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] and in the IBM cloud [2].

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 [4] with no spatial aggregation
Type C Geospatial shape data [4], 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 [5] and published by the Rhein-Neckar Wiki authors [6].

Integration: The integrated table is of type A and comprises the administrative type (District-free City, County City, 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 [7] with data, licensed under the terms and conditions of the DL-DE-BY-2.0 [8].

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 [8] and aggregated to zip-code areas by SUCHE-POSTLEITZAHL.ORG [9].

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 [10], aggregated by SUCHE-POSTLEITZAHL.ORG [11] and hosted by OpenDataSoft [12].

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.

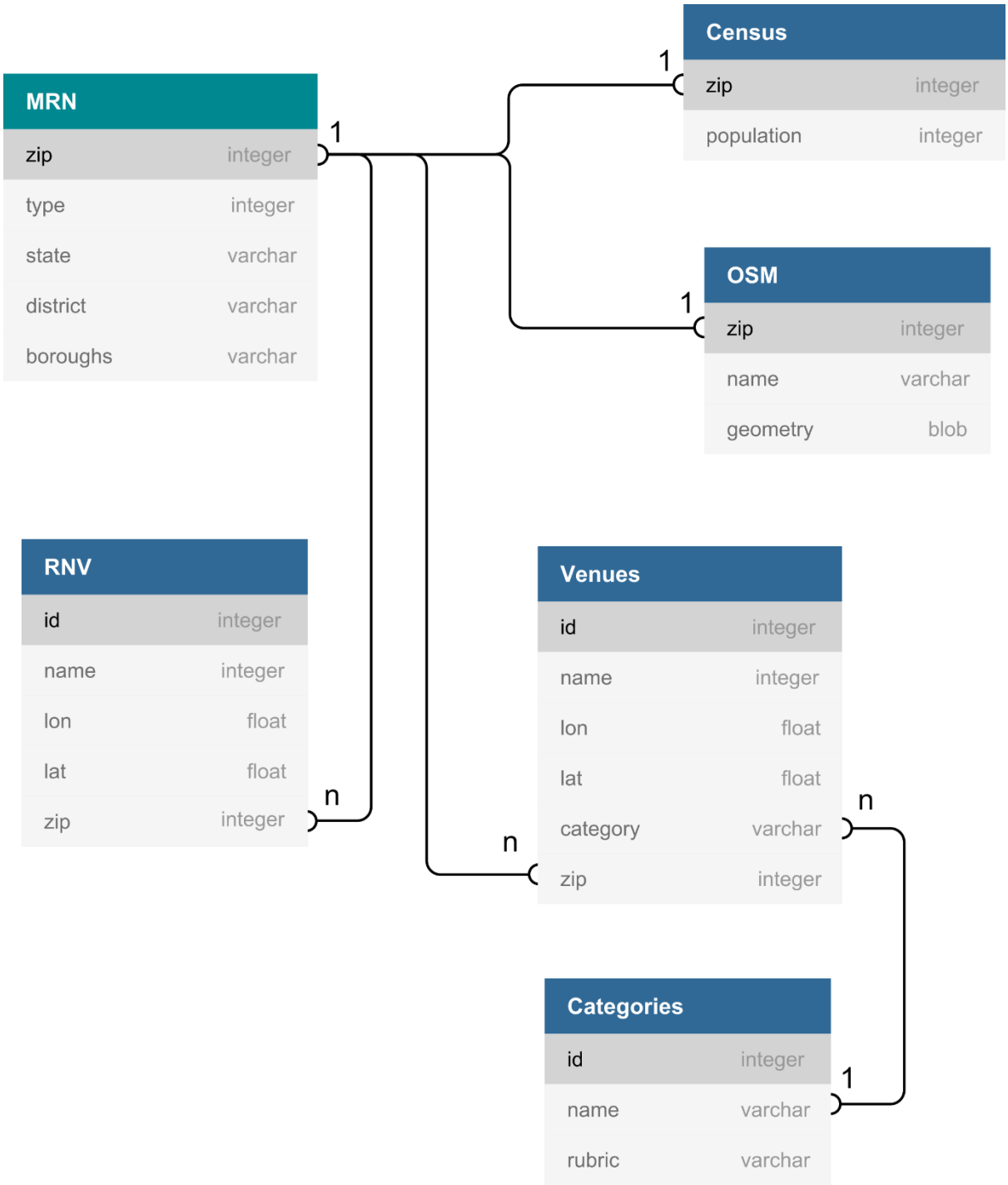
Tables: `Venues`, `Categories`

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 [13].

