An Infrastructure for Spatial Linking of Survey Data

Bensmann, Felix; Heling, Lars; Jünger, Stefan; Mucha, Loren; Acosta, Maribel; Goebel, Jan; Meinel, Gotthard; Sikder, Sujit; Sure-Vetter, York; Zapilko, Benjamin

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Bensmann, F, et al. 2020. An Infrastructure for Spatial Linking of Survey Data. *Data Science Journal,* 19: 27, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2020-027

RESEARCH PAPER

An Infrastructure for Spatial Linking of Survey Data

Felix Bensmann[1], Lars Heling[2], Stefan Jünger[1], Loren Mucha[3], Maribel Acosta[2], Jan Goebel[4], Gotthard Meinel[3], Sujit Sikder[3], York Sure-Vetter[2] and Benjamin Zapilko[1]

[1] GESIS, Leibniz Institute for the Social Sciences, Cologne, DE

[2] Karlsruhe Institute of Technology, Karlsruhe, DE

[3] Leibniz Institute of Ecological Urban and Regional Development, Dresden, DE

[4] German Socio-economic Panel, Berlin, DE

Corresponding author: Felix Bensmann (felix.bensmann@gesis.org)

Research on environmental justice comprises health and well-being aspects, as well as topics related to general social participation. In this research field, among others, there is a need for an integrated use of social science survey data and spatial science data, e.g. for combining demographic information from survey data with data on pollution from spatial data. However, for researchers it is challenging to link both data sources, because (1) the interdisciplinary nature of both data sources is different, (2) both underlie different legal restrictions, in particular regarding data privacy, and (3) methodological challenges arise regarding the use of geo-information systems (GIS) for the processing and analysis of spatial data.

In this article, we present an infrastructure of distributed web services which supports researchers in the process of spatial linking. The infrastructure addresses the challenges researchers have to face during that process. We present an example case study on the investigation of environmental inequalities with regards to income and land use hazards in Germany by using georeferenced survey data of the GESIS Panel and the German Socio-economic Panel (SOEP), and by using spatial data from the Monitor of Settlement and Open Space Development (IOER Monitor). The results show that increasing income of survey respondents is associated with less exposure to land-use-related environmental hazards in Germany.

Keywords: spatial linking; georeferenced survey data; spatial data; environmental justice; research infrastructure; semantic web technologies

1. Introduction

The ongoing increase of environmental challenges and issues, such as urbanisation, accompanied by urban sprawl and soil sealing, loss of biodiversity, climate change and water shortage, raises the pressing need to better understand the correlations between human activities and environmental change (Mayer and Smith, 2019). To address important parts of these challenges, for example, a field of research has evolved focusing on the notion of environmental justice, which comprises health and well-being aspects, as well as topics related to general social participation (Banzhaf et al, 2019). While this research has already been discussed in interplay with politics and civil rights organisations in countries like the USA for a long time, it is only in recent years that it has become visible as part of inequality studies in Germany (Preisendörfer, 2014). Combining survey data and small-scale geospatial data is an emerging topic for social scientists, which enables answering the outlined research questions from an individual-level perspective of survey respondents.

In the geospatial domain, Keenan and Jankowski (2019), showed the research trend on spatial decision science for the last 30 years, where spatial linking of diverse data sources is evident such as – spatial bussiness intelligence (e.g. Bačić and Fadlalla, 2016) or dynamic visualization for participatory decision making with a diverse team of stakeholders. Innovation is still necessary but it is challenging to integrate big geospatial data (Robinson, 2017), processing of large volunteered geoinformation (Goodchild, 2007, Haklay, 2010) and georeferenced social media information (Yan, Schultz, Zipf, 2019, Ilieva, and McPhearson, 2018). There are already some well known infrastructures available from government agencies, non-profits and even commercial entities, that give the public access to geospatial data e.g. OpenGovData, OpenStreetMap, OpenLayer, GoogleEarthEngine (Ilieva, and McPhearson, 2018, Sikder et al. 2019).

However, there are a number of reasons why the necessary integrated use of social science survey data and spatial science data is challenging: (1) the interdisciplinary nature of the two different data sources (Edwards et al, 2011), (2) legal restrictions, especially regarding data privacy (Bluemke, et al 2017; Schweers et al, 2016), and (3) methodological challenges regarding the use of geo-information systems (GIS) for the processing and analysis of spatial data, especially by social scientists (Müller, 2019). An infrastructure which facilitates such an integrated use of social science survey and spatial science data would be a spatial data infrastructure (SDI), defined as a "framework that consists of institutional arrangements, policies, and technologies that would create a conducive environment for the exchange of geographic information related resources in order to create a better information sharing community" (Tumba and Ahmad, 2014, 85). As of yet, an SDI targeting social scientists is only described conceptually (Schweers et al, 2016). In this article, we investigate and develop solutions for addressing the practical challenges and requirements of developing such an infrastructure from a social science, spatial science and computer science perspective.

Throughout the article, we use the example of environmental inequalities with regards to income and land use hazards in Germany, which motivates one research question for the infrastructure. For data analysis, we use georeferenced survey data of the GESIS Panel (GESIS, 2017) and the German Socio-economic Panel (SOEP) (Schupp et al., 2018), and as spatial data, we use data from the Monitor of Settlement and Open Space Development (IOER Monitor) (IOER, 2017). Both the GESIS Panel and the SOEP comprise extensive information from people living in Germany aged between 18 and 70 years. Amongst others, the participants answered questions on their demographic and socio-economic situations (Bosnjak et al., 2018), which makes the GESIS Panel an appropriate data source for our research question. The IOER Monitor contains detailed geographic information on land use in Germany, ranging from indicators about settlement structures up to landscape quality. The infrastructure presented in this article builds the foundation for these kinds of interdisciplinary studies which are otherwise only with large manual effort possible.

The rest of this article is structured as follows: In Section 2, we provide an introduction on the spatial linking of survey data with spatial data and its use in social science survey research and discuss challenges in Germany. We introduce our spatial data source, the IOER monitor, its data and documentation in Section 3. In Section 4, we propose a technical infrastructure for spatial linking which addresses challenges like data privacy and data security among others. As case study, we present an analysis for the environmental inequalities of land use hazards in Germany which has been conducted with the proposed infrastructure in Section 5. We discuss the lessons learnt and conclude in Section 6.

2. Spatial Linking of Survey Data with Geospatial Data

In this section we provide a brief overview on spatial linking in general, the relevance of spatial linking of survey data with spatial data in social science research and the challenges that arise currently in Germany.

2.1. What is Spatial Linking?

Spatial linking describes the technique to combine two or more geospatial datasets into one, and has been used for a long time in disciplines familiar with the use of spatial datasets (Goodchild et al, 1992).[1] This technique is also known as spatial join, a term familiar by users of products of the commercial software provider ESRI, or broader as spatial overlay when two vector datasets are combined Geospatial data are data that contain spatial references in the form of coordinates associated with the observations in the data. Each observation contains at least one coordinate resulting into different geometries such as points (single coordinate), lines and polygons (multiple coordinates), or grids (single coordinates with additional information, e.g., on their geographic resolution). Spatial linking relates the set of coordinates of one dataset to the coordinates of another, resulting in a variety of possible outcomes.

Usually, all spatial linking efforts define one focal dataset to which information of another is merged to, which results in a combined dataset with the same geographic properties as the focal dataset. If the focal dataset are point data and the other one are polygons, the resulting data are point data as well. Vice versa if

[1] Technically, we do not define georeferenced survey data as geospatial data as described here. According to data protection legislation in Germany, personal information such as addresses and corresponding coordinates must not be stored together with the actual survey information. This information is only indirectly related to each other, and linking them requires following a multi-step procedure (Schweers et al, 2016; Jünger, 2019).

the focal dataset are polygons and the other one are points, the resulting data are polygon data. For example, if points are household addresses of people and polygons are neighborhood boundaries we can assign the membership of each household to its neighborhood through spatial linking. Alternatively, the other way around, we can count the number of households within each neighborhood. Accordingly, the type of spatial linking always depends on the geographic structure of the focal dataset.

The algorithms of adding information to the focal dataset not necessarily have simple requests based on, e.g., single points in space. Through the projection into one coordinate space, all geometries can be related to each other, e.g., through distances or intersectional associations. Not only can we apply the membership of each household to a neighborhood, but we can also ask which neighborhoods are within a range of say 2000 meters. If our focal data depict streets as line geometries, we can request which neighborhoods these streets intersect. Generally, there is a multitude of more possibilities of formulating such requests on georeferenced data (e.g., Strobl, 2017) that also depend on research questions within specific scientific disciplines as shown below.

