2020.02 | Patrick Michl

IBM Applied Data Science Capstone Project

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

This project has been created for the partial fulfilment of the requirements of the IBM applied data science certification. All contained results and graphs were created using the Python programming language and are publicly available as a Jupyter notebook on GitHub [1].

Objective

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, statistical aggregation can be derived with respect to arbitrary spatial bounds. For many summary statistics, like population surveys, however, the spatial aggregation topology is predetermined and may only be coarsened by further aggregation. This requires the definition of a (coarsest) aggregation topology.

For this project the (coarsest) aggregation topology is chosen to be given by the zip-code areas, such that any used data source is given in one of the following types:

Type A Fact table with summary statistics, aggregated by zip-code areas (or a subdivision)
Type B Geospatial point data [3] with no spatial aggregation
Type C Geospatial shape data [3], that defines spatial aggregation areas
Type D Additional dimension table for tables of type A, B or C

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 associated cities, current and past events. Due to the wiki principle, the data is validated by the community. The provided information is licensed under the terms and conditions of the CC BY-NC-SA 4.0 [4] and published by the Rhein-Neckar Wiki authors [5].

Integration: The integrated table is of type A and comprises the administrative type (1=District-free City, 2=County City, 3=County Municipality), federal state, district and boroughs of all MRN zip-code areas. 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 open data portal [6] with data, licensed under the terms and conditions of the DL-DE-BY-2.0 [7].

Integration: The integrated table is of type B and comprises an id, a name and the coordinates of all active stops operated by the RNV. Thereupon, the coordinates are used to assign zip-codes based on the OSM data. It is used to summarize the appearances of RNV 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 irregular time intervals. The most recent census, that is provided is the 2011 European Union census. The data is licensed under the terms and conditions of the DL-DE-BY-2.0 [7] and aggregated to zip-code areas by SUCHE-POSTLEITZAHL.ORG [8].

Integration: The integrated table is of type A 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 [9], aggregated by SUCHE-POSTLEITZAHL.ORG [10] and hosted by OpenDataSoft [11].

Integration: The integrated table is of type C and comprises names and the geospatial shapes of the German zip-code areas. It is used to define the spatial aggregation bounds within the individual zip-code areas of the MRN.

Foursquare Labs, Inc.

