Coursera Capstone Report

Coursera Course: Applied Data Science Capstone by IBM

Topic: Recommended cities in Germany to establish new fastfood restaurants

Michael M., 2019-04-06, V1.0

Contents

1. Introduction/Business Problem

An interntional fastfood chain plans to enter the German market. So far, they are not present in Germany.

They engage a data scientist to search for the most promising communities to build new fastfood restaurants.

In detail, they ask for an recommendation on the communities for 2 pilot restaurants.

For logistical reasons, restaurants should be in the same geographical region within Germany.

The decision for the restaurants should be based on

- demographical data,

- as well as analysis of already existing fastfood restaurants.

The analysis especially should look into in how far demographics and existing fastfood restaurants can be related to each other.

As a result, communities should be identfied that according to their demographic specifics currently are served with a number of fastfood restaurants below average.

2. Data

2.1 Demographical Data

The "German Federal Office for Statistics" (DESTATIS) provide on their website statistics on many demographic aspects:

https://www.destatis.de/EN/Home/_node.html

Unfortunatly, most statistics show information on a geographically aggregated level only. Only few statistics allow a drilldown down to the community-level (in German "Gemeinde").

Among those more detailed reports, one could be identified that for each individual community lists:

- Community-ID and Community Name
- area in square-kilometers
- population
- population per square-kilometer
- Geographical longitude and latitude
- and other attibutes that can be omitted for this project.

The report is current as of end 2018, and consequently should provide up to date information. Also, a high level of accuracy can be expected due to its official source.

This is the link to the report in excel-format:

https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/Archiv/GVAuszugQ/AuszugGV4QAktuell.xlsx?__blob=publicationFile&v=3

2.2 Fastfood Restaurants

Foursquare Labs Inc. is a technology enteprise that provides information on many different kinds of venues, mostly related to food, but considering other places of interest as well.
Users of Foursquare's web-app can provide their feedback on venues at any time, as well as suggest new venues. Presumably, this feedback does not provide a totally concise and accurate view on a community's venues, but still should be acceptable, also considering the lack of alternative free data sources.

Venues can be searched by category and geographical data (longitude/latitude and radius) via an API:

https://developer.foursquare.com/docs/api/venues/search

Longitude and Latitude for (the center of) communities will be derived from 2.1 above.
The demographics file above also contains the area in square-km's of a community. From that area the radius can be calculated. A certain level of inacurracy has to be accepted: The search based on radius will draw an exact circle around the provided longitude/latidude, while the actual community borders will be much more irregular. So, some areas might mistakenly in- or excluded.

From Foursquare's "category page", the id for "Fastfood Restaurants" can be retrieved:

https://developer.foursquare.com/docs/resources/categories

For requests made via API, results are provided as a json-file. Per request, a maximum of 50 venues are returned. This needs to be considered when analyzing results. Any resulting number of venues above 49 is unreliable since a higher amount of results just could have been capped by 50.

2.3 Combined data

Results from 2.1 and 2.2 will be joined based on longitude/latitude of the corresponding request to Foursquare. Conseqently, a dataframe will be created that contains for each fastfood-venue the demographical data as described under 2.2.

In a next step, data will be aggegated on Community-level. So, the number of fastfood restaurants per community (plus its demographics) is displayed. This information then will be analyzed regarding dependencies and patterns.

3. Data retrieval, cleaning and limitation

3.1 Demographics file

After importing the relevant demographics-excel-worksheet into a pandas-dataframe, it needs to be cleaned.

First header and footer lines are removed, and irrelevant columns are removed as well. Then meaningful column-names are being created, and data is converted into a numeric format where applicable. Also, a unique row-key is being created from different hierarchical key-elements.

Next, a new column is created, calculating the radius of each community from its surface. Text data in the column "community name" gets removed unnecessary descriptive elements.

The file originally has no clean row orientation. Higher geographical elements (above community, as e.g. state) are represented in a row on their own. Those rows do not have individual entries for demographical attributes. These rows will be excluded in the next step.

