UNIVERSITÄT
LEIPZIG

Toolbox CSS – Spatial Data I

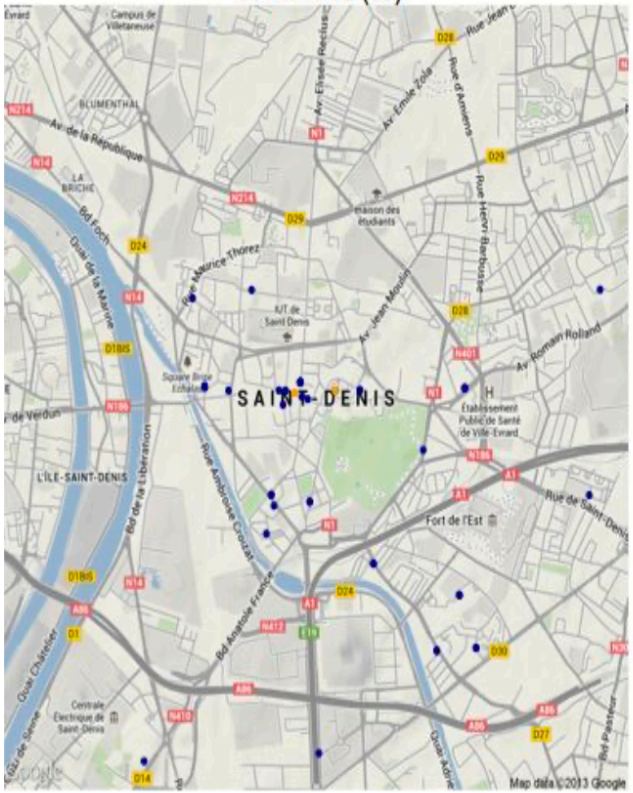
NSG SR 423, 17/12/2024

Felix Lennert, M.Sc.

OUTLINE

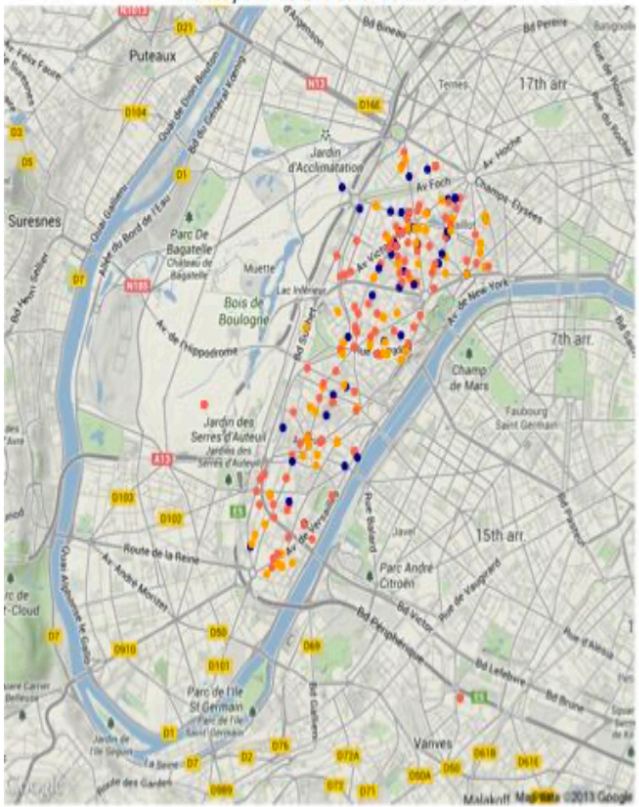
- What is “geospatial data”?
- “place” vs. “space” (Logan 2012)
- Why is it relevant for social sciences?
- Core concepts and terminology
- Moran’s I

Médecins généralistes en exercice libéral
Saint-Denis (93)

