gesis Leibniz-Institut für Sozialwissenschaften

GESIS *Papers* 2017|15

Geocoding and Spatial Linking of Survey Data

An Introduction for Social Scientists

Stefan Müller, Stefan Schweers & Pascal Siegers

GESIS Papers 2017|15

Geocoding and Spatial Linking of Survey Data

An Introduction for Social Scientists

Stefan Müller, Stefan Schweers & Pascal Siegers

GESIS – Leibniz-Institut für Sozialwissenschaften 2017

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Datenarchiv für Sozialwissenschaften
Team National Surveys
Unter Sachsenhausen 6-8
50667 Köln
Telefon: (0621) 1246 - 426
Telefax: (0621) 1246 - 100
E-Mail: stefan.mueller@gesis.org

ISSN: 2364-3773 (Print)
ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Contents

Glossary

In this section we provide definitions and descriptions of central terms used in georeferencing projects. This glossary, however, is not unique. Please also refer to the definitions of terms of the German Data Forum in their final report on Georeferencing Social Science Research Data (RatSWD 2012).

Coordinate Reference System (CRS)
Coordinate Reference Systems (CRS) consist of a definition of the coordinate origins and the earth's curvature. They are necessary to project coordinate points on the designated place on the earth's surface. Thus, they are used within a Geographic Information System (GIS) to display and to process geospatial data. Significant Coordinate Reference Systems for Europe are UTM and ETRS89.

Geospatial Basis Data
Geospatial Basis Data are Spatial Data that yield information on their spatial properties. In contrast to other Spatial Data, they do not necessarily contain information on substantial attributes, but rather on geometries, such as boundaries of administrative units or topographical data. These data often stem from federal agencies and have high relevance for social science survey research as they often match the structure the survey data were sampled..

Geocoding
Geocoding is the process of converting Indirect Spatial References into direct spatial references. Usually, this is done by calculating Geo-coordinates for postal addresses or assigning polygon boundaries to municipality names. Compared with linking data by names or other identifiers, linking geocoded data to other geocoded data based on direct spatial references is straightforward and reliable.

Geo-coordinates
Geo-coordinates consist of a pair of x- and y-values specified either specified in meters or degrees. They represent the distance of the coordinate pair to the origin's coordinate point of the Coordinate Reference System. Being direct spatial references, they can be visualized and processed within a GIS.

Geographic Information System (GIS)
A Geographic Information System is a set of software solutions for digital collections of spatial data. Within a GIS a user can process, edit, analyze and visualize spatial data. Moreover, working on spatial data within a GIS is not limited to locally stored data files, often the particular solution also supports operating on databases or application user interfaces for geospatial data. The most prominent GIS solution for commercial usages is ESRI's ArcGIS, whereas the use of Open Source Software such as QGIS or even R also yields very reliable results.

Georeferencing
Georeferencing is the localization of objects within space by assigning geo-coordinates to these objects. For example, by adding geo-coordinates to distinctive edges of the map, it is possible to georeference the scan of an analogue map. The georeferenced analogue map can then be used regularly as geospatial data for further processing, spatial analysis and visualization. Georeferencing of survey data means to add direct spatial references to the respondents of the survey.

Geospatial Data
Geospatial data are data collected by various specialist disciplines that are projectable within space. Therefore, they are defined by a direct spatial reference. The storage of these data occurs in different data formats that represent different geographical structures, for example:

Vector Data

- Point: Coordinates of a survey respondent
- Line: The road a survey respondent is living in
- Polygon: Boundaries of a municipality a survey respondent is inhabitant of

Raster data

- Measurements within, e.g., subdivision of a country's area that are, in contrast to the size of cities, equally large.

Spatial Analysis
Spatial Analysis is the process of computing statistics and measurements, based on the units' of analysis location in space. A variety of methods are used for such studies, for example, geographic distance calculations or analyzing neighborhood characteristics.

Spatial Linking
Spatial Linking is the assignment of spatial properties or attributes of one spatial object to another spatial object within a GIS. For example, it is possible to link geo-coordinates of survey respondents to administrative boundaries data. Furthermore, knowing the allocation of a respondent of a certain administrative unit enables the addition of further attributes such as the employment rate to the respondents' data.

Indirect and Indirect Spatial References
Indirect spatial references include information on spatial objects that are, not stored as geo-coordinates. This information on spatial objects may be processed names of administrative units or postal addresses within a GIS. Direct spatial references, on the other hand, store information on spatial objects as geo-coordinates. In contrast to data containing indirect spatial references, it is possible to process data containing direct spatial references within a GIS.

1 Introduction

In survey research studying the local contexts of social behavior and social attitudes is in great demand. Researchers consequently try to model the direct living environment of individuals as a relevant predictor of individual behavior or attitudes. Examples are to find in studies of political behavior (Förster, 2017), attitudes towards migration (Klinger, Müller, & Schaeffer, 2017) or the influence of environmental stressors on health (Pedersen, 2015).

Survey data used in studies of human behavior and attitudes, however, often lack a small scale spatial dimension. As a result, researchers have to rely on the attributes of higher level spatial units such as local administrative units (e.g. municipalities or districts) or even whole countries. This dependability often reduces the reliability of the indicators for social contexts and increases the risk of ecological fallacies since the variability of the attributes across individuals within these higher levels units are ignored (Crowder & Downey, 2010).

