Do Restaurants Match Neighbourhoods? Evidence from Yelp

Student:	Niels Wich
Advisors:	Dr. Tobias Böhm, Prof. Dr. Nadine Riedel
Chair:	Institute for Public and Regional Economics
Module:	Project Studies
Term:	Winter 2020/ 2021

Contents

List of Tables

List of Figures

1 Introduction

Over the past decades the digitization of society, and in particular the expansion of the internet, made new data sources available for researchers in the field of regional and urban economics. The merits of these 'big data' sources, in comparison to more traditional data sources like surveys or administrative data, lie in their exhaustive nature and their often more granular structure with respect to time and space. Glaeser et al. (2018b) provide an extensive review of the promises and limitations of those relative newly available data sources for research on urban areas.

One example for the usage of internet data in the literature is Yelp, a review platform for local businesses, around which a small literature has emerged recently. Even though Yelp operates in many countries around the globe the literature has been limited to applications to the US so far. In this paper, I explore the potential of crowdsourced data from Yelp for economic research regarding Germany. To do so, I first compile a dataset based on restaurant listings from Yelp for the city of Berlin and evaluate its suitability as research data source. In a second step I use this restaurant data to investigate whether restaurant features, like type of cuisine or price level, can be used to learn about the socioeconomic composition of neighbourhoods.

The structure of this paper is as follows: In section 2 I give some background information on Yelp and review the existing literature that uses Yelp data. While doing so, I focus on potential peculiarities of Yelp data that should be taken into account in the later analysis. Section 3 describes the data acquisition process, provides a descriptive analysis of the dataset, and a detailed data quality assessment. Subsequently, section 4 tries to shed light on the question whether restaurant features and (socioeconomic) neighbourhood structures match. To achieve this goal I first split my research question into three dimensions and give some theoretical background for each dimension, mostly based on reasoning known from the New Economic Geography literature. To empirically investigate these dimensions, the Yelp data is then linked to a dataset that provides spatial socioeconomic information on a small-scale level. After operationalizing all three dimensions, correlational evidence is presented and evaluated. Section 5 concludes and proposes some extensions.

2 On Yelp and its Research Potential

2.1 About Yelp

Yelp is an online platform that publishes reviews on restaurants and other local businesses all around the world. The company behind the website was found in 2004 and is based in San Francisco (US). Users can search for businesses by location, business category and several other attributes. The search result provides a list of businesses that satisfy the criteria. It includes information on the name of the business, its location, the number

of customer reviews, the average customer rating, a price estimate, and the business categories that it's assigned to. A screenshot of an exemplary search result is presented in figure 1.

Website visitors may also click on one of the list entries to get more detailed information on a local business, in particular to get access to the published business reviews by former customers. Following the crowdsourcing principle, customers can post detailed reviews about local businesses and rate their experience on a scale of one to five stars. The only prerequisite for participating actively in Yelp is to register a free account, which only requires a valid email address. Once published, even persons without a registered account can read the review.

2.2 Yelp Literature Review

There is a small literature that uses data from Yelp for economic research, mainly in the field of regional and urban economics. Here, I'm going to provide an overview how Yelp data has been utilised so far. I put an emphasis on advantages and potential drawbacks that should be considered later on.

Glaeser et al. (2017) explore the potential of Yelp data to measure local economic activity in small time intervals ('nowcasting') and its ability to complement official economic statistics, which are typically published with a large time lag and on a high aggregation level. They exploit changes in the number of restaurants and restaurant reviews over time to predict the number of overall establishments from official statistics on a ZIP-Code level for the US. Besides obtaining in general relatively promising results they find that the estimation accuracy depends on the popularity of Yelp in a region, which in turn depends on factors such as population density and average levels of education and income. There also might be distortions due to listed but actually closed restaurants, which were not removed.

