International Cartographic Association
Association Cartographique Internationale

Semi-Automatic Spatial Classification of Heterogeneous Spatial Open Government Data – Use Case of Germany

Sebastian Meier [a,*]

[a] *Potsdam University of Applied Sciences, Germany, sebastian.meier@fh-potsdam.de*

* Corresponding author

Keywords: Data, Classification, Similarity, Taxonomy, Automation

Abstract:

As part of the research project Open Data Cloud Services (ODCS), we have been trying to overcome some of the limitations introduced through the heterogeneity of spatial open government data (sOGD). In this paper we describe some of the challenges of sODG and one of the tools we built to spatially organize heterogenous sOGD, to make it easier for users to *find data* and *automatically integrate* it into existing data structures and in the future allow for *cross-dataset spatial analysis*.

In Europe and beyond the number of sOGD is continuously growing. Derived from government mapping and planning programs, remote sensing and other sources, open spatial government data provides insights into a broad range of topics from geology to weather phenomena, to socio-demographic developments and much more. The data is provided through a variety of government institutions. The distributed generation and provision of spatial data through these different stakeholders has created an extremely heterogenous data space. Heterogenous regarding formats, data structure (time, space, and other attributes), spatial granularity, classification, down to publication cycles and the documentation of changes over time. While national and international working groups are working towards more standardisation, and e.g., in Europe, particularly the INSPIRE (European Commission, 2022) initiative had a positive impact on the standardisation of open spatial data‘s meta data, as our analysis of all (registered) German open data sets showed, the quality and especially taxonomies still vary strongly between data providers. This heterogeneity presents obstacles for all stakeholders alike from the civil society to researchers to companies trying to build business models on top of open data.

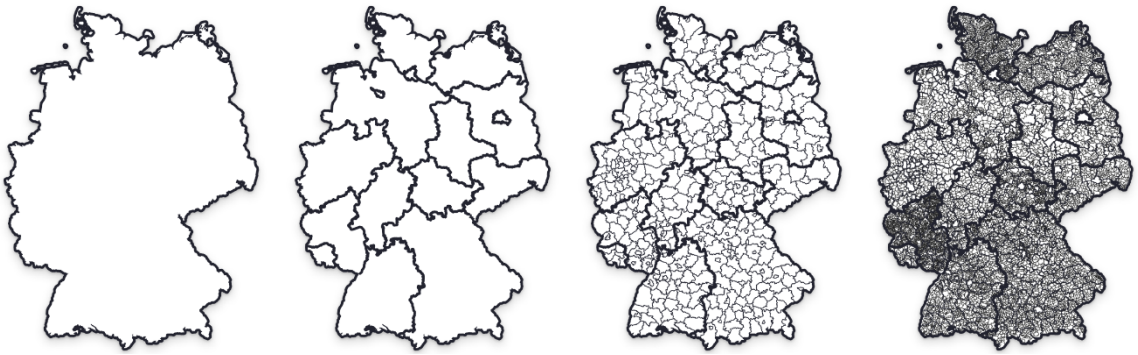

Figure 1. Spatial hierarchy of administrative regions in Germany.

The goal is to classify data sets across a variety of spatial hierarchies. An exemplary hierarchy are administrative regions (see figure 1). Classifying a certain dataset to cover a certain state and its granularity to be on a community level, allows users to compare this data set to other data sets of the same granularity for other states. Other exemplary hierarchies in Germany are statistical districts or various grid systems. Beyond inconsistencies within individual hierarchies, calculating relationships between different hierarchies or even non-hierarchical organizational systems, e.g. urban density areas, systems based on natural features, is a challenge, which we are trying to overcome with our work.

The ODCS infrastructure is comprised of multiple web services for processing OGD. The pipeline for the spatial classification (see Fig. 2) starts with the 1) aggregation of sOGD’s meta data from a variety of sources. The resulting meta data is then 2) homogenized and stored in a central database. Based on the meta data the system then decides if this could potentially be a spatial data set (file format, file name, URL structure, etc.) and then tries to 3) download the actual dataset. The downloaded data is then 4) analysed, if necessary, the analysis and classification process is 5)

Abstracts of the International Cartographic Association, 5, 2022.
European Cartographic Conference – EuroCarto 2022, 19–21 September 2022, TU Wien, Vienna, Austria.
https://doi.org/10.5194/ica-abs-5-59-2022

assisted by a user and the resulting classified data or rather the improved meta data is 6) provisioned to the users of the platform.