Figure 1 depicts an example of spatial linking based on circular buffer areas of sealing of soils data around one specific point. This point represents the focal dataset to which information from the sealing of soils data is added, while substantially sealing of soils is the water and airtight coverage of soils from roads and

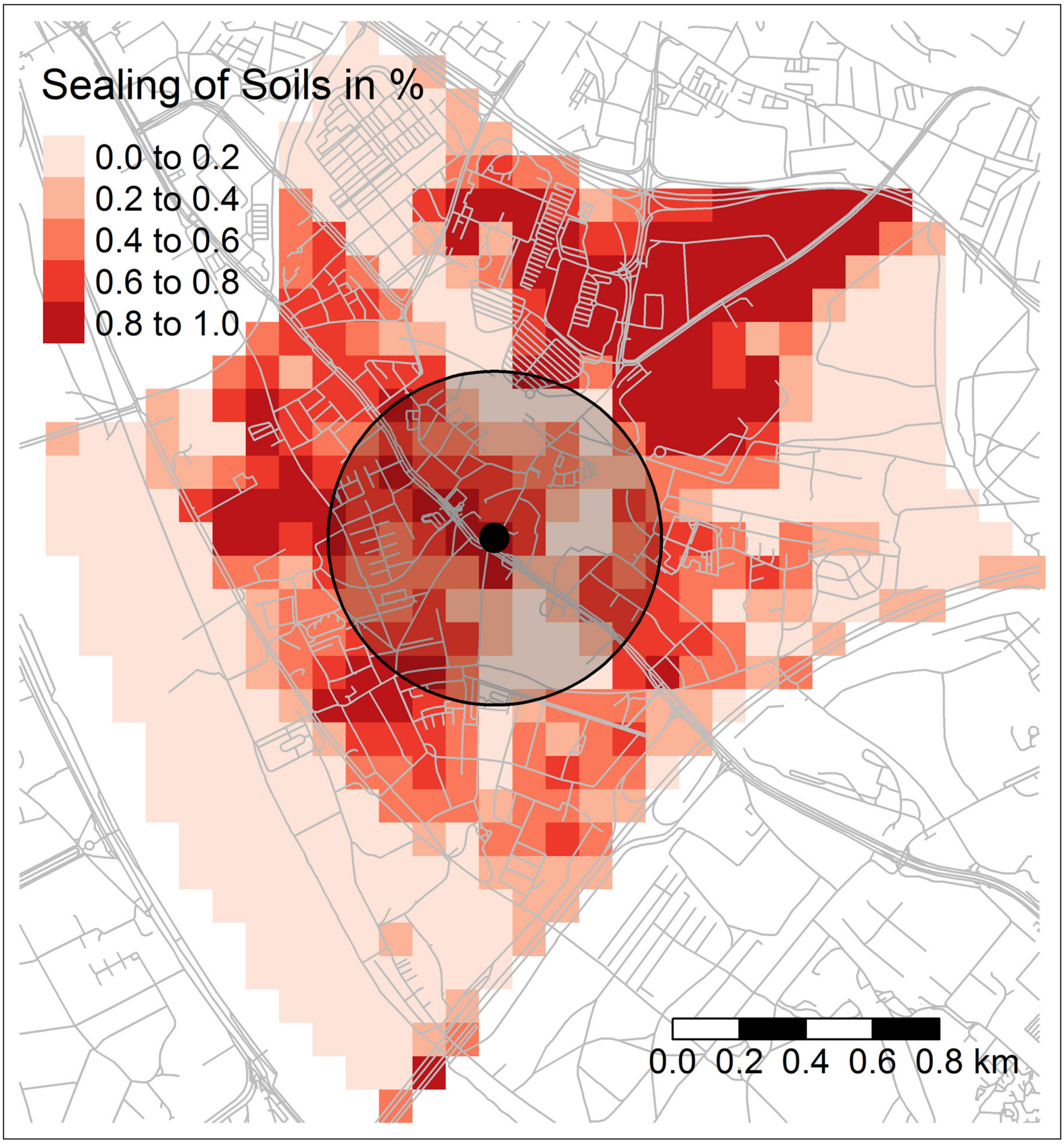

Figure 1: Circular Buffer Areas of Sealing of Soils in the Size of 500m Around a Coordinate.
Data Source: Monitor of Settlement and Open Space Development (IOER, 2017) (Sealing of Soils) and OpenStreetMap (Roads).

buildings. The idea is to gather information, e.g., on mean levels of sealing of soils in a specific area and not only at a certain point. Moreover, these buffer areas can be varied and compared regarding their impact on possible analyses with these data.

2.2. Use in Social Science Survey Research

In the social sciences, survey researchers apply such methods of spatial linking to enrich their existing data with information from auxiliary georeferenced data (Jünger, 2019). For this purpose, they require access to georeferenced survey data and to the auxiliary georeferenced data sources, which is often, however, demanding as such data lack availability or are only available with limited access rights because of data protection considerations, at least in Germany (Schweers, Kinder-Kurlanda, Müller, & Siegers, 2016). Geographic information can increase the risk of re-identifying survey respondents, such that data providers often provide on-site facilities to access the data to control the access and research output (Goebel, 2017). While this is changing, amongst others, with the work we present here, there is also already a variety of research applications that use georeferenced survey data. We witnessed an increasing number of application the subdisciplines of political attitudes research (Klinger, Müller, & Schaeffer, 2017), health (Oiamo, Baxter, Grgicak-Mannion, Xu, & Luginaah, 2015), education (Roos et al., 2013; Weßling, 2016) or social and environmental inequalities (Zwickl, Ash, & Boyce, 2014).

Rather prominently, methods of spatial linking are used in research on social trust or political attitudes. Sluiter, Tolsma, & Scheepers (2015), for example, combined survey data with detailed spatial information on ethnic diversity and investigated whether potential social trust effects vary with the geographic scale. In correspondence to the theory that ethnic diversity negatively affects bonding social capital, they found evidence for such relationships in particular small geographic neighborhoods. However, the findings are mixed and largely depend on the chosen actual geographic scale. Similarly, Förster (2018) detected different levels of political participation by varying the geographic scale of immigrant rates between one km² neighborhoods up to the governmental district level. Thus, what researchers can predict from social science theory and by having access to geospatial information also depends on the geographic scale of the data at hand.

This notion is not different concerning the application we have chosen for this article in the area of environmental justice. Already for a long time, social scientists ask for potential inequalities in exposure to environmental hazards such as air pollution (Crowder & Downey, 2010) or traffic noise (Bocquier et al., 2012). While to considerable extent evidence on this topic stems from the US societal context, extensive work for the German case is missing. We will elaborate on this in Section 2.3. Moreover, the residential segregation structures differ between those societies (Schönwälder & Söhn, 2009), such that research focusing on particular small-scale neighborhoods variations is needed. We will further investigate this issue below at the example of income, and the environmental hazards originating from sealing of soils, road traffic densities as well as industry and trade densities. The study of environmental justice yet represents another prominent application of spatial linking methods in the social sciences.

Besides the potential to formulate new research questions, adding geo information to survey data not only enables researchers to analyze new or simply more information, but it also allows adding upon existing findings. Previous research results can be corroborated rejected or gain more in-depth perspectives. For example, Downey, Crowder, & Kemp (2017) discovered that single-parent mothers in the US are at a higher risk to be exposed to toxic air pollution than two-parent families or even single-parent fathers. Apart from other disadvantages such as socio-economic disparities of single-parent families, these results add another perspective to the general discussion of family structures and inequalities. Without spatially linking survey data and small-scale spatial information this would not have been possible.

2.3. Challenges in Germany

As indicated above using georeferenced data in social science survey research is challenging, starting with the availability of data. Particularly concerning auxiliary data from official authorities and from which researchers can extract information, the situation is complicated. Because Germany is a country with a federal structure consisting of 16 individual states, also multiple statistical offices are in charge to collect data, e.g., on the sociodemographics of their inhabitants. While this data is interesting to be combined with survey data, as a result of the federal structure, researchers face a multitude of data access points. Such a fragmented data landscape often results in data that is unharmonized, erroneous or that is even missing (Schweers et al., 2016).

Likewise, on the side of georeferenced survey data, this also applies. First of all, georeferenced survey data consist mainly of survey data accompanied by the geocoded locations of the participants of the survey. For searching and accessing it, researchers use well known catalogs for survey data. However, often these catalogs lack detailed documentation on georeferences so that researchers cannot be sure if they find all georeferenced survey data available. Prominent research data centers and surveys with georeferenced survey data indeed exist such as the SOEP (Goebel, 2017) or the German Family Panel pairfam (Schmiedeberg, 2015), but finding smaller or even more specialized surveys is difficult. Social scientists who aim to use such data for their research may have to invest some time in researching for suitable data.

Another challenge that directly relates to the issue of data availability is data protection. As mentioned above, georeferenced survey data is sensitive data because it contains information on survey participants' locations which potentially can reveal their identities (Skinner, 2012). According to German data protection legislation, however, disclosing this information is strictly forbidden (Müller, 2019). Therefore, providers of georeferenced survey data usually control the access to this data, e.g., through on-site facilities with specialized and secured work environments.

Moreover, some additional and more specific issues of data protection increase the challenges of using georeferenced survey data. Also according to German data protection legislation, survey information (i.e., the answers to questions in surveys) and personal information (i.e., addresses and coordinates, respectively) must not be stored together. Accordingly, to link survey data and spatial information that was gathered through the use of spatial linking methods, researchers have to follow complicated processes (Goebel, 2017). First, they have to link coordinates with spatial information and second, coordinates are deleted, and only the spatial information is transferred to the corresponding rows of the survey data. Because this process also can get technologically challenging, we developed a solution, which we present in Section 4, to navigate these issues seamlessly and easy for end-users.