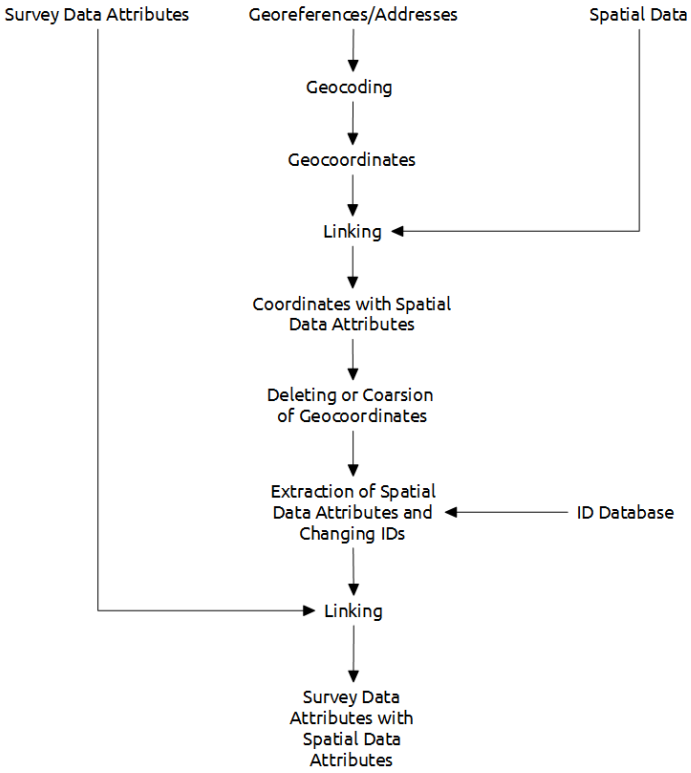
Therefore, to overcome these issues, a trend in the social sciences is to use Geographic Information System (GIS) techniques (Hillmert, Hartung, & Weßling, 2017). The general idea is to geocode available addresses of survey respondents into geo-coordinates and then spatially link them to small-scale spatial data. The resulting linked datasets can then be analyzed without the risk of ecological fallacies. Furthermore, innovative research questions can be addressed by using new data sources such as environmental noise data or by using spatial analyses such as calculating geographic distances to certain points of interest such as nuclear power plants.

This document is a background reference as well as a hands-on to geocoding social science survey data and linking them to available geospatial data attributes. To appropriately hurdle challenges of spatial linking, it contains procedures that transparently and efficiently process three distinct data sources: social science survey data, georeferences and geospatial data. The focus of this paper is to explain these procedures by using example data.

There are different challenges that have to be considered when linking social sciences survey data and geospatial data – some of them are of technical nature, most of them are due to data protection legislation in Germany. Spatial references such as addresses of respondents must not be stored along with the actual survey data attributes (i.e. the answers to the survey questions). At the same time, these spatial references are often used to link exogenous attributes to the survey data attributes. Therefore, researchers have to be careful when using these spatial references so that georeferences and survey data attributes are never stored within one single dataset.

During the project "Georeferencing of Survey Data" (GeorefUm), funded by the German Research Foundation, we developed conceptual and technical solutions for geocoding and linking survey data to spatial data in compliance with the German data protection legislation. The concept as such is simple; Figure 1 gives a schematic overview of this concept.

Figure 1: Spatial Linking of Survey Data and Geospatial Data in Compliance with German Data Protection Legislation



Following this procedure to link survey data to spatial data attributes, we work with the 4 data sources mentioned above, plus one additional database: (1) the data collected from the survey, (2) georeferences, (3) spatial data that contains attributes of the small scale contexts, and (4) an auxiliary ID database for linking the spatial data to the survey data.

(1) The survey data includes a respondent ID variable and the actual survey data attributes.

(2) Georeferences are, in most cases, respondents' addresses containing the street name, house number, zip code and name of the municipality. This information is collected for sampling purposes and is disclosive. Therefore, at least names and addresses of respondents must not be stored together with the survey attributes to prohibit de-anonymization of survey participants. Derogations from this rule require informed consent from survey participants. In addition, these data also contain an ID variable that differs from the ID variable in the survey data.

(3) Spatial data depict data in spatial data formats (see section 3). These are the data from which attributes will be assigned to the survey data later in this process.

(4) The auxiliary ID database contains the respondents' IDs of the survey data (1) as well as from the address data used for geocoding (2). It functions as a key database and is used to merge the spatial data attributes (3) to the survey data (1).

These data sources are stored separated. A best practice approach would be to store them encrypted (for example using the software VeraCrypt: https://veracrypt.codeplex.com/), and at least the key data-

base should be stored in a physically separate location. The data sources come together only on three occasions. At first, it takes place when linking spatial data attributes to the geo-coordinates (after the georeferences were geocoded). After that, when changing the ID variable of the cleaned data[1] with spatial attributes to the ID of the survey data. And lastly, when merging clean data to the survey data. A more detailed explanation of this concept can be found in Schweers, Kinder-Kurlanda, Müller, & Siegers, 2016).

This procedure is in compliance with German data protection legislation as it ensures that at no time the survey data are jointly stored with the georeferences. However, following this approach demands lots of efforts in the planning of the linking process in itself. Furthermore, although it is easy to follow this procedure manually, to document the process transparently script based applications or software solutions might be of interest to facilitate the use of geocoding and spatial linking by non-expert users.

We therefore developed routines in the statistical environment R to facilitate the process following the above concept (for a general introduction to R and spatial data see, e.g., Bivand, Pebesma, & Gómez-Rubio, 2013). These routines are distributed through an R package obtainable at https://github.com/stefmue/georefum. In chapter 3 we show the easy installation of the package within R. The routines have a particular focus on survey data in general and as for the case of spatial data on German Census 2011 data and environmental noise data. Thus, we use these spatial data in this document to demonstrate our routines.

This report is organized as following. In the next section, we give some background of geocoding indirect spatial references. This part is the most problematic step in the process of data linking because it does not only require to comply with data protection regulations but also it cannot be done properly by the user without relying on commercial providers such as Google, Yahoo, or firms specialized in geo-marketing. In the subsequent section, we provide examples of relevant spatial data that can be used to link social science survey data. We introduce German Census 2011 data (i.e. raster data), environmental noise data (i.e. polygon data) and harmonized time series data. Our focus was on data that are easily and openly obtainable. The reason behind this is we wanted to ensure that these data can also be used and processed by any user interested into linking spatial with survey data. In the appendix, we present a larger list of available spatial OpenGovData. Next, we demonstrate routines of spatial linking for usage within the 'georefum' package. The procedure involves simple spatial linking, a method for linking with prior calculated geographic coordinates and a routine for linking with focal analyses results. The last section is a summary of this paper.

[1] With cleaned data we mean data that are free from addresses, address based geo-coordinates or when these information are coarsened.