In a similar paper Glaeser et al. (2018a) employ Yelp data to measure gentrification and quantify associated neighbourhood change, namely changes in housing prices, demographics, and the local business landscape. For this purpose, they link Yelp data to census data for New York City and other US cities. Changes in local business landscape are measured by looking at Yelp listings for different points in time. They find that changes in the number of Yelp establishments and in the number of reviews can indeed predict shifts in housing prices and in the demographic structure of a neighbourhood. Furthermore, they try to investigate whether changes in the local business landscape precede or follow shifts in housing prices and demographics, but the results are rather dim.

In turn, Kuang (2017) does not only focus on the quantity of restaurants and reviews, but instead asks whether the quality of restaurants is capitalized into nearby home values. She uses average star ratings from Yelp to assess the perceived amenity valuation of customers and the price estimate as a proxy for expected restaurant quality (as restaurant

prices are likely to be strongly correlated with the unobserved restaurant quality). By merging Yelp data with official data on property sales in Washington D.C. she concludes that not only the quantity, but also the quality of amenities matters for home values. To include only restaurants that existed at the time of a property sale, she uses the date of establishment provided by Yelp or the date of the first review if the former is not available as a criterion.

A slightly more sophisticated approach to shed light on a related question is chosen by Davis et al. (2019). Their goal is to evaluate the relative importance of preferences as well as spatial and social frictions in determining the restaurant choices of consumers in an urban context: Spatial segregation along demographic lines might imply segregation in consumption, since travel time may play a role in restaurant choice. Social network effects, race specific consumption preferences, and aversion to consume in areas with different demographics could amplify or attenuate this effect. To answer this question, the residential and work locations as well as gender and race of very active Yelp reviewers in New York City were identified from their reviews and profile pictures respectively. Subsequently, consumption patterns were inferred based on restaurant reviews and linked to US census data for information on neighbourhood demographics. Yelp users are more likely to choose restaurants close to their residence or working place, what implies segregated consumption due to segregated residence. But they also tend to prefer restaurants in locations which are similar with respect to the demographic composition of their residence tract (when controlling for the distance to their residence place). The authors acknowledge that their findings might lack external validity, since the studied population of Yelp users potentially differs systematically in consumption patterns from the target population. This notion is reinforced by a comparison of the demographics of the Yelp reviewer sample to official statistics.

A different branch of the literature focuses specifically on the customer reviews and star ratings provided by Yelp. Anderson and Magruder (2012) try to estimate the causal effect of the average star rating on restaurant reservation availability (provided by a different online platform). To circumvent the endogeneity issue that is introduced by the positive correlation between average star ratings and actual restaurant quality, they exploit rounding thresholds of the displayed average star rating to implement a Regression Discontinuity Design (RDD): Yelp displays for each restaurant the average user rating (from 1 to 5 stars) rounded to the nearest half-star. Hence, for a restaurant that falls just below a rounding threshold the displayed average star rating is a half-star lower than for a restaurant just above the rounding threshold, even though their quality levels do not vary systematically. For restaurants in California they find that Yelp ratings have substantial effects on restaurant customer flows. The same research design is employed by Luca (2011), who links Yelp data to restaurant revenue data in Seattle to evaluate the effect of Yelp reviews on restaurant demand. He reports a 5-9% increase in revenue for independent restaurants, if the average star rating increases by one star. For chain

restaurants he finds no effect. In a related paper Luca and Zervas (2016) make use of Yelp reviews as a testing ground, to shed light on the prevalence of and economic mechanisms behind review fraud. They use reviews that were automatically filtered by Yelp for being suspicious or fake to learn about characteristics of fraudulent reviews and subsequently use their findings as a proxy for review fraud.

Finally, Kang et al. (2013) make use of natural language processing techniques to predict the outcome of hygiene inspections in the US based on Yelp reviews. Their goal is to help public policy makers in the allocation of scarce resources. Negative reviews and text segments seem to be good predictors for actual hygiene violations.