In a last step, the demographics-dataframe is limited to relevant communities by 3 criteria:

- Citizens per square km < 1,000. So, rural communities that may have a large surface but only few citizens per square km are ecluded.

- Number of citizens >= 40,000. This excludes villages and other small communities. So, remaining communities are all "Cities".

- Number of citizens <= 400,000. This limit excludes major capitals. Those capitals otherwise might significantly misshape average data. Each capital city would need an individual analysis instead.

3.2 Foursquare restaurants file

The Foursqare-API retrieves fastfood restaurants within defined geographical limits. A first data-scan reveils that 2 communities show 50 restaurants. Since 50 results are the limit for the developer-API, it might actually be more, and these communities would need further focus in case they be selected in later steps. It is about "Oberhausen" and "Mönchengladbach". Unfortunately, they can not easily be excluded via limits in step 3.1 above: Regarding number of citizens, they are with ca. 200,000 citizens just in the middle of applied criteria.

To anticipate already the later clustering effort on those 2 communities. They will be located in clusters 8 and 12. Also box-plot-diagrams will show those outliers. But that should matter only in case any of these clusters be selected for final analysis.

4. Analysis: Statistical correlation of restaurants vs. demographics

4.1. Question

First business question to be solved is: Does the number of fastfood restaurants per community relate statistically to any of the demographical data?

Based on that, it may be possible to draw a regression line that represents the relationship between number of restaurants and most relevant demographical attributes.

The result will allow to identify communities that significantly deviate from the expected average:

Be it that the number of restaurants in the specific community is much higher than expected for the corresponding demographics. In this case the community should not be selected for a new restaurant)

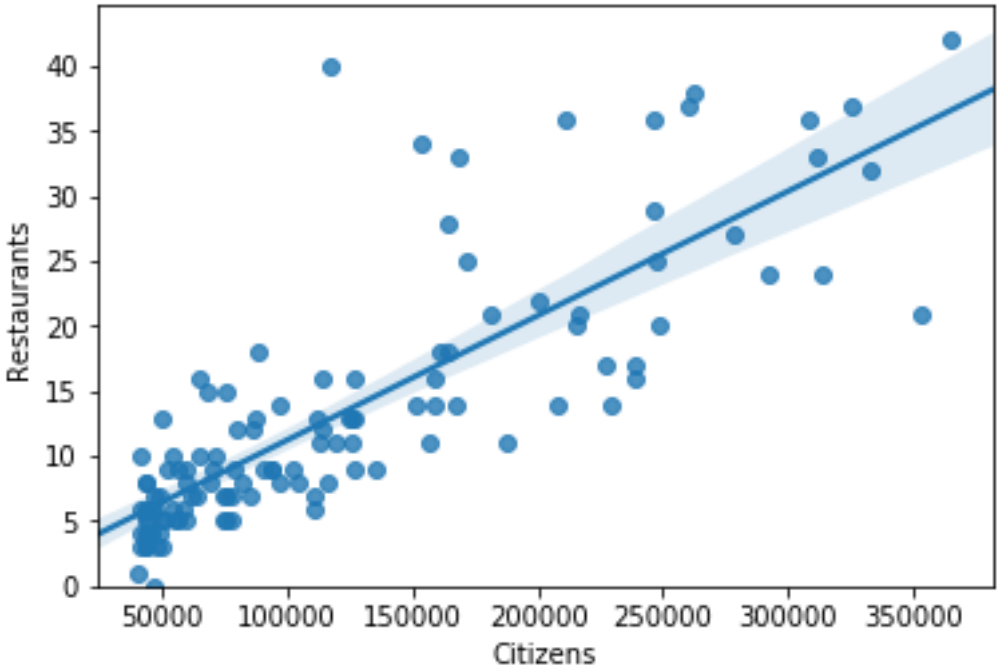
Or be it that that the number of restaurants is much lower than expected according to demographics. In this case the community would be eligible for building a new restaurant.

4.2 Approach/Results

A correlation analysis of shows a pearson coefficient of 0.82 for "number of citizens"related to "number of restaurants, and thereby suggests a high positive linear relationship between both.

The regression line visualizes the relationship of citizens to restaurants, roughly suggesting 1 restaurant per 10,000 citizens.

Fig. 1 Regression line: Citizens vs. Restaurants



In fact, also causational correlationship can be assumed, stating that the more inhabitants are in a community, the more potential restaurant visitors exist, and based on that demand the more restaurants are set up.

(Ideally, the correlationship would have been investigated for subgroups of citizens, esp. grouped by age, but that would have been a too big effort.)

Conseqently, communities with a low "restaurants to citizens ratio" are the most eligible for setup of new restaurants.

So, a new parameter "restaurants to citizens ratio" is calculated and added to the data-set:

("Number of restaurants" + 1) * 1,000 / "number of citizens".

In the calcuation above, "+1" indicates the number of restaurants after a potential new one has been added. A result below 1 should roughly indicate a number of restaurants below average related to number of citizens.

5. Analysis : Geographical clusters

5.1 Question

According to the overall business scenario two new restaurants within the same geographical area shall be established as pilot venues. Conseqently, communities have to be clustered geographically into groups. Official segmentation elements as districts, counties or states may not provide the best solution, since they don't necessarily correspond with proximity of communities selected for this work to each other.

Instead, a "k-means" machine learning algorithm will be applied to group communities according to their geographical position.

5.2 Approach/Results

Latitude and longitude of communities are provided as input parameters to the k-means-algorithm.

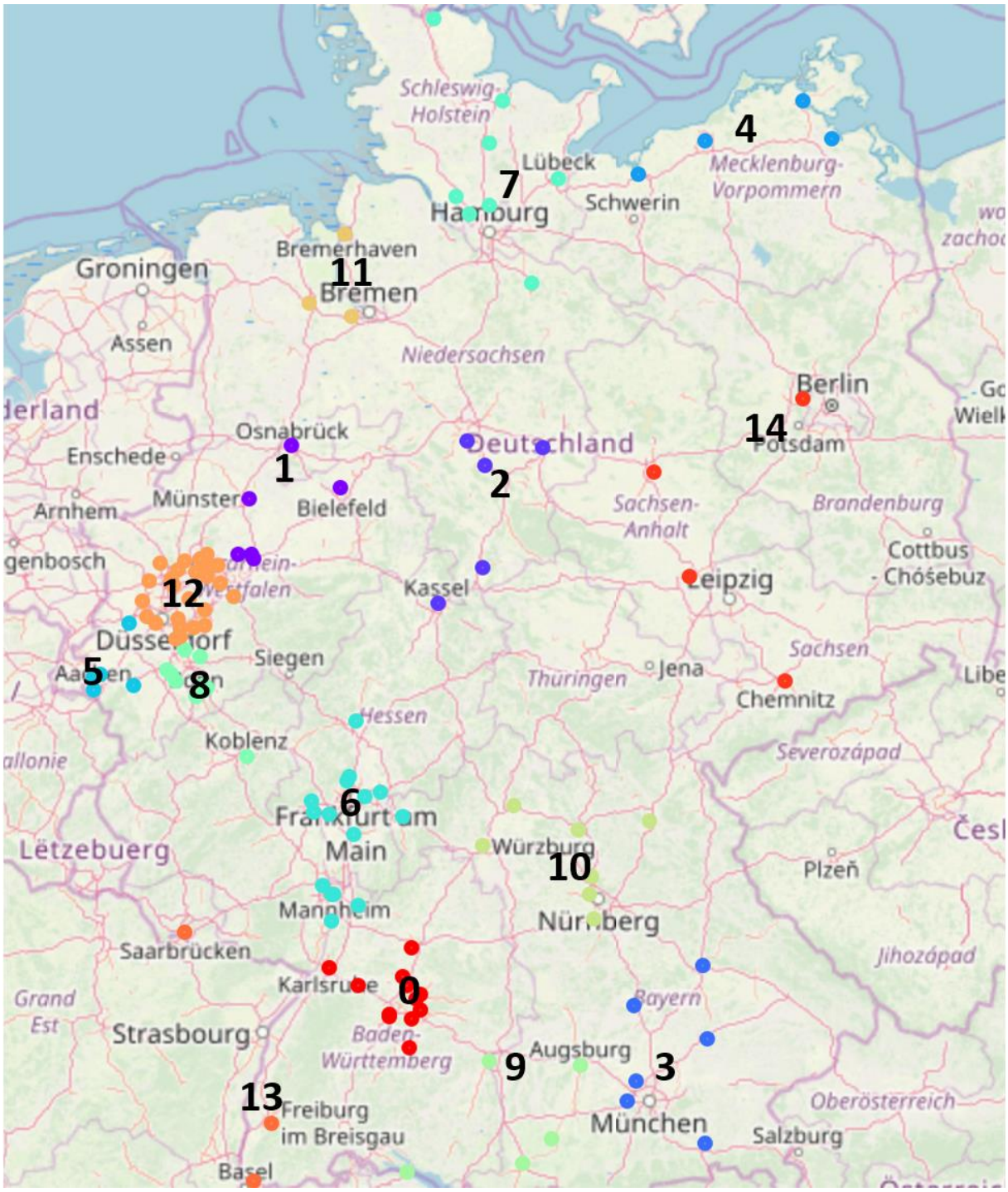
A visual representation on a map allows an evaluation of the quality of the resulting clusters.

K = 10 seems not to provide a granular enough segmentation, while K = 20 seems to cut off even clusters that should be seen as one, and otherwise it does not provide a better segmentation than K

As can be seen on the final map (K = 15), there are clusters that consist of a high number of communities in scope (40,000 < Citizens < 400,000), usually in metro regions around major cities and economic centers, and on the other hand side large areas with few communities in scope only.

A cluster with a high number of communities within a relatively narrow area will best meet the requirement to select 2 communities within the same region.

Fig 2 Clustering of communities related to geographical proximity to each other (k-means)



From a geographical perspective, the majority of clusters may be eligible. Just clusters 4, 7, 13 and 14 should be exluded because they spread only few cities around a very wide area.

6. Analysis: Selection of cluster based on cluster-parameters

6.1 Question

Now, the question is which cluster should be selected among those that from a geographical point of view seem to be eligible.

Criteria that need to be considered are:

Number of communities in a cluster: The higher the number the more options exist to choose 2 of them.

Number of restaurants in each community per cluster: so the variety of communities related to their average data becomes visible.

Restaurants-per-citizens in a community ratio related to cluster: Here, compared to the previous paragraph, the number of restaurants is related to the number of citizens in a community. This seems to be the most relevant, final view.

6.2 Approach/Results

The following table shows an aggregation of raw-data by cluster. A detail view on variety within the clusters provide the 2 box-plots underneath.

Fig. 3 Demographics by cluster

	Citizens	Restaurants	Count_Communities	Rest_to_Cit_Ratio_TOTAL
Cluster_Labels				
0	1168293.0	135.0	12	1.155532
1	989054.0	112.0	6	1.132395
2	711481.0	68.0	5	0.955753
3	507951.0	44.0	6	0.866225
4	369718.0	33.0	4	0.892572
5	692315.0	87.0	5	1.256653
6	1933395.0	229.0	15	1.184445
7	878759.0	98.0	8	1.115209
8	1042725.0	134.0	10	1.285094
9	614696.0	56.0	5	0.911019
10	609555.0	56.0	7	0.918703
11	357628.0	30.0	3	0.838860
12	3391100.0	395.0	25	1.164814
13	459815.0	40.0	3	0.869915
14	768058.0	65.0	4	0.846290

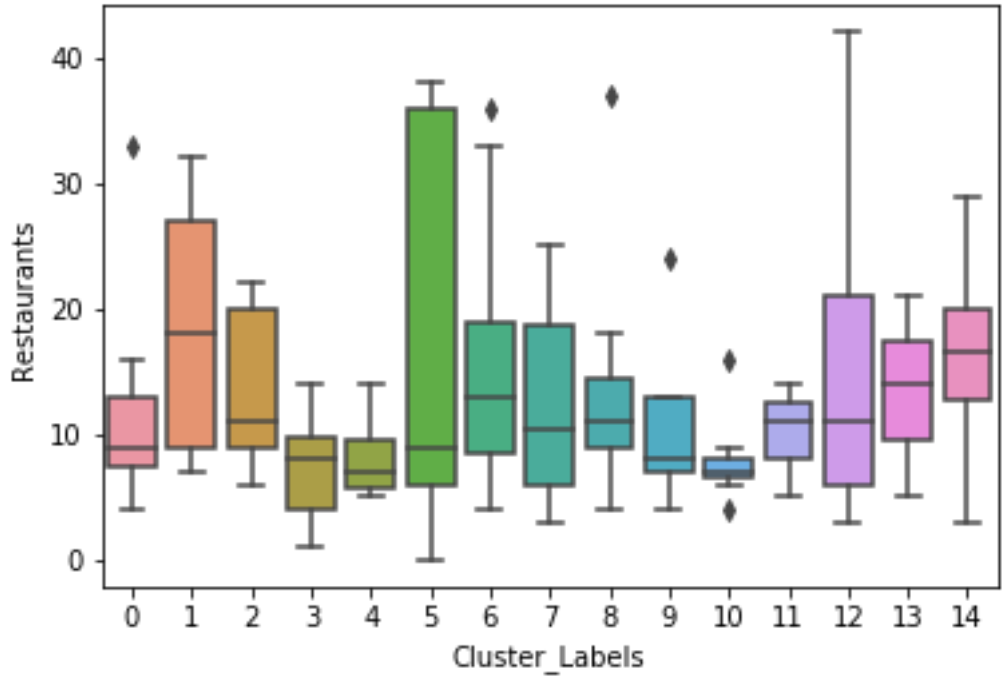
As shown in Fig.3 above, clusters with lowest overall "restaurants-to-citizens ratio" are:

Cluster	Total Restaurants to citizens ratio
11	0.838860
14	0.846290
3	0.866225
13	0.869915
4	0.892572

From a geographical perspective clusters 13 and 14 already have been exlcuded.

So, 3, 4 and 11 are remaining. The following 2 box-plots allow to analyze all clusters in more detail.

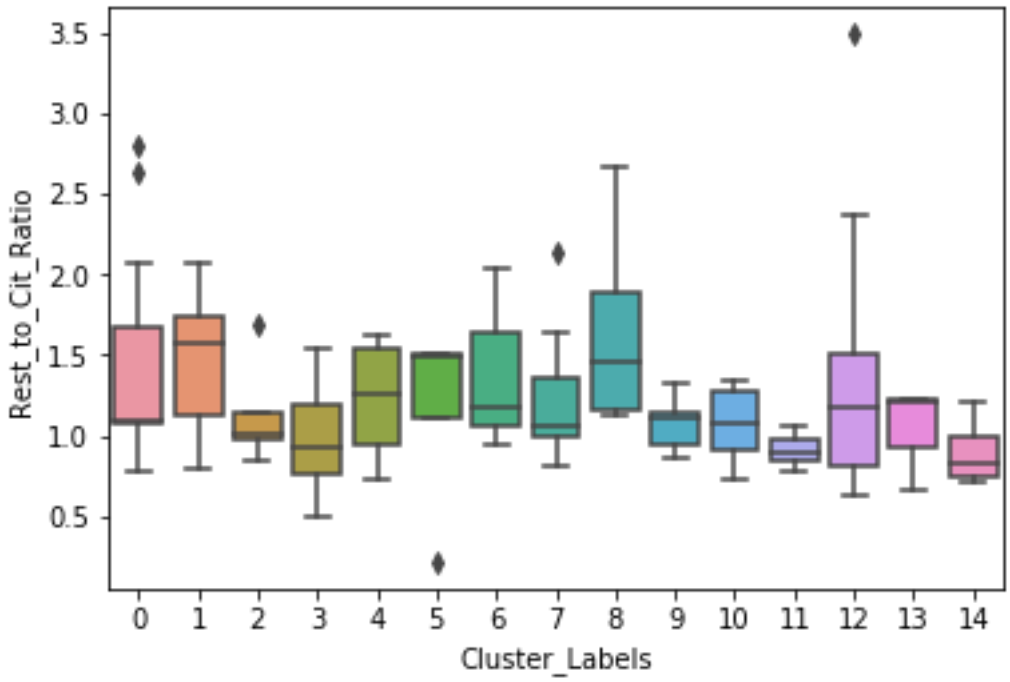
Fig. 4 Restaurants in a community by cluster



Regarding numbers of restaurants in a community per cluster, clusters 3,4, and 11 all show a relatively similar picture. Quartiles have a relatively low expansion around the median. So, all communities are relatively similar to each other.

in the next box-plot (fig. 5), number of restaurants in a community is related to the number of citizens in that community. Here, cluster 4 shows a slightly higher value then 3 and 11. So, 4 might be excluded. Cluster 11 has a very low spread. But when going back to figure 3, for further details, cluster 11 reveals to consist of only 3 community. This is fewer as cluster 3, which is made up out of 6 communities. Consequently, cluster 3 provides more options for the final selection of 2 communities within one cluster.

Fig. 5 Restaurants-to-Citizens ratio by cluster



7. Analysis: Best communities in recommended cluster

7.1 Question

Now, after cluster 3 has been selected, the question is: Which two communities within that cluster should be chosen.

7.2 Approach/Results

Again, the "restaurants to citizens ratio" per community is analyzed. Data is provided in table figure 6. And visualized is that data on the map in figure 7.

Best ratio show city of Germering, and Ingolstadt. Germering has relatively few citizens, but that is made up because just one fastfood-restaurant exists in this city, resulting in a low ratio.

Ingolstadt has more restaurants, but also many more citizens, also resulting in a good ratio.

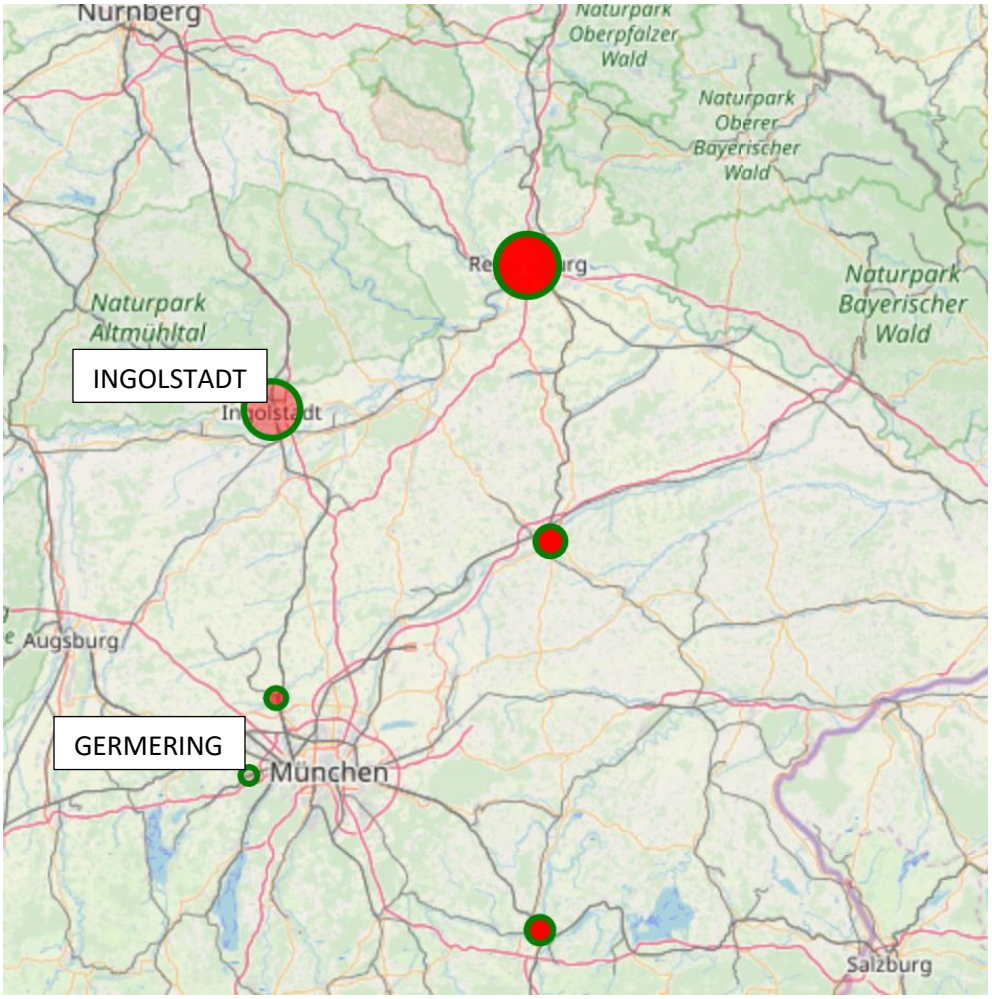
Fig. 6 Details of selected cluster

Community	Square-km	Citizens	Citizens-per-Square-km	Radius-m	Restaurants	Rest_to_Cit_Ratio
Germering	21.60	40285.0	1865.0	2622.0	1.0	0.496463
Ingolstadt	133.35	135244.0	1014.0	6515.0	9.0	0.739404
Dachau	34.96	47255.0	1352.0	3336.0	3.0	0.846471
Regensburg	80.85	150894.0	1866.0	5073.0	14.0	0.994075
Rosenheim	37.22	63080.0	1695.0	3442.0	7.0	1.268231
Landshut	65.83	71193.0	1081.0	4578.0	10.0	1.545096

Relevant data of the table above also can be visualized on the following map.

On the map, size of circle denominates relative community size (number of citizens). And opacity-level shows the relative "restaurants to citizens ratio" where the most transparent communities have the fewest relative number of restaurants, and the most opaque ones the highest number.

Fig. 7 Geographical visualization of selected cluster



8. Conclusion

In a multi-step approach, 2 recommended communities for the establishement of new fastfood restaurants could be identified: Germering and Ingolstadt, both in the state of Bavaria.

To recap steps that led to the result:

In a first step, a linear regression line could be drawn to prove that number of restaurants and number of citizens in a city are usually linearilly related to each other.

Then clusters have been created via k-means-clustering go group communities geographically.

Then, clusters were analyzed regarding overall adequacy, defined by number of restaurants related to citizens below average.

In a last step, cities within the most adequat cluster were investigated.

Fig. 8 Selected city "Ingolstadt"

https://commons.wikimedia.org/wiki/File:Ingolstadt_Panorama_Nord.jpg

Fig. 9 Selected city "Germering"

http://www.total-lokal.de/city/germering/data/82110_50_03_17/index.html

9. Discussion and outlook

Of course, the analysis provided can serve as a first step towards a final recommendation only. Other aspects, that have not been investigated but might matter as well are:

Consumer's buying power in a region: A low coverage with fastfood restaurants might be related to a relative low income and conseqently low demand for restaurants.

Also the opposite might apply: So, low income may correspond to high demand of cheap restaurants. In order to gain a clear picture, fastfood restaurants would need to be put in correlation to other types of restaurants in a region.

Proximity of communities to each other may matter on other aspects than investigated here. While it oviously makes sense to exclude rural communities with a low number citizens, I also excluded major cities (> 400,000 citizens), while existence of those cities usually is related to narrow clusters with many "satellite communities".

Communities that legally are distinct from each other, may be seen from a consumer's perspective as one big entity.

Nevertheless, a first big step has been made: A recommendation for the founding of fastfood restaurants in a defined georaphical area - based on few but meaningful criteria - could be made.