3. Spatial Data

The Monitor of Settlement and Open Space Development (IOER-Monitor) is a research data infrastructure on the topic land use/land cover hosted by the Leibniz Institute for Ecological Urban and Regional Development (IOER). This infrastructure provides comprehensive information on the status of land use changes by means of long-term time series in Germany at a high spatial resolution. The time series currently comprises 13 time periods beginning in 2000, then in 2006 and annually from 2008 onwards. In fact, the IOER- Monitor is enabling an assessment of the sustainable land use development targets on all administrative levels including urban districts. Currently, there are 95 indicators under 10 categories such as: settlement, transport, open space, sustainability, building/material stock, urban sprawl, landscape quality, ecosystem services, renewable energy and risk. Some example indicators are: the proportion of built-up areas, land consumption for transportation, settlement density, residential building stock, soil sealing, green infrastructure and ecosystem services of forest land and so on. The input data for extracting indicator values come from diverse sources such as – basic official geodata, official statistics and open data services.

The Digital Landscape Model (ATKIS-Basis DLM, scale 1: 25,000) is used as one of the important bases in the indicator calculation process, because it provides a detailed geo-topographic description of Germany (e.g. Krüger, et al, 2013, Schorcht et al., 2016). ATKIS-Basis DLM is an object-structured vector database which describes the land use features in high accuracy (redundancy-free and gapless in meshes). It also includes the linear modeled transport infrastructure (i.e. roads, railway). Further input data of the IOER-Monitor are 2D- and 3D-building models (HU-DE, LoD1-DE), the German Land Cover Model (LBM-DE), digitized topographic old maps ((TK25), other geo-data (e.g. protected area, flood areas, Imperviousness High Resolution Layer of the Copernicus Land Monitoring Service) as well as official statistical information (e.g. population, traffic, economy and finance).

Initially, the ATKIS Basis-DLM has to be transformed into a land use map where the linear modules of road and railway features are converted into polygons using a buffering algorithm with the respective width of linear objects. Afterwards, the land use portion under existing area overlays are dissolved according to a priority scheme. A hierarchically structured land use scheme was developed for carrying out the semantic description of the land use (i.e. Krüger, et al, 2013), which differentiates between 35 land use sub-classes under the main classes of settlement, transport and open space. As soon as the annually updated geospatial input data are available, the current land use map is prepared and all indicators are calculated using state-of-the-art GIS technology and applications. Detailed documentation of the methods, data sources and assumptions are available and regularly updated in the monitor metadata sheet for each indicator. An example can be found here.

An interactive, internet-based geoviewer enables the user to obtain digital cartographic visualization, statistics and analyses at a desired spatial scale such as: federal state, regional planning jurisdiction, district, municipality, urban district and INSPIRE-compliant raster grids (raster widths consist of 100 m, 200 m, 500 m, 1000 m, 5000 m and 10000 m). The selected indicator values are displayed as interactive map, table and/or diagram. Comparative functionalities are enabled on selected spatial scales. Domain experts can access the display capability of the maps in the GIS environment via Web Map Service (WMS) geoservices; but can also import data directly via Web Coverage Service (WCS) or Web Feature Service (WFS) geoservices. Currently, the services of the IOER Monitor are used in development policy formulation, public administration, urban and spatial planning, the economic sector, science/education and even in private personal projects. All data and services are freely available under DE-CC-BY-2.0. The IOER-Monitor is registered in the repository r3data.org.

In the context of the presented infrastructure for geospatial linking, IOER-Monitor geoservices were added which can be used to link respondents to their environment with regard to land use by means of a WPS service using on-the-fly calculations on the geodata.

4. Technical Infrastructure for Geospatial Linking

Designing a system taking the interests of users and data providers as well as legal considerations into account is a challenging task. In this section, we first detail the most relevant preconditions and the requirements of a system which allows for spatial linking of survey data. Thereafter, we present our infrastructure design which meets these requirements.

4.1. Background

The presented technical infrastructure has been developed in the project Social Spatial Science Research Data Infrastructure (SoRa) which is funded by the Deutsche Forschungsgesellschaft (DFG) in Germany and executed by the partners GESIS, IOER, KIT and SOEP. The objective of the project is to build spatial and social science research data infrastructures and to extend them in order to enable interoperability and data integration, as well as conformity to international standards. The initial implementation is guided by the interests of our partner data providers GESIS Panel and SOEP respectively their customers which are social scientists working in Germany or with German-based survey and spatial data. In a follow-up project, we plan to increase the number of data sources from the social sciences and from the sptial sciences. However, technically the infrastructure is not restricted to these domains.

The SoRa project draws upon research on environmental justice for its example use case and to demonstrate the use of the whole infrastructure. It depicts an adequate example because this research requires access to environmental data, mostly on a rather small-scale and for a large proportion of people.

The **spatial data** provided through the spatial data infrastructure of the IOER Monitor fulfills this demand as it contains land use indicators on a granular grid of 100m by 100m. The *IOER-Monitor* offers the user more than 85 spatial indicators in different time periods, based on vector and raster data. The underlying geodata is subject to comprehensive pre-processing. Manual random sampling and semi-automatic controls, based on the basic data used, serve to ensure the quality of the results. The aim is to provide a basis for assessing the sustainability of land use development and closely related issues (Meinel & Schumacher, 2011). **Survey data** is used from the SOEP and the GESIS Panel. Both comprise extensive information from people living in Germany aged between 18 and 70 years. Amongst others, the participants answered questions on their demographic and socio-economic situations (Bosnjak et al., 2018), which makes the GESIS Panel an appropriate data source for our research question.

4.2. Prerequisites

In this section, we will introduce the prerequisites and requirements to implement the spatial linking task for which the proposed infrastructure is designed.

First, we present the properties of the data which is held by the different stakeholders. The main stakeholders are the survey data providers and the spatial data provider, in our use case GESIS, SOEP, and IOER. The dimensions to be considered in the spatial linking task are the temporal and the spatial dimension, i.e. the date when the survey was conducted and the locations the participant is associated with at that date. Analogously, temporal and spatial dimensions are relevant for the spatial data, i.e. the date when the indicator was measured and for which location. Furthermore, the spatial data may be provided on different levels of granularities, i.e. spatial resolution. Considering these dimensions the data integration may be performed along the spatial and the temporal dimension following two main approaches:

1. Static integration: For each survey participant, all available spatial indicators are determined once and stored along the survey data.
2. Dynamic integration: Given a request, the spatial indicators are computed on demand for a selection of participants.

Both approaches yield advantages and disadvantages. The main advantage of a static integration approach is its performance and data availability. Once the data integration process is complete, the data can be provided to the researchers and the computation needs to be performed just once. However, its main disadvantage is the static nature of the resulting dataset. Ahead of the integration, it needs to be determined which indicators should be integrated on which level of granularity and for which timeframes.

Dynamic integration refers to the integration upon a request. As a result, the data is not precomputed which alleviates the shortcomings of the static approach. However, the dynamic approach requires the availability of all data sources whenever the data is demanded. In contrast to the static approach, it does not pose any restrictions on the degrees of freedom when it comes to the integration as it allows for any combination of temporal dimension and spatial resolution. Consequently, it provides more possibilities for exploring the data. According to the argumentation outlined above, the proposed infrastructure follows a dynamic approach.

Both types of stakeholders, survey data providers and spatial data providers are concerned with data privacy issues. Often there exists a trade-off between providing data in an easy accessible way and ensuring responsible use and protection of sensitive information. Conditions on handling and processing the data are typically specified in the data providers' terms of use. However, it is very cumbersome to compile a common terms of use document that satisfies the requirements of all connected data providers. Users have to sign the terms of use of the data providers individually and comply to their conditions. Our approach to this is to create an environment that applies a high level of privacy that suffices all participants while reducing the cost for creating and maintaining such an environment.

According to the characteristics of the data being processed within the infrastructure, it is necessary to consider any data privacy issues which may arise from combining data of these sources. As the goal is a spatial linking of the data, it is necessary to associate the participants of the survey with coordinates. For instance, these coordinates may represent the location of the participants apartment or work place. A central aspect of anonymizing survey data is avoiding data consumers to uniquely identify the participants. However, adding the spatial dimension, the chance of uniquely identifiable participants increases. Since it is possible to identify persons individually by this information, coordinates are considered to be of sensitive nature (see e.g. article 4 of European General Data Protection Regulation (European Parliament, Council of the European Union. (2016)). The infrastructure needs to avoid creating or exposing such sensitive data or, if absolutely needed, handle it in appropriate ways. In particular, it must not leak any such information through user querying, which has implications on the options we have to exploit both survey data and spatial data.

- Users cannot be allowed to select or filter participant records based on geographic locations, because they might be able to single out individuals living in areas with unique properties.
- Users cannot be allowed to extend records with data that can be connected to a specific location, for the same reason.

For security purposes data providers handling such sensitive data typically store data and coordinates on different servers. Thus access to multiple systems to resolve the mapping from a participant to a coordinate is required. As a result, the user of the infrastructure can be served without granting access to the coordinates directly, but only to spatial information associated with the coordinates of the participant. For instance, when assuming that a researcher is looking for data about population density in the area of living of a given set of participants, the procedure usually requires to get the coordinates from the participants and then look up the population density indicator values for each of it. However, by the data provider‘s terms and conditions users must not be able to associate participant's data with their coordinates. Thus, the coordinates are not handed to the user client (SoRa-App) to allow this lookup. Instead, the provider arranges for the indicator lookup by itself and then returns solely the indicator data. Furthermore, not only the access to the data needs to be enforced but also any data transfer that might occur between the components of the infrastructure need to be secured. Therefore, any communication via the Internet is performed by default using the Transport Layer Security (TLS) protocol, which ensures secure data transmission.