To put it in a nutshell, using Yelp data for economic research has the advantage of a fine spatial and temporal resolution, while it might also have major drawbacks like non-representativeness and imprecision. The reviewed papers showed that Yelp data for the US can be valuable for empirical economic research, while there are no known applications for other countries yet. Hence, it seems to be a natural question to ask whether Yelp can also be a beneficial data source for research on other countries as well.

3 The Dataset

In this section I provide an overview of a novel dataset on restaurants in Berlin (Germany) that was collected from Yelp by means of webscraping. After a short explanation of the data acquisition process, I describe the dataset and review the data quality. I decided to use Berlin as a testing ground, since it's the most populous city in Germany and therefore is likely to have a large number of restaurants. Additionally, following Glaeser et al. (2017), it seems to be a reasonable assumption that Yelp adaption is higher in large cities.

3.1 Data Acquisition

I conducted the data collection throughout the 40th week of the year 2020 by using the RSelenium package within the R environment. The maximum number of restaurants listed for a single search request is limited to 240 by Yelp. Hence, to obtain an at most exhaustive picture of all restaurants in Berlin listed on Yelp, I started search queries for each available restaurant category separately. Additionally, I partitioned the query by city districts in case a restaurant category contained more than 240 listed restaurants. Since I'm not interested in the detailed restaurant reviews I only collected the information provided by the restaurant listings as seen in figure 1.

Processing of the raw data was done with the stringr package in R by using regular expressions to detect contents. To get the geographic coordinates (longitude and latitude) from the provided address information I employed the Google Maps API via the geocode function of the ggmap package in R.

3.2 Description of the Dataset

Overall, the dataset contains information on 9,817 restaurants in Berlin listed on Yelp. The variables which were extracted directly from the search results are the restaurant name, location information (precise address and city district), number of reviews, price estimate, and restaurant categories. Table 1 provides descriptive statistics for the variables.

Each restaurant can be assigned to up to four categories that are subsequently translated into four distinct variables. In the whole dataset the number of unique categories amounts to 288. Table 2 displays the most frequently occurring restaurant categories and the distribution of restaurants across the city districts. 1,943 restaurants did not receive a single review so far. 50% of all restaurants in the dataset have at most 4 reviews, whereas the mean review number is 17.37 indicating a right-skewed distribution.

The price estimate ranges from one up to four €-signs, where a larger number of €-signs indicates a higher price estimate. Pricing information is only available for 5,768 restaurants of which the vast majority has either one (2,111) or two (3,000) €-signs. In contrast to the other features of the search result, the price estimate is not a part of the crowdsourcing process, but is rather calculated by Yelp itself in an opaque way. Unfortunately, it's also not clear how to interpret the €-signs quantitatively, as no official explanation by Yelp could be found. Kuang (2017) indicates the following interpretation for the $-signs on Yelp.com that is likely to carry over: "Price range is the approximate cost per person for a meal including one drink, tax and tip. $ = under $10; $$ = $11-$30; $$$ = $31-$60; $$$$ = above $61."

Figure 2 (left) displays the spatial distribution of the restaurants based on their coordinates. One can clearly see that there is a high density of restaurants in the centre of Berlin, whereas the outskirts are sparsely populated. To illustrate this, I calculated the shortest distance of each restaurant to Berlin main station. The choice of Berlin main station is motivated by its central location, but could be easily replaced by other locations. Figure 2 (right) presents a kernel density estimate of the distances: Roughly 75% of the restaurants are located in a range of 7.5 km around Berlin main station. This supports the notion of a strong spatial concentration of restaurants in the city centre provided by the previous visualisation.