2 Geocoding of Indirect Spatial References

Conducting spatial linking and spatial analysis requires geo-coordinates. If direct spatial references are not collected during the interview (for example using GPS applications during the interview) the first step to spatially link survey data to geospatial data is the conversion of indirect spatial references into direct spatial references, a procedure called geocoding. Indirect spatial references are, e.g., addresses, names of municipalities or zip codes. They work by names or numbers as identifiers to link them to other data. In contrast, direct spatial references consist of geo-coordinates, whose processing can take place in Geographic Information Systems (GIS). While they also work by numbers as identifiers, they are explicitly related to a certain point or a certain set of points in space. Therefore, it is possible to map direct spatial references within space.

The most common types of direct spatial references are points or polygons. A point, for example, can be the location of a respondent's dwelling. A polygon can depict the boundaries of the municipality, where a respondent lives. Other kinds of spatial structures that are not of interest in our particular case of geocoding are lines and raster data. (cf. chapter 3). Finally, as geospatial data can also contain attribute fields for every geo-coordinate defined within the data; they can be used as a link to assign the attribute fields to the geo-coordinates of the survey respondents.

A large range of different geocoding services exist. These services are used to converse indirect spatial references such as addresses into direct spatial references. In general, users of these services transmit indirect references to these services and request direct references for each of them. The technical solutions, however, may vary, ranging from working with Application Programming Interfaces (API) to web interfaces that can be fed with spreadsheet data containing the indirect references.

Prominent examples of commercial geocoding services are Google Maps or Yahoo. One drawback of these services is that providers of commercial geocoding services often save the requests to their services. The storage of such sensitive data like respondents' locations, fail to comply with German data protection legislation and therefore causing usage problems. There is a lack of information from the commercial supplier's side, about the usage of the data they obtain and save through requests for their service. Commercial providers of geocoding services are therefore often ineligible to be used in research projects with sensitive information on respondents. Providers of geo-marketing services also offer services which comply with data protection legislation, but these are costly for research projects.

In contrast, the geocoding service of the Federal Agency for Cartography and Geodesy (BKG) in Germany offers a lot of advantages compared to the commercial service providers. Most important, the BKG converts addresses into geo-coordinates without storing the addresses transmitted. Moreover, the spatial data on addresses which are used by the BKG are more current and accurate than the data from commercial services. In consequence, the quality of the geocoding is higher. For this reason, the GeorefUm project used BKG geocoding service.

A major inconvenience of the BKG geocoder, however, is that it is only available for organizations that are founded by the federal government of Germany. Researchers at Universities that are founded by

the federal states have no free access to the services of BKG. GESIS aims at offering an interface between academic research and services for high-quality geocoding services of the BKG. [2]

Let us briefly elaborate the mechanics of the geocoding service of the BKG. First, a user a user uploads a Comma Separated Value (CSV) file to the interface of the service. This file contains the postal addresses that should be converted into geo-coordinates. See an example of the structure of this CSV file in Table 1. Secondly, the user chooses the relevant field information (street, house number, zip code, etc.) and the coordinate reference system in which the calculated geo-coordinates will be stored. Finally, after starting the geocoding process, the user receives an overview of the process and can save the results in a new CSV file.

Table 1: Structure of CSV File Requested for Geocoding

Respondent ID	Street	House Number	Postal Code	Town
1	Example Street	1	12345	Example City

Although the utilization of this service is safe and straightforward, a set of arrangements have to be made. The uploaded CSV file should only contain addresses and an ID variable for survey respondents, but no further information such as names and survey data attributes. Moreover, the name of the CSV should not allow drawing any conclusions to the project title of the study as an inference from the submitted addresses to the results of the corresponding service should not be possible. For example, the study name or project titles should not be mentioned in any of the files transmitted to BKG. These arrangements should minimize the risk of re-identifying survey respondents by attackers of the BKG geocoding service or the own computer infrastructure.

In addition to the geo-coordinates, the BKG geocoding service provides measures of the quality of geocoding for each address. They are displayed either directly on the map or can be later analyzed in the downloaded CSV file. The BKG itself considers a quality of geocoding measurement of at least 96 % percent as good, which also means that the addresses do not have to be corrected. As for the case of measurements below the threshold of 96 % a manual correction of erroneous addresses and their subsequent and a new geocoding can lead to an overall good set of direct spatial references.

[2] On request, GESIS functions as an interface to the geocoding service of the BKG so that users can geocode their indirect spatial references through us. In general, the geocoding service of the BKG is available to all federal institutions in Germany for usage (for detailed documentation of the service see: http://www.geodatenzentrum.de/docpdf/geokodierungsdienst.pdf. Using the BKG geocoding service presupposes signing an agreement.

3 Exemplary Acquisition and Harmonization of Spatial Data

The second step for linking survey data to spatial data is the acquisition of relevant spatial data and its preparation for spatial linking procedures. The GeorefUm projected conducted a systematic research for available administrative spatial data. Among others, data from the federal statistical agencies were reviewed regarding their utility and relevance for social science research. As exemplary use cases, data on environmental noise and aggregated data from the 2011 German Census were acquired and harmonized. This section describes the data and introduces procedures to work with these data. A list of potential data sources for spatial data from official statistics in Germany is in the appendix of this paper.

First, let us install the 'georefum' package mentioned earlier. To perform the installation, please apply the following command. Please note, that the 'devtools' package must be installed to install R packages from GitHub; furthermore, the install process may take a while as the package contains a lot of data.

```
devtools::install_github("stefmue/georefum")
```

When completed, the package should be loaded into the R environment with *library(georefum)*.

3.1 Environmental Noise Data

Data on exposure to environmental noise are collected in all member countries of the European Union (EU) based on the Environmental Noise Directive (2002/49/EC) of the EU (European Parliament 2002). In Germany, federal states or municipalities publish the results as maps in the form of simple PDFs, geospatial and geoPDFs, or interactive web services. Federal states and municipalities deliver the raw data collected for this purpose to the Federal Agency of Environment (UBA), which stores them in a data repository run by the European Environment Agency (EEA) and its European Environment Information and Observation Network (EIONET). This repository, called the Central Data Repository (CDR), collects, among other data, all environmental noise data deposited by the EEA member countries, including Germany, under a Creative Commons License (CC-BY). Figure 2 is a map showing how the actual data appear.

