Culinary Clustering

Tobias Splith

E-mail: `tobias.splith@web.de`

Abstract. The location of venues in cities is all but arbitrary. Especially culinary venues like restaurants, cafés and ice cream shops tend to form clusters. One might think that the high venue density in these clusters results in a high competition between the venues. But it might also be possible that the venues profit from each others proximity. A customer who has eaten in a restaurant might want to drink a coffee in the cafe next door. Or a person who has eaten in an Asian restaurant on Monday might visit the French restaurant across the street on Thursday. This work will analyze the composition of clusters formed by culinary venues. It will focus on venues recommended by the foursquare api in the city of Berlin.

Date: 10.09.2020

1. Introduction

In cities it can be often observed that specific venues are close to each other and form spatial clusters. E.g. restaurants and bars can often be found in the downtown- or in the university-area of a city, museums and parks might also be clustered in the city center, commodity stores are clustered in malls etc. This work will focus on the city of Berlin and on ”culinary” venues, i.e. restaurants, cafés, bakeries, bars and so on. These venues will be clustered based on their location and the composition of these spatial clusters will be analyzed. From this I will try to find out if specific venues may profit from being in the direct neighborhood of other specific venues. E.g. it might be possible that a pizza store tends to be more successful if it is in the direct neighborhood of a bar and can, therefore, more often be found in a cluster with one or more bars. On the other hand it might be possible, that a certain venue does not work well in a specific cluster. These would be valuable information for an entrepreneur who is interested in opening a new venue or for a city that is interested in developing a specific area.

2. Data

This work will use three data-sets:

The free version of the simplemaps world cities database (https://simplemaps.com/data/world-cities). This database contains all prominent cities of the world, their population and their location.

The categories endpoint of the foursquare api. This endpoint returns a .json file that contains all the categories available in the foursquare api. The data is organized in a tree-like structure, where e.g. the food venue category contains the categories ”Afghan Restaurant”, ”African Restaurant”, ”American Restaurant”, ”Asian Restaurant” and many more. The category ”Asian Restaurant” again contains many Categories like ”Burmese Restaurant”, ”Cambodian Restaurant”, ”Chinese Restaurant” , etc. and the category ”Chinese Restaurant” again contains a large number of even more specific categories. An overview of the categories of the foursquare api can be found here: https://developer.foursquare.com/docs/build-with-foursquare/categories/

The explore endpoint of the foursquare api. This endpoint returns 100 recommended venues in a given radius around a specific location. An algorithm that returns all recommended venues of a city from foursquare is described in the Data-Notebook of this work: https://nbviewer.jupyter.org/github/TobiSpl/Coursera_Capstone/blob/master/Data.ipynb

3. Methods

This work will be composed of 4 Steps:

- obtain information of successful venues from multiple large cities in Europe.
- compare the composition of culinary venues in these cities

- cluster the spatial data for the culinary venues in Berlin with a density-based spatial clustering algorithm (DBSCAN)
- analyze the composition of the clusters and obtain trends with a principal-component analysis and k-means clustering.

A detailed description of the data analysis can be found in the Data Analysis-Notebook (https://nbviewer.jupyter.org/github/TobiSpl/Coursera_Capstone/blob/master/Capstone.ipynb)

3.1. Parameter Selection for DBSCAN

In order to find spatial clusters of culinary venues in Berlin the density based clustering algorithm DBSCAN is used. This algorithm has two mandatory parameters: The first one ϵ or `eps` describes the radius in which a specific number of points must be found in order to produce a cluster. The second ”min_samples” is the minimum number of points we need to produce a cluster.

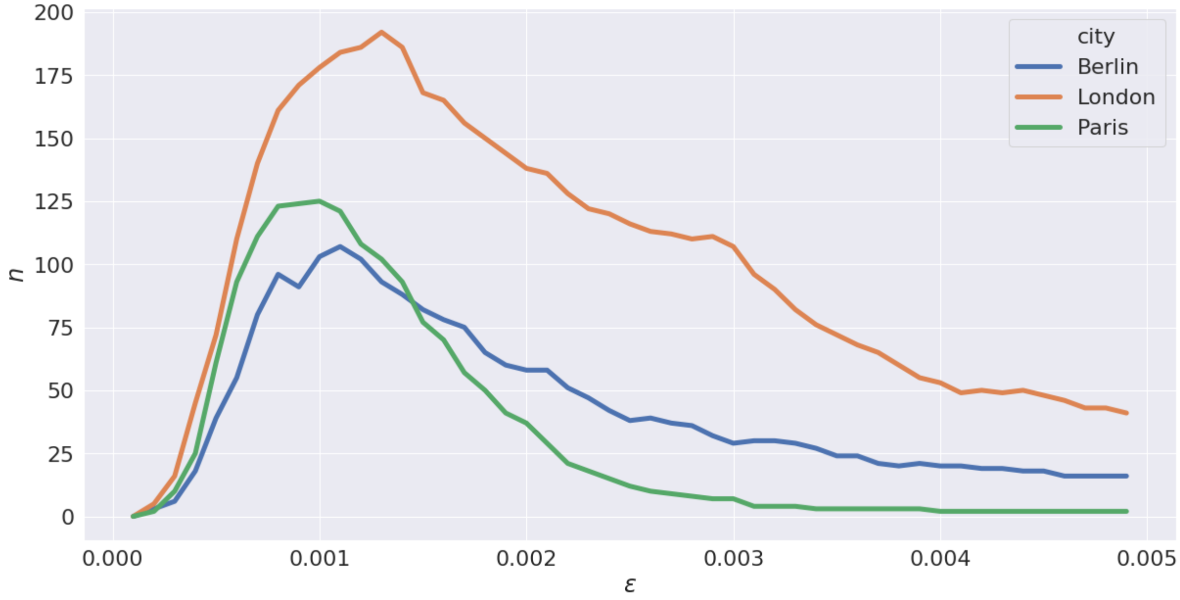

Figure 1. Number of Clusters obtained with DBSCAN as a function of the ϵ-parameter.

Since we like to analyze the composition of the clusters we will avoid clusters that are composed of fewer than 7 venues. For a small eps-value only a small number of clusters will be found, since the density needed to produce a new cluster is quite high. For large eps-values the clusters tend to merge together, reducing again the total number of clusters. In figure 1 the number of clusters obtained with DBSCAN are shown for three cities as a function of the ϵ-parameter. For Berlin the maximum number of spatial clusters is obtained with $\epsilon = 0.0011$ which corresponds to a radius of $r \approx 120\,\mathsf{m}$.

3.2. Principal-Component Analysis

“PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on” wikipedia. It can show which features in a dataset are the most important to describe the whole dataset.

Due to one-hot encoding procedure of the 369 culinary categories of the foursquare api we have a 369-dimensional dataset. With the help of PCA the features in this dataset with the highest variance can be identified. PCA would yield 369 new features, which are perpendicular linear combinations of the original features and are ordered with descending importance. Each of this new features n_i is calculated from the original features o_j with the equation

$$n_i = c_{i,1}o_1 + c_{i,2}o_2 + ...c_{i,j}o_j + ... ,$$

where $c_{i,j}$ are the coefficients that transform the old features into the PCA-features.

We are only interested in the 10 original features that have the highest influence to the most important 30 new features. We pick the first 30 components of our pca and identify the 10 original features that have the highest absolute coefficients $c_{i,j}$ for each of these categories. Furthermore, every category where the contribution to the variance multiplied with the coefficient $c_{i,j}$ is below 0.5% is dropped. This way the following 31 important categories were obtained:

Asian Restaurant, Bakery, Bar, Bistro, Breakfast Spot, Burger Joint, Café, Cocktail Bar, Coffee Shop, Dessert Shop, Doner Restaurant, French Restaurant, German Restaurant, Hotel Bar, Ice Cream Shop, Indian Restaurant, Italian Restaurant, Korean Restaurant, Middle Eastern Restaurant, Nightclub, Pizza Place, Pub, Restaurant, Seafood Restaurant, Steakhouse, Sushi Restaurant, Trattoria/Osteria, Turkish Restaurant, Vegetarian / Vegan Restaurant, Vietnamese Restaurant, and Wine Bar.

3.3. Parameter Selection for Kmeans

The Kmeans-clustering algorithm has on mandatory parameter which is the number of clusters. On the one hand clusters with very high numbers of items (here the items are the spatial clusters) are not favorable because the significant information will be overlaid with a large amount of noise in these clusters. On the other hand if we choose a to large the number of clusters a lot of clusters will only have 1 or two items and show now significant trend, i.e. a large the number of clusters overfits the data. Note, that it is okay to have some clusters with very few items, since these clusters contain the ”exotic” compositions (compositions that are unusual) and will reduce the noise in the other clusters. In table 1 the size of the clusters is shown for different values of the number of clusters. We compromise on a value of 16 clusters where we have no cluster with more than 20 items and only 6 clusters with less then 5 items.

number of clusters	sizes of clusters
4	49, 19, 11, 28
5	3, 24, 27, 12, 41
6	16, 5, 47, 1, 27, 11
7	24, 1, 11, 3, 37, 1, 30
8	10, 30, 15, 24, 4, 11, 1, 12
9	11, 19, 13, 25, 3, 11, 1, 10, 14
10	10, 34, 4, 11, 2, 1, 10, 9, 3, 23
11	17, 16, 3, 5, 10, 19, 3, 1, 18, 12, 3
12	10, 33, 4, 11, 1, 1, 10, 9, 3, 23, 1, 1
13	10, 34, 3, 11, 1, 1, 10, 9, 3, 21, 1, 1, 2
14	9, 11, 1, 2, 21, 3, 13, 17, 6, 1, 1, 5, 6, 11
15	10, 33, 3, 11, 1, 1, 12, 8, 3, 19, 1, 1, 2, 1, 1
16	5, 16, 16, 2, 6, 2, 12, 12, 5, 1, 3, 14, 1, 1, 6, 5
17	5, 16, 16, 2, 6, 2, 10, 13, 5, 1, 3, 13, 1, 1, 6, 6, 1
18	5, 16, 16, 2, 6, 2, 8, 13, 2, 1, 3, 14, 1, 1, 6, 5, 1, 5
19	2, 14, 1, 2, 1, 8, 17, 11, 1, 5, 14, 6, 1, 10, 2, 9, 1, 1, 1

Table 1. Influence of the `n_clusters`-parameter on the size of the clusters obtained with kmeans-clustering in Berlin.

4. Data Analysis and Results

4.1. Category Composition of Cities

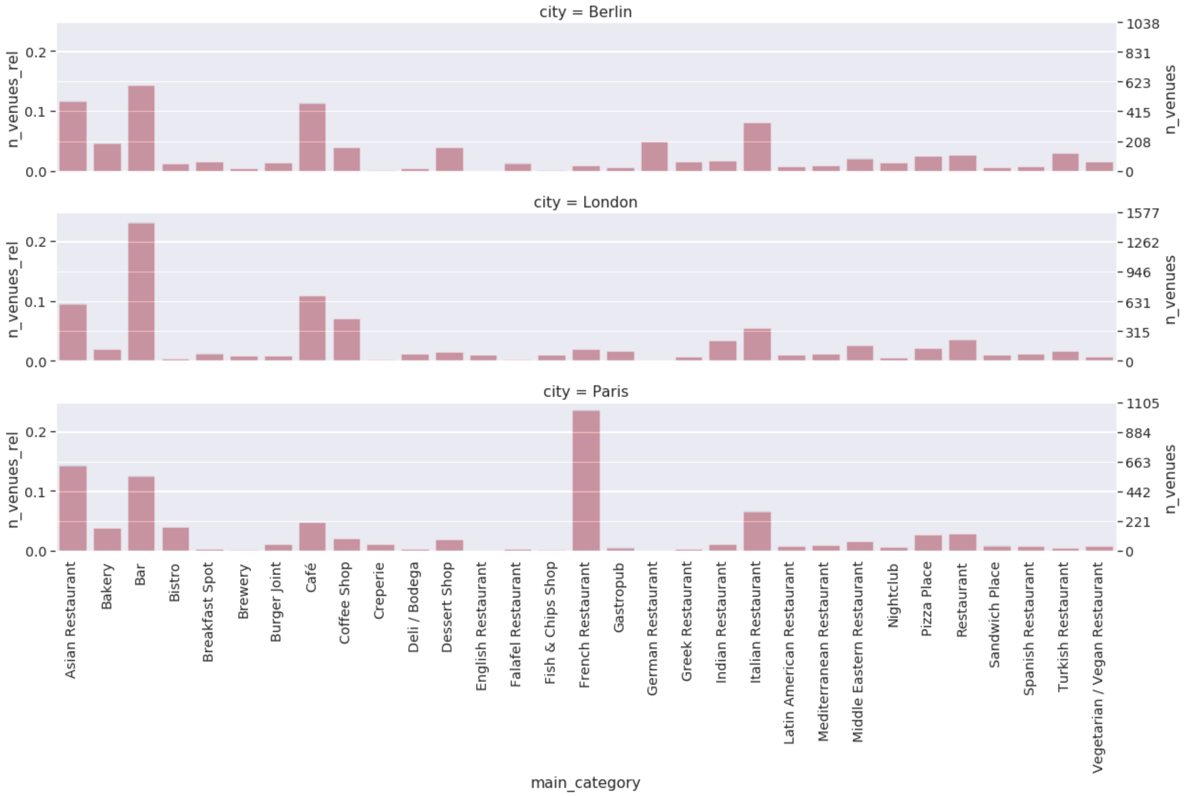

Figure 2. Category composition of the cities Berlin, London and Paris

In figure 2 the categorical composition of Berlin, London and Paris are shown. As we can see, the composition of the cities in some categories is quite unique. For example, the relative number of French Restaurants in Paris and the number of Bars in London are significant above the relative number of similar venues in the other cities. Surprisingly, the relative number of Cafés in Paris is below the average of the other cities.

The relative number of venues in other categories, like Asian Restaurants, Italian Restaurants, Pizza Places, etc. seem to be rather similar. For each city we obtained a unique composition or ”culinary fingerprint”.

4.2. Spatial Clustering

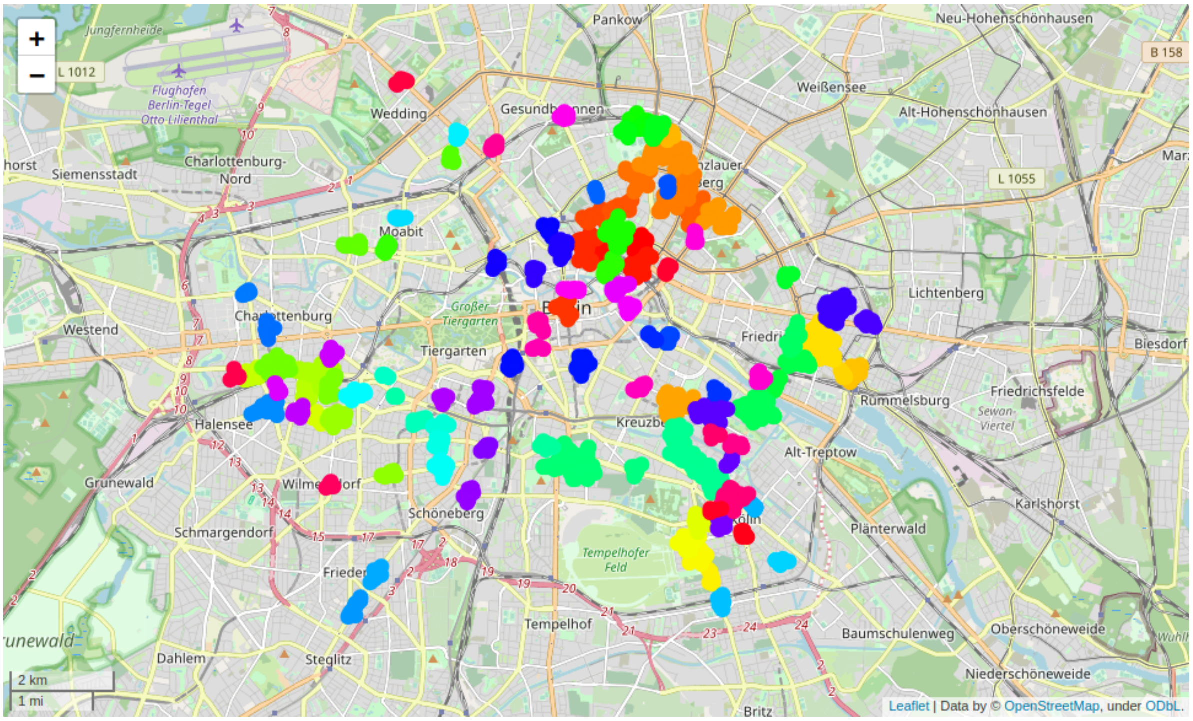

Figure 3. Spatial clusters obtained with the DBSCAN algorithm in Berlin

In figure 3 the venues of the spatial clusters are shown on the map of Berlin. The venues are colored according to their cluster label. The DBSCAN algorithm returned 107 spatial clusters for Berlin.

Of course a lot of recommended venues in Berlin are not included in these spatial clusters. Figure 4 shows the difference between the composition of the unclustered data and the average composition of Berlin shown in figure 2. For Asian Restaurants, Bars, Coffee Shops and Vegetarian/Vegan Restaurants the difference between their relative number in the unclustered data and their relative average number is significantly smaller than zero. Therefore, these venues are more often located in the clusters. On the other hand, the same difference is significantly larger than zero for bakeries,

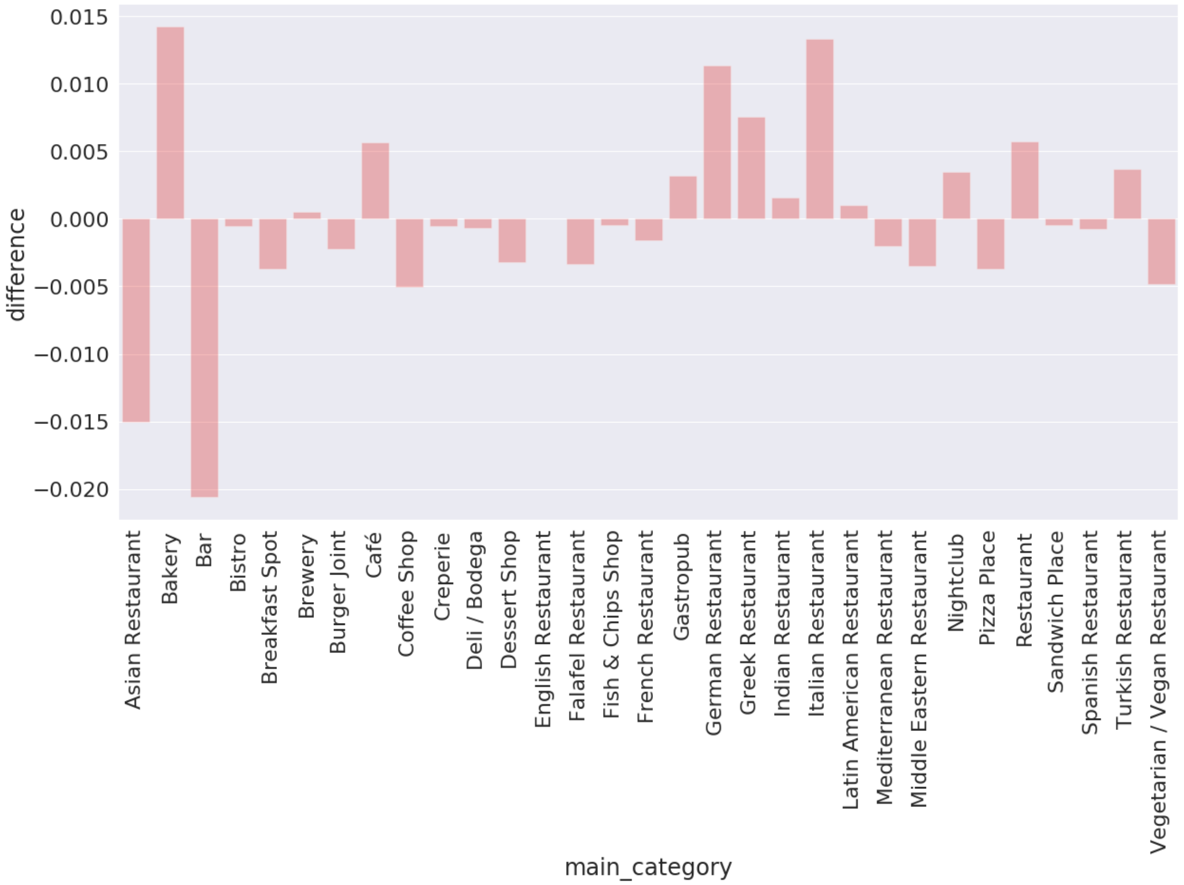

Figure 4. Difference between the composition of the venues not in spatial clusters and the mean category composition of Berlin.

cafés, German restaurants, Greek restaurants, Italian restaurants and venues with the unspecific 'Restaurant' label. These venues are more often outside of clusters. For the bakeries these findings are not surprising, since bakeries can often be found in residential areas, where no venue clusters are to be expected. The inverse could be assumed for bars, i.e. bars tend to be outside of residential areas and, therefore, more often clustered. The difference between Asian restaurants and German, Greek and Italian restaurants in the distribution is an interesting find of this study. It shows that Asian restaurants in Berlin tend to be part of spatial culinary clusters while the other restaurants tend to be more evenly distributed over the City.

4.3. Kmeans-Clustering

Kmeans-Clustering is used in order to find patterns in the composition of the spatial clusters. The difference between the composition of a cluster and the average composition of all clustered data for the 31 most important categories identified with PCA will be analysed for the largest five Kmeans-clusters.

In figure 5 the composition difference for the Kmeans-cluster 7 is shown. This

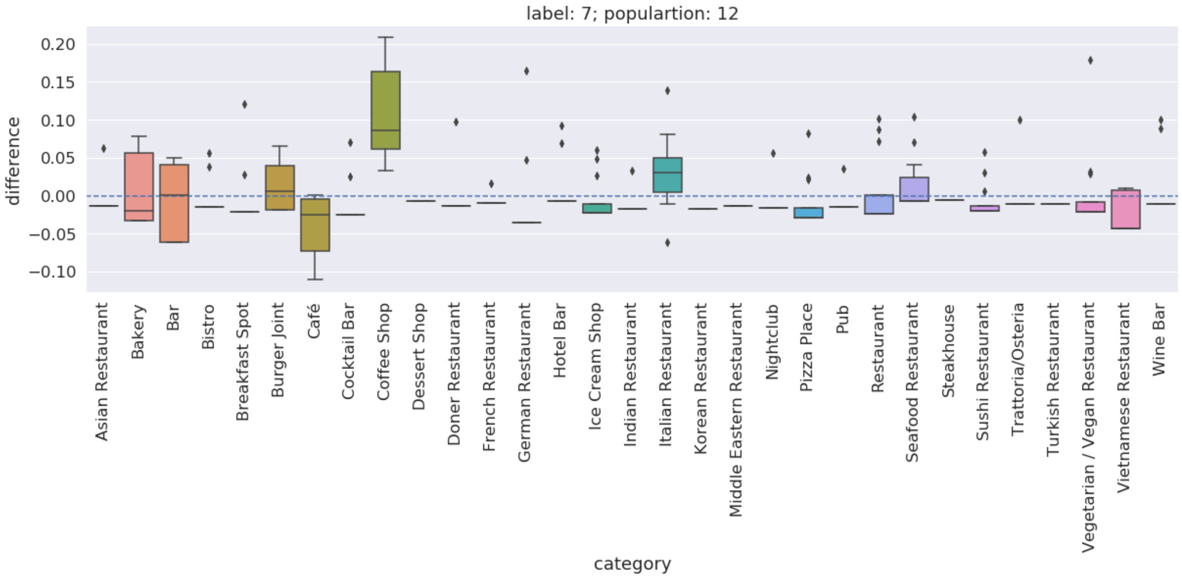

Figure 5. Difference between the composition of the kmeans-cluster with the cluster-label ”7” and the mean category composition of spatial clusters in Berlin.

cluster includes 12 spatial clusters and is characterized by a large number of coffee shops and Italian restaurants, while the number of cafes and bakeries is below average.

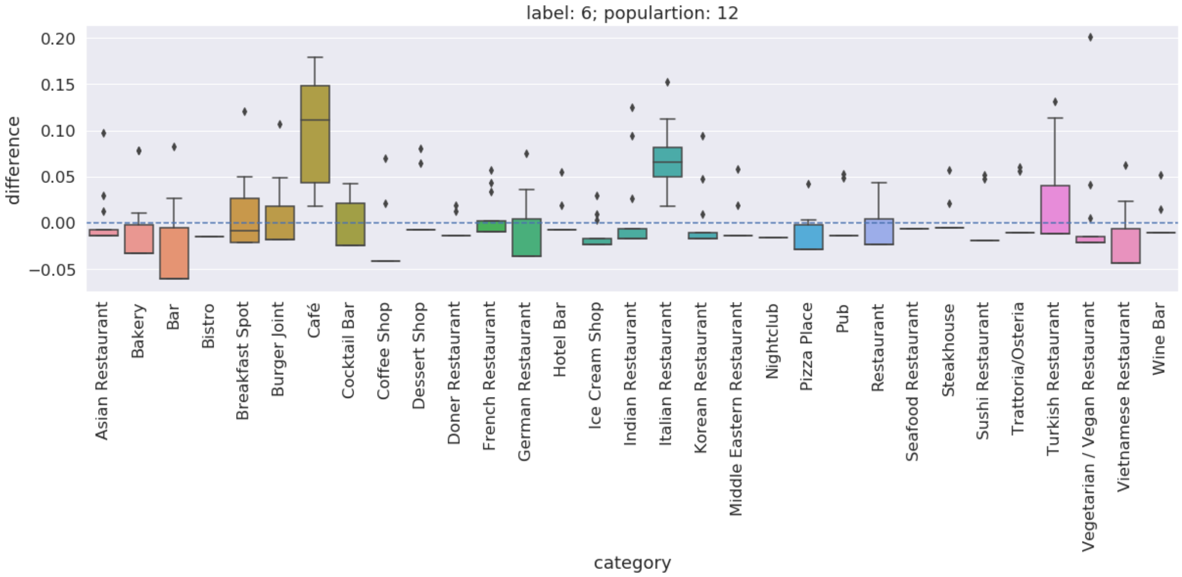

Figure 6. Difference between the composition of the kmeans-cluster with the cluster-label ”6” and the mean category composition of spatial clusters in Berlin.

Figure 6 shows the composition difference for the Kmeans-cluster 6. This kmeans-cluster consists of 12 spatial clusters. We can see, that the number of cafés and Italian restaurants is significantly above average, while the number of bars and Vietnamese restaurants is significantly below average.

Figure 7 shows the composition difference for the Kmeans-cluster 2, which is

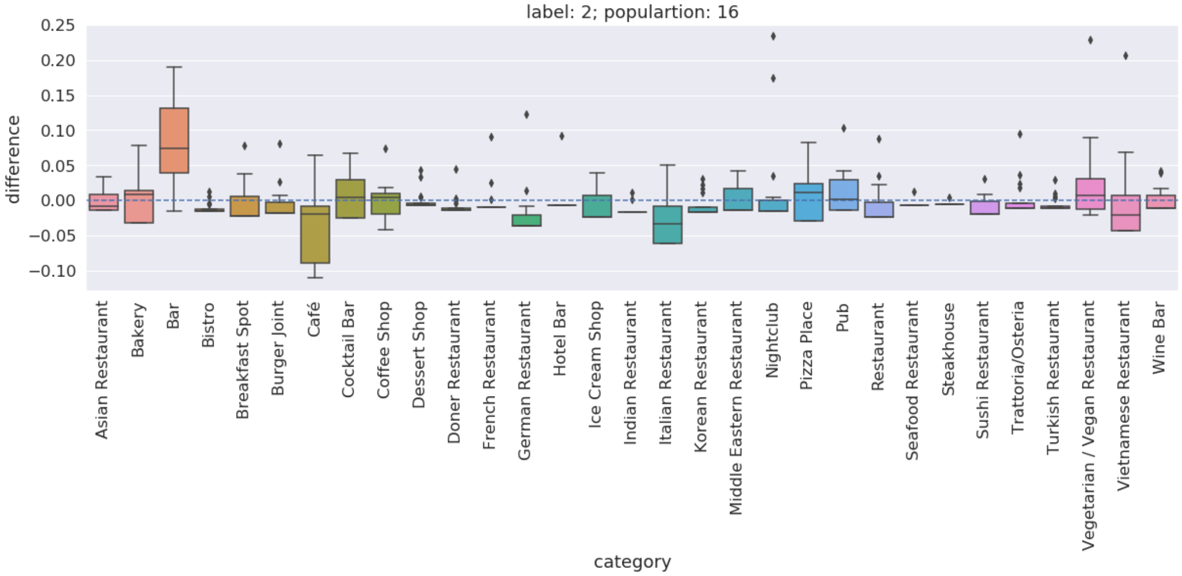

Figure 7. Difference between the composition of the kmeans-cluster with the cluster-label ”2” and the mean category composition of spatial clusters in Berlin.

composed of 16 spatial clusters. This cluster has an above average number of bars and pizza places and a below average number of cafés and Italian restaurants. The large outliers in the ”Nightclub” categorie are indicating that a parts of this cluster might be focused on nightlife venues.

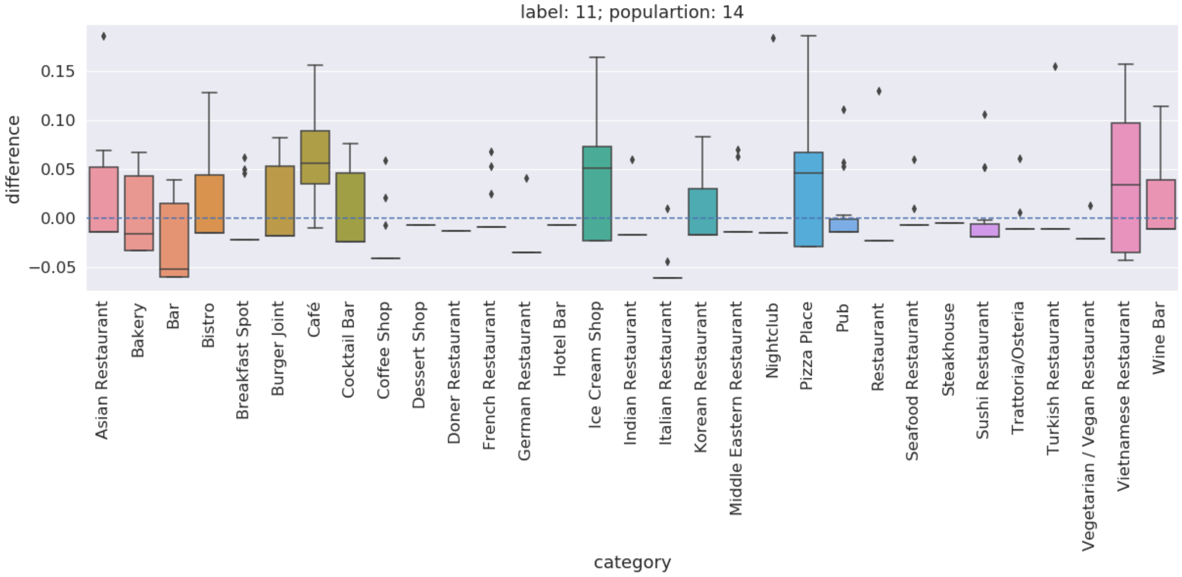

Figure 8. Difference between the composition of the kmeans-cluster with the cluster-label ”11” and the mean category composition of spatial clusters in Berlin.

In figure 8 the composition difference for the Kmeans-cluster 11 is shown. This kmeans-cluster of 14 spatial clusters is characterized by a above average number of cafes, ice cream shops, pizza places and Vietnamese restaurants, and a below average number of bars and Italian restaurants. This could be a promising composition for a successful cluster.

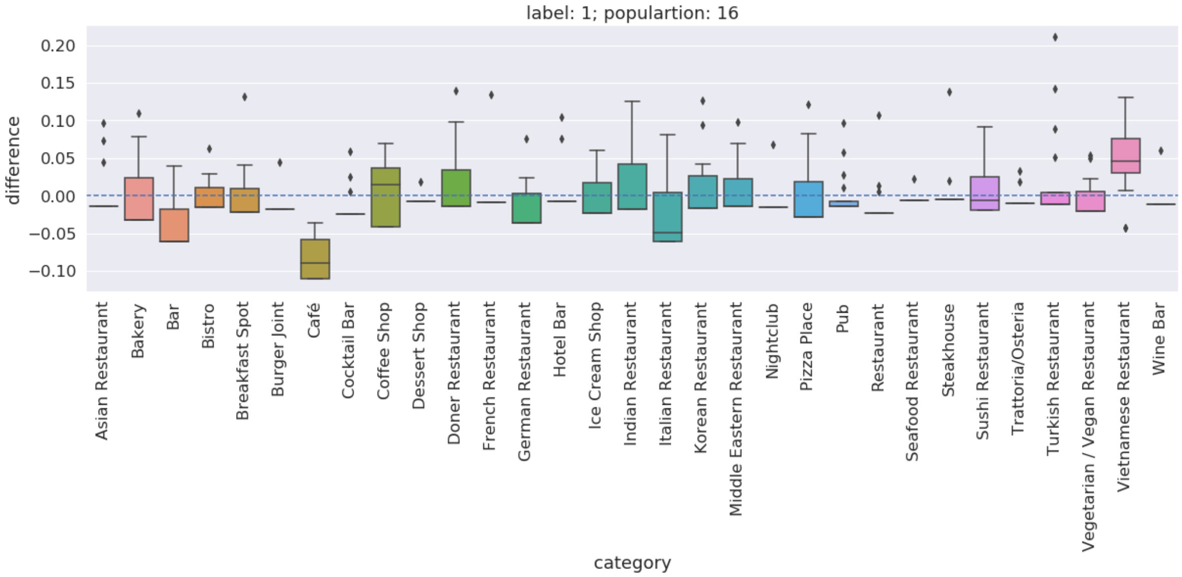

Figure 9. Difference between the composition of the kmeans-cluster with the cluster-label ”1” and the mean category composition of spatial clusters in Berlin.

Figure 9 shows the composition difference for the Kmeans-cluster 1, which is composed of 16 spatial clusters. This kmeans-cluster shows a high variance in many categories. Only the low number of bars and cafés and the high number of Vietnamese restaurants seem to be significant, indicating, that the later might not be a good combination with the former.

In figure 10 the positions of the spatial clusters which are members of the 5 largest Kmeans-clusters are shown. While there are the most ”2”-clusters (nightlife) in the south-east of Berlin, the ”7”-clusters seem to be centered around the north-east of town. All other clusters are more or less evenly distributed.

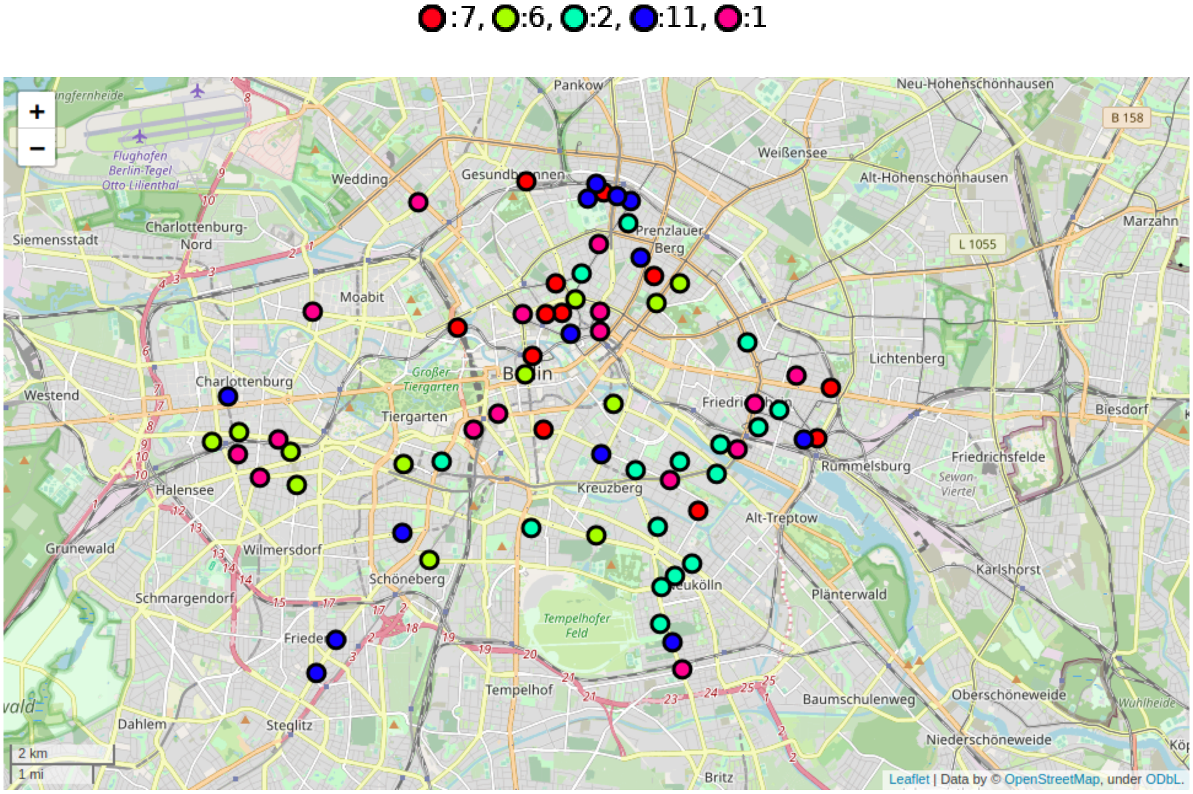

Figure 10. Position of the spatial clusters which are members of the five largest Kmeans-clusters.

5. Conclusion

In this work all venues recommended by foursquare in the cities of Berlin, Paris and London were obtained. From this data the culinary fingerprints i.e. the venue category composition of these cities were identified.

The data obtained for Berlin was studied further: At first we identified spatial venue clusters with the location data of the venues. The venues that were not included in any clusters were studied and the category composition of these venue group was compared with the average composition of Berlin. It can be clearly seen, that Asian restaurants, bars, coffee shops and vegetarian/vegan Restaurants can more often be found in spatial clusters, while bakeries, cafés, German restaurants, Greek restaurants, Italian restaurants and venues with the unspecific “Restaurant” label are more evenly distributed in the city of Berlin.

The composition of each spatial cluster was analyzed further. This was done with principal-component analysis (pca) and kmeans-clustering. The pca yields 31 categories that have the highest contribution to the variance of the data-set. The kmeans-clustering was applied and the composition of the resulting clusters of the spatial clusters was compared in these 31 categories. While the data is quite noisy we can clearly see that some categories are more often found next to each other then other categories. E.g. the density of restaurants seems to be significantly lower in areas with a high density of bars and vice versa. Cafés on the other hand seem to be present in all combinations of categories. Multiple restaurant categories have a above average occurrence in a single kmeans-cluster, indicating that these restaurants might benefit from each others proximity.