3.3 Data Quality Assessment

A major concern when using crowdsourced data for research purposes is the potentially non-representative nature of the data. It's very likely that the collected data does not represent the stock of restaurants in Berlin perfectly due to both overcoverage and undercoverage of restaurants. Overcoverage may occur, if restaurants that do not exist any longer are still listed on Yelp. On the other hand, undercoverage would be present if recently established or less well known restaurants are not (yet) listed on Yelp. Whether this phenomenon introduces systematic distortions to econometric analysis depends on the extent, the

research question at hand, and the underlying causes: If undercoverage and overcoverage occur completely at random this might not be a severe issue, as it solely adds some random noise to the data that may be compensated by sample size. On the other side, more severe distortions can be expected if the probability that a listed restaurant is non-existent or an existing restaurant is not listed depends on (unobserved) variables. One could think for instance about the degree of competition in a specific market segment or the popularity of a restaurant as driving factors for sample selection.

Whereas undercoverage is rather difficult to evaluate with the data at hand, it's feasible to get an idea of the magnitude of overcoverage. Even though a verification of every single restaurant in the dataset is desirable, this would be extremely cumbersome. For this reason, I only took a random subsample (with replacement) of 100 observations from the full dataset and verified the existence of the restaurants by searching the internet.

It turns out that only 75 of the 100 restaurants in the subsample (due to the small size of the subsample relative to the full sample no restaurant was sampled multiple times) did exist as of the point in time of the analysis. Taking into account the introduced sampling error, a 95% confidence interval has a lower bound of 65.3% and an upper bound of 83.1%. Hence, it can be concluded with great certainty that there is indeed a substantial amount of overcoverage in the data. It seems like that restaurants, which were added once to Yelp, don't get removed reliably when shut down. One reason for this might be a low Yelp adaption in Germany.

Another, more indirect, way to assess the data quality is to have a look at the number of users and usage patterns over time. A more active community of contributors may result in better data quality, as newly established restaurants are more likely to be added quickly and shut down restaurants to be removed faster. Unfortunately, there is no direct way to quantify the usage patterns since Yelp does not publish any official statistics about them. A potentially useful proxy can be attained by looking at search volumes of popular internet search engines. Luckily Google, arguably the most widespread search engine, provides an easy accessible analytics tool for this task, called Google Trends. It allows to compare the search volume of a search term within a geographic region across time and also to quantify the relative popularity of a query in a region compared to other regions.

The popularity measure for interregional comparisons adjusts for the population size of a country, by relating the number of queries for a specific term to the overall number of queries in that country. Subsequently, the popularity measure is normalized to 100 for the country with the highest popularity to allow for a straightforward interpretation. By looking at the time frame from January 2010 until September 2020, it becomes evident that in Germany the popularity of the term 'Yelp' is only roughly 4% of the popularity in the US. In fact, even the second country in the list (Canada) has a popularity of only 24% compared to the US, supporting the notion that Yelp adaption is by far the highest in the US. [1]

[1] Data source: Google Trends.

Figure 3 shows the search volume in the mentioned time frame as a time series of monthly observations for Germany and the US. Note that for each country the search volume is normalised to 100 in the month with the highest volume. Therefore, it's not valid to compare the index values between the countries in absolute value. It can be seen that Yelp popularity in Germany increased rapidly in the second half of 2013 and stayed on that level until mid of 2016. Since then however, the popularity of Yelp declined steadily and is today approximately on the same level as before the rapid increase. The observed pattern coincides with the (rather subjective) observation that most reviews for restaurants in Berlin date back to the above mentioned time frame. A similar trend can be observed for the US, even though Yelp adaption began earlier and changes in search volume occurred more smoothly over time.

This observation has profound implications for data quality: First, overcoverage might be mainly introduced by restaurants that were shut down only in recent years, since the removal probability declined proportionally to the popularity of Yelp. For the same reason, recently established restaurants are less likely to be listed on Yelp what introduces undercoverage. Second, the minor Yelp adaption in Germany relative to the US could be a reason for poorer data quality in total and explains why research with Yelp data has been limited to the US so far.

