A spatial weights matrix W derived from the zoning system in Fig. 1.2: the case of a binary first order contiguity matrix

	1	2	3	4	5	6	7	8	9
1	0	1	1	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	1	1	0	1	1	1	0	1	0
4	1	0	1	0	0	1	1	0	0
5	0	1	1	0	0	0	0	1	0
6	0	0	1	1	0	0	1	1	1
7	0	0	0	1	0	1	0	0	1
8	0	0	1	0	1	1	0	0	1
9	0	0	0	0	0	1	1	1	0

Figure: Fischer and Wang 2011: 9

Fun facts about W

- W is the 23rd letter in the alphabet - which can't be a coincidence
- The W matrix has dimensions n times n, where n is the number of areas covered
- W does not have to be filled with 0s and 1s
- As alluded to, it does not have to be based on common borders or geographic distance
- Neumayer and Plümper (2016: 175, Footnote 1) stress that W should not be called "weights" but "connectivity" matrix because rows do not need to / should not sum up to 1

→ Last but not least, you can use it to publish articles with the title "**W**" (Neumayer and Plümper 2016)!

Five things to consider about W (Neumayer and Plümper 2016)

- Use direct measures of connectivity instead of geographic proximity/contiguity to capture causal mechanisms
- Don't row-standardized W. Just don't.
- Transforming W (such as taking the log of the distance) involves serious assumptions
- Spatial effects are not necessarily uni-dimensional. This means multiple Ws.
- Spatial effects are not necessarily uni-directional

Five things to consider about W (Neumayer and Plümper 2016)

- Direct measures of connectivity → beyond contiguity, based on theory, avoiding misspecification bias
- Don't row-standardized W → homogeneity of total exposure to spatial stimulus
- Transforming W → Set irrelevant cases to zero, log of distance is a strong assumption, row-standardization alters rankings, use distance bands
- Uni-dimensional → multiple W's, additively or in interaction
- Uni-directional → positive and negative dependencies, define two W's for each group

Why W is WICHTIG

- The choice of W heavily affects the results and is always a bit arbitrary - sorry, cough - needs theoretical justification (Fischer and Wang 2011: 8, 21)
- Playing around with alternative Ws is recommended

→ "...careful, accurate, powerful specification of **W** ... is of crucial empirical, theoretical, and substantive importance..." (Franzese and Hays 2008: 757).

First overview of software and R packages

- First choice: R or dedicated GIS → going for R here
- R packages sp, spdep, maptools, rgdal...: Developed by leading figures in geoscience (Bivand, Pebesma...), allow reading point/polygon data, plotting as well as spatial modelling
- For enhanced visualization, but not from geoscience: ggplot2 for advanced graphs such as in the Economist (`http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html`)

→ We will elaborate on this in more detail

Geographic data and how to use it

- Attributes and geographic information: Typically, these come bundled together in shapefiles
- Further attribute data can be easily merged using area identifiers
- Format geoJSON: → `https://github.com/julianbernauer/geospat/blob/master/data/vorfaelle.geojson`

→ We'll cover this in today's application

Alternative software

- GeoDa: Free software for spatial analysis, see also further information such as Python-based geospatial multilevel-modelling, see `https://spatial.uchicago.edu/`
- QGis: Open source, see `https://www.qgis.org/de/site/`
- ArcGIS: A bit more commercial, see `https://www.esri.de/produkte/arcgis`
- Aside: Geographic data can be used for real-time georeporting in case of disaster!