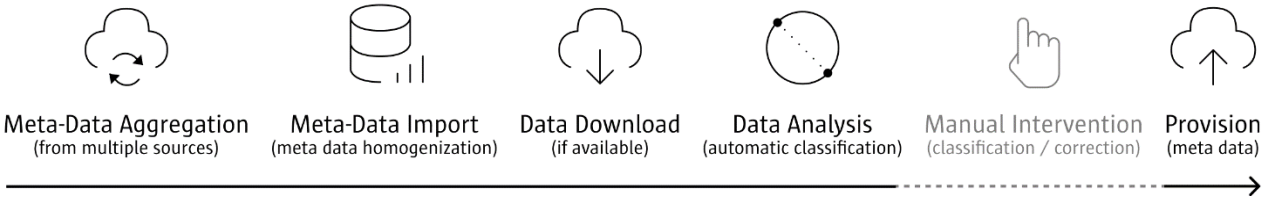

Figure 2. ODCS data processing pipeline for semi-automatic spatial classification.

As described above, the classification should provide improved meta data, to serve users a more precise search and, for future applications, allow for automatic spatial analysis across multiple data sources. To build the classification we primarily rely on spatial hierarchies and relationships between data sets (unconnected, touching, intersecting, containing). To identify those relationships, data sets are analysed threefold: 1) The bounding box of the dataset is calculated and compared to existing bounding boxes in the database. Compared to the next two analysis steps this is fast and works for raster as well as vector datasets. For vector datasets we 2) proceed to generate a "union" from the dataset and then compare the union to other already classified datasets. Exact matches are quite rare (even for datasets from the same provider), therefore, we use buffers on the existing data sets as well as the to be classified data set. As figure 3 illustrates, one and the same border can exist in a variety of geometries. Differences resulting from simplification, generalisation, precision, or actual errors in the data. The differences are in part so big, that we implemented a graphical user interface for human assistance, to manually classify "outliers".

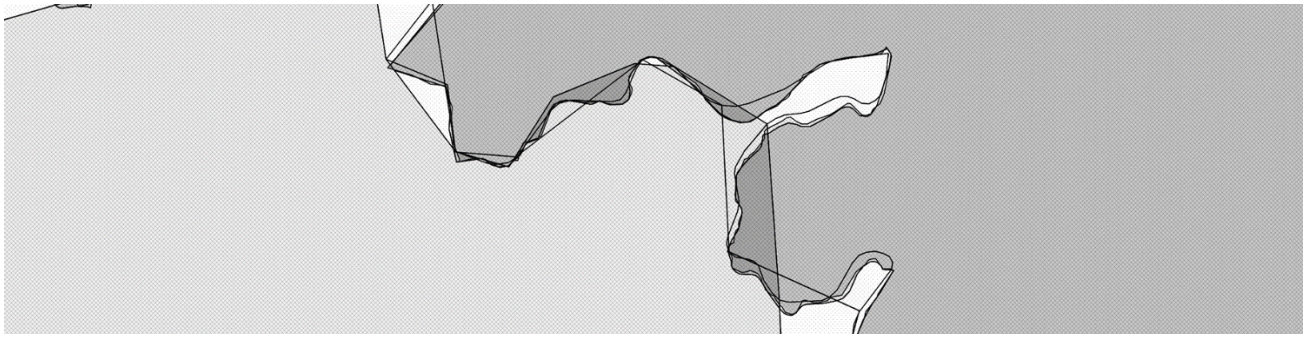

Figure 3. Border between two federal states in Germany, described by a variety of different spatial representations.

If a "good enough" match cannot be found, the next best matches are stored in the database and are then provided to the user to choose from. After the dataset's "union" is analysed, we then 3) split the features up into individual polygons, to again compare those to the already classified polygons in our database. Data sets or rather their polygons that could not be classified are added to the manual classification queue and are then classified by the user. Through this semi-automatic process, the library of polygons quickly grows and increases the automation over time. The system also runs a thematic classification based on meta data and data set attributes, which goes beyond the scope of this paper.

Due to the limitations of the abstract, we focused on an overview of the method, rather than the technologies used to perform the analysis. In short: GDAL, OGR2OGR, NodeJS, Python, PostgreSQL + PostGIS.

While the above-described use case was developed for Germany's OGD portals, the technique is not limited to German spatial data sets, as all applied methods are not language-depended, but only rely on spatial geometries. Even the meta-data homogenization, which is not described in depth in this paper, mostly relies on language-independent methods.

Acknowledgements

The Open Data Cloud Services received funding from the BMWK's IGP program. All resulting applications are published under open-source licenses on GitHub: https://github.com/opendatacloudservices. The graphics in this paper use the Iconoir icon set by Luca Burgio, available under MIT license.

References

European Commission (2022) INSPIRE Knowledge Base. https://inspire.ec.europa.eu/ Accessed on 2022-04-2

https://doi.org/10.5194/ica-abs-5-59-2022