2020 | Patrick Michl

IBM Applied Data Science Capstone Project

Segmenting and clustering postal code areas
in the Metropolitan region Rhine-Neckar (MRN)

Subject and Application

This project uses geospatial and location data to explore the Metropolitan region Rhine-Neckar in Germany. The focus is on the cartographic presentation of open data from various providers to highlight local specificities within the region.

Metropolitan region Rhine-Neckar

The Rhine-Neckar region comprises the major cities of Mannheim, Ludwigshafen am Rhein and Heidelberg, their surrounding areas, the more rural Neckar-Odenwald district and the Southern Palatinate. Since this area is largely identical with the core area of the historic electoral Palatinate, close socio-cultural ties exist despite the current division into three federal states. Due to this historically grown strong regional ties the proficiencies clustered into spatially condensed hot spots, e.g. for industry, arts, shopping facilities, recreation and education.

Cartographic representation

The goal of this project is to highlight regional specificities of the Rhine-Neckar region by applying statistical analysis of geospatial and location data. This allows a rough overview of individual strengths within the region. In this purpose for a more distinctive presentation simple clustering algorithms are incorporated.

Description of Data

Preliminary Requirements

A fundamental question, that arises with the statistical analysis of geospatial data concerns the choice of an appropriate spatial aggregation. While for data with complete geospatial information, like point data, the statistics can be derived with respect to arbitrary aggregations, based on previous summary statistics, like population, the specified spatial aggregations are predetermined and may only be coarsened by further aggregation. Hence for this project the geospatial data is required to either be given in one of the following formats:

- **Type A** Geospatial multipoint data (ISO/TC 211) that defines spatial aggregation boundaries
- **Type B** Geospatial point data (ISO/TC 211) with no spatial aggregation
- **Type C** Summary data for a spatial aggregation by postal code regions
- **Type D** Summary data for a spatial aggregation that refines postal code regions

OpenStreetMap (Table: OSM)

OpenStreetMap is a collaborative project to create a free editable map of the world. The geodata underlying the map is considered the primary output of the OSM project. The creation and growth of OSM has been motivated by restrictions on use or availability of map data across much of the world, and the advent of inexpensive portable satellite navigation devices.

The acquired data is of **type A**, comprises geospatial multipoint data for all German zip-code regions and is used to define the spatial aggregation boundaries. The data is licensed under the terms and conditions of the Open Database License (link), aggregated by SUCHE-POSTLEITZAHL.ORG (link) and hosted by OpenDataSoft (link).

Rhein-Neckar Wiki (Table: MRN)

The Rhein-Neckar Wiki is a free knowledge database for the Rhine-Neckar metropolitan region. It collects information about the associated cities, current and past events in and around them. The special feature of the Rhine-Neckar-Wiki is the Wiki principle, which is also used by Wikipedia. It allows both free and advertising-free access to the information, as well as the own participation and involvement without prior knowledge.

The acquired data is of **type D**, comprises all zip-code regions within MRN region and is used to restrict the OSM data to the MRN. The data is licensed under the terms and conditions of the CC BY-NC-SA 4.0 (link) and published by the Rhein-Neckar Wiki (link).

Rhein-Neckar-Verkehr GmbH (Table: RNV)

The Rhein-Neckar-Verkehr GmbH (RNV) is the most important traffic alliance in the metropolitan region Rhein-Neckar. It operates the suburban railway, tram and bus routes in Mannheim, Heidelberg and Ludwigshafen. On their open data portal (link) The RNV provides an interface as well as numerous data packages around public transport

The acquired data is of **type B**, comprises information about all active stops operated by the RNV and is used to summarize transportation frequencies within the zip-code regions of the MRN. Therefore, any stop is assigned to it's surrounding the zip-code region. The data is licensed under the terms and conditions of the dl-de-by-2.0 (link) and collected and published by the Rhein-Neckar-Verkehr GmbH.

Federal Statistical Office of Germany (Table: Census)

The federal statistical office of Germany provides geospatial population data for. This data is collected in a national population census, which is held at unregular intervals. The most recent census, that is provided by the federal statistical office is the 2011 European Union census.

The acquired data is of **type C**, comprises population data for all zip-code areas in and is used to incorporate demographic information in the data analysis. The data is licensed under the terms and conditions of the dl-de-by-2.0 (link) and aggregated to zip-code regions and published by SUCHE-POSTLEITZAHL.ORG (link).

Foursquare Labs, Inc. (Table: Foursquare)

Foursquare is an American provider for location data collected via billions of check-ins. The company rose to prominence by popularizing the concept of real-time location sharing and checking in. The data is provided via a REST API and therefore always reflects the current data stock provided by Foursquare.

The acquired data is of **type C**, comprises location data, given by the venues of different categories, and is used to summarize the frequencies of these respective categories within the individual zip-code regions. The data is licensed under the terms and conditions of the Foursquare License Agreement (link).

Overview of the Tables

The following tables represent the data schema after collection and data transformation

OSM	
zip	**Integer**
name	String
geometry	ISO/TC 211

MRN	
zip	**Integer**
state	String
district	String
boroughs	Set of Strings

RNV	
id	**Integer**
name	String
lon	Float
lat	Float

Census	
zip	**Integer**
population	Integer

Foursquare	
zip	**Integer**
arts	Integer
college	Integer
event	Integer
food	Integer
nightlife	Integer
outdoors	Integer
professional	Integer
residence	Integer
shop	Integer
travel	Integer