Figure 2: Left: A cut out from a map of the city of Cologne with buildings and streets; Right: the same cutout added with a layer of road traffic noise data

Note: Data source: OpenStreetMap and EIONET Central Data Repository (CDR)

3.2 German Census 2011 Data

The German Census 2011 data were collected within the 2011 European Union census regulation (763/2008). These data aim at monitoring basic demographic compositions of the population on a small spatial scale. In Germany, some attributes from the Census are available on 1 km² aggregated grid cell level covering the whole territory of the Federal Republic of Germany. The attributes contain information on the size of population (absolute number of inhabitants, in persons), the mean age of population (in years), the proportion of residents under the age of 18 (in %), the proportion of inhabitants over the age of 65 (in %), proportion of inhabitants with foreign citizenship (in %), the mean size of household (in persons), the vacancy rate (in %), the living space per person (in m²), and the living space per dwelling (in m²). To download these data[3] the following line can be executed:

```
georefum::download_census_1km()
```

After downloading, the data are stored as comma separated values data files (CSV). To use them in spatial analyses they have to be converted into a spatial data format. Because the data has gridded data attributes, the choice of a raster file format is appropriate. R already offers means to perform these operations. The package georefum can be used for the German Census data in particular by applying the following command:

```
georefum::census_rasterize()
```

The rasterized data can be stored as separated raster data files on hard disk, or they can be returned as an R object into the R global environment. The hard disk approach might be advantageous when the files are going to be processed within another GIS related program. There is already a ready to use

[3] There are also size of population data on smaller 100x100m grid cells as well as categorized census attributes data available. They can be downloaded by using the *georefum::download_census_100m* and *georefum::download_census_1km_cat()* functions. However, be warned: Downloading the Census 100x100m data takes a while since the volume of data is large.

version of the data delivered with the georefum package; they can be loaded into the R workspace by using the following command:

```
data(census.attr)          # 1km² grid data
data(census.attr.cat)      # 1km² grid data (categorical version)
data(census.Einw.100m) # 100x100m inhabitants grid data
```

Figure 3 illustrates the structure of these data. It shows two cut-out maps of the city of Cologne. The left hand side depicts an ordinary city map with displayed roads and buildings. To the figure on the right hand side a layer of the German Census data with the proportion of immigrants was added. As it can be seen, it is composed of differently colored but uniform and rectangular grid cells – these are the 1 km² raster cells. The color denotes the amount of immigrants in the grid cells ranging from low levels (light yellow) to higher levels (deep red).

Figure 3: Left: cutout from a map of the city of Cologne with buildings and streets; Right: Same cutout added with a layer of attributes on the amount of immigrants in 1km² grid cells

Note: Data source: OpenStreetMap and German Census 2011

The two use cases for preparing spatial data for linking with survey data show that the operations differ in complexity depending on the quality of the spatial data acquired for linking. Harmonization of data from different data sources is a burdensome task. Unfortunately, the federal structure of government in Germany leads to a fragmented production of spatial data. The use of the data would be tremendously facilitated if pooled datasets would be published at least for data that exists for the entire territory of Germany.

Many sources of spatial data, however, only cover particular administrative units that are responsible for data collection (e.g. health care facilities in particular municipalities or districts). This kind of data is more appropriate for studies on subnational level (e.g. studies of single cities of districts).

4 Harmonized Time Series

The analysis of time series is often challenged by non-comparable units of measurement. One of the reasons is because of territorial reforms the boundaries of municipalities and administrative districts change. In Germany, especially after the reunification in 1990, there were continuous reforms to reduce the number of independent municipalities in the new East German federal states, e.g., 1994 and 2007 in the federal state Saxony-Anhalt. The reforms affect the shape of boundaries but also names and IDs of municipalities and administrative districts. A lot of research questions in the social sciences, however, require comparable areas (i.e. with stable boundaries) (Ackermann & Traunmüller, 2014).

Fortunately, the problem of changing territorial units can be addressed by using georeferencing techniques to create harmonized time series. Harmonized time series are time series that are based on a reference year for administrative units to which all other year-unit combinations are comparable. For example, the information on administrative units of the year 2010 can be used to assign them to preceding years. In the following, we will explain the harmonization of municipalities by georeferencing using the example of the German General Social Survey (GGSS) with the reference year 2010 (GESIS - Leibniz-Institut für Sozialwissenschaften, 2011).[4]

As the addresses of the respondents of the GGSS have to be deleted due to data protection legislation, we based the unit for the harmonization process on municipalities. The information on municipalities in each one of the GGSS refer to the date of sampling, for example, as for the case of the 2010 GGSS to June 30, 2009. To use georeferencing techniques for harmonizing these municipalities to a particular reference year, we needed georeferenced information about the boundaries of these municipalities for each sampling year as shapefiles.[5]

One of the difficulties of the available shapefiles is that they refer to the beginning or the end of each corresponding year. In the case of the GGSS 2010 information on municipalities, however, refer to June 30, 2009. For this reason, all changes during a particular year have to be tracked using the change lists of territorial units in Germany offered by the Federal Statistical Office and the Statistical Offices of the Federal States[6]. As it will be described in the following, this approach guarantees that all municipalities are correctly assigned to even if territorial reforms occurred within a particular year.

This tracking routine is straightforward. In the case of the GGSS, a simple R script comparing the list of municipalities in the GGSS with the list of municipalities subject to modification during a certain year documents when changes occurred to municipalities in the GGSS (Step 1). For example, for all municipalities of the GGSS 2010 the script checks whether there was a territorial reform between July 1 and December 31, 2009. The routine then harmonizes the affected municipalities to the date of December 31, 2009. This step is necessary because the municipalities have to be comparable to the shapefiles with which the whole time series of the GGSS is harmonized in a later step.

---

[4] Since these procedures required specialized routines on specific data that are not generalizable, we renounced offering them as a ready to use R functions in the 'georefum' package. They can be obtained, however, by request to the authors of this paper.

[5] Shapefiles of high quality for German municipalities are provided by the BKG at http://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div?gdz_spr=deu&gdz_akt_zeile=5&gdz_anz_zeile=0&gdz_user_id=0 (Shapefiles from recent years can be found under the button "Archiv")

[6] https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/NamensGrenzAenderung/NamensGrenzAenderung.html

In the next step the harmonization of German municipalities data takes place (Step 2). For this purpose, the shapefiles of all German municipalities of each corresponding GGSS survey year are spatially linked to the shapefile of all German municipalities of the reference date December 31, 2009. The result is a reference list for each survey year municipality combination and to which municipality each pair belongs to the date of December 31, 2009. After the municipalities are harmonized on the aggregated level they can be harmonized on the individual respondents' level.