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 [14] 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 [15] 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 `MRN.population`.

Foursquare group frequency, area density, per capita density and rubric ratio

The Foursquare tables `Venues` and `Categories` are left outer joined on the 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 `Venues.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 in the integer attribute `MRN.venue_count`. At this foundation, the area density, the per capita density and the rubric ratio are respectively derived by a division of `MRN.[RUBRIC]_count` through `MRN.area`, `MRN.population` and `MRN.venue_count` and stored within the float attributes `MRN.[RUBRIC]_density`, `MRN.[RUBRIC]_supply` and `MRN.[RUBRIC]_ratio`.

Data Cleansing

Whereas the RNV data may be assumed to be objective, the Foursquare data depends on the cooperation of its users, and therefore may be assumed to be subjective. It is therefore reasonable to evaluate the Foursquare data in terms of a joint probability density function (PDF) and thereupon to detect and filter anomalies within the data by a threshold of a minimal probability, e.g. given by a p-value.

Density-model based anomaly detection for venues per population

The detection of anomalies within the venues per population requires an estimation of the joint PDF of venues and population. A good starting point to gather model assumptions about this joint PDF is therefore to visually search for patterns within joint representations [Figure 1].

Figure 1: Venues against population: (Top-Left) linear scale, (Top-Right) linear-log scale, (Bottom-Left) log-linear scale, (Bottom-Right) log-log scale

Whereas the linear scale and the semi-linear scales do not reveal apparent coherences between venues and population, the log-log scale indicates a strong (non-linear) correlation. Although this observation is not sufficient to imply a power law relationship [16] it can be used to model the joint PDF.

Based on the assumption that the non-linear correlation is subject to an even distribution in log-scales, the joint PDF is modelled as follows: For the log-transformed data the principal components are derived. Due to the strong correlation in log-scales, the first describes the non-linear relationship between venues and population and the second describes the local deviation from this relationship. This deviation is modelled as a normal distribution with respect to data projected to the second principal component [Figure 2].

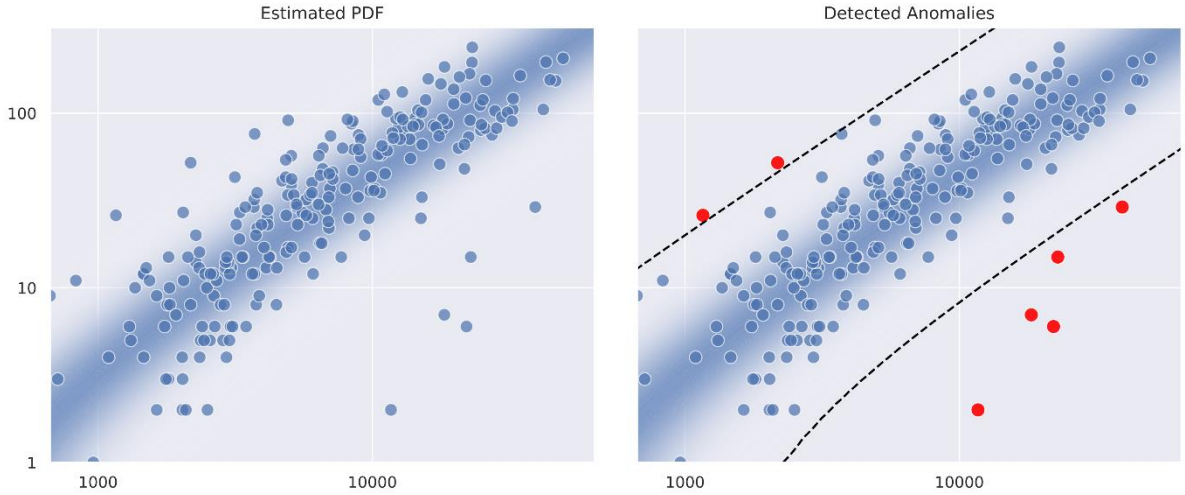


Figure 2: Estimated PDF of venues and population: (Left) Estimated Density in logarithmic scale, (Right) Points with a p-value < 0.05 are detected as anomalies (red)