Based on these mostly unsatisfactory results, the question arises what can be done in order to improve the representativeness of the dataset. Whereas undercoverage can hardly be mitigated, it might be feasible to detect non-existent restaurants based on their observable characteristics. In particular it seems plausible that listed restaurants with no or only a few reviews are more likely to be actually permanently closed. The number of reviews is likely to be a strong predictor for restaurant popularity, what in turn increases the probability of removal from Yelp in case of a shut down. It also seems reasonable to assume that popular restaurants don't have to shut down in the first place. Moreover, the availability of a price estimate might be informative about the existence of a restaurant in the sense that restaurants with a price estimate are more likely to exist. Due to the previously described sketchy nature of the price estimate the mechanism behind this prior assumption is however tentative.

To estimate the probability that a listed restaurant actually exists, I used standard logistic regression models estimated by maximum likelihood. Table 3 summarises the results of three different specifications. Model 1 only includes the number of reviews as predictor, whereas model 2 additionally includes the availability of a price estimate as a dummy variable. As expected, the number of reviews has a highly significant positive effect in both specifications. Rather unforeseen, the availability of price information has a significant negative impact on the probability in Model 2 what opposes my prior assumption. Model 3 is driven by the idea that the probability of an existing restaurant is not increasing monotonously in the number of reviews. Instead there could be a sharp cut-off between restaurants with no reviews and restaurants with at least one review.

Therefore I included a dummy variable indicating whether a restaurant has at least one review, the price information dummy, and an interaction term between both. It turns out that only the review number dummy has a significant (positive) impact on the estimated probability.

In terms of accuracy all three models perform only slightly better than a non-informative predictor, that would predict 75% of all outcomes correctly by assigning each observation to 'exists'. Since my main goal is to identify non-existent restaurants it's important to focus particularly on false positive predictions (restaurants that do not exist, but get classified as existent). In this realm, model 1 is able to detect 44% of all non-existent restaurants in the subsample correctly (specificity). On the other hand it identifies 93% of all truly existing restaurants in the subsample correctly (sensitivity). Model 3 has identical performance metrics but includes non-significant parameters and model 2 has worse predictive capabilities, such that I stick to the parsimonious model 1.

A specificity of 44% is still not a great achievement – 56% of all non-existing restaurants in the subsample are classified incorrectly. One possible remedy is to increase the classification threshold above 0.5 to predict more of the truly non-existent restaurants correctly. Of course, this comes at the cost of more misclassified existent restaurants. Given that the majority of all restaurants actually exist, the loss in sample size would be quite substantial. The so called receiver operating characteristic (ROC) curve displayed in figure 4 illustrates this trade-off. It depicts sensitivity and specificity for each feasible classification threshold. It is clear that the optimal choice of the classification threshold is a difficult, partly subjective, task and it can be questioned whether this attempt has a net benefit at all. Based on the ROC curve, one reasonable choice of the classification threshold would be 0.562, what results in an improved specificity of 60% whereas sensitivity decreases only moderately to 88%. A further increase of the classification threshold would lead to substantial reductions in sensitivity while increasing specificity only slightly. Additionally, the chosen threshold yields the same accuracy as the default threshold, such that the choice can be evaluated as indeed optimal. When using model 1 with the optimised probability threshold as a classifier for the whole sample, sample size reduces to 67.3% (i.e. 6603 observations). As the number of reviews is the only predictor in the logistic regression model, the produced decision rule relates to a removal of all restaurants with less than two reviews.

There are two more potential strategies to increase the representativeness of the data: One way could be to link it to data that was collected during the period in which Yelp enjoyed its highest popularity in Germany. Additionally, it would be thinkable to exploit the dates of reviews (if available) to infer the existence of restaurants in a specific time frame. Since I didn't collect this information this option remains hypothetical. Due to the high share of restaurants with almost no reviews this would be also a very crude approximation.