To achieve this, we need first a link between the list of municipalities created in step 1 and one to the harmonized reference list from step 2 (Step 3). As a consequence, all municipalities are then harmonized to the changes during a survey year of the GGSS and to the reference date December 31, 2009. Subsequently, this harmonized information can be linked to the survey data of the GGSS in each year by using the actual area identifiers as the key variable.

The result of all these steps is a GGSS data files containing harmonized area information. This linking does not only include original area information at the time of the interview but also harmonized area information within each interview year and to the reference date of December 31, 2009. The information now is referring to the same time point; and time series analyses are applicable. These analyses enable the tracking of changes in social attitudes over the whole time period covered by the GGSS data which was especially problematic for the East German federal states.

5 Spatial Linking and Spatial Analyses

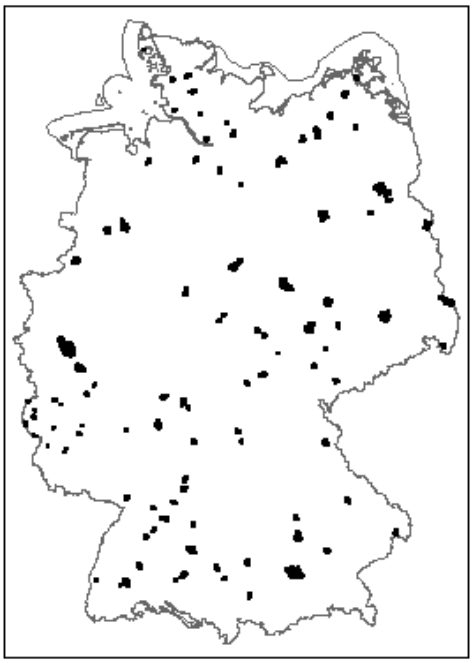
The third step for linking survey data to spatial data is the actual spatial linking. Three distinct datasets are needed to comply with the legislation of German data protection, (see also Figure 1):

(1) The survey data with an individual identifier and all other survey attributes; (2) a dataset containing the same identifier as in the survey dataset in addition to another second identifier used for geocoding of indirect spatial references; and (3) a dataset containing the second identifier and the geocoordinates. The package 'georefum' offers three commands to simulate these datasets:

```
georefum::simulate_coordinates()
georefum::simulate_surveydata()
georefum::create_key_database()
```

Because we do not want to distribute real addresses from GGSS participants, in a first step we simulate address coordinates from fictional respondents from a fictional survey. This simulation imitates the sampling process of large population surveys such as the German General Social Survey. It initially creates a random sample with N = 100 of all municipalities in Germany. Then it generates a random sample of coordinates with N = 100 within each of the drawn municipalities. We receive a random sample of coordinates of in total N = 10,000 consequently. Moreover, an ID for each fictional respondent (i.e. geo-coordinate) is added as well. In Figure 4 the distribution of these coordinates in Germany can be seen. It is noticeable that the coordinates are clustered and not equally distributed over the area of Germany. The sampling process of our fictional survey is now completed.

Figure 4: Distribution of Simulated Respondents' Coordinates Clustered in German Municipalities



Note: Data Source: Federal Agency for Cartography and Geodesy (BKG)

In the third step, we create variables with varying values for 10,000 rows; thus, we simulate our fictional survey data.[7] Furthermore, the second ID variable which was just created in the second step is added to each of the corresponding rows in the fictional survey data set. Our survey is now completed and as we already presented the spatial data in the preceding section all data that are needed to link our data is setup.

Eventually, the third step denotes the creation of the ID database. This move is a very crucial step because it ensures that we are able to link the survey data with geospatial data in the future process. It simply takes the two ID of the coordinates and the survey data - and saves them into a new dataset. We will come back to this data later.

Please note that all simulation procedures we just completed can be skipped if using own real survey data is intended. Only the geo-coordinates with a distinct ID, the survey data with second distinct ID and an ID database that contains both of these IDs are needed. It is necessary, however, is to name the variables of the data according to the documentation of the individual spatial linking procedures.

5.1 Simple Spatial Linking

There are different methods of linking survey data to spatial data attributes. They range from the simple assigning of spatial data attributes to coordinates in a particular location to more sophisticated techniques that calculate distances or create neighborhood clusters which then are joined to the coordinates. Nevertheless, the routine of linking, as described in section 1, is the same for the different methods. However, they all assume that the addresses are already geocoded; we did this in the previous subsection.

For a simple linking of spatial data attributes, the coordinates are linked to the spatial data attributes by using the coordinates as a key to merge the attributes. Both data sources can be described within the same space and a same coordinate reference system, while the coordinate data source is the focal data to which the attributes are joined. Now, if the links (i.e. the coordinates) are matching, the corresponding attribute value is assigned to the focal data. Regardless of the format of the spatial data (polygons, rectangle grid cells, etc.) the attributes which are assigned during the process of linking are sharing the same shape as the focal data – in our case the shape of simple spatial points.

Thereafter, the coordinates have to be deleted or coarsed, i.e. aggregated to higher spatial level units to prepare for the linking of the spatial attributes of the survey data. How much coarsening of the data is needed to minimize the risk of disclosure depends on the characteristics of both the attributes from the spatial data and the survey data. There are no simple rules of thumb to determine risks of deanonymization. After performing this step, the ID database is used to change the ID that remained within the data during the whole process of linking to the ID of the survey data. This is necessary to link the spatial data attributes to the survey data. Finally, the spatial data attributes are added to the survey data by applying simple merging functions which exist in all relevant statistical software packages on the market, e.g., R, Stata or SPSS.

[7] Admittedly, it might be unrealistic to achieve a response rate of 100%. For convenience reasons, however, we did not account for fake refusals of our fake survey. At the same time, in general, linking geospatial data attributes to a gross sample is an exciting endeavor as sample quality and non-response can be analyzed in the context of geospatial information.