4.3. Architecture

The main components of the architecture and the interactions between them are shown in **Figure 2**. The main components comprise

1. the SoRa-App component, where the survey data is selected and a spatial linking request is sent;
2. the Survey Data Discovery component, enabling discovery and search of relevant survey data based on its metadata;

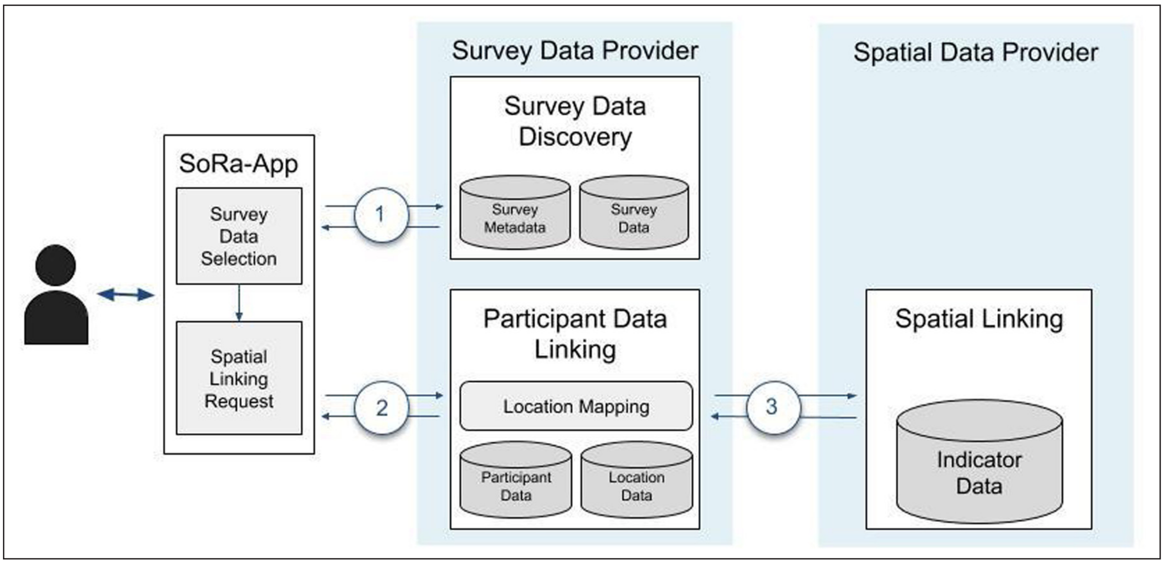

Figure 2: Architecture of the SoRa-Infrastructure.

3. the Participant Data Linking component, where the participants' coordinates can be resolved;
4. the Spatial Linking component, where the actual spatial linking is conducted.

In the first step (1) the user can explore the available survey data at the Survey Data Discovery component. Thereafter, the spatial linking request is defined and issued at the Participant Data Linking Component (2). The Location Mapping process retrieves the participants' location data and requests the corresponding indicator values at the Spatial Linking component (3). The spatial linking request is completed once the Location Mapping process returns the linked data.

Survey Data Selection

The first step for spatial linking of survey data is the survey data selection. The SoRa-App enables the user to search for and discover the survey relevant to the user's research question. The foundation of the survey data discovery is the metadata about the survey. To facilitate the deployment of this component at various survey data providers, the survey metadata is described using a well-defined vocabulary, namely the DDI-RDF Discovery Vocabulary (DISCO) (Bosch et al., 2015). DISCO is a subset of the Data Documentation Initiative (DDI Alliance, 2018) standard (DDI) which is used to document metadata about surveys and other observational methods in the social, behavioral, economic, and health sciences. The subset allows for discovering this type of data as well as its metadata (Bosch et al., 2012). At the same time DISCO is based on the Resource Description Framework (RDF) (Schreiber & Yves, 2014), a World Wide Web Consortium (W3C) standard for data interchange on the web. It is a schema-less graph-based data model which allows for merging data from various data sources. This allows for providing a common interface to discover relevant data regardless of the data provider. DISCO allows for describing the most fundamental concepts such as studies, variables, questions and representations. As an example, the data for a variable to assess the gender of the participants is shown in **Figure 3**. The graph shows the RDF representation of the variable "gender". The variable is associated with a question to measure the variable and representation for the variable. Several answers are associated with the corresponding representation.

Each provider may extend the core survey metadata with his own vocabulary. For instance, additional terms may be used to represent relationships between variables, such as the alteration of a variable over time. As a result, each survey data may extend the survey data discovery functionality. For the concrete use case addressed in this article, survey metadata from the GESIS Panel and SOEP have been modelled using DISCO for their use in the developed infrastructure.

Indicator Selection

The SoRa-App is also responsible for indicator selection. Metadata describing the indicators is provided by the Spatial Linking component (in RDF and Java Script Object Notation (JSON) (ECMA, 2017)), such that the SoRa App can display this information to support the selection of the best-suited indicator. The metadata

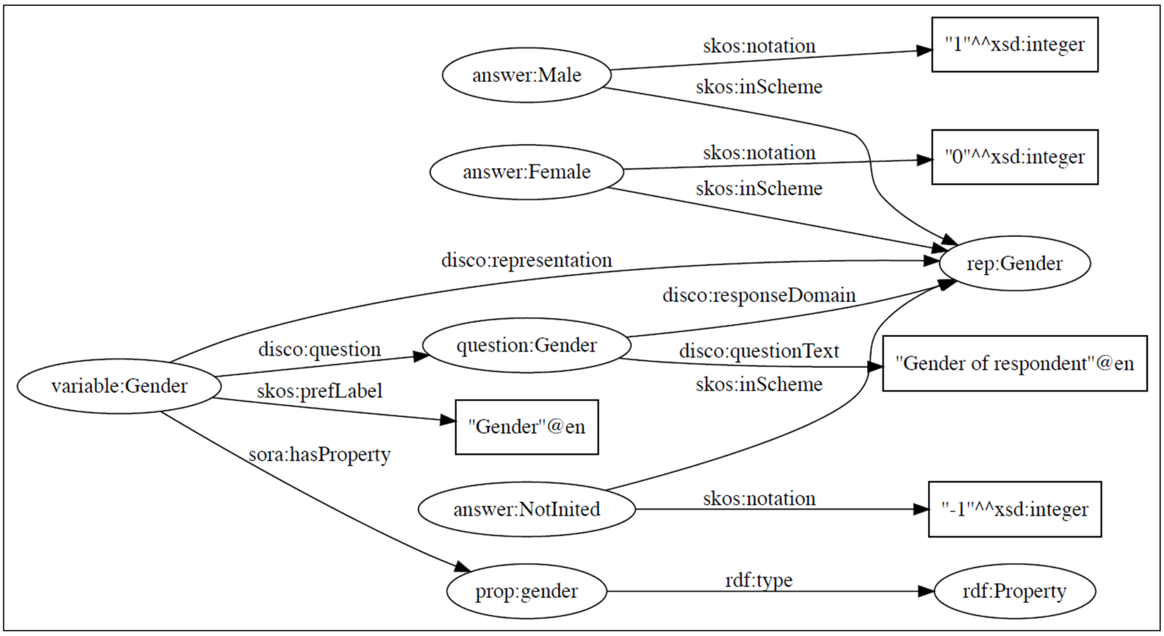

Figure 3: An Example on the Metadata of a Variable.

includes information on the methodology and interpretation of the indicators. In order to improve findability in the user interface, the indicators are sorted into ten groups ranging from traffic-related indicators to environmental indicators.

Spatial Linking Request

After selecting the survey data relevant to the user's question, the SoRa-App allows the user to specify how the selected data should be spatially linked. Therefore, users can manage their spatial linking *requests* in the app. This includes the creation and execution of new requests, viewing statuses of requests and eventually downloading the results. A spatial linking request includes the participant identifiers from the selected survey data, a location mapping function and an indicator parameterization. Next, we define the components of such a request.

Definition (Indicator Parameterization)
An indicator parameterization is a 3-tuple $\rho = (i, y, b)$ with

- i a unique identifier
- $y \in \mathbb{N}$ a year
- $b \in \mathbb{N}$ a buffer value specified in meters

The corresponding indicator function ι maps a coordinate and a given indicator parameterization to the indicator value.

Definition (Indicator Function)
Given an indicator parameterization ρ, the indicator function ι maps a coordinate $c = (x,y) \in R^2$ to an indicator value with $\iota: \rho \times R^2 \rightarrow R$.

The participants' request, provides the core functionality of the app as it allows to retrieve the indicator values for a set of participants.