4 Do Restaurants Actually Match Neighbourhoods?

In this section I demonstrate how the previously described dataset can be linked to geospatial data for further analysis. As an example I'm using the RWI-GEO-GRID dataset (Breidenbach and Eilers, 2018; RWI and microm, 2019) , which is shortly described first. The subsequent analysis is meant as a preliminary empirical examination of the relationship between neighbourhood structure and restaurant features. It has a purely correlational nature, i.e. I don't make strong causal claims. However, I provide some heuristic theoretical reasoning for expected and observed associations.

4.1 The RWI-GEO-GRID Dataset

The RWI-GEO-GRID dataset (Breidenbach and Eilers, 2018; RWI and microm, 2019) offers aggregated socioeconomic data for Germany on a 1 x 1 km raster level. In total, Germany can be partitioned into approximately 361,000 1 x 1 km grid cells of which around 220,000 are covered by the dataset. Each grid cell contains information on the demography (e.g. age structure, sex, and ethnicity) of the grid dwellers, the household and neighbourhood structure (e.g. household composition and unemployment rate), and mobility (car segments). Data is available for the year 2005 and the years 2009 to 2017. Originally the data is provided by microm, a marketing service contractor who uses individual level data from various sources (mainly from companies acting in data intensive industries) to aggregate the dataset.

In contrast to administrative data, the granular structure of the RWI-GEO-GRID dataset allow to uncover heterogeneity within administrative boundaries, which makes it a useful complement for my dataset. Overall, the administrative area of Berlin consists of 1014 grid cells, where cells at the administrative borders only cover a fraction of the area. From these grid cells the RWI-GEO-GRID dataset contains information on 934 cells (for the year 2017). The missing grid cells are either completely uninhabited or very sparsely populated and subsequently anonymised.

4.2 Sample Construction and Spatial Aggregation

As discussed in the previous section, the issue of non-existent restaurants could likely be attenuated by removing restaurants with very few reviews, what comes at the cost of a substantial reduction in sample size. However, in the given context this might not be a beneficial strategy to increase data quality: I use the RWI-GEO-GRID data of 2017 for my analysis, what is roughly the point in time when the popularity of Yelp started to decline in Germany. On the basis of the data at hand it's impossible to say when a restaurant actually shut down, such that it's not unlikely that it still existed in 2017. Hence, the full dataset might come closest to the true stock of restaurants in 2017. Since the spatial distribution of socioeconomic attributes is arguably relatively stable over time, using the

RWI-GEO-GRID data from previous years would unlikely alter the results substantially. Based on these considerations I stick to the full sample here. Due to the nature of the analysis the results are likely not to be sensitive to this aspect anyway.

In order to link my Yelp data to the RWI-GEO-GRID dataset, it is necessary to aggregate individual observations at the grid level. Practically this means that for each restaurant in the dataset the respective grid cell is identified what then allows for aggregation of restaurant features. Those units can then be directly linked to the RWI-GEO-GRID dataset. In total, the 9,817 restaurants are distributed across 597 grid cells, i.e. 417 grid cells do not contain any restaurants.

A drawback of this straightforward aggregation approach is that restaurants that are located close to the border of a grid cell are only taken into account for a single grid cell, even though they might be informative for neighbouring grid cells as well. This limitation could potentially be mitigated by weighting observations in the aggregation process relative to their distance to a grid cell. Alternatively, it would be feasible to build larger areas by merging several neighbouring grid cells, either by building uniform composite grid cells (e.g. 4 x 4 km grid cells built from four 1 x 1 km grid cells) or by accounting for neighbourhood structure (e.g. small level administrative areas). All those methods have some merits, but also disadvantages. For the sake of simplicity I stick to the simple aggregation approach here.

