2020.01 | Patrick Michl

IBM Applied Data Science Capstone Project

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

Objective

This project has been created for the partial fulfilment of the requirements of the IBM applied data science certification in 2020.

The 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.

Data Integration

A fundamental design parameter for the statistical analysis of geospatial data regards the spatial aggregation topology. For data with complete georeferencing, like point data, the statistics can be derived with respect to arbitrary spatial aggregations. For many summary statistics, like population surveys, however, the spatial aggregation topology is predetermined and may only be coarsened by further aggregation. Hence it is required to define a coarsest aggregation topology before data acquisition.

For this project the coarsest aggregation topology is chosen to be given by the zip-code areas, such that any used geospatial data is required to either be given in one of the following formats:

- **Type A** Geospatial polygon data (RFC 7946) that defines spatial aggregation boundaries
- **Type B** Geospatial point data (RFC 7946) with no spatial aggregation
- **Type C** Summary data for a spatial aggregation by zip-code areas
- **Type D** Summary data for a spatial aggregation that refines zip-code areas

Rhein Neckar Wiki

Table: `MRN`

Description: The Rhein-Neckar Wiki is a free knowledge database for the metropolitan region Rhine-Neckar. It collects information about the associated cities, as well as current and past events in and around them. The wiki principle allows free access to the information, the own participation and involvement without prior knowledge. The provided information is licensed under the terms and conditions of the CC BY-NC-SA 4.0 [1] and published by the Rhein-Neckar Wiki authors [2].

Integration: The integrated data is of type D and comprises all MRN zip-code regions by the administrative type, federal state, district and a set of assigned boroughs. It is used as the primary source for an administrative definition of the MRN.

Rhein-Neckar-Verkehr GmbH

Table: `RNV`

Description: The Rhein-Neckar-Verkehr GmbH (RNV) is the most important traffic alliance in the metropolitan region Rhein-Neckar. It operates suburban railways, trams and bus routes in Mannheim, Heidelberg and Ludwigshafen. The RNV provides an interface as well as numerous open data packages around public transport [3]. The data is licensed under the terms and conditions of the dl-de-by-2.0 [4] and collected and published by the Rhein-Neckar-Verkehr GmbH.

Integration: The integrated data is of type B and comprises information about all active stops operated by the RNV. It is used to summarize the appearance frequencies of active stops within the individual zip-code areas of the MRN.

Federal Statistical Office of Germany

Table: `Census`

Description: The federal statistical office of Germany provides geospatial population data for Germany. 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 data is licensed under the terms and conditions of the dl-de-by-2.0 [4] and aggregated to zip-code areas by SUCHE-POSTLEITZAHL.ORG [5].

Integration: The integrated data is of type C and comprises population data for all zip-code areas of the MRN. It is used to incorporate demographic information into the data analysis.

OpenStreetMap

Table: `OSM`

Description: 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 data is licensed under the terms and conditions of the Open Database License [6], aggregated by SUCHE-POSTLEITZAHL.ORG [7] and hosted by OpenDataSoft [8].

Integration: The integrated data is of type A, comprises geospatial polygon data for all German zip-code areas and is used to define the spatial aggregation boundaries of the MRN.

Foursquare Labs, Inc.