For a p-value < 0.05 two groups of anomalies are to be distinguished:

The **upper group** (the red points on the lop-left corner in the right graph of Figure 2) are areas, that by far exceed the statistical expectations regarding the venue supply (venues per population). In the order of increasing p-values this group consists of the zip-code areas {*Kallstadt*} (Bad Dürkheim) and {*Zeiskam*} (Germersheim). The question arises whether these areas have been selected by chance and therefore due to statistical errors within the data. However, the regional and cultural background suggests, that the selection is reasonable: Kallstadt and Zeiskam are both located within the Palatinate wine-growing region, the largest wine-growing region in Germany. This region not only has a thriving tourism, but also a distinct pub culture. Indeed, the high pub density in Kallstadt has already been subject of public interest [17].

The **lower group** (the red points on the bottom-right corner in the right graph of Figure 2) are areas, that fall below the statistical expectations regarding the venue supply. In the order of increasing p-values this group consists of the zip-code areas {*Friedrichsfeld*} (Mannheim), {*Abenheim*, *Rheindürkheim*, *Ibersheim*, *Herrnsheim*} (Worms), {*Neuostheim/Neuhermsheim*, *Lindenhof*} (Mannheim), {*Wieblingen*, *Pfaffengrund*} (Heidelberg), {*Bensheim*} (Bergstraße) and {*Rheinau*} (Mannheim). Apart of Bensheim, it is apparent, that all other boroughs are suburbs. This encourages the conclusion, that the population in these areas is rather attracted by the urban centres than by local venues. Although the explanation for Bensheim seems more complicated, the underlying situation could be very similar. Due to the short distance to Frankfurt and the direct connection via the Autobahn A5, it is evident that a significant proportion of the population in Bensheim are commute workers. This might also cause, that many aspects of the professional and cultural life in Bensheim relocate to Frankfurt and the surrounding area.

Density-model based anomaly detection for venues per area

The detection of anomalies within the number of venues per area requires an estimation of the joint PDF of venues and area. Thereby once again a good starting point is to gather model assumptions about this PDF by visually inspecting joint representations [Figure 3].

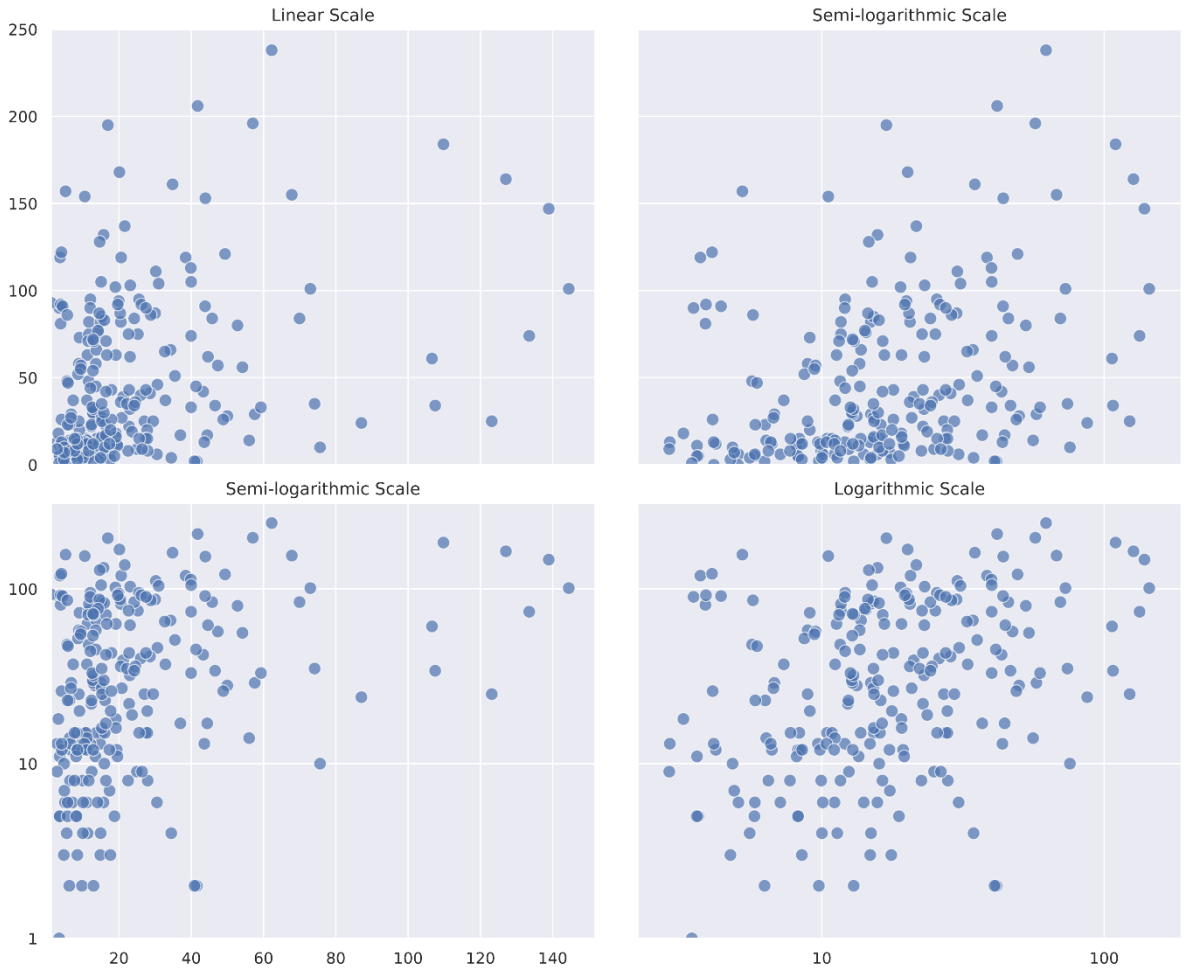


Figure 3: Venues against area: (Top-Left) linear scale, (Top-Right) linear-log scale, (Bottom-Left) log-linear scale, (Bottom-Right) log-log scale

The linear scale and the semi-linear scales do not reveal apparent coherences. In difference to the venues per population scatter plot, however, also the log-log scale hardly seems to allow the assumption of a correlation. This is not surprising by considering a broad range of population densities that directly affects the venue per area distribution. Due to the multiplicative composition of the area, however, the log-log scale still seems as a natural choice to model the joint PDF of venues and area.

Based on the assumption that the non-linear correlation is subject to an even distribution in log-scales, the joint PDF is modelled as follows: At first for the log-transformed data the principal components are derived. Due to the linear correlation in log-scales, the first principal component describes the non-linear relationship between venues and areas and the second describes the local deviation from this relationship. This deviation is modelled as a normal distribution with respect to the projection of the data to the second principal component [Figure 4].

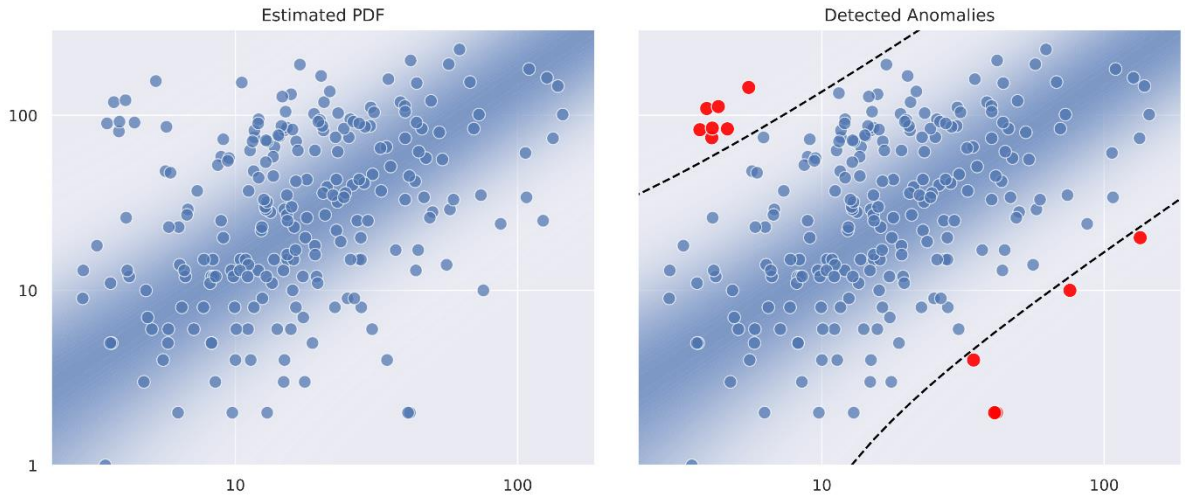


Figure 4: Estimated PDF of venues and area: (Left) Estimated density in logarithmic scale, (Right) Points with a p-value < 0.05 are detected as anomalies (red)

For a p-value < 0.05 two groups of anomalies are to be distinguished:

The **upper group** (the red points on the lop-left corner in the right graph of Figure 4) are areas, that by far exceed the statistical expectations regarding the venue density (venues per area). In the order of increasing p-values this group consists of the zip-code areas {*Innenstadt (östlicher Teil)*} (Mannheim), {*Innenstadt (westlicher Teil)*} (Mannheim), {*Bahnstadt*, *Bergheim*, *Weststadt*} (Heidelberg), {*Mitte*, *West*} (Ludwigshafen), {*Schwetzingerstadt/Oststadt*} (Mannheim), {*Süd*} (Ludwigshafen), {*Neckarstadt-Ost/Wohlgelegen*} (Mannheim). Essentially, these zip-code areas form the centres (and neighboring areas) of the MRN city triangle Mannheim, Ludwigshafen and Heidelberg. In the course of a continuous urban redensification within these areas about the last century the high venue densities are an expected result.

The **lower group** (the red points on the bottom-right corner in the right graph of Figure 4) are areas, that fall below the statistical expectations regarding the venue density. In the order of increasing p-values this group consists of the zip-code areas {*Ibersheim*, *Abenheim*, *Herrnsheim*, *Rheindürkheim*} (Worms), {*Rosenberg*} (Neckar-Odenwald-Kreis), {*Elmstein*} (Bad Dürkheim), {*Gossersweiler-Stein*, *Völkersweiler*, *Waldrohrbach*, *Rinnthal*, *Waldhambach*, *Silz*, *Münchweiler*, *Eußerthal*, *Albersweiler*, *Ramberg*, *Wernersberg*, *Dernbach*} (Südliche Weinstraße) and {*Schönbrunn*} (Rhein-Neckar-Kreis). For these zip-code areas two characteristics are significant: Large areas and low population densities, when compared to other zip-code areas within the MRN. Therefore, also in this case the selection seems conclusive, such that it can be assumed that these are not statistical outliers but actual regional specificities.

References

[1] P. Michl, „Segmenting and clustering postal code areas in the Metropolitan region Rhine-Neckar,“ 02 2020. [Online]. Available: https://github.com/fishroot/IBM-Applied-Data-Science-Capstone/blob/master/final-assignment.ipynb.

[2] P. Michl, „Segmenting and clustering postal code areas in the Metropolitan region Rhine-Neckar,“ 02 2020. [Online]. Available: https://tinyurl.com/tet2qvb.

[3] „Creative Commons BY 4.0,“ [Online]. Available: https://creativecommons.org/licenses/by/4.0/.

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

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

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

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

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

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

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

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

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

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

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

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

[16] A. C. e. al., „Power-Law Distributions in Empirical Data,“ 2009. [Online]. Available: https://doi.org/10.1137/070710111.

[17] M. Hörnle, „Hängt die Pfalz den Odenwald ab?,“ *Rhein-Neckar-Zeitung*, 2020.

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