Attacks on refugees

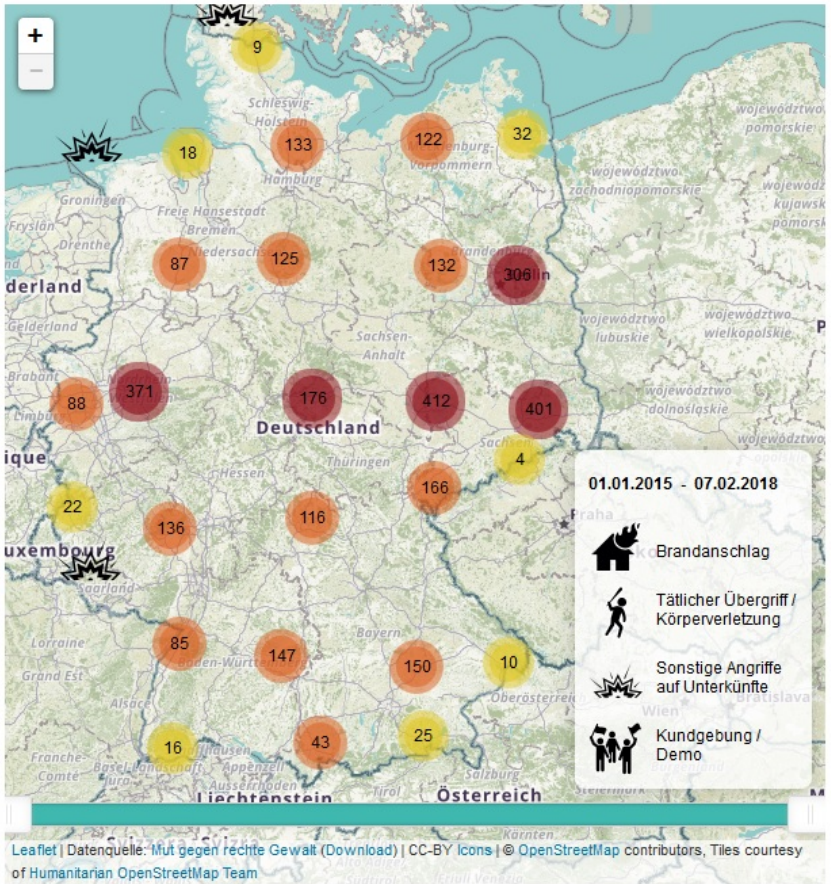
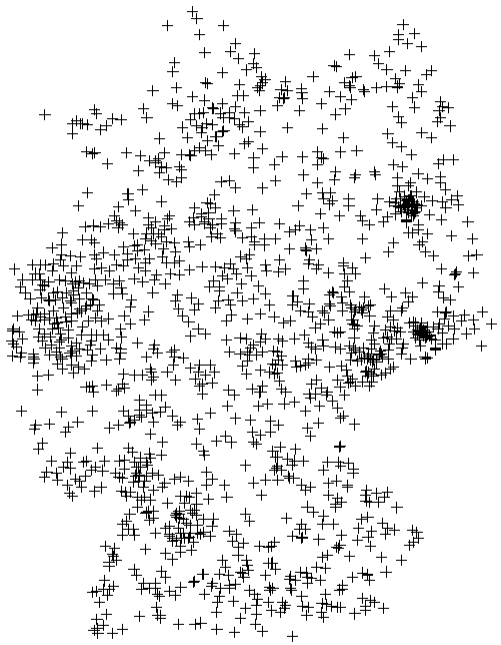


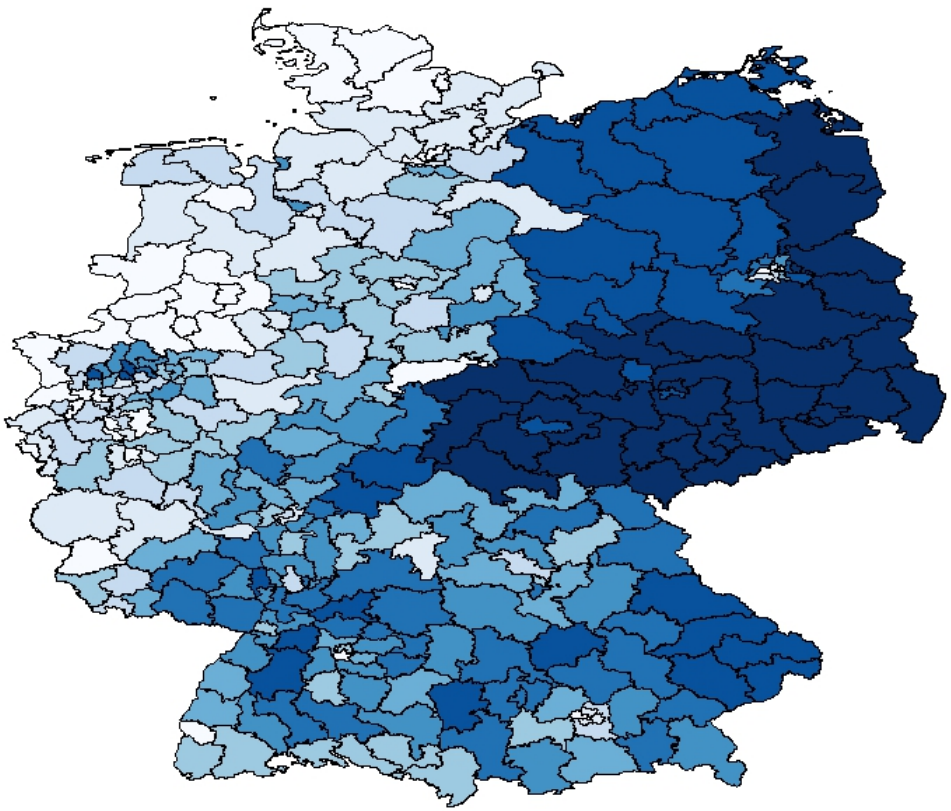
Figure: Attacks on refugees in Germany 2015-2017(?)