Definition (Spatial Linking Request)
A spatial linking request is a 3-tuple $R = (P, \alpha, \rho)$ with

- $P = \{p_1, \ldots p_n\}$ a set of participant identifiers
- $\alpha: P \rightarrow R^2$ a mapping from each participant to a corresponding coordinate
- ρ an indicator parameterization

A spatial linking request is issued at the Participant Data Linking component of the survey data provider (c.f. (2) in **Figure 2**). In the request the coordinates are explicitly stated, as the Survey Data Discovery component merely provides the participant identifiers. The actual coordinates are (typically) not stored with the survey data, i.e. the participant identifiers and may not be revealed to the user. The *Location Mapping* process as part of the Participant Data Linking component executes function α for a given participant identifier and thus, retrieves the coordinates for a participant. Since there are potentially several locations associated with a participant, such as the household location and the work location, the function α specifies which type of location should be retrieved. The Participant Data Linking component sends the coordinates with the indicator parameterization to the Spatial Linking component (c.f. (3) in **Figure 2**) which returns the indicator values according to the parameters. Thereafter, the Location Mapping step is *reversed* and the coordinates are replaced by the participant identifiers of the request and merged with the indicator values. The outcome is a *Spatial Linking Result* R* which combines the request and the resulting spatially linked participant data.

Definition (Spatial Linking Result)
A spatial linking result is a tuple R* = (R, P*) with P*

- R a spatial linking request, and
- $P^* = \{p_1^*, \ldots, p_n^*\}$ with $p_i^* = (p_i, v_i)$ the spatially linked participant data where p_i the participant identifier and v_i the indicator value according to the indicator parameterization ρ of R for participant p_i at the location specified by α in R.

The Spatial Linking Result is returned to the SoRa-App and is joined with the previously selected survey data. The resulting data set may be retrieved by the user for further analysis.

Spatial Linking Component

The Spatial Linking Component allows for retrieving the high-resolution raster data of the *IOER-Monitor* indicators (c.f. Section 3). Internally, a *Web Processing Service (WPS) (*Mueller, M., & Pross, B., 2014*)* is used to access the indicator values for a given coordinate. Such WPSs are generally applied to process geodata or to refine input data based on predefined standards of the *Open Geospatial Consortium (OGC)* (Müller & Portele, 2005). In our case, the WPS enables linking of coordinates to the corresponding indicator value.

Interacting with a WPS can be relatively complex. Thus, in order to facilitate the access to the indicators, the spatial linking service of the WPS is also made available via a RESTful-Application Programming Interface (API) (Fielding, 2000). The API receives an indicator parameterization and a set of coordinates as an *Indicator Request* and returns indicator values for the coordinates according to the parameterization.

Definition (Indicator Request)
An indicator request is a tuple W = (C, ρ) with

- $C = \{c_1, \ldots, c_n\}$ a set of coordinates with $c_i = (x_i, y_i, e_i)$ where x_i and y_i are longitude and latitude and e_i the coordinate reference system
- ρ an indicator parameterization

The indicator values are available as raster data. Depending on the buffer value specified in the indicator parameters, the raster size is chosen and for each coordinate the corresponding raster value is determined. Since all indicator raster maps are available in the *EPSG-3035* coordinate system, deviating input values must be transformed into this base coordinate system. In case a buffer value $b > 0$ is specified, for each coordinate c all pixels within the circular area with diameter b around c are selected. From these pixels the indicator value is average to retrieve the *buffered* indicator value. The result of the spatial linking process is an Indicator Result.

Definition (Indicator Result)
An indicator result is a tuple W* = (C*,ρ) with $C^* = \{c_1^*, \ldots, c_n^*\}$ a set of linked coordinates with $c_i^* = (c_i, v_i)$ where c_i a coordinate and v_i the indicator value according to the indicator parameterization ρ.

5. Case Study: Income and Environmental Hazards

To answer the exemplary research question about environmental inequalities with regards to income and land use hazards in Germany (see 2.2), we draw on the three aforementioned sources of data (see Section 3 for details).

We hypothesize that with higher incomes people experience less exposure to land use. This assumption is based on longstanding research, that found that economic resources, such as income, provide vehicles to afford to live in neighborhoods with higher quality (Li et al., 2020; McAvay, 2019). In the analyses, we use three indicators which use for exploring environmental inequalities we justify below: sealing of soils which we already have seen above, traffic density, and industry & trade density. We will present them below in more detail.

All data sources were spatially linked using circular buffer methods by using the approach described in Section 4.3, whereas the georeferenced survey data embodies the focal data. Accordingly, the analysis data exist in the same data format as ordinary survey data but are enriched with additional attributes of sealing of soils. Moreover, the linking was conducted for each survey dataset separately to be able to evaluate potential differences between the two surveys and their impact on the results.

For the purpose of this article, the results are limited to exemplarily show some first steps of analyzing the data. We are aware and note below that further investigations may provide more in-depth analysis of the mechanisms of environmental inequalities, as they are common in the literature (Glatter-Götz et al, 2019; Best and Rüttenauer, 2018; Downey and Kemp, 2017). However, as far as the German case is concerned, there are not yet many studies that combine individual-level information from survey data with environmental information from geospatial data to investigate environmental inequalities. Nevertheless, we should be cautious of overinterpreting the results of the estimated relationship between income and land use below. Green spaces, for example, can also be part of a debate of gentrification (Pearsall and Eller, 2020), so that further studies should also take factors of urbanization and segregation into account (Rüttenauer, 2019). The focus of this article, therefore, is to describe the infrastructure facilitating access to such new data to encourage further studies.

5.1. Measures

The following measures have been used in this case study.

5.1.1. Survey Measures

Income The GESIS Panel contains a measure for household income as categorized values with collapsed income ranges (**Table 1**). In contrast to open questions of income, this measure comprises fewer missing values and its distribution is normal, which is convenient for the estimated models. The same categorization was used for the data of the SOEP in order to create a harmonized measure of income between both survey datasets. We refer to this measure of household income shortly as income below.

Sociodemographic and Contextual Controls Apart from the central survey measures we also control for the gender of the participants (0 = male; 1 = female), their age in years (18–70), their highest achieved education (low, middle, and high). Moreover, we also control for the number of inhabitants in the one km^2 neighborhoods of each participant as we suspect that the higher the population density is the higher is the amount of environmental hazards.

5.1.2. Geospatial Data Measures

We use three different indicators: sealing of soils, traffic density, and industry and trade density. For all indicators, we extracted the grid cell value of the data which contain 100 m by 100 m raster data as well as 500 m, 1000 m, and 2000 m buffers. In the following, we describe the content of the three indicators before we proceed to the analysis of the combined data.

Sealing of Soils is the air and water-tight coverage of land through buildings and roads. A consequence of too high amounts of sealing of soils is that water cannot seep away and no exchange of air between soils and the environment takes place. Moreover, sealing of soils can be considered as an inverse measure of green areas such as parks, forests or other recreational areas. As such green areas affect people's lives positively,

Table 1: Income Ranges.

Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Income range	0–299	300–499	500–699	700–899	900–1099	1100–1299	1300–1499	1500–1699	1700–1999
Class (cont.)	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
Income range (cont.)	2000–2299	2300–2599	2600–3199	3200–3999	4000–4999	5000–5999	6000–9999	≥10000	

e.g., through their impact on stress processing (Thompson et al., 2012) or health (World Health Organization, 2016), a lack of green areas is an environmental hazard and subject of an ongoing debate in the environmental justice literature (Kabisch & Haase, 2014; Wolch, Byrne, & Newell, 2014).

Traffic Density Road traffic and its consequences affect health as well. This effect on health holds true for noise originating from road traffic (Babisch, 2014; Sørensen et al., 2012; Sygna, Aasvang, Aamodt, Oftedal, & Krog, 2014), and for air pollution that vehicles emit (Pedersen, 2015). Therefore, we use an indicator that measures the density of roads because it combines both of these risk factors for health.

Industry & Trade Density We hypothesize that exposure to industry and trade facilities poses a twofold risk. First, noise and air emissions of the facilities and vehicles frequenting the facilities have a direct effect on people living near these facilities (Marques & Lima, 2011). Second, large areas of industry and trade facilities correlate negatively with free spaces such as green areas (Wolch, Byrne, & Newell, 2014). For this reason, we use an indicator of the IOER monitor that comprises information on areas used, amongst others, for retail and service, docks, waterworks, refinery, waste removal facilities, or waste disposal sites.

By comparing the descriptive statistics of all variables in **Table 2**, we observe some slight differences between the GESIS Panel and the SOEP after each dataset was linked to geospatial data measures. While it is interesting to see if these differences affect the estimations below, it is also crucial to control for them in the analyses.

5.2. Analysis Strategy

Our analysis is based on a model that predicts land use hazards based on the survey participants' income. Thus, all other variables are held constant, whereas the income of the participants varies. These results are

Table 2: Descriptive Statistics of all Variables Used in the Analysis.