We will now show how the described procedure can be applied with the 'georefum' R package on German Census 2011 data and environmental noise data. All single steps are already included within the corresponding functions so that only the file paths to the individual data sources have to be adjusted.

As for the case of linking census data to survey data the following command can be used:

```
georefum::census_linking_simple()
```

To use the same logic for environmental noise data, we can use the following command:

```
georefum::cdr_linking_simple()
```

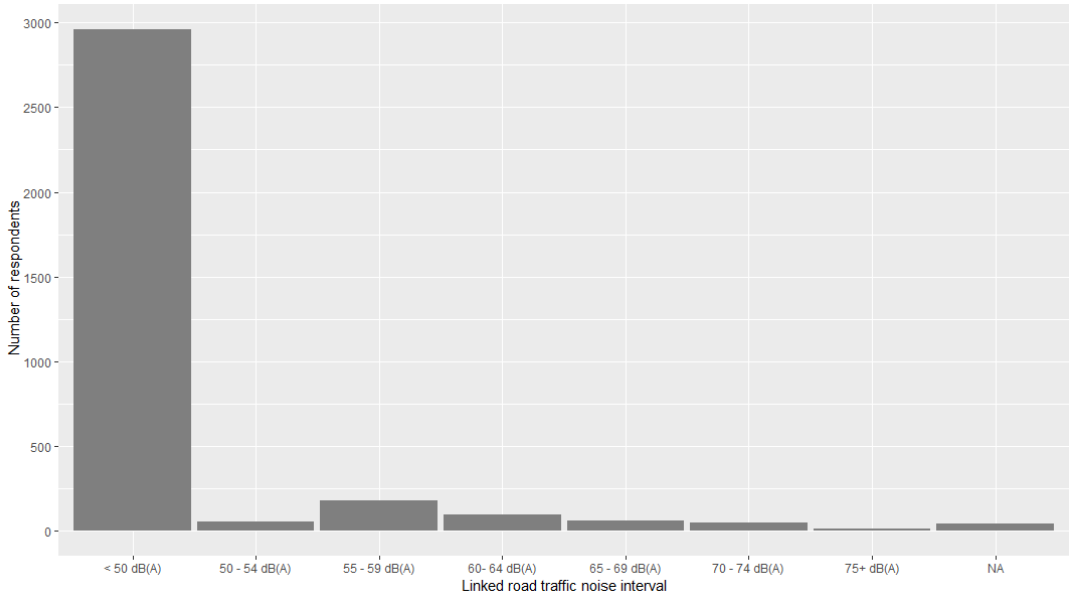
We only use a small proportion of the environmental noise data because they are large and require computational power with a lot of memory. We do not want to assume this kind of setup for any of our users. The command would run several days to be completed on the whole dataset. In addition, as outlined above, the environmental data that can be freely acquired contain erroneous files. Therefore, they also demand manual work for correction.

This is the simplest case of linking survey data. It is based on matching coordinates between respondents' addresses and the attributes of geospatial data that are projected on the same map. At the same time, it is possible to extend this approach. More complex samples of spatial linking are therefore shown in the following two subsections.

5.2 Geographic Distance Calculations

For the environmental noise data, the simple linking of the attributes to the survey data leads to a skewed distribution of the spatial attributes. This is because only a small proportion of the respondents experience severe exposure to road traffic noise at their residence. Figure 5 shows the distribution of road traffic noise exposure linked to the German General Social Survey 2014. Only a small proportion of respondents locations were assigned to a corresponding road traffic noise value. Consequently, using statistical procedures such as regressions lack statistical power.

Figure 5: Distribution of Road Traffic Noise Attributes linked to the German General Social Survey



Note: This information was gathered from spatial datasets containing categorical values of dB(A) road traffic noise measured at the respondents' dwelling. Part of the noise definition were that values lower than 50 dB(A) were considered as being non-existing noise. Thus, values ranging from 0 up to 50 dB(A) are put into one category. At the same time, it is the most commonly occupied category.

To overcome this issue, we designed alternative measures of exposure such as geographic distance calculations. They link the spatial data attributes not on the simple overlaying geo-coordinates level but on results of the spatial analysis. In the simplest case, geographic distance calculations consist of distance measuring between two points. But we also can apply it to different combinations of spatial geometries.

Concerning road traffic noise data we can measure distances between points (i.e. the respondents' address) to the nearest edge of a polygon (i.e. the road traffic noise data) with a certain noise level attribute, for instance, 65 decibels. We can thereafter use these measurements in meters as proxy measurements for actual noise exposure that was not covered by the methods of environmental noise mapping, e.g. because the road where the respondents are dwelling was not defined as a main road. The advantage of this approach is that we receive values for all respondents of a sample and do not lack statistical power.

In the 'georefum' package the following command can be applied in order to achieve a geographic distance calculation for linking the results to survey data:

```
georefum::cdr_linking_distances()
```

The further parameters of this command follow the same logic as the simple linking procedures in the previous section. We derive the attributes, in this case geographic distances, we link them to coordinates of the respondents, we delete the coordinates, and finally, they are linked to the survey data attributes.

Other research questions that are not connected to data problems of simple linking but also use geographic distance calculations are relevant in studies of political behavior or accessibility of public facilities. For instance, some scholars studied whether the distance to poll sites does have an influence on voting behavior or whether the distance to kindergartens does have an effect on professional childcare (Biedinger, Kolb, & Klein, 2015).

5.3 Focal Analyses

Other applications in empirical social science research focus on extended neighborhood characteristics in comparison to the direct one. There already exists vast literature, however, the definition of the neighborhood often remains vague, or they cannot properly be operationalized due to the limited availability of small-scale spatial data and geocoded survey data. One exception are the German Census data available on 1 km² grid cells described in section 3.2. Since these characteristics indeed are small scale and also are evenly distributed over the whole area of Germany, they can be used to build comparable neighborhoods for all respondents of a survey.

We use so-called focal analyses to operationalize these neighborhoods. Because the grid cells of the German Census data are all designed in the same way, every simple calculation can be applied to each of the grid cells. Focal analysis uses one grid cell as a focal point (hence the name) and calculates statistics on surrounding grid cells. For instance, the attribute of the focal point can then be compared to the attributes of surrounding grid cells. This procedure is illustrated in Figure 6.