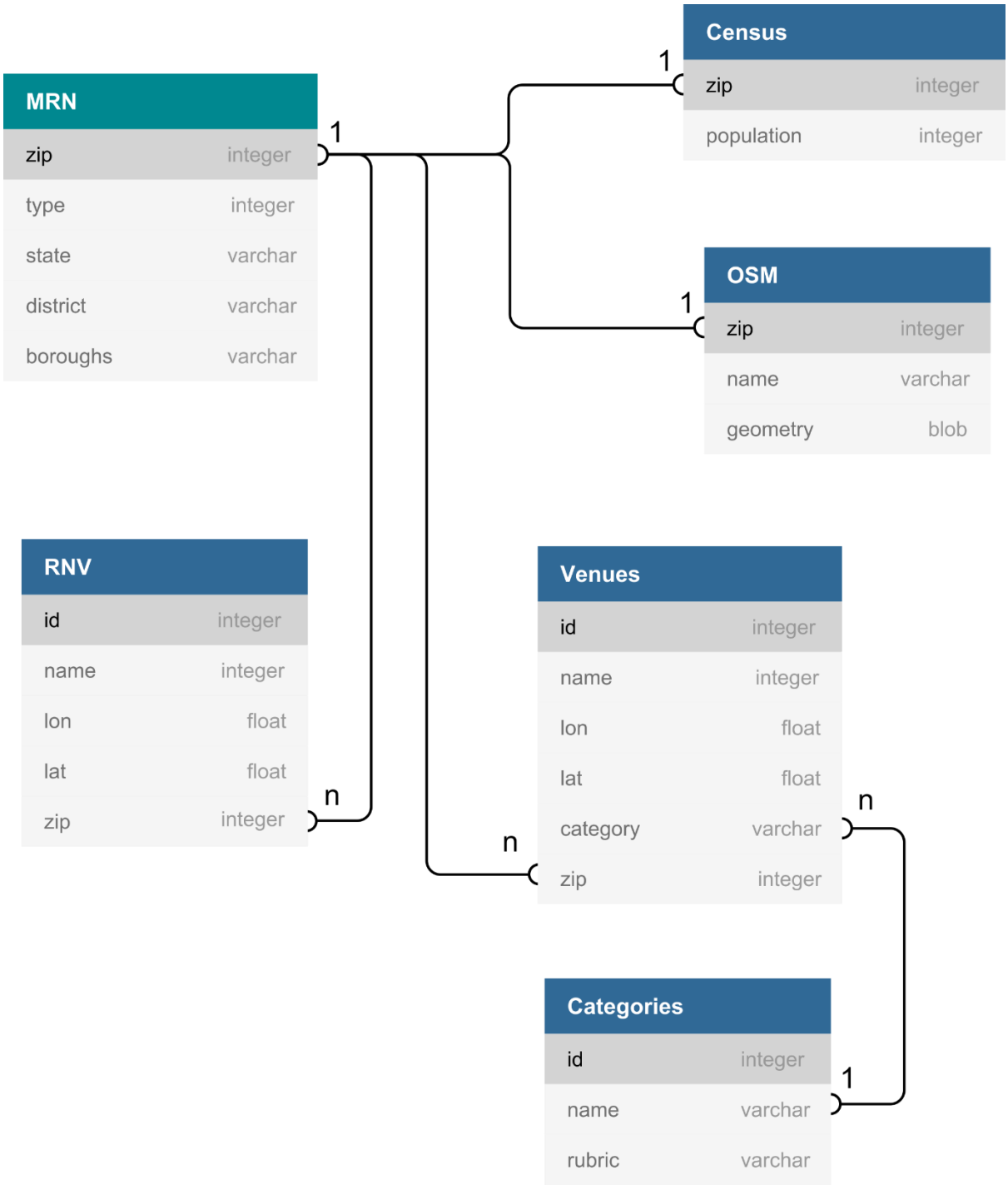
Tables: `Venues`, `Categories`

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. 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 [12].

Integration: The integrated tables are of type C and D and comprise location data, given by the venues of different categories and the category types. They are used to summarize the appearances of venues of the respective types 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 **counts** of certain objects within the individual zip-code areas of the MRN. These can be calculated by grouping tables of the objects by zip-codes and subsequently to take the group sizes. However, in order to derive comparable summary statistics, it is required to apply a further normalization of these count variables. This normalization can be carried out in several ways:

(1) The **area density** averages a count variable by the enclosed area of the spatial aggregation boundaries. This approach provides spatial densities of object distributions.

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

(3) The **rubric ratio** measures the frequency of a rubric within a larger set of outcomes. It thus equals the conditional probability that the outcome of a categorical variable can be found within a given rubric under the precondition of the occurrence of an outcome.

Area estimation

In accordance to the OGC standard [13] 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 [14] to derive the areas for all simply connected polygons, which in turn are summed up to those 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 count, area density and per capita density

The table `RNV` is grouped by its attribute `zip` and 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, area density, per capita density and rubric ratio

The Foursquare tables `Venues` and `Categories` are left outer joined on the respective attributes `Venues.category` and `Categories.name`, which provides the string attribute `Venues.rubric`. This attribute comprises the values: 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 number of objects of each rubric is stored within a respective integer attribute `MRN.[RUBRIC]_count`. Thereupon also the sum of all venues per zip-code area is stored within the integer attribute `MRN.venue_count`. At this foundation, the area density, the per capita density and the group ratio are derived by a division of `MRN.[RUBRIC]_count` through `MRN.area`, $\texttt{MRN.population} \cdot 10^{-3}$ and `MRN.venue_count` and stored within the float attributes `MRN.[RUBRIC]_density`, `MRN.[RUBRIC]_supply` and `MRN.[RUBRIC]_ratio`.

Data Cleansing

Density-Model based anomaly detection for venue data

Whereas the RNV data may be assumed to be free of inconsistencies, the Foursquare data depends on the voluntary cooperation of its users, and therefore is prone to errors. It is therefore reasonable to evaluate the Foursquare data in terms of their probabilities, and thereupon to detect and filter anomalies within the data. Thereby it must be considered that a density estimation of the total number of Foursquare venues is more reliable, than density estimations of the individual rubrics.

In logarithmic scales a strong linear correlation between the number of Foursquare venues and the population per zip-code area appears.