	GESIS Panel				**SOEP**			
	Mean	**SD**	**Minimum**	**Maximum**	**Mean**	**SD**	**Minimum**	**Maximum**
Survey Measures								
Age	46.69	13.37	18.00	73.00	48.96	16.74	18.00	100.00
Gender	0.52	0.50	0.00	1.00	0.54	0.50	0.00	1.00
Education	2.21	0.77	1.00	3.00	1.90	0.76	1.00	3.00
Income	11.71	2.83	1.00	17.00	11.16	3.15	1.00	17.00
Geospatial Measures								
Sealing of Soils 100 m	0.47	0.24	*	*	0.51	0.24	*	*
Sealing of Soils 500 m	0.32	0.20	*	*	0.36	0.20	*	*
Sealing of Soils 1000 m	0.25	0.18	*	*	0.29	0.19	*	*
Sealing of Soils 2000 m	0.19	0.16	*	*	0.22	0.17	*	*
Traffic Density 100 m	15.67	7.53	*	*	15.86	7.33	*	*
Traffic Density 500 m	10.20	3.60	*	*	10.77	3.61	*	*
Traffic Density 1000 m	08.09	3.21	*	*	8.69	3.25	*	*
Traffic Density 2000 m	6.45	2.77	*	*	6.93	2.86	*	*
Industry and Trade Density 100 m	02.04	9.59	*	*	1.86	9.11	*	*
Industry and Trade Density 500 m	4.84	7.73	*	*	5.10	7.47	*	*
Industry and Trade Density 1000 m	5.33	6.22	*	*	06.05	6.72	*	*
Industry and Trade Density 2000 m	5.16	4.93	*	*	06.09	5.51	*	*
Number of Observations	2598				24136			

Data Source: Georeferenced GESIS Panel (GESIS, 2017) and Georeferenced Socio-economic Panel (Schupp et al., 2018).
* Minimum and maximum values deleted due to data protection.

presented graphically as predicted values. In order to prevent predicting values of sealing of soils over 100% or below 0% we use a logit transformation:

$$y_{logit} = \ln\frac{y}{1-y}$$

This transformation bounds the predicted values of the dependent variables to a range that lies between 0% and 100% after re-transforming it back to the original scale with the following formula:

$$\hat{y} = \frac{e^{\hat{y}_{logit}}}{1+e^{\hat{y}_{logit}}}$$

Finally, we estimate all models with cluster-robust standard errors. Using them in a survey sampling setting is generally recommended (Abadie, Athey, Imbens, & Wooldridge, 2017), but we also argue that different residential policies between municipalities may produce additional clustering among participants who live in the same municipality. Accordingly, the cluster-robust standard errors are based on the participants' municipality.

5.3. Results

Can we observe land use hazard exposure of survey participants as a function of their income? We use prediction plots to answer this question based on our data. These plots yield the predicted value of each land use hazard indicator on the y-axis in relation to the income of participants on the x-axis. Moreover, within each of these plots different slopes and their corresponding 95% confidence intervals depict the predicted values for different geographic sizes of the indicators (100 m, 500 m, 1000 m, and 2000 m).

Figure 4 shows the results for the estimates based on the GESIS Panel. With regards to sealing of soils in particular, we can observe a distinct decreasing slope of sealing of soils if participants' income is increasing. In comparison to the lowest-income group, the predicted value for sealing of soils of the highest-income group ~14% smaller. However, the overall exposure and the steepness of the slopes get smaller with increasing sizes of geographic neighborhoods. While this pattern remains significant with regards to the other two indicators, the effect of income is only small.

These results are corroborated with the data of the SOEP displayed in **Figure 5**. Moreover, as the sample size of the SOEP is larger than the GESIS Panel the estimates are more precise and, in the case of sealing of soils, more pronounced. In comparison to the lowest-income group, the predicted value for sealing of soils of the highest-income group is ~19% smaller, which is a difference of 5% to the GESIS Panel. The predictions for the other two land use indicators are similar to those of the GESIS Panel analysis.

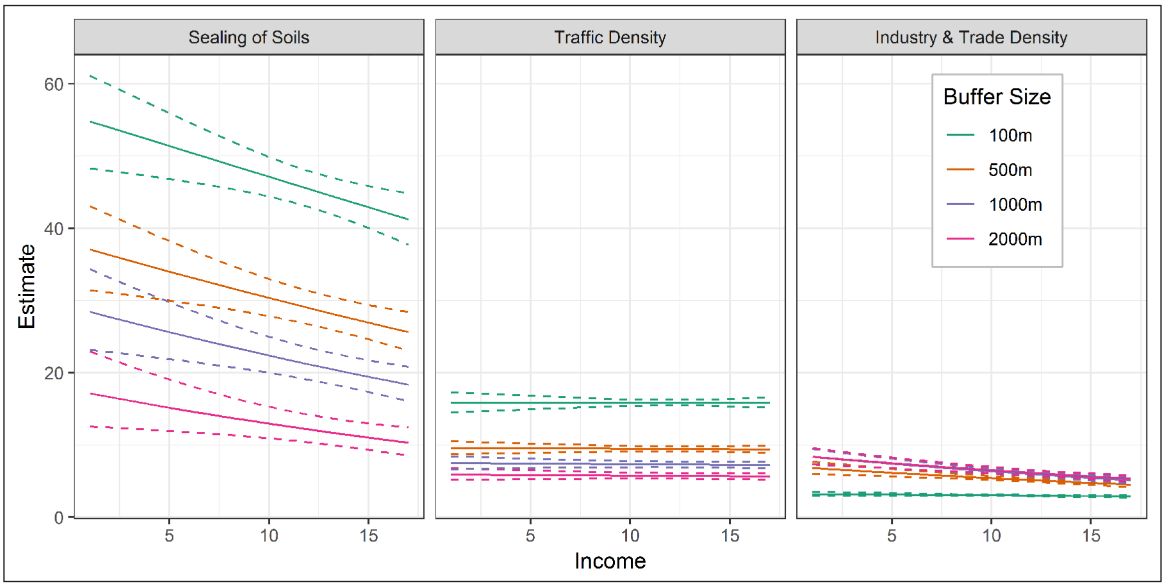

Figure 4: Results of the Linear Predictions as a Function of Income and the Three Land Use Indicators Across Different Geographic Scales for the GESIS Panel (N = 2598).
Data Source: Geoferenced GESIS Panel and Monitor of Settlement and Open Space Development.

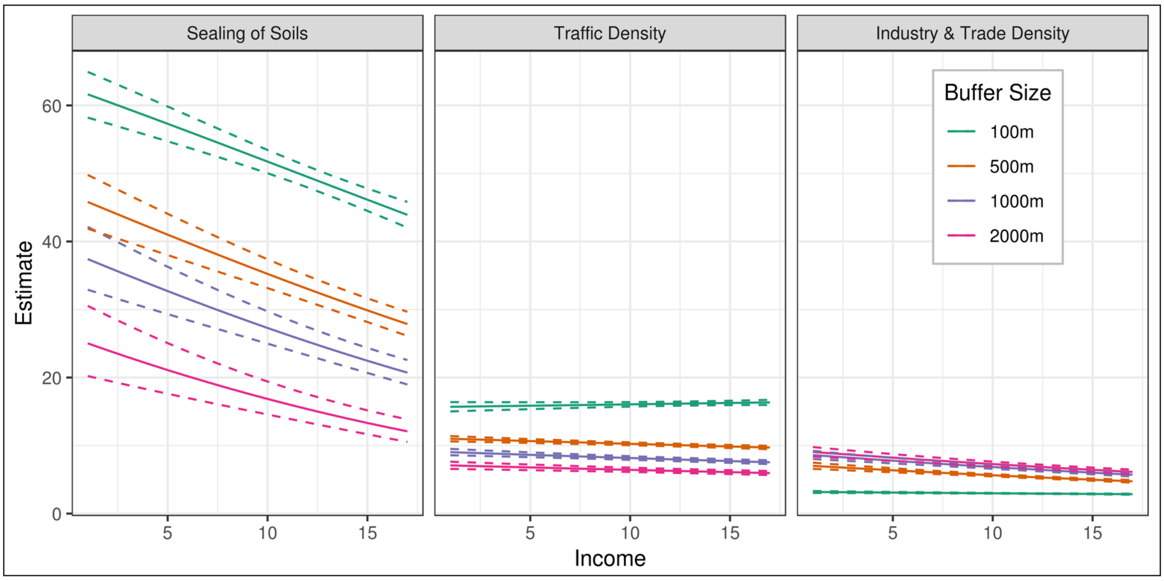

Figure 5: Results of the Linear Predictions as a Function of Income and the Three Land Use Indicators Across Different Geographic Scales for the SOEP (N =24136).

Data Source: Georefererenced Socio-economic Panel (doi:10.5684/soep.v33.1) and Monitor of Settlement and Open Space Development.

Overall, the analysis demonstrates the use of combining the data across different research data infrastructures. Using similar but slightly different population samples replicates similar findings: generally, income reduces the exposure to land use hazards, mainly with regards to sealing of soils. Such findings can and should be extended, e.g., by taking into account other sociodemographic characteristics such as ethnic minority status (Zwickl et al., 2014). Environmental justice research is a longstanding research field, considering factors of segregation (Miller, 2019), ethnicity (Crowder and Downey, 2010), health (Dreger et a, 2019), and political participation (Banzhaf et al, 2019). The results here are a starting point corroborating existing findings from the aggregate-level on the individual-level while controlling individual-level attributes. It would be interesting to see how future studies adapt the presented infrastructure and add up to such research questions.