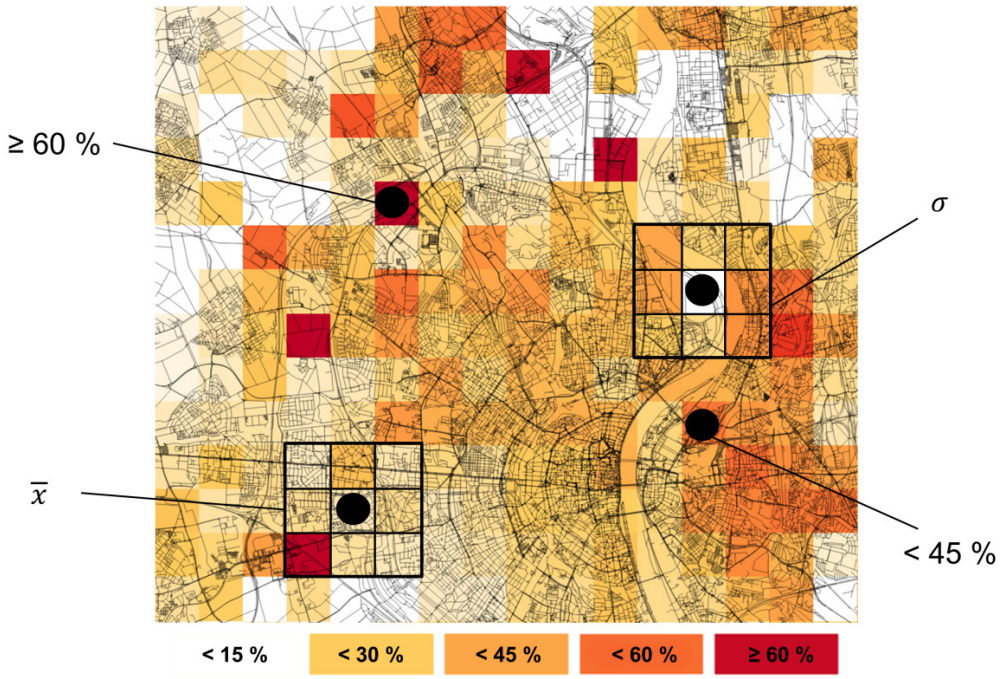
There is no standard way of applying focal analyses; it depends on the research questions. Whether which descriptive statistic should be used or whether which adjacent neighborhood grids should be incorporated is up to theory or empirical evidence as well. A flexible approach to the modelling of living environments therefore seems to be appropriate.

This being said, the 'georefum' package provides a function that offers an intuitive and flexible way to create neighborhoods basing on focal analyses. It heavily depends on the *focal()* function of the 'raster' package in R. In addition, it wraps this function in the same procedures as in the other functions of linking:

```
georefum::census_linking_focal()
```

In this example, we build a mean statistic of the surrounding neighborhood of a focal point. Note that we do not include the attribute value of the focal grid cell itself; it therefore depicts the mean value of the surrounding grid cells. In combination with simple linking procedures, both resulting attributes can comparably be analyzed, e.g., to explore distinct effects of direct neighborhoods and adjacent ones. As this procedure is abstract, we displayed this relationship between both of these measures in Figure 6.

Figure 6: Visualization of Simple Spatial Linking of Raster Data (Single Points) and Focal Analyses with Raster Data (Enclosed Points)



Note: Data source: OpenStreetMap and German Census 2011

The figure demonstrates that surrounding grid cells are adjacent to the focal grid cell. The information on these grid cells is used to calculate a mean value which can then be compared to the focal grid cell attribute value. This application is relevant for research questions in which hypothesis based on relations between direct neighborhoods and adjacent neighborhoods are tested (Legewie & Schaeffer, 2016). For example, in migration research differences in the amount of immigrants in a certain area are of high interest (Klinger et al., 2017). Other examples may apply to voting behavior research.

6 Conclusion

Geocoding and spatial linking of survey data are in great demand. This document serves as a beginner reference for social scientists by defining the central terms and giving examples for georeferencing projects. Also, we introduced the R package 'georefum' which provides functions to spatially link survey data to German Census 2011 data and environmental noise data. We hope that both this document and the 'georefum' package serve as a starting point for users' linking projects.

At the same time, the world of geospatial data is rich and advanced. Providers of geospatial data are working in diverse fields; available geospatial data contain information on a large range of different measures. Therefore, this short document cannot serve as an exhaustive introduction for social scientist into the world of georeferencing and geospatial data. The contents of the available data are too diverse, data formats are differing severely, and the stakeholders within the field are ranging from commercial providers to federal agencies.

Having said this, in light of an increasing amount of available geospatial data we presume that the focus on geospatial data within the social sciences and related fields will be more and more enlarged. Our project "Georeferencing of Survey Data" therefore not only developed tools for spatial linking matters – presented here –, but also developed expertise and service for consulting and advice. Besides the R package, GESIS will provide open access for academic researchers to the geocoding facilities of the BKG to avoid using web features that are critical for data protection issues or commercial service providers that are expensive. Other advances are going to develop dedicated data infrastructures to analyze linked data of social science survey data and data from the spatial sciences. Furthermore, the major conferences in the social sciences not only recently provide specialized sessions for georeferenced survey data. Thus, it's worthwhile to gain skills in spatial analysis.

7 References

Ackermann, K., & Traunmüller, R. (2014). Jenseits von Schwerkraft und H?llenfeuer. Nicht-lineare Kontexteffekte auf den Zusammenhang von religi?ser Gruppenzugeh?rigkeit und individuellem Wahlverhalten bei f?nf Bundestagswahlen. *Politische Vierteljahresschrift*, *55*(1), 33–66. https://doi.org/10.5771/0032-3470-2014-1-33

Biedinger, N., Kolb, J.-P., & Klein, O. (2015). Ethnische Unterschiede bei der Wahl des Kindergartens: Wer wählt den nächstgelegenen Kindergarten? Presented at the 3. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Bochum.

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-7618-4

Crowder, K., & Downey, L. (2010). Interneighborhood migration, race, and environmental hazards: modeling microlevel processes of environmental inequality. *AJS; American Journal of Sociology*, *115*(4), 1110–1149.