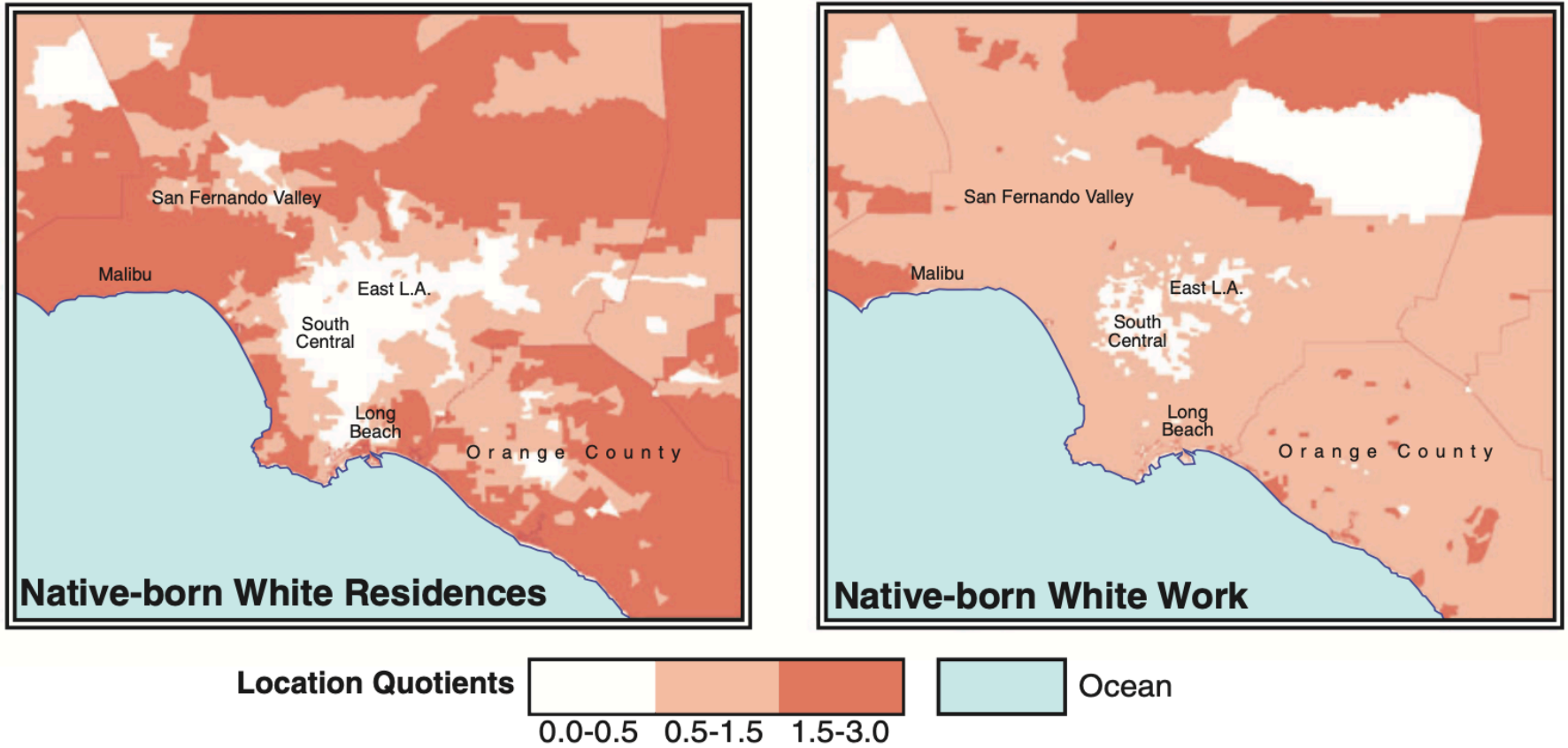


Secteur d'exercice
- secteur 1
- secteur 2

Médecins généralistes en exercice libéral
Paris, XVIè arrondissement



Secteur d'exercice
- Non conventionne
- secteur 1
- secteur 2


San Fernando Valley
Malibu
East L.A.
South Central
Long Beach
Orange County
Native-born White Residences
San Fernando Valley
Malibu
East L.A.
South Central
Long Beach
Orange County
Native-born White Work
Location Quotients
0.0-0.5
0.5-1.5
1.5-3.0
Ocean

https://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html

GEOSPATIAL DATA – DEFINITION

- Information about location and shape of geographic features
- Contains more:
  - Geographic position (where?)
  - Attributes (what?)

- Example: school's location (where) plus meta data (number of students, year built, funding, racial composition)

PLACE VS SPACE

- Place: Socially constructed locations with history and meaning
  - Neighborhoods, cities, regions
  - Characteristics, demographics, identity
  - Social and symbolic boundaries
- Space: Explicit focus on location and relative position
  - “where things are or where they happen”
  - How they relate to other locations
  - Distance and proximity relationships
  - Critical concept(s): **distance** – proximity, exposure, access
  - In our case: how does this affect people

WHY DOES IT MATTER?

- “Everything happens somewhere” (Logan 2012)
  - All social action is embedded in place
  - A person’s location can affect all kinds of outcomes
- Hence, these relationships matter:
  - Distance affects social relationships
  - Proximity influences exposure (to, e.g., behaviors) and access (to, e.g., education, health care)
  - Geographic clustering reveals patterns

WHY DOES IT MATTER?

- Example: Public transportation usage (Baum-Snow & Kahn 2000)
- Who uses public transportation?
  - Mainly homeowners, college-educated, non-African American residents
  - Can you come up with a story for why this could be?

WHY DOES IT MATTER?

- Example: Public transportation usage (Baum-Snow & Kahn 2000)
- Who commutes via mass transit?
  - Mainly homeowners, college-educated, non-African American residents
  - Can you come up with a story for why this could be?

  => Real reason: mass transit was mostly built in suburban areas where these people live

  => Without spatial data, measuring availability of mass transit (i.e., distance to station), conclusions might have looked very differently

WHY IS IT IN THIS COURSE?

It has become easier to acquire and repurpose these data

- More data:
  - Census data; administrative records
  - Crowdsourced data (OpenStreetMap)
  - Satellite imagery
  - But also new data sources: Mobile location data, Social networks
- Better tools
  - Mapping tools in R (*sf*, *ggmap*, *spdep*, *spatialreg*)
  - APIs
  - Computer vision

=> Including these things into our research provides a new angle and facilitates a more realistic representation of the (social) world

=> … and the figures just look dope as hell

DISTANCE

First Law of Geography: “everything is related, but near things are more related than distant things” (Tobler 1970)

“the most basic source of homophily is space: We are more likely to have contact with those who are closer to us in geographical space” (McPherson et al. 2001)

- Distance can have certain meanings:
  - Exposure (to nearby things)
  - Access (to resources)
- Problem: how do we measure distance adequately based on, e.g., survey data

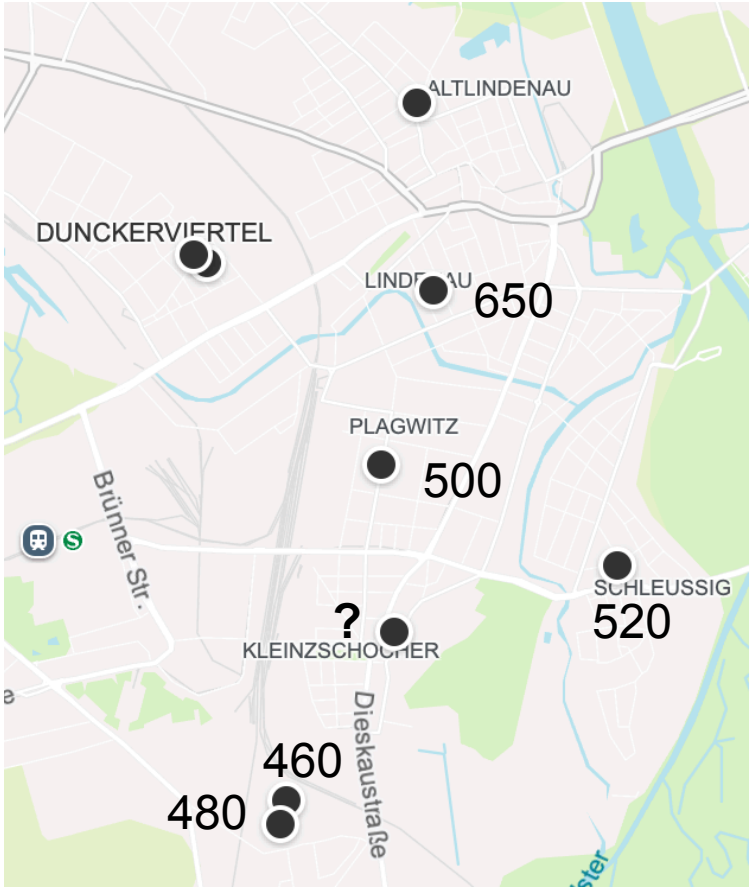
DISTANCE

- Problem: how do we measure distance adequately based on, e.g., survey data
  - “Egocentric neighborhoods” – connecting people’s location’s characteristics to the people’s characteristics – if there is good data
    - E.g., built environment (street access etc.) around a person’s home and their travel behavior (Frank et al. 2004)
  - If there is insufficient data: assume a person is living at the centroid of their unit of measurement (e.g., zip code, county, city) – simplifying assumption

DISTANCE

Refinement: “kriging”

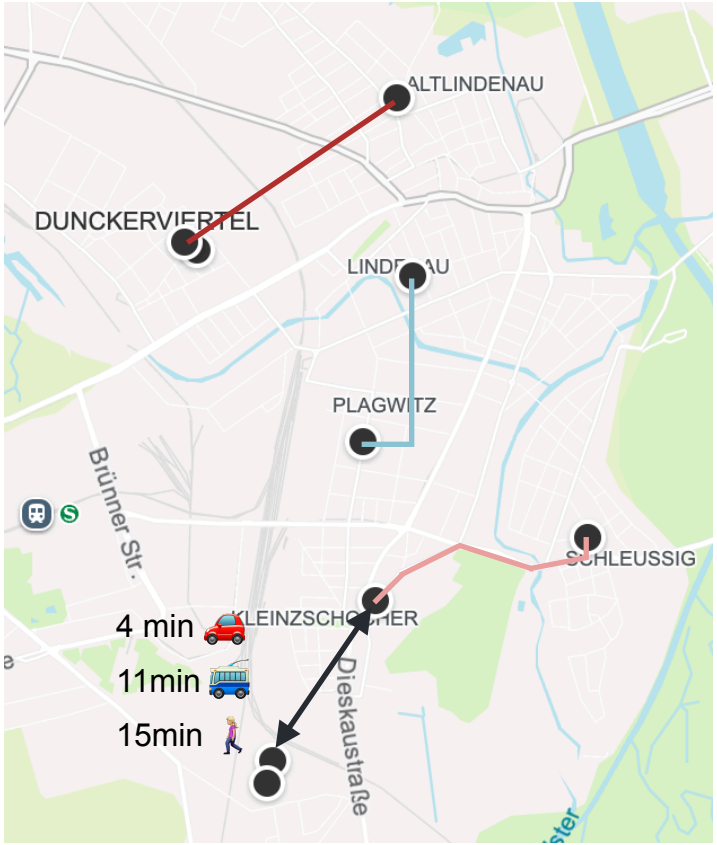
- simulate the distribution of people at every point in space – example: house prices at different locations
- Then you just treat it as if you had full data



Immobilienscout24 – fictitious prices (unfortunately)

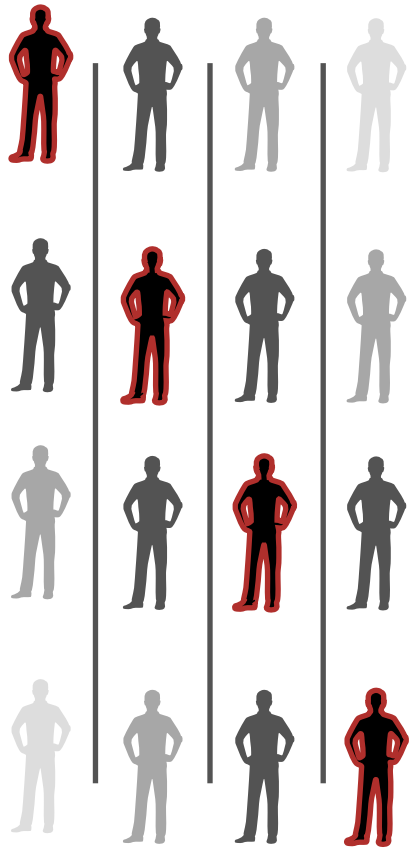
DISTANCE

- Also: how do we measure distance?
  - Euclidean: as the crow flies —
  - Manhattan (north/south, east-west) —
  - Road distance —
  - Actual driving/cycling/walking/public transport time (e.g., Google Maps API) ⟷ => Can also take into account physical boundaries (rivers, train tracks, etc. **– they can make for exciting natural experiments!**)
- Once we have this: how much more do closer places matter?



DISTANCE

- Once we have this: how much more do closer places matter? – DECAY function $f(d_{ij})$ – first law of geography
  - Problem: tricky to know the actual function
  - Solutions:
    - find realistic, empirical estimates (e.g., interview people, measure decay – e.g., when looking at segregation, measure percentage of co-ethnic residents at different cut-off distances)
    - try different models (linear, exponential, cut-offs), use them all, compare results



DISTANCE

- Another problem: distance from where
  - Usual assumption: place of residence
  - But: other places might matter as well – work place, school place

DISTANCE – CONCLUSION

- These are all things to bear in mind when thinking about distance
  - Depends very much on data availability
  - No obviously superior option, but be transparent about the potential drawbacks of your choice

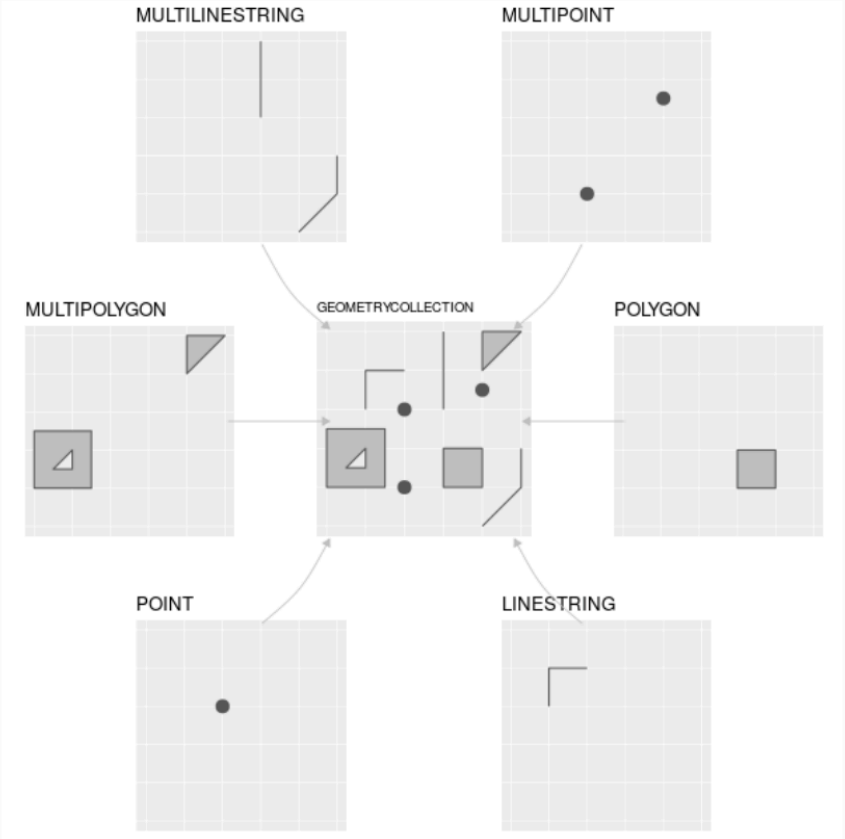
TECHNICAL ASPECTS

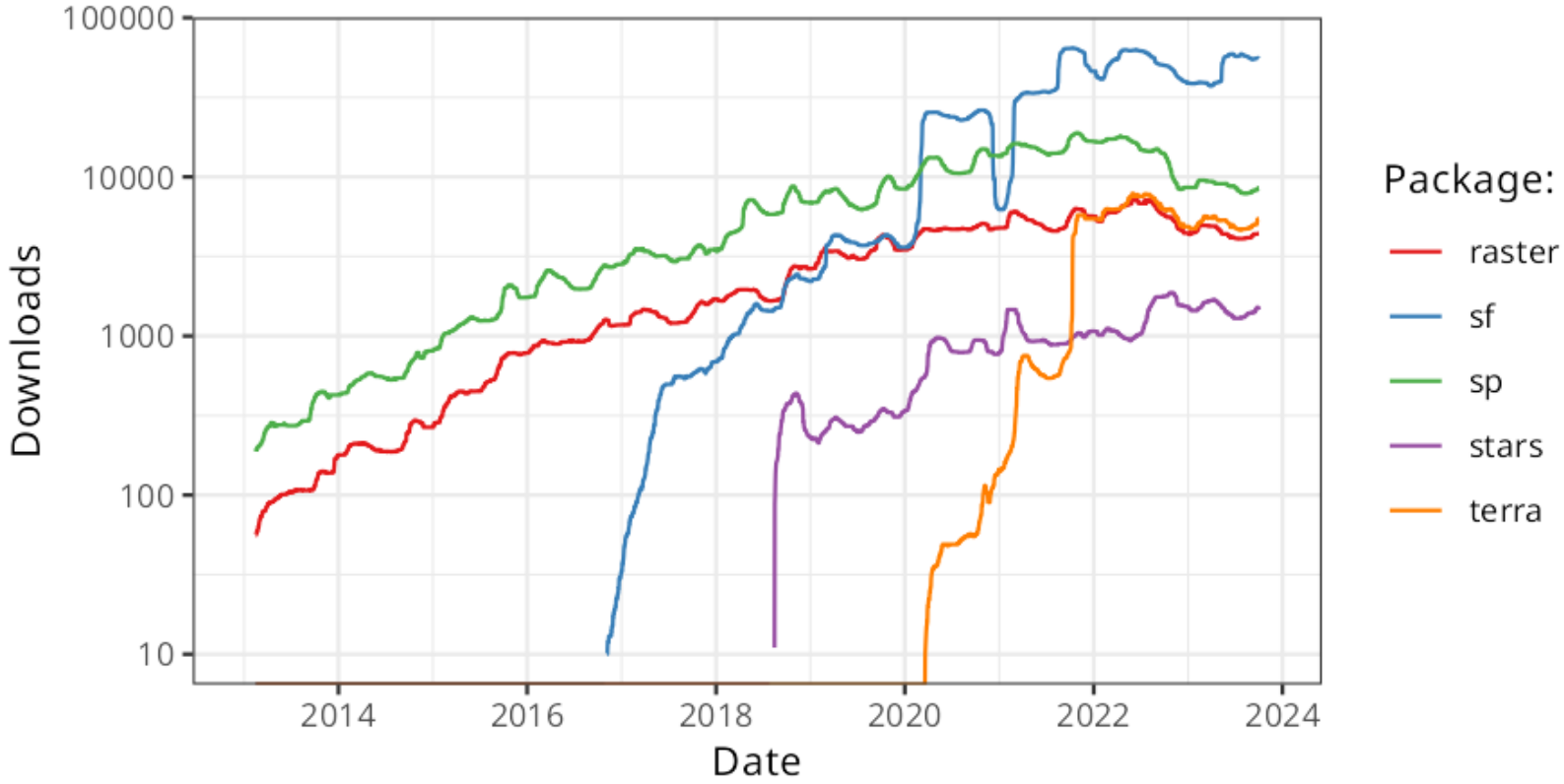
- How is this implemented from a technical perspective?
  - vector model (points, lines, polygons)
    => we will work with this and the *sf* package (Pebesma 2018)
  - raster model (dividing the surface into even cells)
    => natural sciences usually work with this

  => data can be converted as well

TECHNICAL ASPECTS

sf = simple features – supports different geometry types

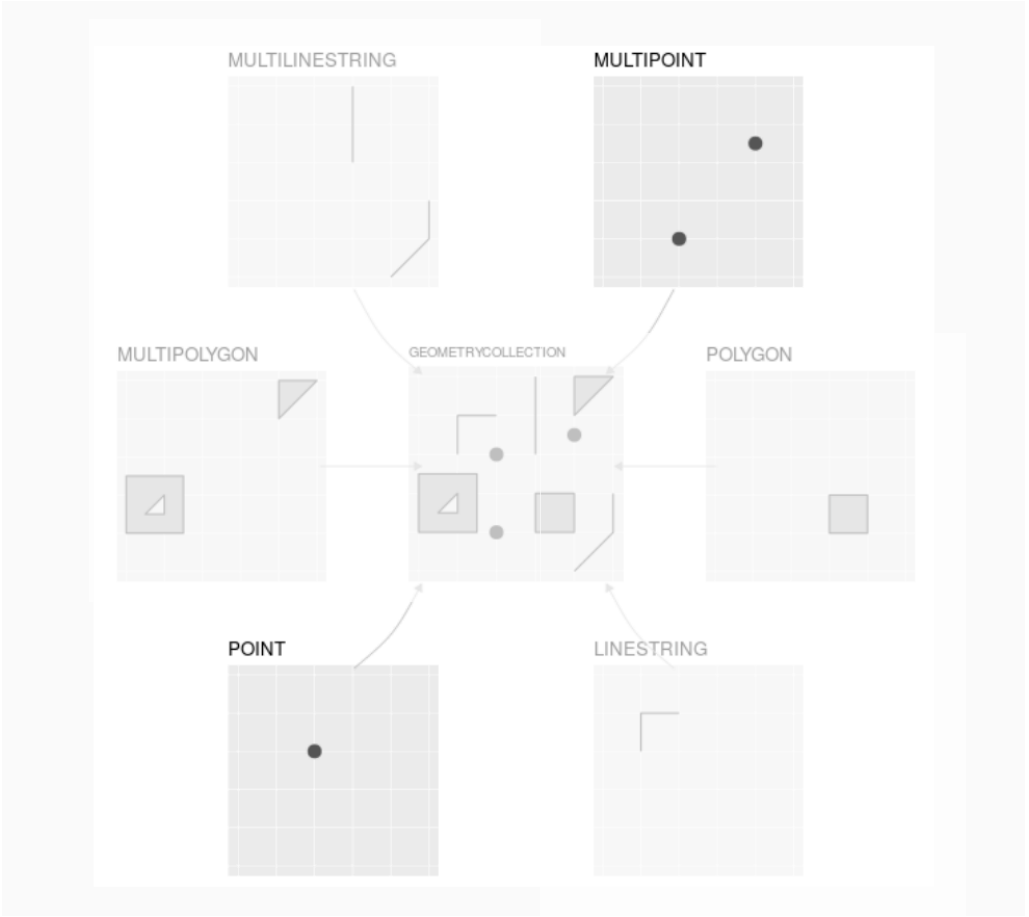



100000
10000
1000
100
10
Downloads
2014
2016
2018
2020
2022
2024
Date
Package:
raster
sf
sp
stars
terra

TECHNICAL ASPECTS

Points: specific locations, signified by x, y coordinates

Examples: schools, crime incidents

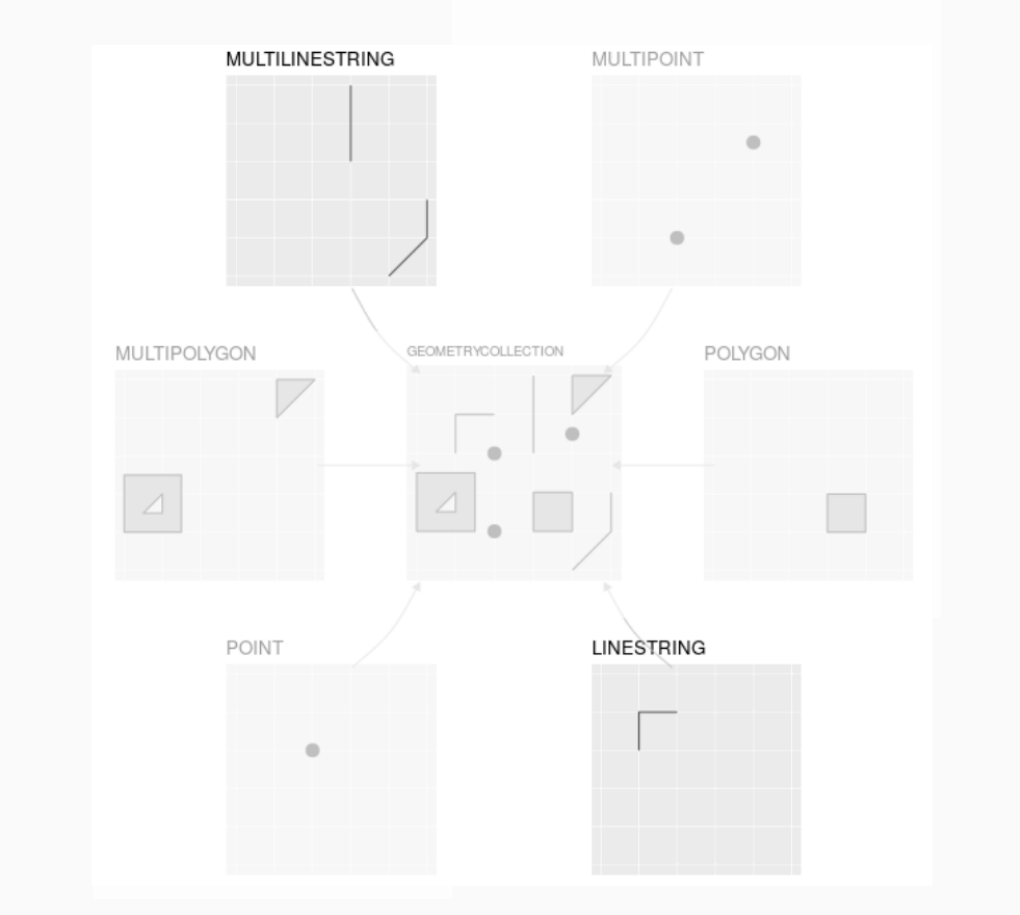


TECHNICAL ASPECTS

Linestring: networks and paths

Examples: roads, rivers, boundaries

Can be, for instance, used to measure distance along these line strings



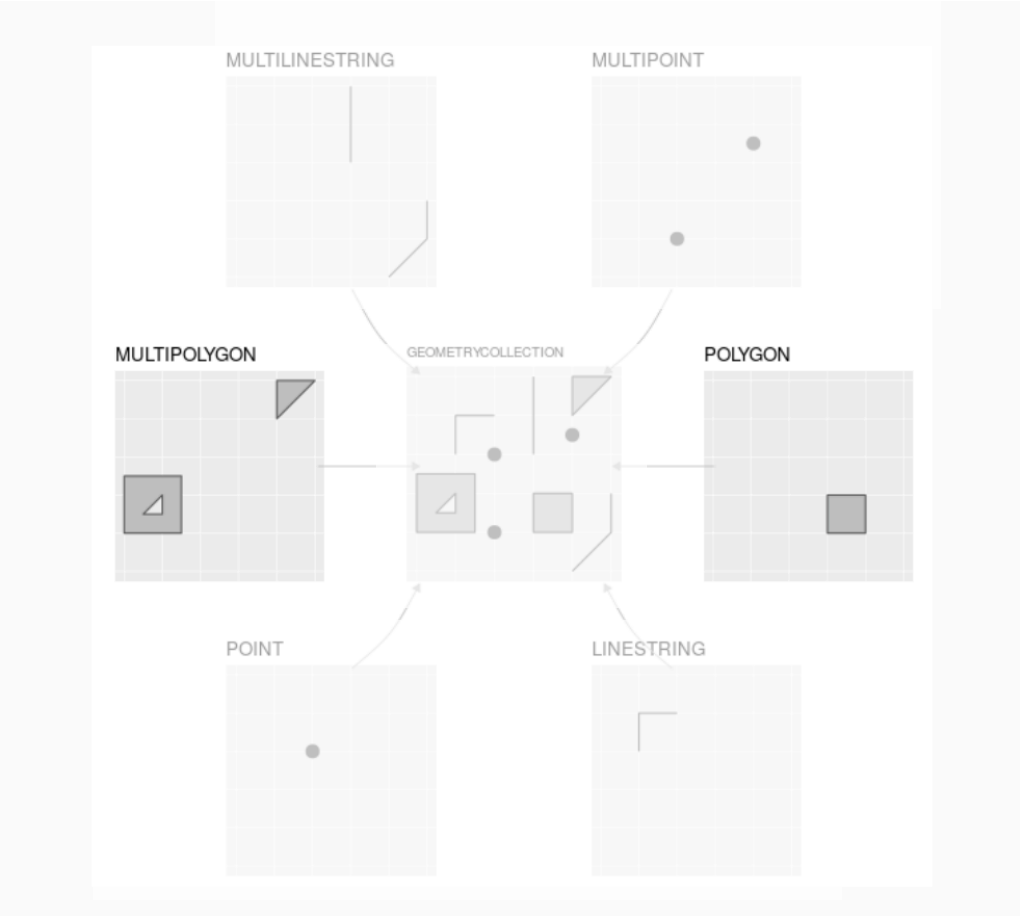
TECHNICAL ASPECTS

Polygons: areas and regions

Examples: census tracts, counties
=> shape files

To measure distances between two polygons: use their centroid

Also: see whether *points*/ *linestrings* are *within* a certain region (*polygon*)

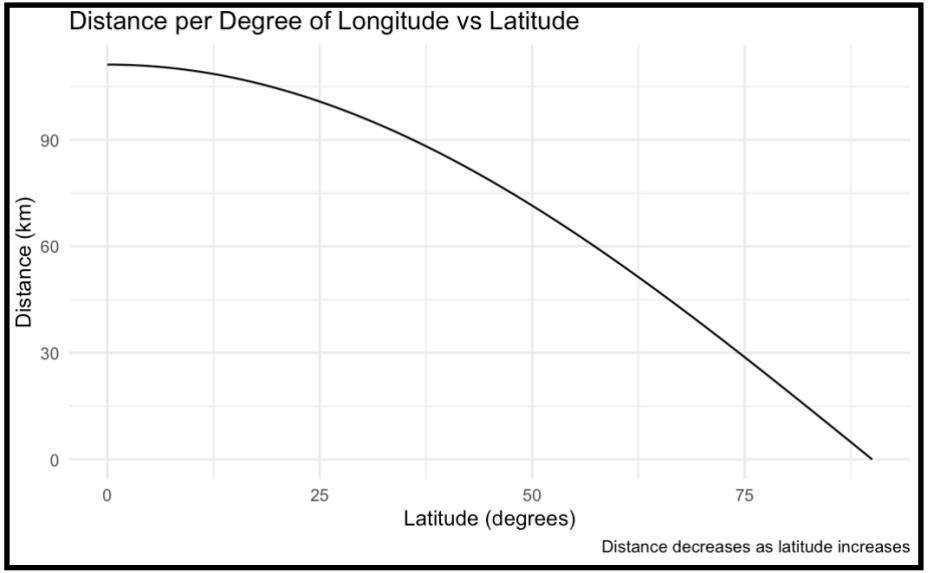


TECHNICAL ASPECTS – PROJECTION

TECHNICAL ASPECTS – PROJECTION

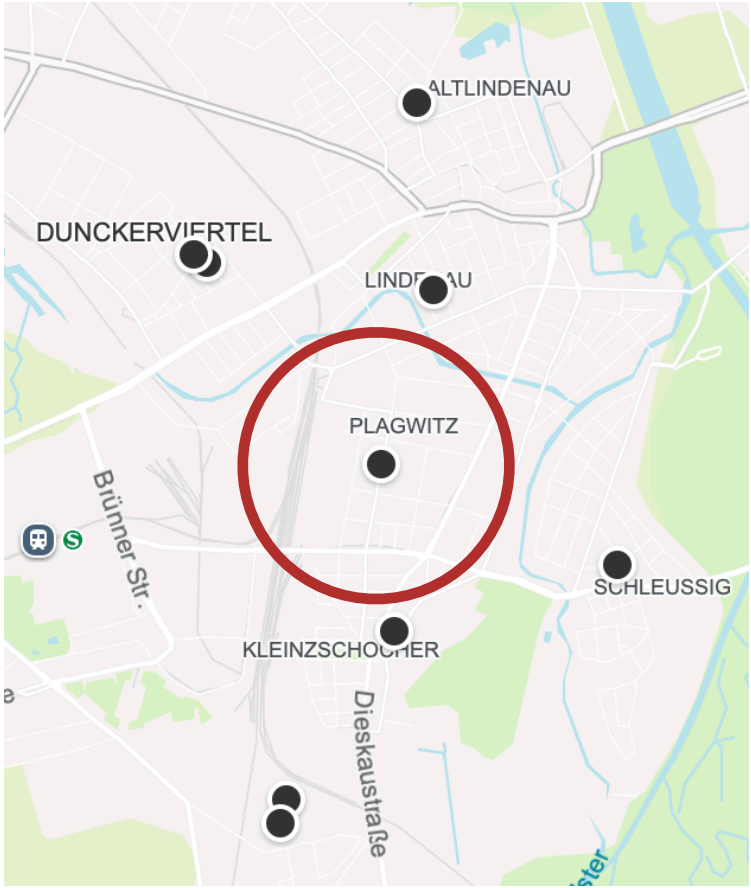
- The earth is not flat (if you disagree, please leave this room quietly)
- Coordinates as we know them (lat/long) are angular measurements
- As we saw before: Canada is overrepresented
  => the distance in km between lat=60 and lat=61 is smaller than between lat=0 and lat=1

Fix: project to local coordinate systems (find adequate projection here: https://epsg.io/; see examples here: https://aditya-dahiya.github.io/visage/geocomputation/crs_projections.html)



TECHNICAL ASPECTS – BUFFERS

Sometimes you also want to know whether something is **within** a certain distance – this is what buffers are for

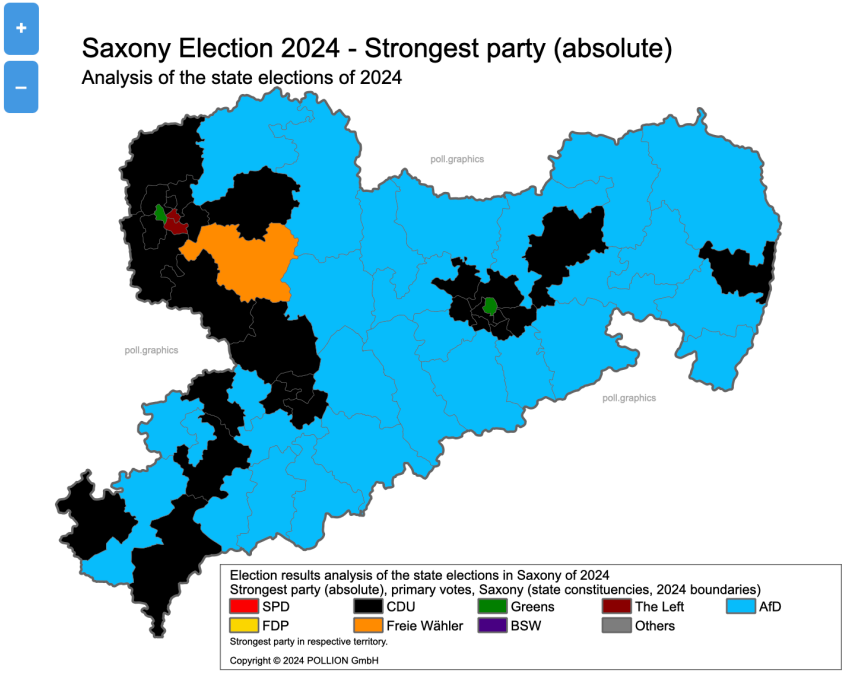


TECHNICAL ASPECTS – MODIFIABLE AREAL UNIT PROBLEM

The way we draw our boundaries may affect our outcome drastically

E.g., vote shares on quarter, city, commuting zone level

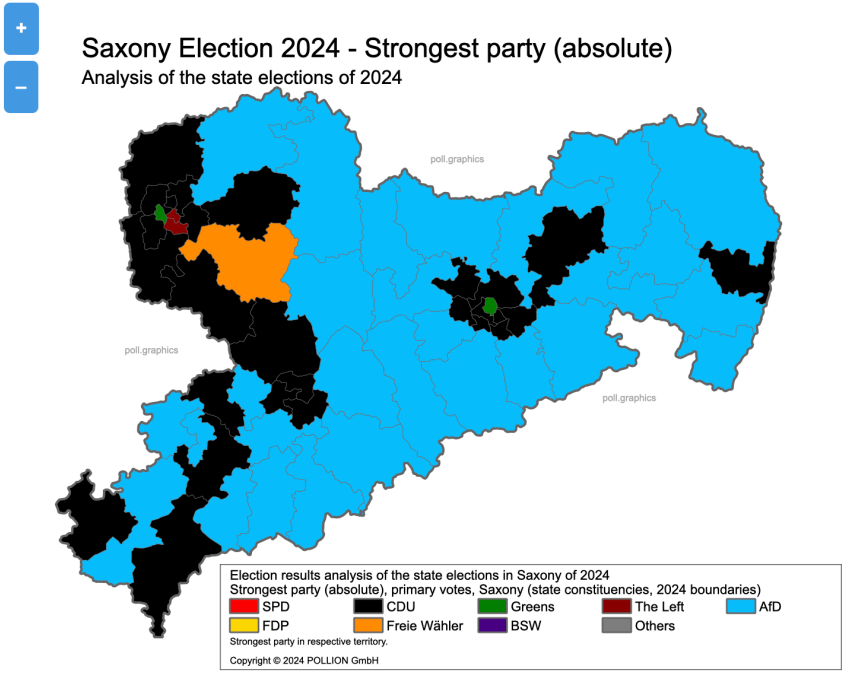
=> Multiple ways to address this, zero ways to fully solve this



TECHNICAL ASPECTS – MODIFIABLE AREAL UNIT PROBLEM

1. Use the smallest unit possible
2. Then aggregate from there to get an idea of the extent of the issue
3. If data allow for it: use distance-based measures and decay functions instead of fixed boundaries
4. Multilevel models can incorporate multiple levels of aggregation at the same time (more on this next session)
5. Sensitivity analysis: run the same model using different areal units

=> Important: be transparent about it, bear in mind how your choice of unit might change your results



TECHNICAL ASPECTS – AUTOCORRELATION/MORAN'S I

First Law of Geography: "everything is related, but near things are more related than distant things" (Tobler 1970)

=> Spatial Autocorrelation – things are more similar (autocorrelated) due to proximity

We want to describe how values of the same variable co-vary based on their location (e.g., race in neighborhoods – segregation, mean income in neighborhoods)

TECHNICAL ASPECTS – AUTOCORRELATION/MORAN'S I

Measure: Global Moran's I

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

N = number of spatial units

W = sum of all spatial weights

$w_{i\ j}$ = spatial weight between i and j

x_i = value at location i

$\bar{x}$ = mean value

TECHNICAL ASPECTS – AUTOCORRELATION/MORAN'S I

How much do these two points "matter" for each other – distance

How different are these two points from the mean respectively => does this for all possible locations => gets large if there are more extreme values

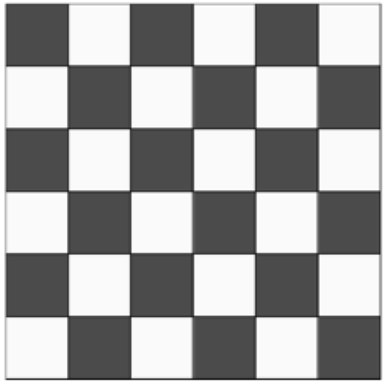
$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Adjustment by number of observations and connectivity

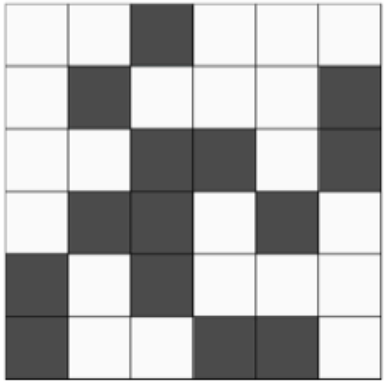
Normalizes by overall variation

TECHNICAL ASPECTS – AUTOCORRELATION/MORAN'S I

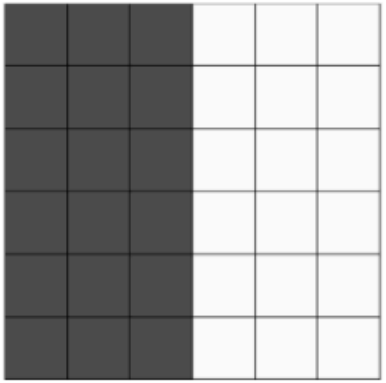
Negative spatial autocorrelation



No spatial autocorrelation



Positive spatial autocorrelation



TECHNICAL ASPECTS – AUTOCORRELATION/MORAN'S I

Measure: Moran's I

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

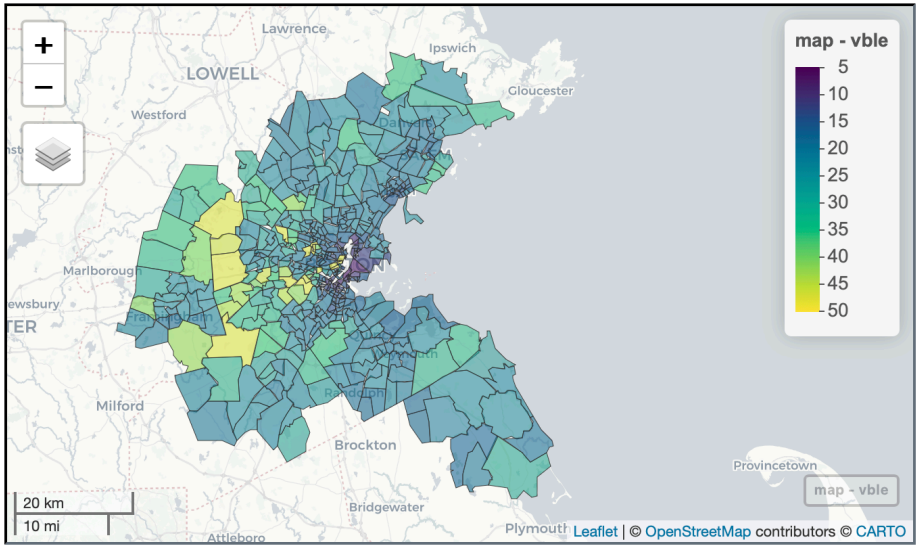
N = number of census tracts in Boston

W = sum of all spatial weights

$w_{i\,j}$ = spatial weight between census tract i and census tract j

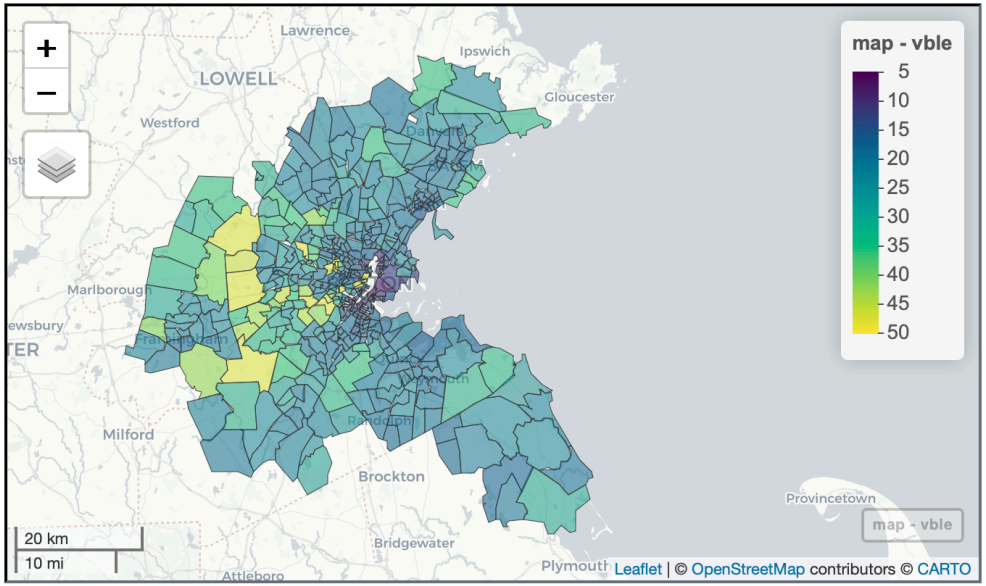
x_i = median house price at census tract i

$\bar{x}$ = mean value of median house prices overall



https://www.paulamoraga.com/book-spatial/spatial-autocorrelation.html

TECHNICAL ASPECTS – AUTOCORRELATION/MORAN'S I



https://www.paulamoraga.com/book-spatial/spatial-autocorrelation.html

Moran's I = 0.63

TECHNICAL ASPECTS – LOCAL MORAN'S I

Global Moran's I gives us a measurement for the entirety of the units

Local Moran's I gives us an estimate *per unit*

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2/n} \sum_{j \in N_i} w_{ij}(x_j - \bar{x})$$

n = total number of spatial units

$w_{i\,j}$ = spatial weight between i and j

x_i = value at location i

x_j = values all neighboring units

x_k = values of all units

$\bar{x}$ = mean value

TECHNICAL ASPECTS – LOCAL MORAN'S I

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2 / n} \sum_{j \in N_i} w_{ij}(x_j - \bar{x})$$

"how much does x_i differ from the mean" – z-standardized value of x_i

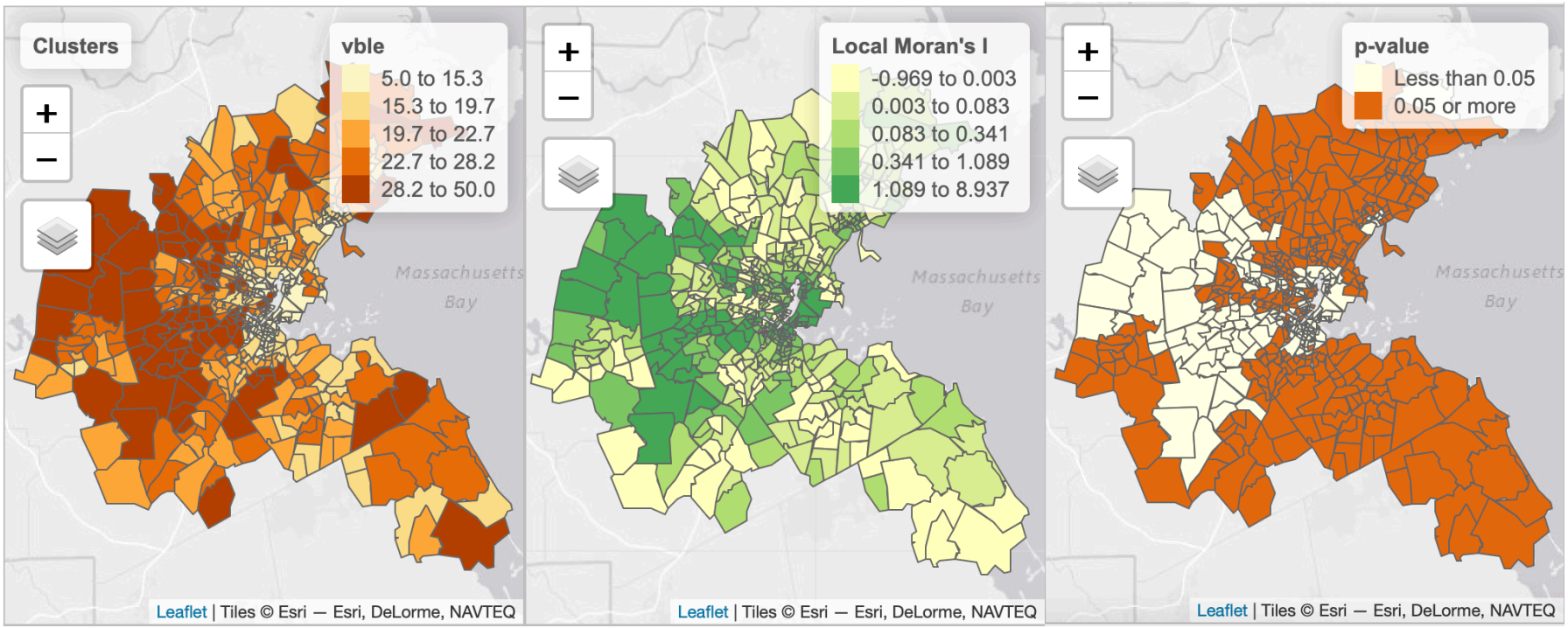
"how much do x_i's neighbors differ from the mean" – z-standardized value of x_j

TECHNICAL ASPECTS – LOCAL MORAN'S I

Goal: classification into

- high-high (hotspot): significant p, high Local Moran's I surrounded by high local Moran's Is, high value in general
- high-low (outlier): high surrounded by low
- low-low (cold spot): low surrounded by low
- low-high (outlier): low surrounded by high
- insignificant (unclear pattern)

TECHNICAL ASPECTS – LOCAL MORAN'S I



https://www.paulamoraga.com/book-spatial/spatial-autocorrelation.html

CONCLUSION – FOR NOW

- Space adds a new way to control for factors that our classic models arguably cannot account for (i.e., distance, presence of certain things)
- We can ask new questions
- But also need to bear in mind new pitfalls
- And, most importantly: maps look darn good in papers

UNIVERSITÄT
LEIPZIG

MERCI

Felix Lennert
Institut für Soziologie

felix.lennert@uni-leipzig.de
www.uni-leipzig.de