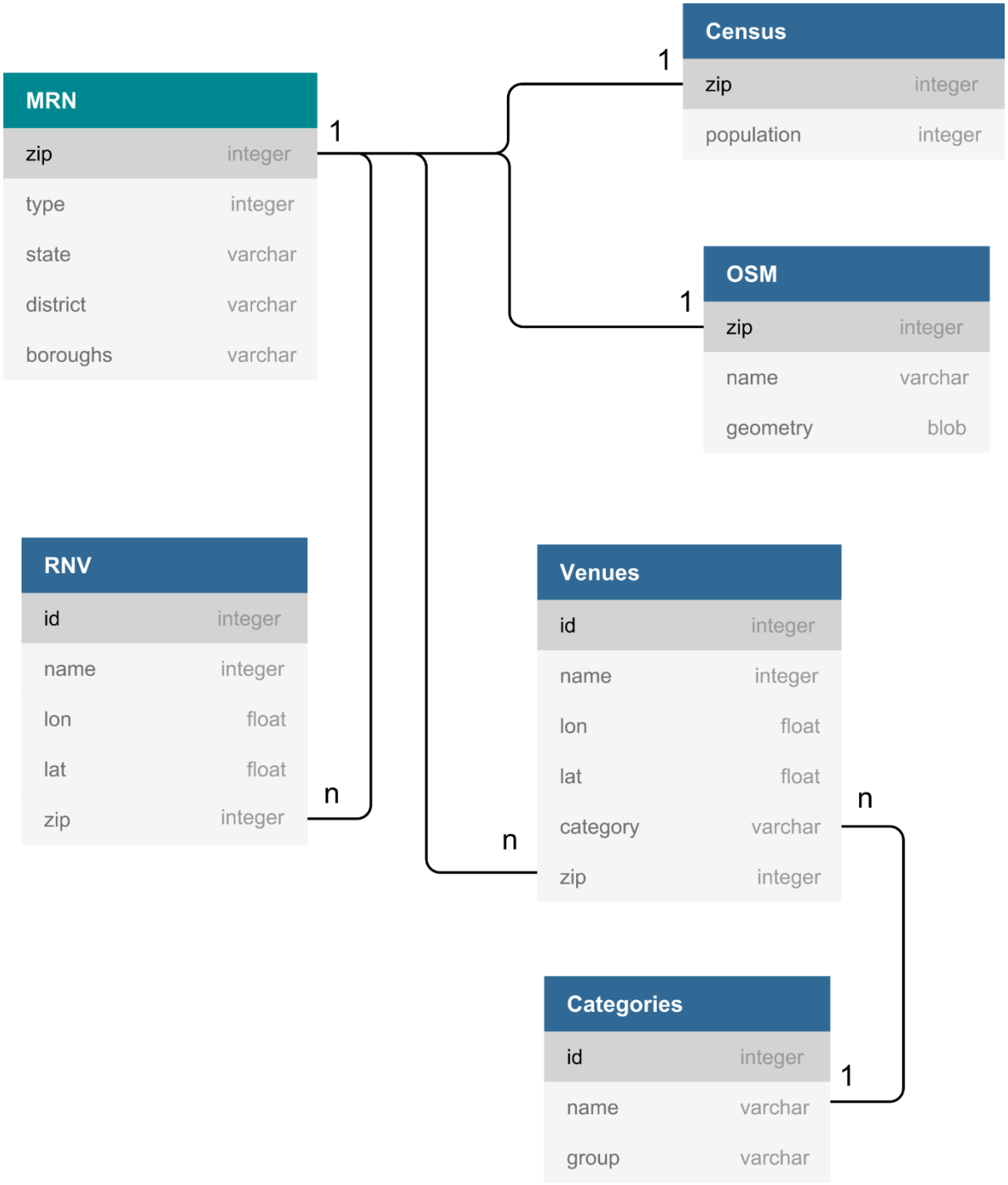
Tables: `Venues`, `Categories`

Description: 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 API and therefore reflects the current data stock provided by Foursquare. The data is licensed under the terms and conditions of the Foursquare License Agreement [9].

Integration: The integrated data is of type C, comprises location data, given by the venues of different categories, and is used to summarize the appearance frequencies of venues of the respective categories within the individual zip-code areas of the MRN.

Data Integration Overview

The following Entity-Relationship Model represents the data schema after the ETL process:



Data Transformation

Apart of the geospatial information, the quantitative information comprises the **frequencies** of certain appearances within the individual zip-code areas of the MRN. These respectively can be extracted grouping. However, in order to derive comparable summary statistics, it is required to apply a normalization, which in turn is feasible through different approaches:

(1) The **per area density** averages the aggregated sum by the enclosed area of the spatial aggregation boundaries. This approach provides spatial densities like the population density.

(2) The **per capita density** averages the aggregated sum by the total population within the spatial aggregation boundaries. In this way summary statistics are obtained, that describe the appearance frequency with respect to the population and therefore a kind of supply.

Area estimation

In accordance to the OGC standard [10] the attribute `OSM.geometry` contains encoded multipolygons with a polar coordinate reference system (EPSG:4326, e.g. used by GPS satellite navigation). The area estimation therefore requires a preceded transformation into Cartesian coordinates. This allows the subsequent application of the Shoelace formula [11] to derive the areas for all simply connected polygons, which in turn are summed up to the multipolygons, that describe the zip-code areas. Finally, these estimates are stored in the float type attribute `MRN.area`.

Population density estimation

First, the tables `MRN` and `Census` are left outer joined on their common attribute `zip`, which provides the integer attribute `MRN.population`. Thereupon the quotient of the attributes `MRN.population` and `MRN.area` is stored within the float type attribute `MRN.population_density`.

RNV stop frequency, per area density and per capita density

The table `RNV` is grouped by its attribute `zip`. Afterwards the sizes of the groups are stored within the integer attribute `MRN.rnv_count`. Thereby the NULL values are initialized by zeros. At this foundation, the float attributes `MRN.rnv_density` and `MRN.rnv_supply` are derived by a respectively division of `MRN.rnv_count` through `MRN.area` and $\texttt{MRN.population} \cdot 10^{-3}$.