Förster, A. (2017). Diversity vs. Democracy? Neighbourhood Ethnic Heterogeneity and Electoral Turnout in Germany. *In Review*.

GESIS - Leibniz-Institut für Sozialwissenschaften. (2011). ALLBUS/GGSS 2010 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2010). GESIS Data Archive. https://doi.org/10.4232/1.10760

Hillmert, S., Hartung, A., & Weßling, K. (2017). Dealing with space and place in standard survey data. *Survey Research Methods*, *Special Issue: Uses of Geographic Information Systems Tools in Survey Data Collection and Analysis*(forthcoming).

Klinger, J., Müller, S., & Schaeffer, M. (2017). Der Halo- Effekt in einheimisch-homogenen Nachbarschaften: Steigert die ethnische Diversität angrenzender Nachbarschaften die Xenophobie? *Zeitschrift Für Soziologie*, forthcoming.

Legewie, J., & Schaeffer, M. (2016). Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict. *American Journal of Sociology*, *122*(1), 125–161. https://doi.org/10.1086/686942

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. https://doi.org/10.3390/ijerph120303246

Schweers, S., Kinder-Kurlanda, K., Müller, S., & 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. https://doi.org/10.1080/15420353.2015.1100152

8 Appendix

Here we provide information on potential and publically available spatial data sources that can be linked to geocoded survey data. Although this list is extensive, we cannot guarantee that it is also exhaustive. Because the spatial data landscape is very fragmented (Schweers et al. 2016), it's hard to be sure having captured all available information. We believe, however, since our scope was to find data that are relevant for the social sciences, that this list provides good support for finding appropriate data. We briefly describe what data can be acquired in each source and most importantly, how and where it can be obtained.

European Environmental Agency

What?

- Modern presentation of environmental topics such as noise or air pollution in different map applications

How?

- The data can be displayed in interactive maps
- Some indicators can be downloaded as datasets for further GIS processing

Where?

https://www.eea.europa.eu/data-and-maps

European Environment Information and Observation Network Data Repository (CDR)

What?

- A large range of environmental indicators gathered in context of EU obligations (e.g., noise and air pollution)
- Any EU member country deposit their data in single repositories on the site
- Data availability for whole country extents are therefore sometimes very fragmented

How?

- The data can directly be downloaded as shapefiles or other data formats
- Login is not always required

Where?

https://cdr.eionet.europa.eu/

Eurostat

What?

- Extensive European database for regional statistics of EU member states, including information on population, healthcare or economy
- Smallest regional units are NUTS3 regions

How?

- The data can directly be downloaded in different table formats

Where?

http://ec.europa.eu/eurostat/

Federal Agency for Cartography and Geodesy (BKG)

What?

- Offers an extensive set of different services for spatial data application ranging from geocoding or routing services up to detailed maps for administrative boundaries in Germany

How?

- Not all services are free before registration as the BKG is service provider for other German federal agencies and not the general public
- However, they also provide access to open data, e.g. for landscape and terrain models, topographic maps or general maps

Where?

- *http://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div?gdz_spr=eng&gdz_akt_zeile=5&gdz_anz_zeile=0&gdz_user_id=0*

Federal Agency for the Environment (UBA)

What?

- Data Portal for important environmental indicators, e.g. air pollution, for the whole extent of Germany

How?

- An extensive range of the environmental indicators can be visualized through their map application
- Some of the data can be downloaded, however, not all of them are available on a small spatial scale

Where?

http://www.umweltbundesamt.de/daten/umweltindikatoren (As of the data June 16, 2017, the English version was under construction: http://www.umweltbundesamt.de/en/data)

Geoportal NRW

What?

- Large library of spatial data on the population, environment, infrastructure, etc. for the State North Rhine-Westphalia

How?

- The data can be displayed in the map application

- The application also offers download services for some of the data to process them in own GIS software

Where?

https://www.geoportal.nrw/

German Federal Statistical Office (Republic and state)

What?

- Regional database for Germany that contains detailed results of official statistics, e.g. population, labor market, healthcare, etc.
- Data are not available as spatial data formats
- The smallest available regional unit are municipalities

How?

- The tables can be downloaded from the pages of the database
- Convenient functions for registered users such as saving of compiled tables

Where?

https://www.regionalstatistik.de/genesis/online/

German Historical GIS (HGIS)

What?

- Historical Shapefiles for the German Reich 1848 – 1914
- Additional data for single regional units or other countries of the world can be found in the Open Geoportal of the Harvard Library

How?

- The data can directly be downloaded as shapefiles or raster data
- No login required

Where?

http://calvert.hul.harvard.edu:8080/opengeoportal/

German Spatial Data Infrastructure (GDI-DE)

What?

- Maps for a large set of different indicators ranging from alternative energy sources over noise pollution caused by railways to water networks
- They are offered for the whole extent of Germany, but for different geographical scales ranging from single points up to the German States
- New data sources are continuously added

How?

- The maps can be locally stored as pictures or KML; they can even be accessed via a WMC web service
- No login required

Where?

http://www.geoportal.de/DE/Geoportal/Karten/

GovData Portal

What?

- Data for a large set of different indicators concerning the population, environment, economy, etc.
- However, only a few are offered for the whole extent of Germany (e.g. data that are also available through GDI-DE)
- New data sources are constantly added

How?

- The data can directly be downloaded
- The collection of different data formats is very diverse
- No login required

Where?

https://www.govdata.de/

IOER Monitor of the Leibniz Institute of Ecological Urban and Regional Development (IOER)

What?

- Amongst others, vast library of land use indicators for the whole extent of Germany

How?

- The indicators can be visualized through their map application
- They also can be integrated into GIS software and also be downloaded

Where?

http://maps.ioer.de/detailviewer/raster/

Indicators and Maps for Spatial and Urban Development (INKAR) of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBR)

What?

- Large set of different indicators concerning spatial and urban development such as residential structure or labor market statistics for the whole extent of Germany
- The indicators are available for various administrative units such as Kreise, Municipalities or the Federal States of Germany

How?

- The indicators can be requested by using an interactive application

Where?

http://www.inkar.de/ (Unfortunately, only in German)