6. Conclusion

For several investigations in social science research, there is the necessity for spatial linking in order to enrich survey data with information from auxiliary georeferenced data. However, challenges exist regarding data access, data privacy and data security among the spatial linking itself. Hence, there is a need to support researchers from the social sciences and related domains that work with survey data in accessing and linking these data sources. In this article, we presented a spatial social science research infrastructure which addresses these challenges and supports users during the spatial linking process. We used this infrastructure for investigating environmental inequalities of land use hazards in Germany and could show that increasing income of survey respondents is associated with less exposure to environmental hazards originating from land use in Germany.

However, in order to increase the benefit for users, we plan to support them in their statistics tools like R where they actually analyse the data. Also, the involvement of multiple stakeholders increases the organizational effort, in particular when aspects of data privacy and data security have to be addressed among multiple stakeholders, i.e sensitive data like georeferenced survey data must not be accessible outside of its data provider. In the infrastructure setup at hand, it has to be ensured that processing involving this data is executed on-site. The SoRa-App is currently under development as a prototype in the SoRa project. We plan to make the prototype publicly available in 2020.

Software Information

R version 3.5.2 (2018-12-20) – "Eggshell Igloo"
With packages

- “magrittr” 1.5: general piping of procedures
- “tidyverse” 1.2.1: data wrangling, e.g., procedures from “dplyr” (0.8.0.1)
- “sf” 0.7-2: geospatial data manipulations, e.g., buffers and spatial joins
- “raster” 2.8-19: extraction of raster values
- “velox: 0.2.0: fast loading of huge raster files
- “estimatr” 0.1.4: estimation of cluster robust standard errors regression models
- “car” 3.0-2: inverse logit function
- “boot” 1.3-20: re-transforming inverse logit scales
- “ggplot2” 3.1.0: visualization of results
- “tmap” 2.2: Creating the buffer map

QGIS Desktop 3.0.3
https://qgis.org/en/site/forusers/download.html.

Data Accessibility Statement

- GESIS. (2017). GESIS Panel – Extended Edition. GESIS Data Archive, Cologne, ZA5664 Datenfile Version 19.0.0. https://doi.org/10.4232/1.12742 Available on-site
- IOER. (2017). Monitor of Settlement and Open Space Development. Leibniz Institute of Ecological Urban and Regional Development. Available for Download https://monitor.ioer.de
- SOEP V33.1: doi:10.5684/soep.v33.1, Available as SUF from FDZ SOEP
- SOEP Hauskoordinaten, Available on-site at FDZ SOEP

Funding Information

The work presented in this article has been funded by the German Science Foundation (DFG) under the project number 316669855.

Competing Interests

The authors have no competing interests to declare.

References

Abadie, A, Athey, S, Imbens, GW and **Wooldridge, J.** 2017. *When should you adjust standard errors for clustering? (No. w24003)*. National Bureau of Economic Research. DOI: https://doi.org/10.3386/w24003

Babisch, W. 2014. Updated exposure-response relationship between road traffic noise and coronary heart diseases: a meta-analysis. *Noise and Health*, 16(68): 1. DOI: https://doi.org/10.4103/1463-1741.127847

Bačić, D and **Fadlalla, A.** 2016. Business information visualization intellectual contributions: An integrative framework of visualization capabilities and dimensions of visual intelligence. *Decision Support Systems*, 89: 77–86. DOI: https://doi.org/10.1016/j.dss.2016.06.011

Banzhaf, HS, Ma, L and **Timmins, C.** 2019. Environmental Justice: Establishing Causal Relationships. *Annual Review of Resource Economics*, 11(1): 377–398. DOI: https://doi.org/10.1146/annurev-resource-100518-094131

Best, H and **Rüttenauer, T.** 2018. How Selective Migration Shapes Environmental Inequality in Germany: Evidence from Micro-level Panel Data. *European Sociological Review*, 34(1): 52–63. DOI: https://doi.org/10.1093/esr/jcx082

Bluemke, M, Resch, B, Lechner, C, et al. 2017. Integrating Geographic Information into Survey Research: Current Applications, Challenges and Future Avenues. *Survey Research Methods*, 11(3): 307–327. DOI: https://doi.org/10.18148/srm/2017.v11i3.6733

Bocquier, A, Cortaredona, S, Boutin, C, David, A, Bigot, A, Chaix, B, Verger, P, et al. 2012. Small-area analysis of social inequalities in residential exposure to road traffic noise in Marseilles, France. *The European Journal of Public Health*, 23(4): 540–546. DOI: https://doi.org/10.1093/eurpub/cks059

Bosch, T, Cyganiak, R, Wackerow, J and **Zapilko, B.** 2012, September. Leveraging the DDI model for linked statistical data in the social, behavioural, and economic sciences. In *International Conference on Dublin Core and Metadata Applications*, 46–55.

Bosch, T, Cyganiak, R, Wackerow, J and **Zapilko, B.** 2015. DDI-RDF Discovery Vocabulary. Retrieved April 2, 2019, from http://rdf-vocabulary.ddialliance.org/discovery.html.

Bosnjak, M, Dannwolf, T, Enderle, T, Schaurer, I, Struminskaya, B, Tanner, A and **Weyandt, KW.** 2018. Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS panel. *Social Science Computer Review*, 36(1): 103–115. DOI: https://doi.org/10.1177/0894439317697949

Crowder, K and **Downey, L.** 2010. Interneighborhood migration, race, and environmental hazards: Modeling microlevel processes of environmental inequality. *American Journal of Sociology*, 115(4): 1110–1149. DOI: https://doi.org/10.1086/649576

DDI Alliance. 2018. Document, Discover and Interoperate. Retrieved April 2, 2019, from https://www.ddialliance.org/. DOI: https://doi.org/10.1111/jomf.12355

Downey, L, Crowder, K and **Kemp, RJ.** 2017. Family structure, residential mobility, and environmental inequality. *Journal of Marriage and Family*, 79(2): 535–555. DOI: https://doi.org/10.1111/jomf.12355

Dreger, S, Schüle, S, Hilz, L, et al. 2019. Social Inequalities in Environmental Noise Exposure: A Review of Evidence in the WHO European Region. *International Journal of Environmental Research and Public Health*, 16(6): 1011. DOI: https://doi.org/10.3390/ijerph16061011

ECMA. 2017. Standard ECMA-404 The JSON Data Interchange Syntax. Retrieved April 2, 2019, from https://www.ecma-international.org/publications/standards/Ecma-404.htm.

Edwards, PN, Mayernik, MS, Batcheller, AL, et al. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5): 667–690. DOI: https://doi.org/10.1177/0306312711413314

European Parliament, Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved April 2, 2019, from http://data.europa.eu/eli/reg/2016/679/oj.

Fielding, RT and **Taylor, RN.** 2000. *Architectural styles and the design of network-based software architectures* (Vol. 7). Doctoral dissertation: University of California, Irvine.

Förster, A. 2018. Ethnic heterogeneity and electoral turnout: Evidence from linking neighbourhood data with individual voter data. *Electoral Studies*, 53: 57–65. DOI: https://doi.org/10.1016/j.electstud.2018.03.002

GESIS. 2017. GESIS Panel – Extended Edition. GESIS Data Archive, Cologne, ZA5664 Datenfile Version 19.0.0. DOI: https://doi.org/10.4232/1.12742

Glatter-Götz, H, Mohai, P, Haas, W, et al. n.d. Environmental inequality in Austria: do inhabitants' socioeconomic characteristics differ depending on their proximity to industrial polluters? *Environmental Research Letters*. DOI: https://doi.org/10.1088/1748-9326/ab1611

Goebel, J. 2017. SOEP 2015 -Informationen zu den SOEP-Geocodes in SOEP v32. DIW.

Goodchild, M, Haining, R and **Wise, S.** 1992. "Integrating GIS and spatial data analysis: problems and possibilities." *International journal of geographical information systems*, 65: 407–423. DOI: https://doi.org/10.1080/02693799208901923

Goodchild, MF. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4): 211–221. DOI: https://doi.org/10.1007/s10708-007-9111-y

Haklay, M. 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design*, 37(4): 682–703. DOI: https://doi.org/10.1068/b35097

Ilieva, RT and **McPhearson, T.** 2018. Social-media data for urban sustainability. *Nature Sustainability*, 1(10): 553–565. DOI: https://doi.org/10.1038/s41893-018-0153-6

Jünger, S. 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 – Leibniz-Institut für Sozialwissenschaften. DOI: https://doi.org/10.21241/ssoar.63688

Kabisch, N and **Haase, D.** 2014. Green justice or just green? Provision of urban green spaces in Berlin, Germany. *Landscape and Urban Planning*, 122: 129–139. DOI: https://doi.org/10.1016/j.landurbplan.2013.11.016

Keenan, PB and **Jankowski, P.** 2019. Spatial Decision Support Systems: Three decades on. *Decision Support Systems*, 116: 64–76. DOI: https://doi.org/10.1016/j.dss.2018.10.010

Klinger, J, Müller, S and **Schaeffer, M.** 2017. Der Halo-Effekt in einheimisch-homogenen Nachbarschaften. *Zeitschrift für Soziologie*, 46(6): 402–419. DOI: https://doi.org/10.1515/zfsoz-2017-1022

Krüger, T, Meinel, G and **Schumacher, U.** 2013. Land-use monitoring by topographic data analysis. *Cartography and Geographic Information Science*, 40(3): 220–228. DOI: https://doi.org/10.1080/15230406.2013.809232