Source: `http://mut-gegen-rechte-gewalt.de/chronik-karte`, downloaded 7 February 2018

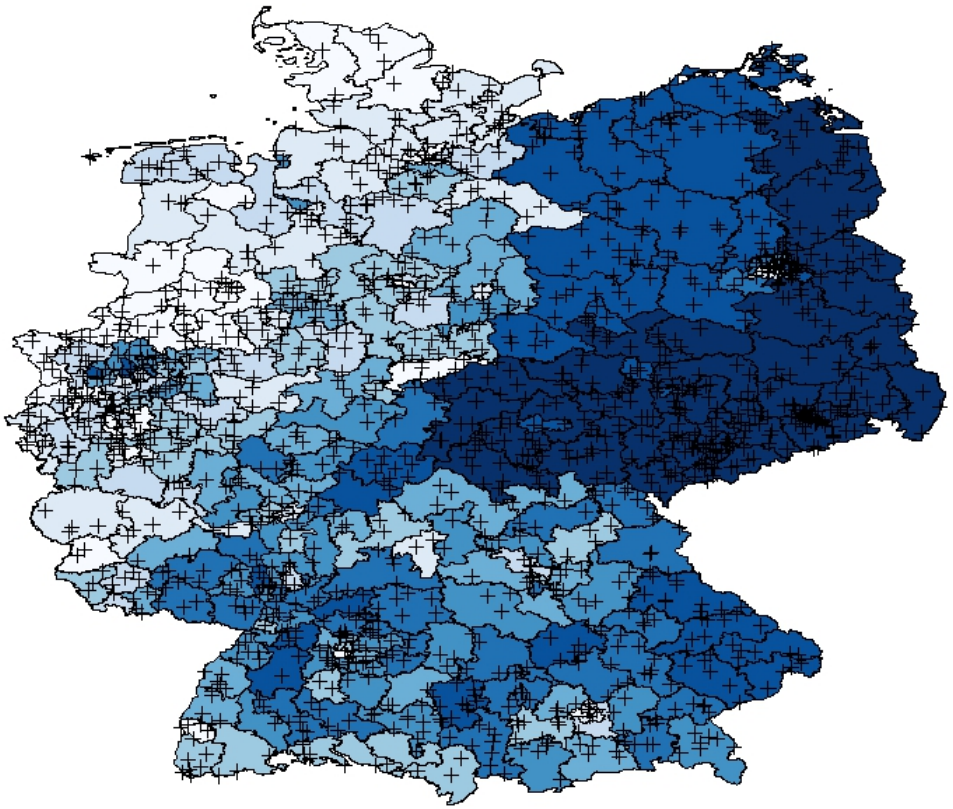
Attacks on refugees as points



AfD vote in 2017 German federal election



AfD vote and attacks on refugees



Motivation for the running example

1. Imagine you wonder about the coincidence of attacks on refugees and the AfD vote
2. You like to investigate the spatial patterns, control for alternative explanations and consider possible endogeneity

Discussion

Do you see causal links between populism and attacks?

Jäckle and König 2017, WEP

1. Strength of "extreme and populist right-wing parties"
2. Geodetic distance of the closest attack
3. Cumulative attacks in district and number of attacks in Germany in previous week
4. Socio-cultural controls: attitudes towards foreigners...
5. Socio-cultural context: East Germany
6. Contact: share of foreigners
7. Social disintegration: voting abstention
8. Material-economic: unemployment, education...
9. Other: population density...

My hunch

1. Use real geographic W with distance bands to model contagion between attacks
2. Endogeneity: Instrument AfD vote share / structural equation model
3. Halo effect of share of foreigners via separate W
4. ...

Conclusions

Take-away messages

- Spatial proximity can be defined in many ways - geographic or not
- W is powerful, use its flexibility
- Outlook: Spatial data inherently violates standard regression assumptions → Also substantively interesting!

Thank you for your attention!

From points to areas to graphs: Dirichlet partition

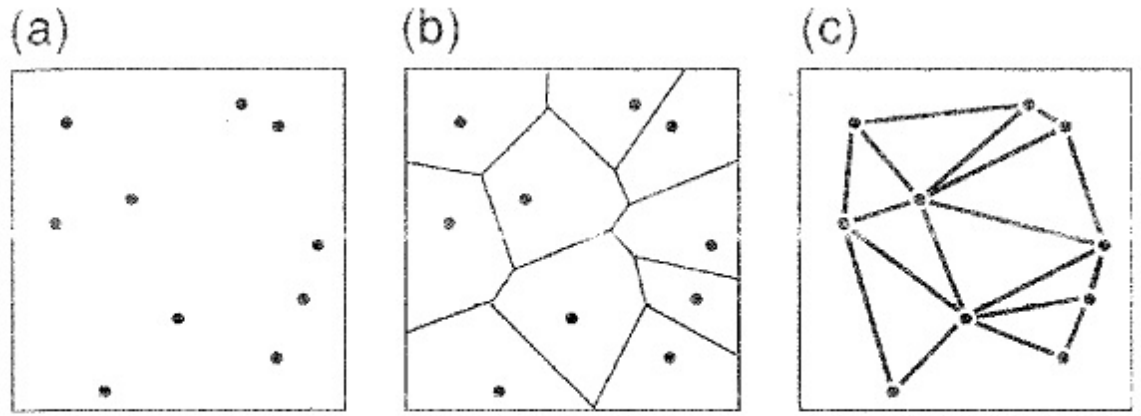


Figure 2.10 Neighbours defined using a Dirichlet partition

Figure: Haining 2003: 81