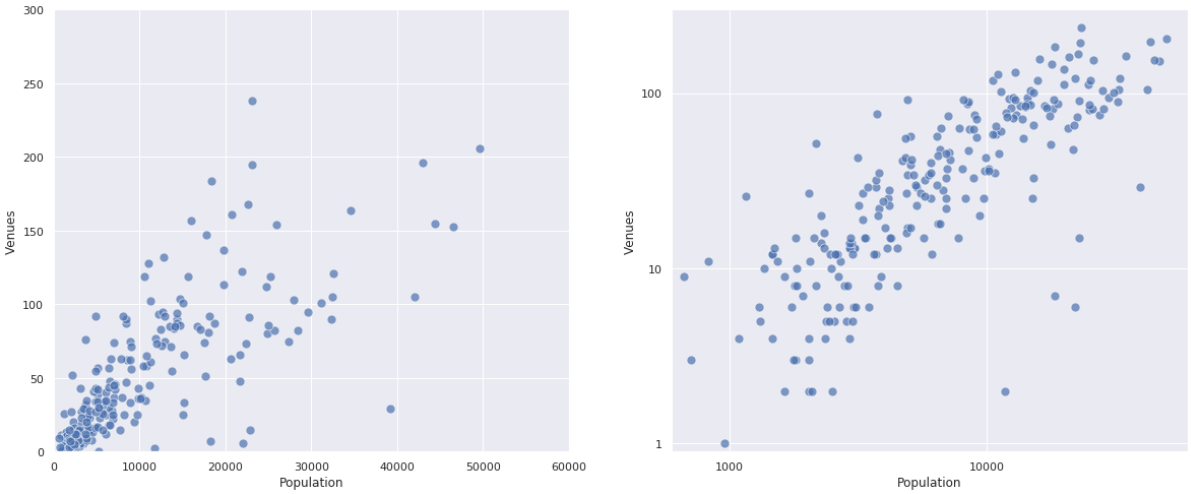


Figure 1: Number of venues, plotted against the population.
Left: Linear scales, Right: Logarithmic scales

Based on the assumption that this trend will continue to some extent for an increasing population, the probability density is modelled as follows: For the log-transformed data the two principal components are derive. Whereas the first describes the relationship between log(*venues*) and log(*population*), the second describes the deviation from this relationship. Thus, the probability density is modelled as a normal distribution with respect to the second principal component.

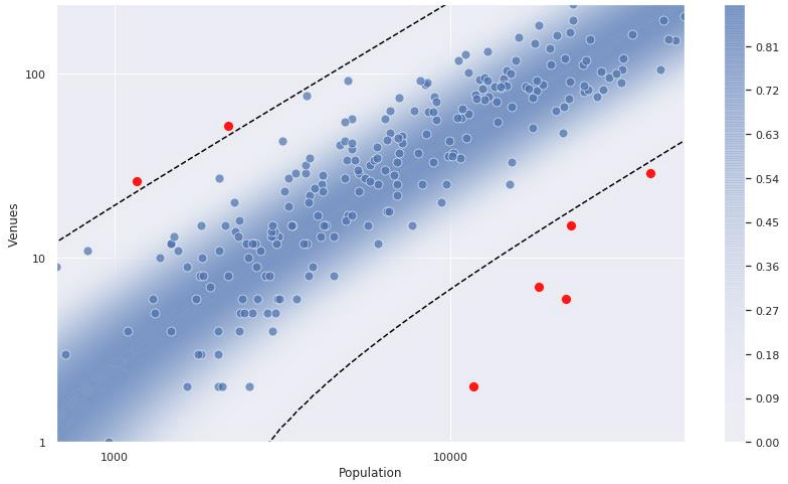


Figure 2: Joint probability density of venues and population.
Points with a p-value < 0.05 are detected as anomalies (red)

Feature Engineering

Discretization of continuous features

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 [15]. Finally, the location dissemination is derived by “1 - *distance_correlation*(*X*, *Y*)” and stored within the float type attribute `MRN.dissemination`.

Data Visualization

References

[1] [Online]. Available: https://github.com/fishroot/IBM-Applied-Data-Science-Capstone/blob/master/final-assignment.ipynb.

[2] [Online]. Available: https://creativecommons.org/licenses/by/4.0/.

[3] [Online]. Available: https://tools.ietf.org/html/rfc7946.

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

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

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

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

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

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

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

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

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

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

[14] [Online]. Available: https://en.wikipedia.org/wiki/Shoelace_formula.

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