Li, J, Auchincloss, AH, Rodriguez, DA, et al. 2020. Determinants of Residential Preferences Related to Built and Social Environments and Concordance between Neighborhood Characteristics and Preferences. *Journal of Urban Health*, 97(1): 62–77. DOI: https://doi.org/10.1007/s11524-019-00397-7

Marques, S and **Lima, ML.** 2011. Living in grey areas: Industrial activity and psychological health. *Journal of Environmental Psychology*, 31(4): 314–322. DOI: https://doi.org/10.1016/j.jenvp.2010.12.002

Mayer, A and **Smith, EK.** 2019. Unstoppable climate change? The influence of fatalistic beliefs about climate change on behavioural change and willingness to pay cross-nationally. *Climate Policy*, 19(4): 511–523. DOI: https://doi.org/10.1080/14693062.2018.1532872

McAvay, H. 2019. Socioeconomic status and long-term exposure to disadvantaged neighbourhoods in France. *Urban Studies*. DOI: https://doi.org/10.1177/0042098019882338

Miller, S. 2019. Park Access and Equity in a Segregated, Southern U.S. City: A Case Study of Tallahassee, FL. *Environmental Justice*, 12(3): 85–91. DOI: https://doi.org/10.1089/env.2018.0026

Mueller, M and **Pross, B.** 2014. OGC® WPS 2.0 Interface Standard. Retrieved April 2, 2019, from http://docs.opengeospatial.org/is/14-065/14-065.html. DOI: https://doi.org/10.1177/0042098009104575

Müller, M and **Portele, C.** 2005. GDI-Architekturmodell. In Bernhard, L and Fritzke, J (eds.), *Geodateninfrastruktur: Grundlagen und Anwendung*, 83–92. Heidelberg: Herbert Wichmann Verlag.

Müller, S. 2019. Räumliche Verknüpfung georeferenzierter Umfragedaten mit Geodaten: Chancen, Herausforderungen und praktische Empfehlungen. In: Jensen, U, Netscher, S and Weller, K (eds.), *Forschungsdatenmanagement Sozialwissenschaftlicher Umfragedaten*, 211–229. Grundlagen Und Praktische Lösungen Für Den Umgang Mit Quantitativen Forschungsdaten. Opladen, Berlin, Toronto: Verlag Barbara Budrich. DOI: https://doi.org/10.2307/j.ctvbkk1p8.15

Oiamo, TH, Baxter, J, Grgicak-Mannion, A, Xu, X and **Luginaah, IN.** 2015. Place effects on noise annoyance: Cumulative exposures, odour annoyance and noise sensitivity as mediators of environmental context. *Atmospheric Environment*, 116: 183–193. DOI: https://doi.org/10.1016/j.atmosenv.2015.06.024

Pedersen, E. 2015. City dweller responses to multiple stressors intruding into their homes: Noise, light, odour, and vibration. *International journal of environmental research and public health*, 12(3): 3246–3263. DOI: https://doi.org/10.3390/ijerph120303246

Preisendörfer, P. 2014. Umweltgerechtigkeit. Von sozial-räumlicher Ungleichheit hin zu postulierter Ungerechtigkeit lokaler Umweltbelastungen. *Soziale Welt*, 65(1): 25–45. DOI: https://doi.org/10.5771/0038-6073-2014-1-25

Robinson, AC, Demšar, U, Moore, AB, Buckley, A, Jiang, B, Field, K, Sluter, CR, et al. 2017. Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *International Journal of Cartography*, 3(sup1): 32–60. DOI: https://doi.org/10.1080/23729333.2016.1278151

Roos, LL, Hiebert, B, Manivong, P, Edgerton, J, Walld, R, MacWilliam, L and **de Rocquigny, J.** 2013. What is most important: Social factors, health selection, and adolescent educational achievement. *Social Indicators Research*, 110(1): 385–414. DOI: https://doi.org/10.1007/s11205-011-9936-0

Rüttenauer, T. 2019. Bringing urban space back in: A multilevel analysis of environmental inequality in Germany. *Urban Studies*, 56(12): 2549–2567. DOI: https://doi.org/10.1177/0042098018795786

Schmiedeberg, C. 2015. *Regional Data in the German Family Panel (pairfam).*

Schönwälder, K and **Söhn, J.** 2009. Immigrant settlement structures in Germany: General patterns and urban levels of concentration of major groups. *Urban Studies*, 46(7): 1439–1460. DOI: https://doi.org/10.1177/0042098009104575

Schorcht, M, Krüger, T and **Meinel, G.** 2016. Measuring Land Take: Usability of National Topographic Databases as Input for Land Use Change Analysis: A Case Study from Germany. *ISPRS International Journal of Geo-Information*, 5(8): 134. DOI: https://doi.org/10.3390/ijgi5080134

Schreiber, G and **Yves, R.** 2014. RDF 1.1 Primer. Retrieved April 2, 2019, from https://www.w3.org/TR/rdf11-primer/. DOI: https://doi.org/10.1016/j.ecolecon.2014.09.013

Schupp, J, Goebel, J, Kroh, M, Schröder, C, Bartels, C, et al. 2018. Sozio-oekonomisches Panel (SOEP), Daten der Jahre 1984–2016. Version: 33.1. SOEP – Sozio-oekonomisches Panel. *Dataset*. DOI: http://doi.org/10.5684/soep.v33.1

Schweers, S, Kinder-Kurlanda, K, Müller, S and **Siegers, P.** 2016. Conceptualizing a spatial data infrastructure for the social sciences: An example from Germany. *Journal of Map & Geography Libraries*, 12(1): 100–126. DOI: https://doi.org/10.1080/15420353.2015.1100152

Sikder, S, Herold, H, Meinel, G, Lorenzen-Zabel, A and **Bill, R.** 2019. Blessings of open data and technology: e-learning examples on land use monitoring and e-mobility. *Conference Proceedings of the STS*

Conference Graz 2019. Critical Issues in Science, Technology, and Society Studies, 6–7 May 2019. Graz: TU Graz, 2019, S. 402–414. DOI: https://doi.org/10.3217/978-3-85125-668-0-22

Skinner, C. 2012. Statistical Disclosure Risk: Separating Potential and Harm. *International Statistical Review*, 80(3): 349–368. DOI: https://doi.org/10.1111/j.1751-5823.2012.00194.x

Sluiter, R, Tolsma, J and **Scheepers, P.** 2015. At which geographic scale does ethnic diversity affect intra-neighborhood social capital? *Social science research*, 54: 80–95. DOI: https://doi.org/10.1016/j.ssresearch.2015.06.015

Sørensen, M, Andersen, ZJ, Nordsborg, RB, Becker, T, Tjønneland, A, Overvad, K and **Raaschou-Nielsen, O.** 2012. Long-term exposure to road traffic noise and incident diabetes: a cohort study. *Environmental health perspectives*, 121(2): 217–222. DOI: https://doi.org/10.1289/ehp.1205503

Strobl, C. 2017. Dimensionally extended nine-intersection model (de-9im). *Encyclopedia of GIS*, 470–476. DOI: https://doi.org/10.1007/978-3-319-17885-1_298

Sygna, K, Aasvang, GM, Aamodt, G, Oftedal, B and **Krog, NH.** 2014. Road traffic noise, sleep and mental health. *Environmental research*, 131: 17–24. DOI: https://doi.org/10.1016/j.envres.2014.02.010

Thompson, CW, Roe, J, Aspinall, P, Mitchell, R, Clow, A and **Miller, D.** 2012. More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. *Landscape and urban planning*, 105(3): 221–229. DOI: https://doi.org/10.1016/j.landurbplan.2011.12.015

Tumba, AG and **Ahmad, A.** 2014. Geographic Information System and Spatial Data Infrastructure: A developing societies' perception. *Universal Journal of Geoscience*, 2(3): 85–92.

Weßling, K. 2016. *The influence of socio-spatial contexts on transitions from school to vocational and academic training in Germany* (Doctoral dissertation, Eberhard Karls Universität Tübingen).

Wolch, JR, Byrne, J and **Newell, JP.** 2014. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landscape and urban planning*, 125: 234–244. DOI: https://doi.org/10.1016/j.landurbplan.2014.01.017

World Health Organization. 2016. Urban green spaces and health: a review of evidence. Copenhagen, Denmark: World Health Organization.

Yan, Y, Schultz, M and **Zipf, A.** 2019. An exploratory analysis of usability of Flickr tags for land use/land cover attribution. *Geo-Spatial Information Science*, 22(1): 12–22. DOI: https://doi.org/10.1080/10095020.2018.1560044

Zwickl, K, Ash, M and **Boyce, JK.** 2014. Regional variation in environmental inequality: Industrial air toxics exposure in US cities. *Ecological Economics*, 107: 494–509. DOI: https://doi.org/10.1016/j.ecolecon.2014.09.013

How to cite this article: Bensmann, F, Heling, L, Jünger, S, Mucha, L, Acosta, M, Goebel, J, Meinel, G, Sikder, S, Sure-Vetter, Y and Zapilko, B. 2020. An Infrastructure for Spatial Linking of Survey Data. *Data Science Journal,* 19: 27, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2020-027

Submitted: 05 September 2019 **Accepted:** 09 June 2020 **Published:** 02 July 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS