Data Linking II:

Georeferenced Survey Data

Stefan Jünger,
GESIS Data Archive

CESSDA Training Day
27-28 November 2019, Cologne

cessda.eu @CESSDA_Data

cessda

gesis Leibniz Institute for the Social Sciences

Data Linking II: Georeferenced Survey Data

Stefan Jünger, GESIS Data Archive

November 29, 2019

CESSDA Training Days 2019 | November 27-28, Cologne

Leibniz Association

Content

Introduction

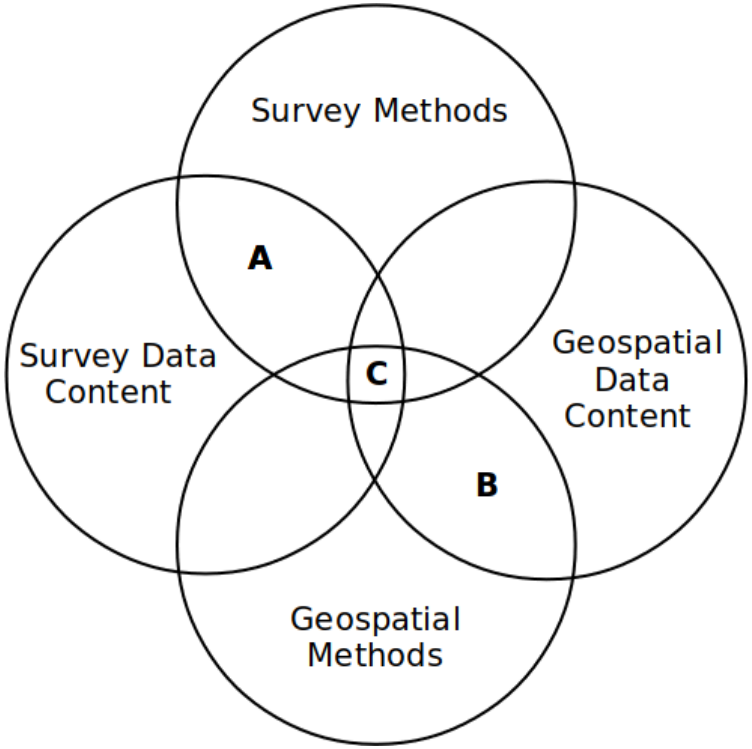
- Data Types
- General Concepts

Challenges

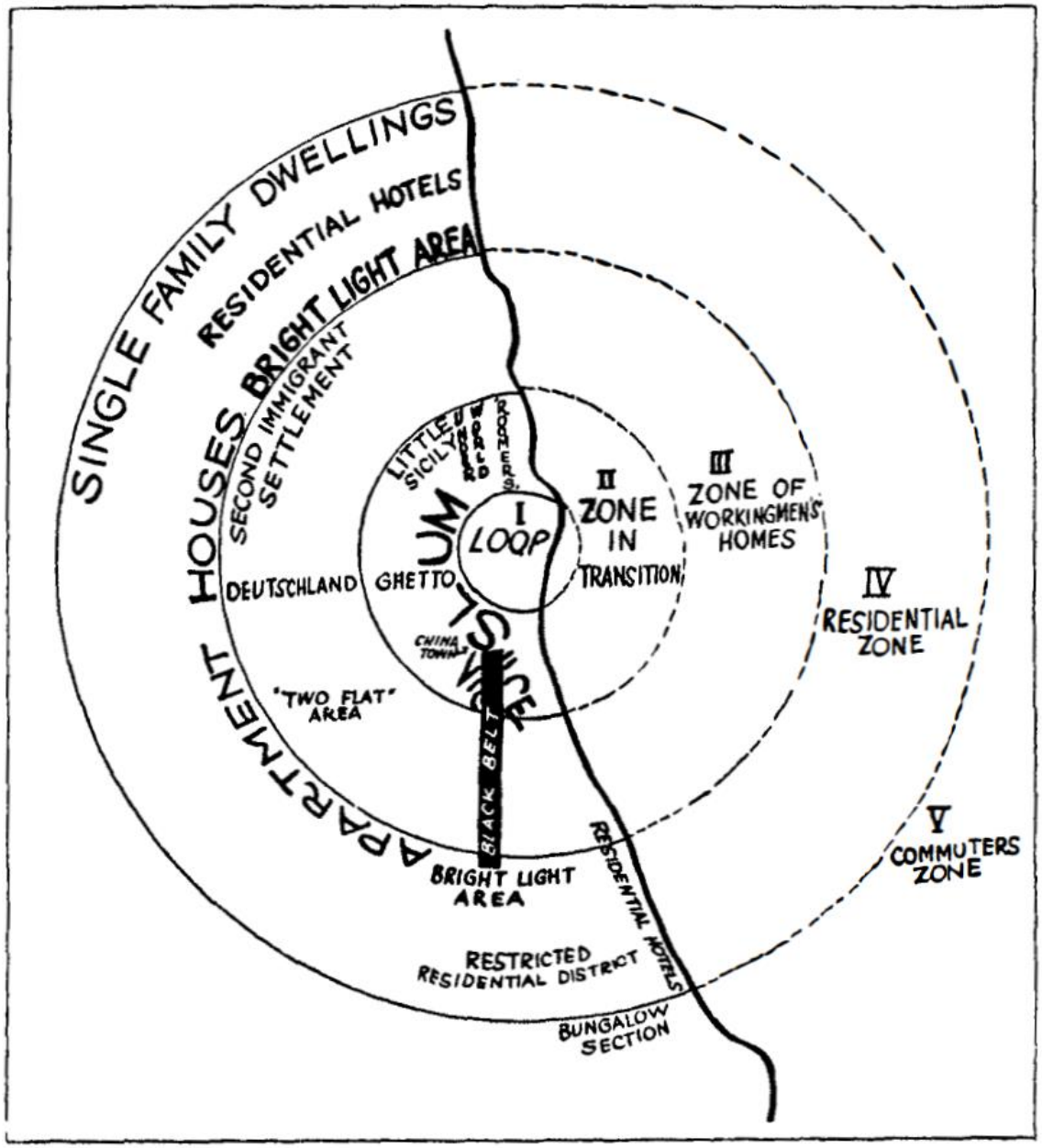
- Technological and organizational
- Data Privacy

Spatial Linking & Applications

Extra: Documentation



Jünger, 2019



Park et al., **1925**

Space & Place – an old hat?

Space has been a topic already for a long time

- Chicago School
- Incorporated in classic theories, e.g., Allport's Contact Theory (1952)

Subject of Urban Studies and regional sociology

- e.g., study of Gentrification
- Social mobility

From the past to (almost) today

Qualitative and theoretical work is manifold

- A lot of regionally limited studies
- Difficult to translate to

...quantitative research

- Space often defined by data
  - e.g., administrative borders
- Neighborhood as 'container'

Today

Increased amount of available data

- Quantitative and on a small spatial scale

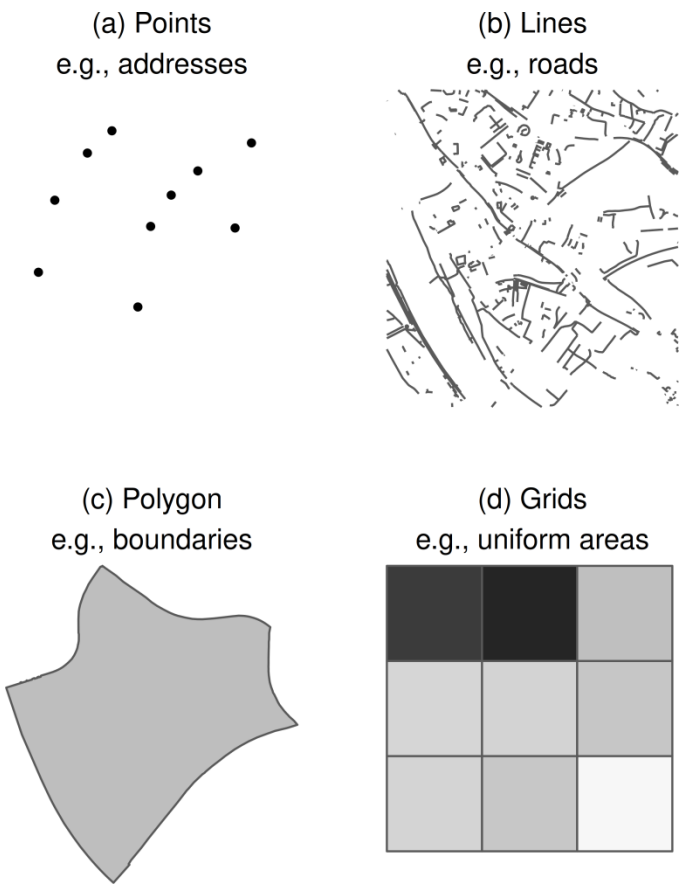
Better tools

- Personal computer with enough horse power
- Standard software, such as R, can be used as Geographic Information System (GIS)

More applications which use georeferenced data

Introduction

Georeferenced Data



Jünger, 2019

Data with a direct spatial reference → **Geo-coordinates**

- Contain information about geometries
- Optional: Content in reference to the geometries

Georeferenced Survey Data

Survey data enriched with geo-coordinates

- Or other direct spatial references
- I'll stick with geo-coordinates, however

ALLBUS

GLES German Longitudinal Election Study

Panel gesis

...

SOEP

Prerequisite: Geocoding

Indirect spatial references have to be converted into direct spatial references

- Addresses to geo-coordinates
- Necessary to project data in space

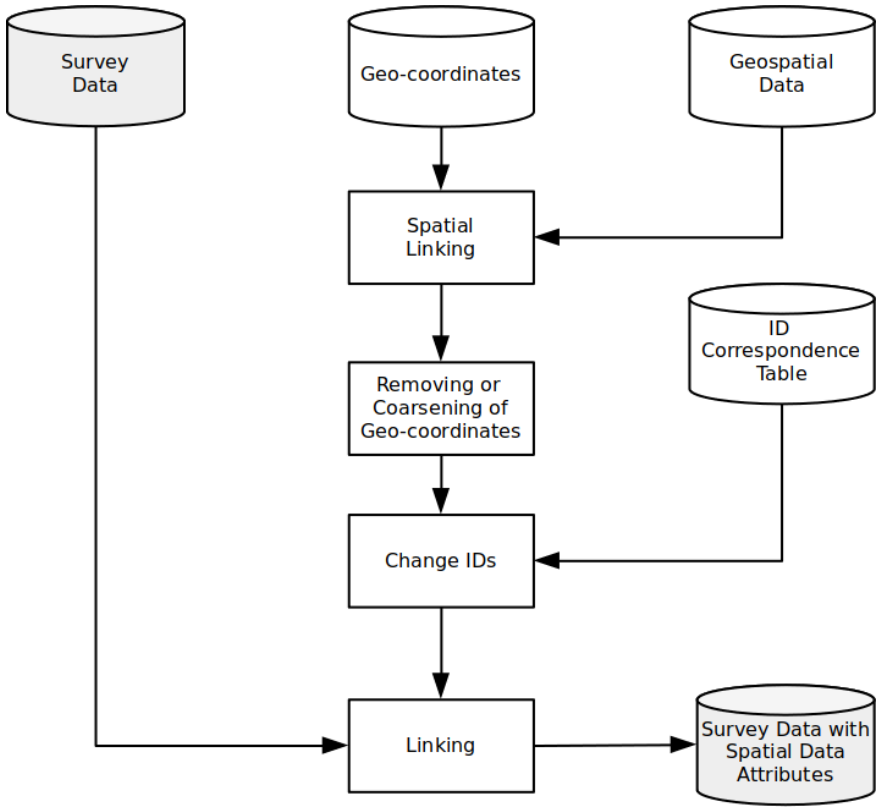
Different service providers can be used

- e.g., Google, Bing, OSM
- In Germany: Federal Agency of Cartography and Geodesy (BKG)

GeorefSurvData are no Geospatial Data

We must not store geo-coordinates and survey data in one dataset

- Differences to geospatial data
- More complicated workflow to work with (see Challenges)



Jünger, 2019

Geospatial Data

Essentially georeferenced data as defined before

- Information about geometries and related information

Can be projected jointly in one single space

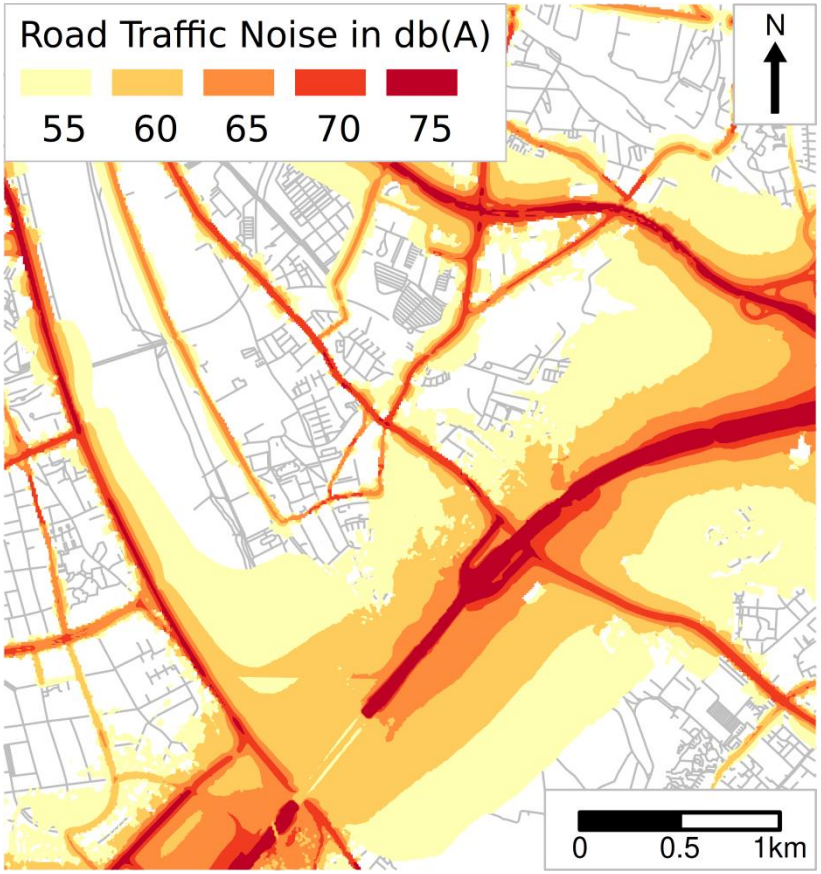
- Allows data linking and extraction of substantial information

This is why they can serve as auxiliary information, i.e., context data, for survey data!

Example I: Road Traffic Noise

Information on sound pressure levels in dB(A) for all main roads in Germany

- Collected in correspondence with EU Environmental Noise Directive
- Also information on rail, air, and industry noise

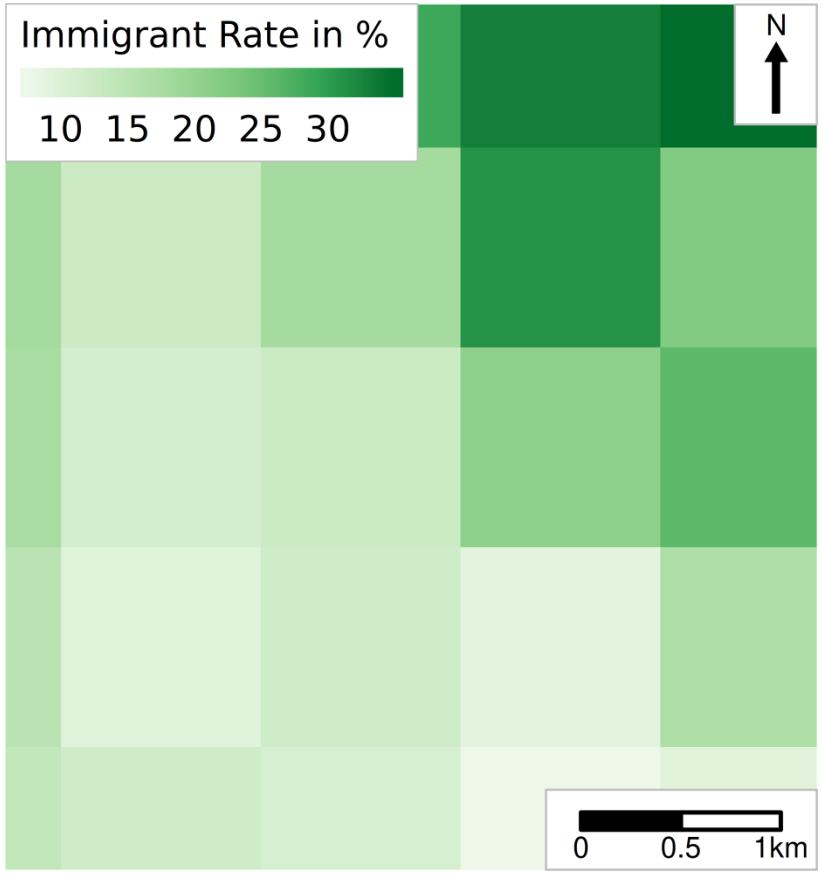


German Environmental Agency / EIONET Central Data Repository, 2016, and OpenStreetMap / GEOFABRIK, 2018 / Jünger, 2019

Example II: Immigrant Rates

Information immigrant rates in 1km² neighborhoods in Germany

- Collected in correspondence with 2011 European Census Regulation
- Also information on other sociodemographics

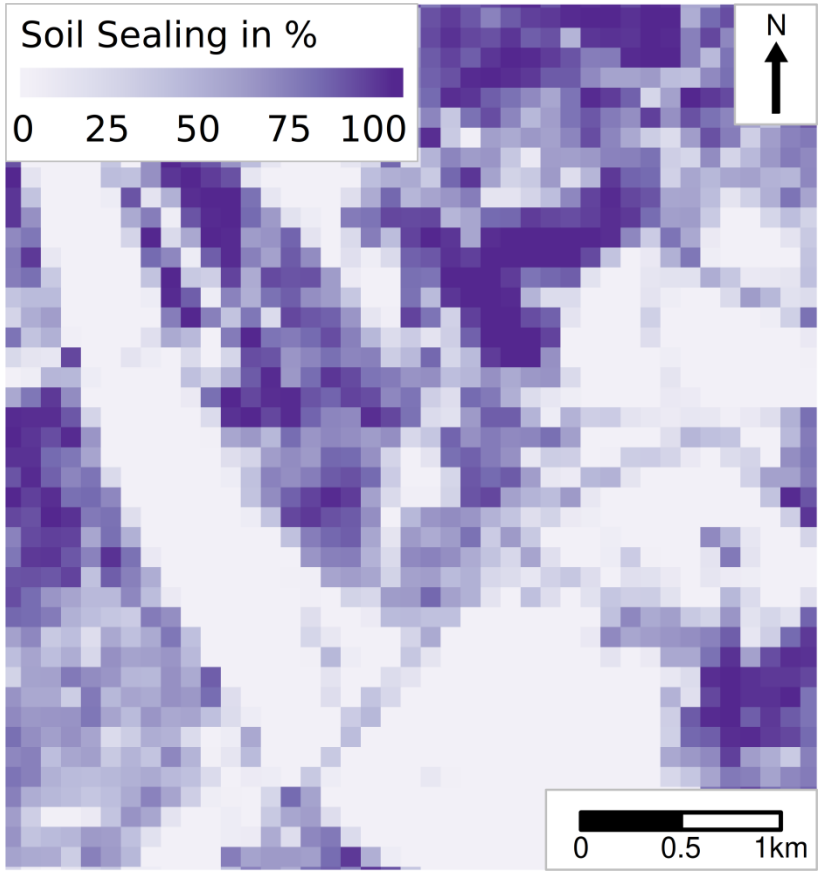


Statistical Offices of the Federation and the Länder, 2016 / Jünger, 2019

Example III: Soil Sealing

Information on air and water tight coverage of soils in 100m X 100m grid

- Part of the Monitor of Settlement and Open Space Development (IOER Monitor)
- Yes, there are even more land use indicators



German Environmental Agency / EIONET Central Data Repository, 2016, and OpenStreetMap / GEOFABRIK, 2018 / Jünger, 2019

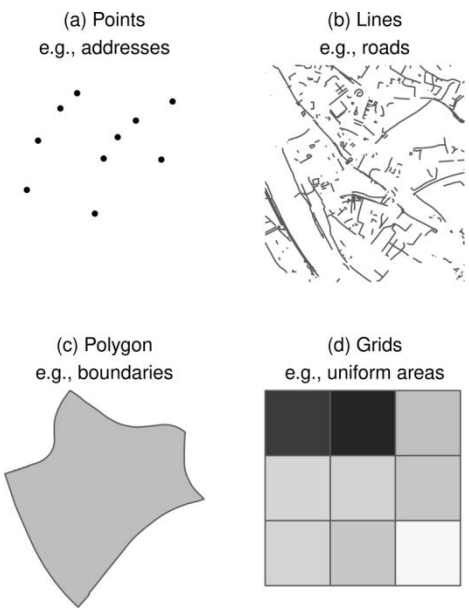
Data Specifics

Formats

- Vector data (points, lines, polygons)
- Raster data (grids)

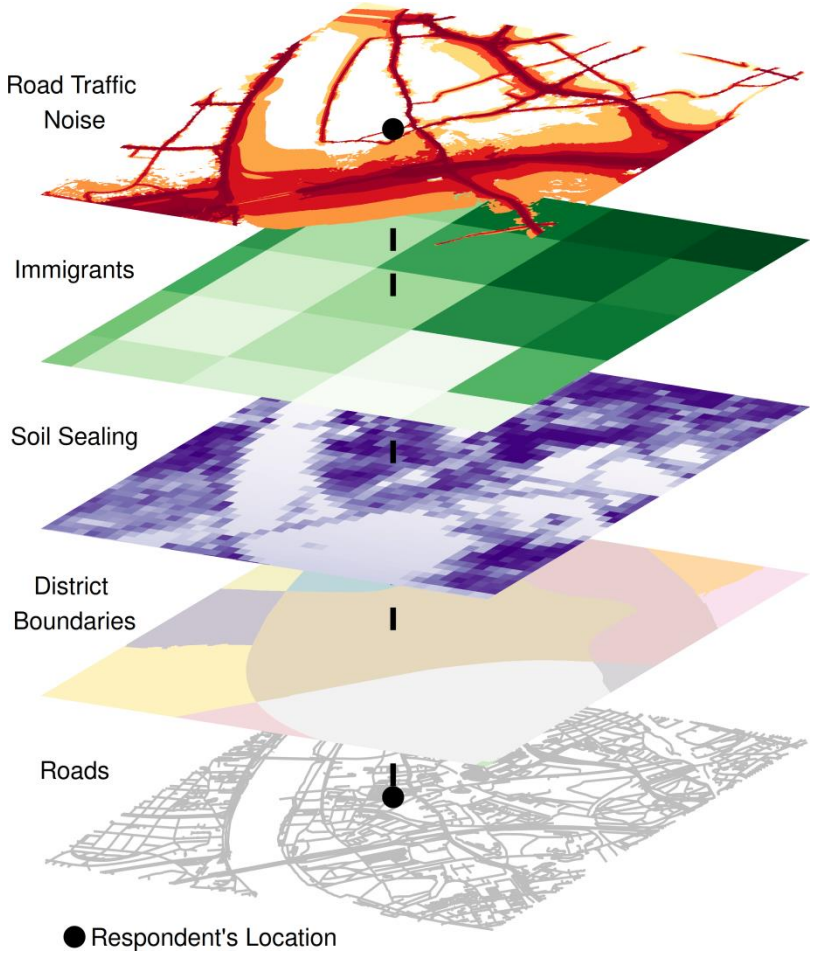
Coordinate reference systems

- Allow the projection on earth's surface
- Differ in precision for specific purposes
- Must match in order to conduct...



Jünger, 2019

Spatial Linking



OpenStreetMap / GEOFABRIK, 2018, City of Cologne, 2014, Leibniz Institute of Ecological Urban and Regional Development, 2018, Statistical Offices of the Federation and the Länder, 2016, and German Environmental Agency / EIONET Central Data Repository, 2016 / Jünger, 2019

Challenges

Data Availability

Geospatial Data

- Often de-centralized distributed
- Fragmented data landscape, at least in Germany

Georeferenced Survey Data

- Primarily survey data
- Depends on documentation
- Access difficult due to data protection restrictions

Technical Procedures

Geocoding

- Reasonable automated procedure
- But differ in quality and access rights
- High risk for data protection

GIS procedures

- Requires exploiting specialized software (you'll learn about QGIS today)
- Can get complex and resource intensive

Data Protection

That's one of the biggest issues

- Explicit spatial references increase risk of re-identifying anonymized survey respondents
- Can occur during the processing of data but also during the analysis

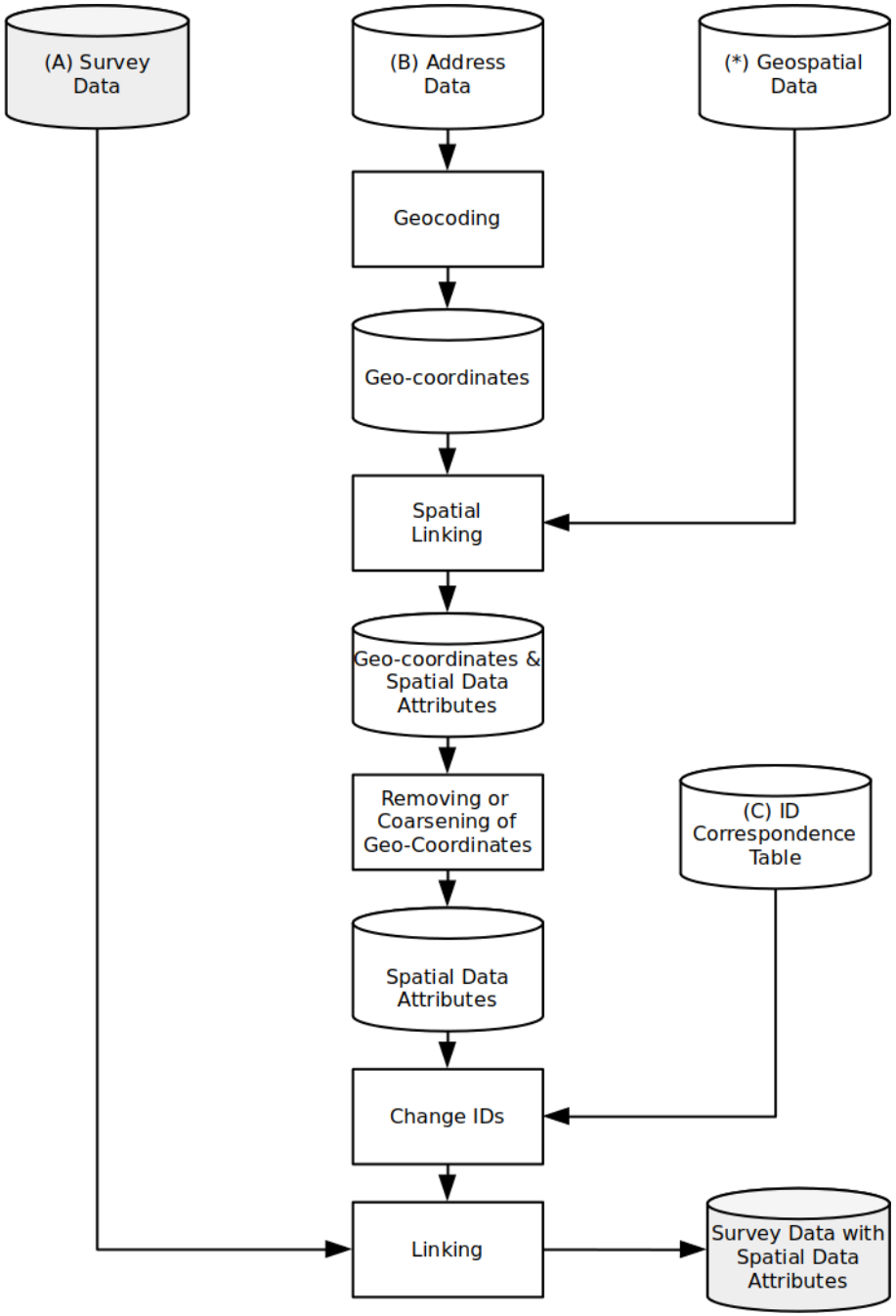
Affects all phases of research and data management!

Legal Regulations

Storing personal information such as addresses in the same place as actual survey attributes is not allowed in Germany

- Projects store them in separate locations
- Can only be matched with a correspondence table
- Necessary to conduct data linking

gesis Leibniz Institute for the Social Sciences



Jünger, 2019

Distribution & Re-Identification Risk

Data may still be sensitive

- Geospatial attributes add new information to existing data
- May be part of general data privacy checks, but we may not distribute these data as is

Safe Rooms / Secure Data Centers

- Control access
- Checks output

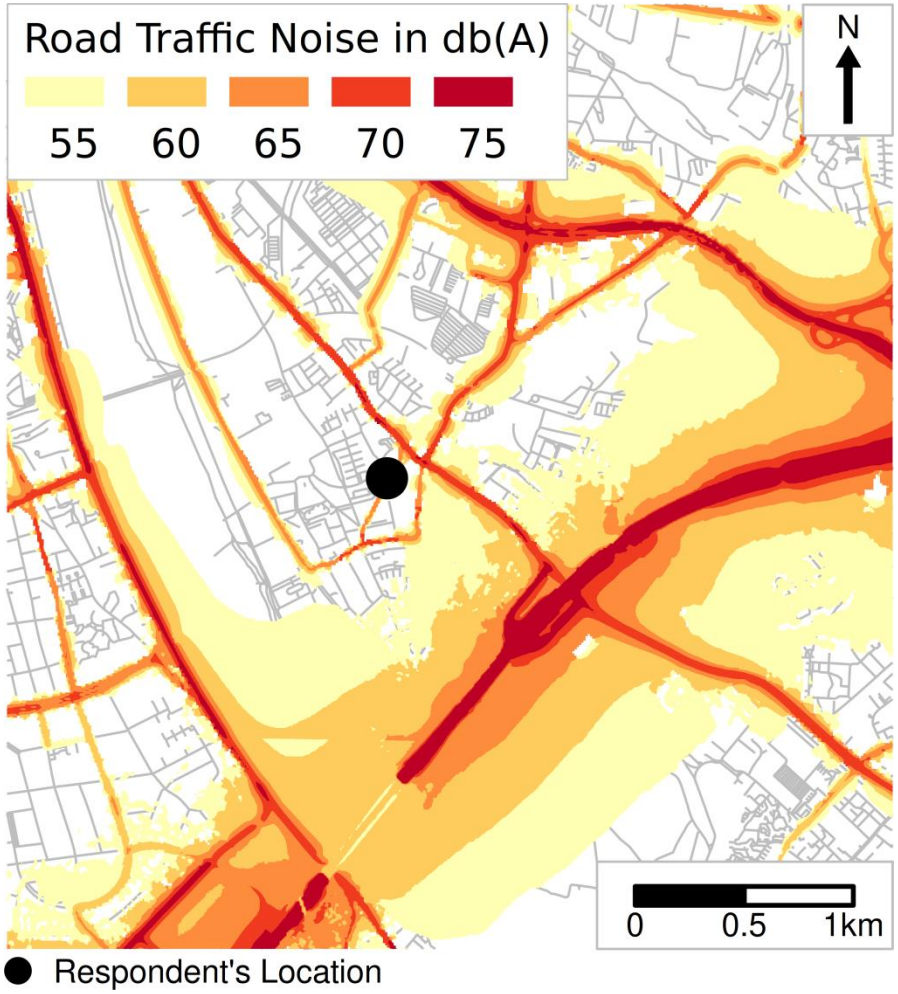
Spatial Linking & Applications

Flexible Tool of Spatial Linking

As georeferenced data are projected in one single space they can be related to each other

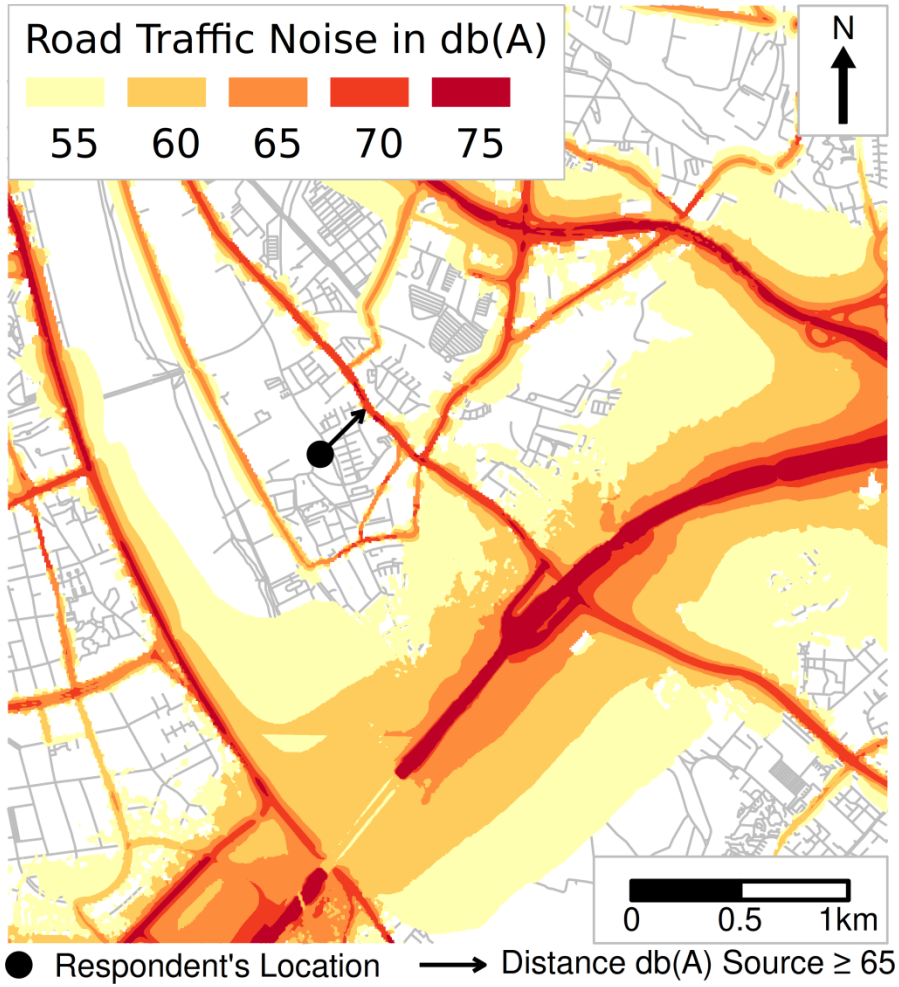
- 1:1
- By proximities, e.g., distances
- Combination, e.g., spatial buffers

1:1 Location



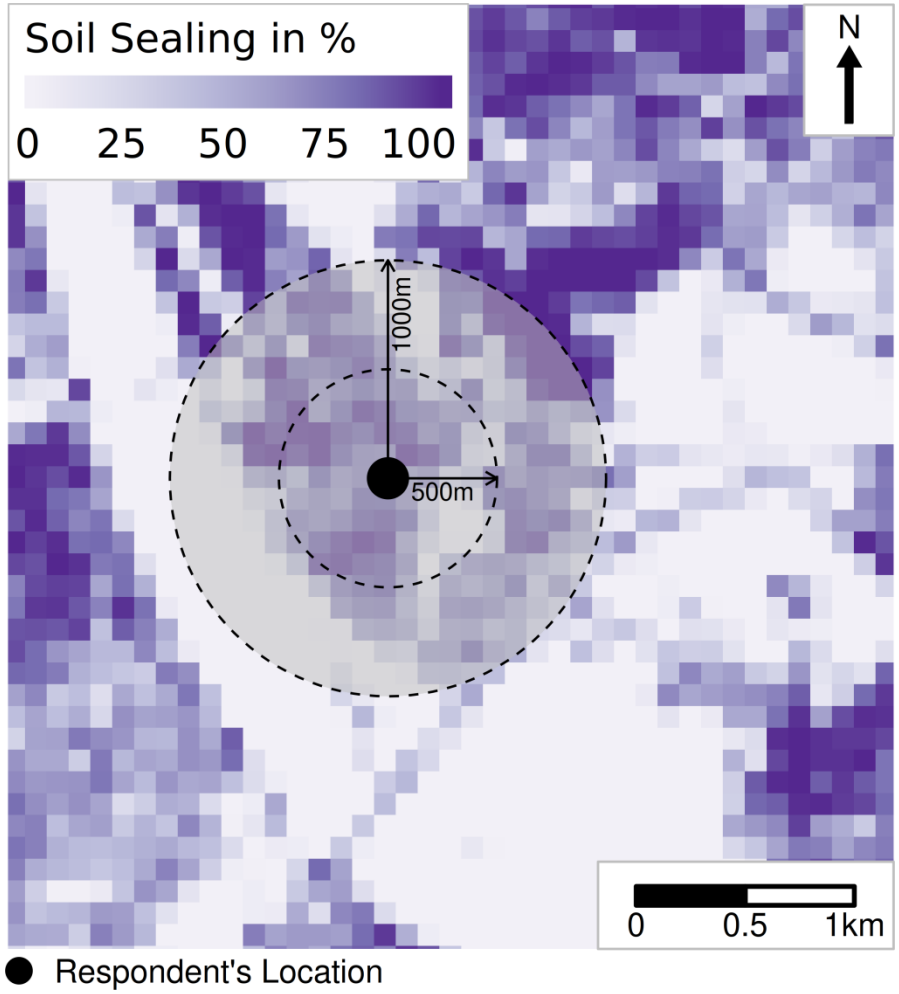
German Environmental Ageny / EIONET Central Data Repository (2016) and OpenStreetMap / GEOFABRIK (2018) / Jünger, 2019

Distances



Statistical Offices of the Federation and the Länder, 2016 / Jünger, 2019

Spatial Buffers



Leibniz Institute of Ecological Urban and Regional Development, 2018 / Jünger, 2019

Analysis

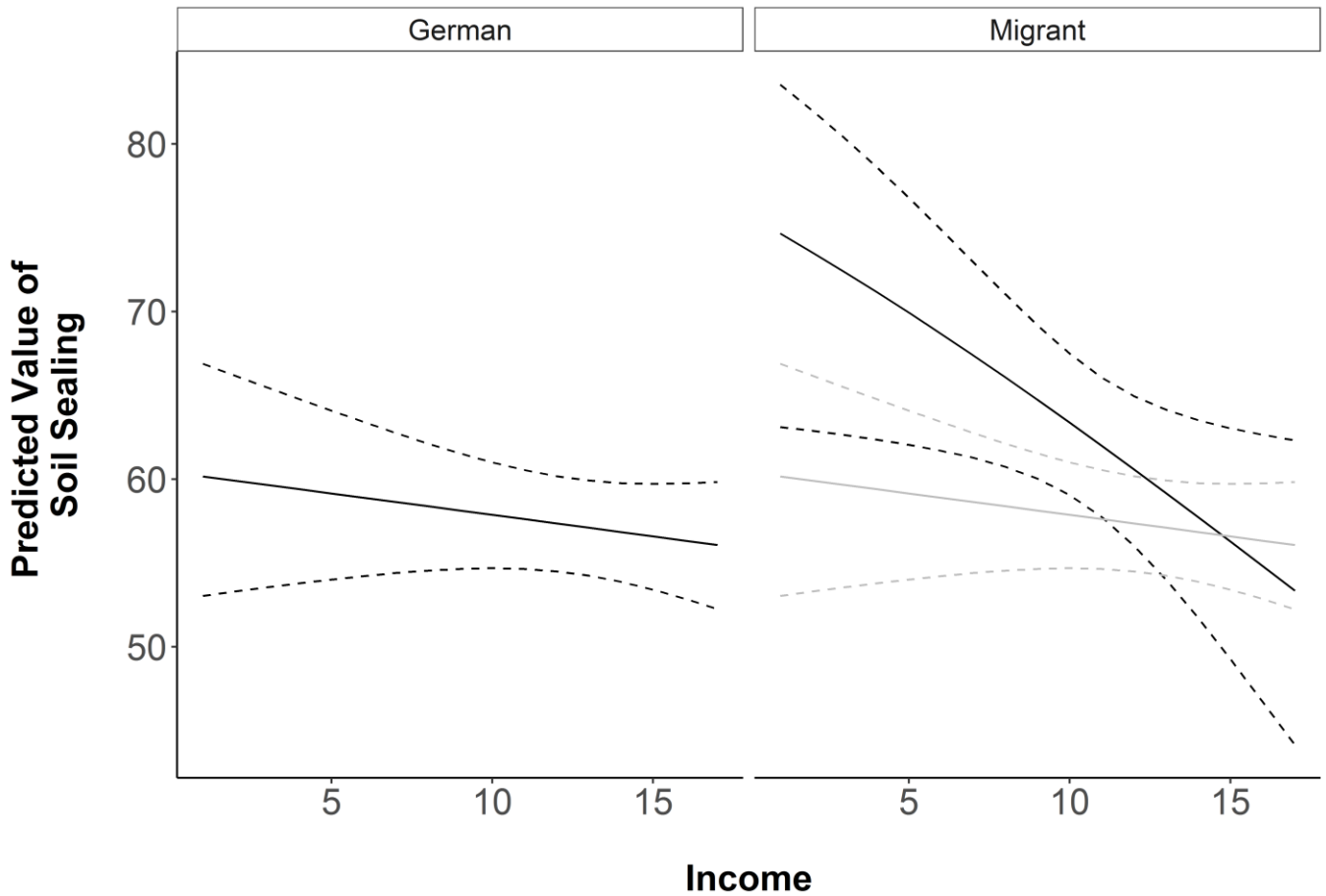
Special analysis techniques not necessarily needed

- As geospatial data are really small clustering because of that not a big problem

Data, however, may be endogenous

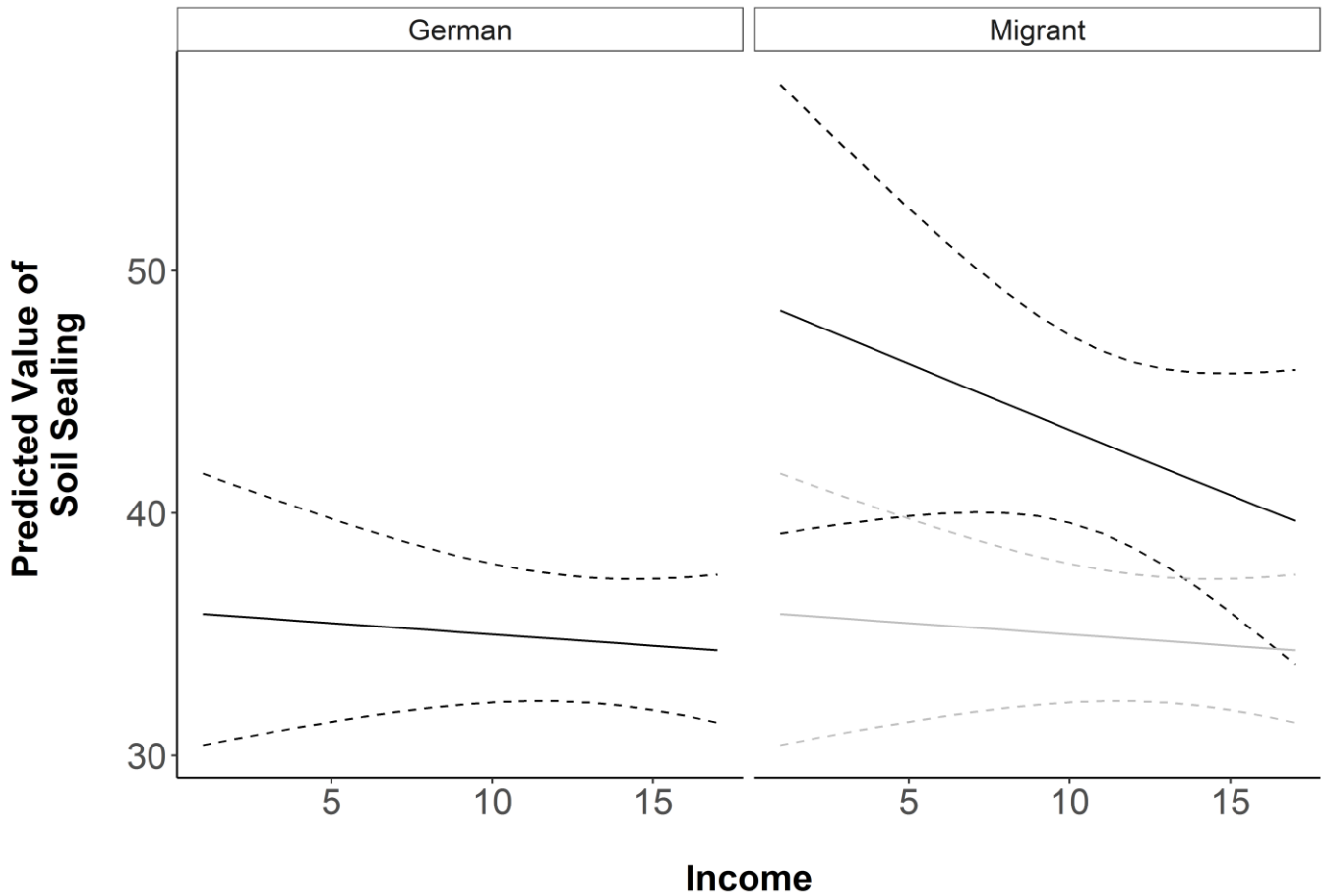
- OLS or Multilevel Models do not fix that!
- Fixed Effects or Spatial Lag Models may be more appropriate

In Brief: Environmental Inequalities



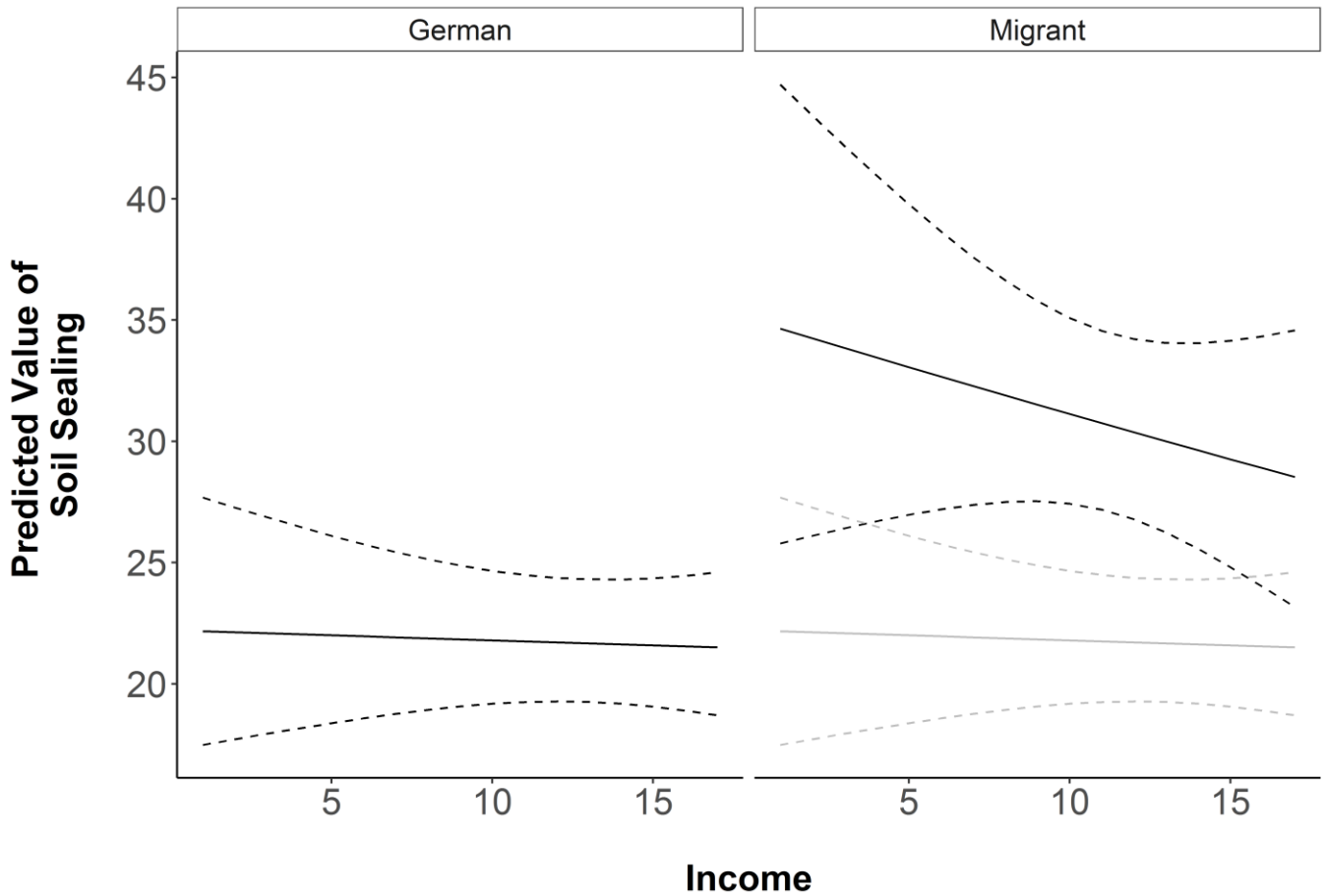
Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

In Brief: Environmental Inequalities



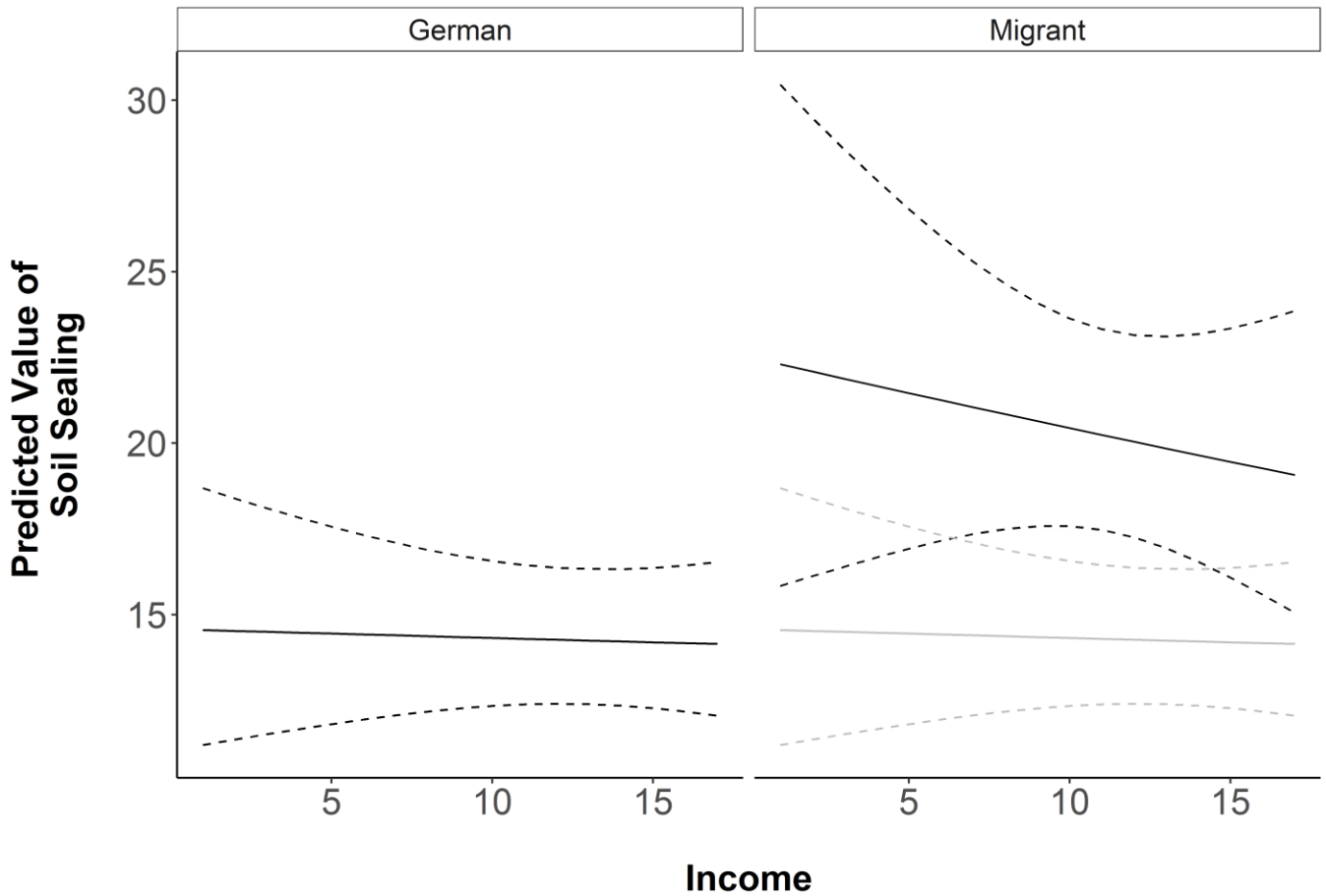
Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

In Brief: Environmental Inequalities

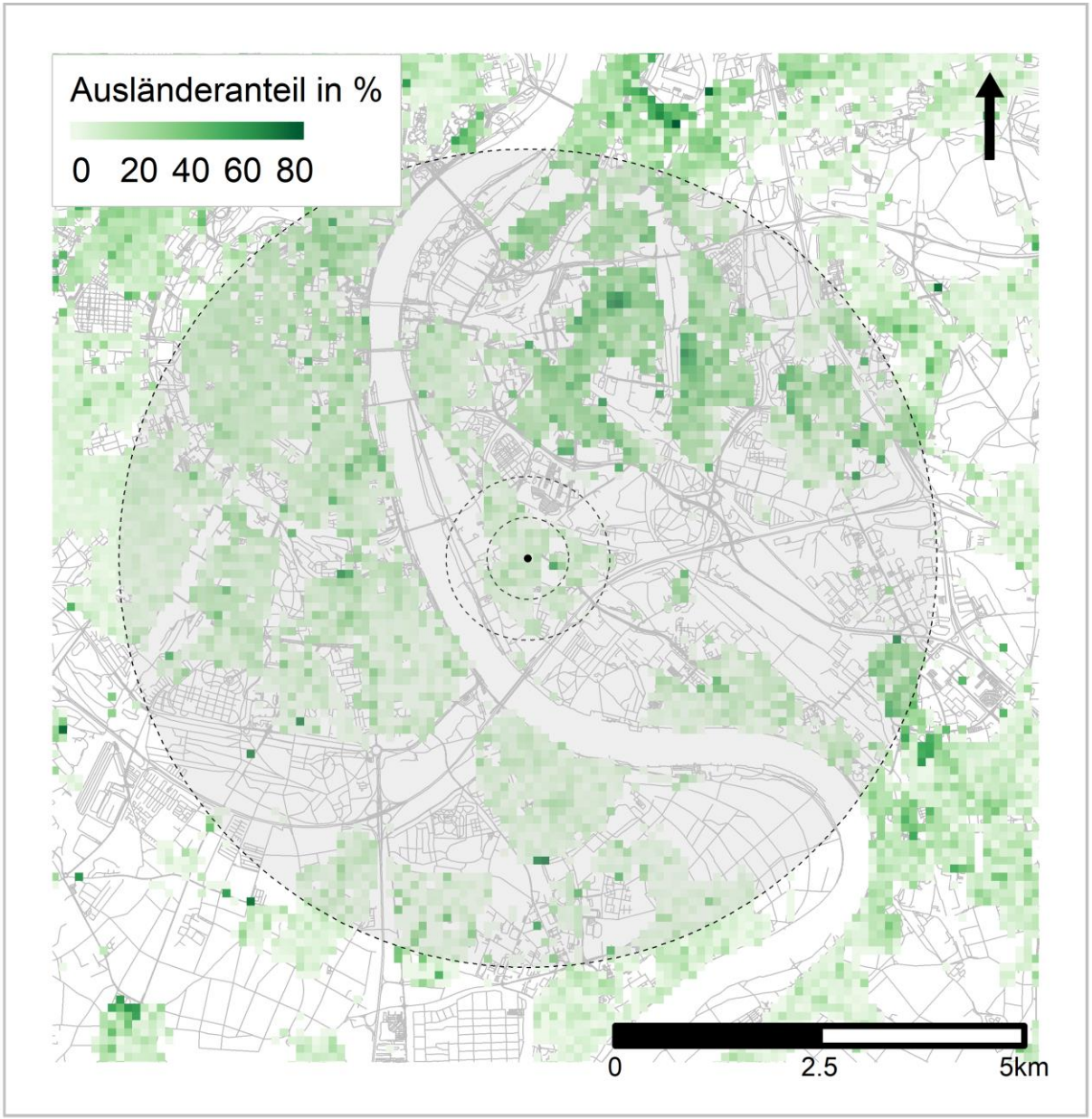


Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

In Brief: Environmental Inequalities

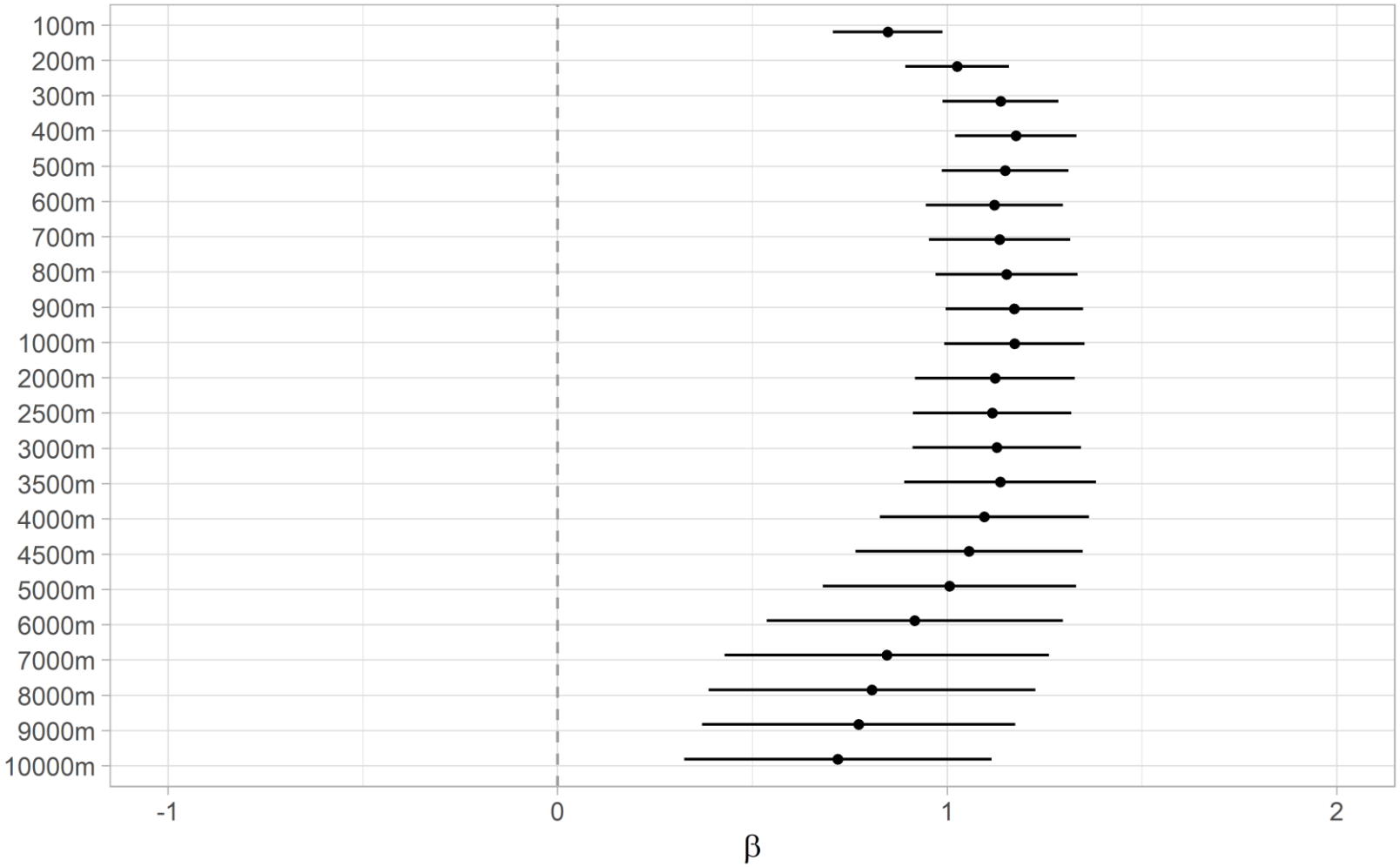


Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852



Datenquellen: OpenStreetMap / GEOFABRIK (2018), Stadt Köln (2014), und Statistische Ämter des Bundes und der Länder (2016)