Foursquare group frequency, per area density and per capita density

First, the Foursquare tables `Venues` and `Categories` are left outer joined on the attributes `Venues.category` and `Categories.name` which provides the string attribute `Venues.group`. This attribute contains the Foursquare group for each venue: Arts & Entertainment, College & University, Event, Food, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. Thereupon the table `Venues` is grouped by the attribute `zip` and for each group the numbers of appearances of are stored within respective integer attributes `MRN.[GROUP]_count`. At this foundation, the float attributes `MRN.[GROUP]_density` and `MRN.[GROUP]_supply` are respectively derived by a division of `MRN.[GROUP]_count` through `MRN.area` and $\texttt{MRN.population} \cdot 10^{-3}$.

Areal location dissemination

Apart of their frequencies the geospatial location data in the tables `RNV` and `Venues` may also be aggregated by their areal dissemination. In the first step the tables `RNV` and `Venues` are concatenated to obtain more samples per zip-code area. Afterwards the coordinates of the resulting table are transformed to Cartesian coordinates. This allows a derivation of the distance correlation within each zip-code area [11]. Finally, the location dissemination is derived by “1 - *distance_correlation*(x, y)” and stored within the float type attribute `MRN.dissemination`.

Feature Engineering

...

Data Visualization

Generic Features

Area	**Attribute: MRN.area**
Population density	**Attribute: MRN.population_density**

RNV: Active Stops

Frequency	**Attribute: MRN.rnv_count**
Density	**Attribute: MRN.rnv_density**
Per Capita	**Attribute: MRN.rnv_supply**

Foursquare: Arts & Entertainment

Frequency	**Attribute: MRN.arts_count**
Density	**Attribute: MRN.arts_density**
Per Capita	**Attribute: MRN.arts_supply**

Foursquare: College & University

Frequency	**Attribute: MRN.college_count**
Density	**Attribute: MRN.college_density**
Per Capita	**Attribute: MRN.college_supply**

Foursquare: Event

Frequency	**Attribute: MRN.event_count**
Density	**Attribute: MRN.event_density**
Per Capita	**Attribute: MRN.event_supply**

Foursquare: Food

Frequency	**Attribute: MRN.food_count**
Density	**Attribute: MRN.food_density**
Per Capita	**Attribute: MRN.food_supply**

Foursquare: Nightlife Spot

Frequency	**Attribute: MRN.nightlife_count**
Density	**Attribute: MRN.nightlife_density**
Per Capita	**Attribute: MRN.nighlife_supply**

Foursquare: Outdoors & Recreation

Frequency	**Attribute: MRN.outdoors_count**
Density	**Attribute: MRN.outdoors_density**
Per Capita	**Attribute: MRN.outdoors_supply**

Foursquare: Professional & Other Places

Frequency	**Attribute: MRN.professional_count**
Density	**Attribute: MRN.professional_density**
Per Capita	**Attribute: MRN.professional_supply**

Foursquare: Residence

Frequency	**Attribute: MRN.residence_count**
Density	**Attribute: MRN.residence_density**
Per Capita	**Attribute: MRN.residence_supply**

Foursquare: Shop & Service

Frequency	**Attribute: MRN.shop_count**
Density	**Attribute: MRN.shop_density**
Per Capita	**Attribute: MRN.shop_supply**

Foursquare: Travel & Transport

Frequency	**Attribute: MRN.travel_count**
Density	**Attribute: MRN.travel_density**
Per Capita	**Attribute: MRN.travel_supply**

References

[1] [Online]. Available: https://creativecommons.org/licenses/by-nc-sa/4.0/.

[2] [Online]. Available: https://rhein-neckar-wiki.de/Postleitzahlen.

[3] [Online]. Available: https://opendata.rnv-online.de/.

[4] [Online]. Available: https://www.govdata.de/dl-de/by-2-0.

[5] [Online]. Available: https://www.suche-postleitzahl.org/download_files/public/plz_einwohner.xls.

[6] [Online]. Available: https://www.openstreetmap.org/copyright.

[7] [Online]. Available: https://www.suche-postleitzahl.org/downloads.

[8] [Online]. Available: https://public.opendatasoft.com/explore/dataset/postleitzahlen-deutschland.

[9] [Online]. Available: https://foursquare.com/legal/api/licenseagreement.

[10] [Online]. Available: https://www.opengeospatial.org/docs/is.

[11] [Online]. Available: https://rsaiconnect.onlinelibrary.wiley.com/doi/abs/10.1111/pirs.12451.