Schätzung des subjektiven Ausländeranteils mit Pufferzonen im OLS Modell



Datenquelle: ALLBUS 2016; Standardisierte Regressionskoeffizienten mitsamt 95% Konfidenzintervall basierend auf Cluster-robusten Standardfehlern; alle Modelle kontrollieren Alter, Geschlecht, Bildung, Einkommen, Erwerbstätigkeit, Hauseigentum, Links-Rechts-Einstufung, Ost-West, Gemeindegrößenklasse; N = 1348

Extra: Documentation

based on content prepared by
Borschewski, Förster, Jünger, and Zenk-Möltgen

Documenting Georeferenced Social Science Survey Data with DDI

A CESSDA Metadata Office and CESSDA Training Video Tutorial

Kerrin Borschewski, André Förster, Stefan Jünger and Wolfgang Zenk-Möltgen

Based on:
Jünger, Stefan; Borschewski, Kerrin; Zenk-Möltgen, Wolfgang. 2019. Documenting Georeferenced Social Science Survey Data: Limits of Metadata Standards and Possible Solutions. *Journal of Map & Geography Libraries.* https://doi.org/10.1080/15420353.2019.1659903

metadata-office@cessda.eu
cessda.eu @CESSDA_Data

Use Case: Data Types & Sources

- Areal information supports analysis of social behavior

Social Science Survey Data

German General Social Survey (GGSS) (2014)

1st Geospatial Data Source

EIONET Central Data Repository (2012) – road traffic noise data

2nd Geospatial Data Source

German Census Data (2011) – immigrant rates

Spatial Linking (simplified)

Data Source: German Environmental Agency, EIONET Central Data Repository (2016) OpenStreetMap / GEOFABRIK (2018)

Data Source: German General Social Survey (2014)

Data Source: Statistical Offices of the Federation and the Länder (2016)

ID	Survey Question 1	...	Survey Question k	Road Traffic Noise (polygon data)	Immigrant Rates (1km² grid)
1	5	...	“maybe”	55	8.90
...	...	...	...	...	...
n	2	...	“yes”	75	34.78

Table 1: Structure of the Survey Data Enriched with Road Traffic Noise Measurements and Immigrant Rates

Metadata Standards

Social Science Survey Data	Geospatial Data	Georeferenced Survey Data linked to different sources of Geospatial Data
Data Documentation Initiative (DDI)	ISO 19115	???

Documentation Issue

- Data = geospatial data information from 2 different sources linked to social science survey data

- Data stem from different sources
- Data have different formats + different geographic structures

- Currently: no solution for this documentation issue

Workaround 1

- Split dataset logically into different studies (using DDI element StudyUnit)

No non-DDI files needed + DDI compliant

Lack of description for combined dataset

Workaround 2

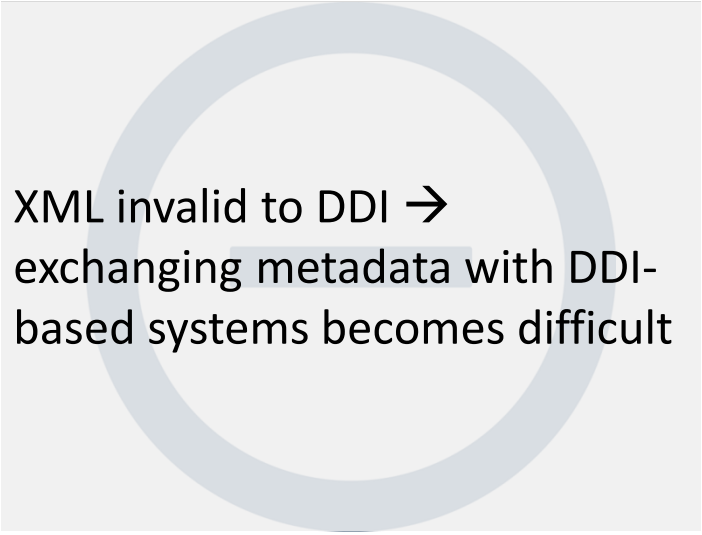
- Use ISO 19115 for geospatial data description and reference files in DDI at the variable level

Documentation in standards made for data types + DDI compliant

No semantic integration of attributes into DDI + non-DDI files needed

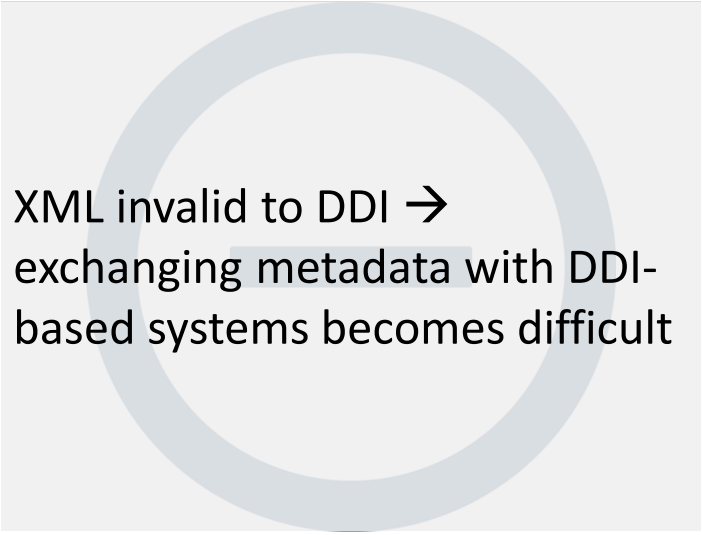
Workaround 3

- Apply options for geographic structure description (from study level) to variable level



Workaround 4

- Use options to describe geographic structures in DDI (at study level) + reference (illegitimately) from variable level



Workaround Choice

- Need for valid DDI instance?
- How many datasets involved?
- How many actors + stakeholders involved in processing of data/metadata?
- Metadata exchanged with other actors using DDI standard?

Summary

Emerging field of research

- Data has to be produced, curated and distributed
- ...and to be documented

Role of data infrastructures

- Support researchers
- Make themselves familiar with these new data

References

Allport, Gordon W. (1954). The Nature of Prejudice. Cambridge, Massachusetts: Addison-Wesley Publishing Company.

GermanEnvironmentalAgency/EIONETCentralDataRepository (2016). RoadTraffic Noise 2012 Shapefiles. Retrieved November 30, 2016, from https://github.com/stefmue/georefum/blob/master/data/cdr.road.lden.dat.rda

GESIS - Leibniz Institute for the Social Sciences (2015). ALLBUS/GGSS 2014 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2014). GESIS Data Archive. https://doi.org/10.4232/1.12209

Leibniz Institute of Ecological Urban and Regional Development (2018). Soil Sealing. Monitor of Settlement and Open Space Development. http://monitor.ioer.de/cgi-bin/wcs?MAP=S40RG_wcs

Jünger, Stefan (2019). Using Georeferenced Data in Social Science Survey Research: The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel. GESIS-Schriftenreihe 24. Köln: GESIS. https://doi.org/10.21241/ssoar.63688

Jünger, Stefan; Borschewski, Kerrin; Zenk-Möltgen, Wolfgang (2019). Documenting Georeferenced Social Science Survey Data: Limits of Metadata Standards and Possible Solutions. Journal of Map & Geography Libraries. https://doi.org/10.1080/15420353.2019.1659903

OpenStreetMap / GEOFABRIK. (2018). Governmental District Shapefiles. Retrieved September 28, 2018, from https://download.geofabrik.de/europe/germany/nordrhein-westfalen/koeln-regbez-latest-free.shp.zip.

Park, Robert E., Ernest W. Burgess, and Roderick D. McKenzie (1925). The City. Suggestions for Investigation of Human Behavior in the Urban Environment. Chicago and London: University of Chicago Press.

Statistical Offices of the Federation and the Länder (2016). Immigrant Rates. German Census 2011. Retrieved November 06, 2016, from https://github.com/stefmue/georefum/blob/master/data/census.attr.rda.

Q & hopefully A

stefan.juenger@gesis.org
@StefanJuenger

gesis Leibniz Institute for the Social Sciences

Leibniz Association