4.3 Theoretical Background

To shed some light on the question, whether restaurants match neighbourhoods I focus on three potential dimensions:

First, to what extent can the spatial distribution of restaurants be linked to the distribution of inhabitants? Theoretically, a positive association can be expected since restaurants might prefer to locate in densely populated areas to have access to a large market since traveling is costly. On the other hand, city dwellers may prefer to locate close to restaurants for the same reason. This reasoning is known from the New Economic Geography literature as a home market effect. Densely populated areas might also provide more rental opportunities for restaurants. In contrast, a disassociation of restaurant and population density can be the result of segregated housing and consumption areas within a city.

Second, does racial segregation within a city coincide with segregation of matching restaurants (i.e. restaurants that offer food that can be associated with a specific ethnicity)? Again, some sort of home market effect comes into play: Ethnicities may have restaurant preferences, which are associated to their cultural heritage. Hence, it could pay for restaurants to agglomerate in areas with a matching ethnic composition and vice versa. Restaurant owners belonging to a certain ethnicity might in turn have a preference for running a restaurant close to their own residence. If the ethnicity of a restaurant owner

and the type of restaurant matches (what is common), this would also imply a spatial concentration of restaurants based on racial segregation. On the other hand, there could be market crowding effects and a love for variety among ethnicities that could turn out to be dispersion forces. This dimension is also closely related to the paper by Davis et al. (2019).

Third, is the price level of a restaurant associated with the purchasing power of close by city dwellers? In a similar manner as in the previous dimension one could argue that income groups might have different restaurant preferences and additionally different budget constraints, while facing segregated residence. This would imply a spatial concentration of restaurant price levels, given that restaurants adjust their prices to the budget constraints of closely residing inhabitants.

4.4 Variable Construction and Methodology

To answer these questions with the data at hand it's necessary to choose appropriate variables of the RWI-GEO-GRID dataset and to construct grid level features from the Yelp dataset. Subsequently, I link each pair of variables by simple linear regression models. Descriptive statistics of all variables are depicted in table 4.

For the first dimension I simply add up the number restaurants within each grid cell as an explanatory variable for the number of inhabitants within that grid cell. To account for the right-skewed distribution of restaurants across the grid cells, I use the log of the number of restaurants in the regression instead.

The implementation of the second dimension is somewhat more complex. I decided to focus on the groups of inhabitants that have an ethnic background from non-European Islamic countries and from (South-) East Asia. Both represent relatively large ethnic groups in Berlin and both cultural spheres have well established restaurant cultures in Germany (not to say they are homogenous). My approach is to identify restaurant categories from Yelp that match the associated ethnic group. In the case of an Islamic ethnic background there are 13 matching restaurant categories, inter alia the categories Kebab, Turkish, and Lebanese. For the (South-) East Asian ethnicity there are 18 matching Yelp categories, such as Vietnamese, Sushi, and Chinese. The full list of matching restaurant categories is appended to table 5. Given that each restaurant can be assigned to more than one category on Yelp, a restaurant is identified as 'matching' if it is assigned to at least one of the categories. Next I calculate the share of matching restaurants relative to the total number of restaurants for both ethnicities in each grid cell. While this approach helps to control for the general spatial concentration of restaurants, it comes at the cost of very imprecise measurement in grid cells with few restaurants. E.g., if a grid cell is only populated by a single restaurant the share of matching restaurants is either one or zero and therefore strongly led by coincidences. To account for this, I only include grid cells with a certain minimum number of restaurants in my analysis. A larger minimum number

yields a more meaningful estimate of the share of matching restaurants, while reducing sample size. Since the choice of a minimum number of restaurants could have potentially a big impact on the results I present them for three different minimum values, namely ≥ 10, ≥ 20, and ≥ 30. Finally, I link the share of matching restaurants to the share of the respective ethnicity in a grid cell from the RWI-GEO-GRID dataset.

To shed light on the last dimension I rely on the price estimates provided by Yelp. I calculate the average price estimate within each grid cell and subsequently relate it to the unemployment rate of that grid cell. Again, results for all three minimum number of restaurants values are presented, since fewer observations imply a higher variance of the estimator.

4.5 Results

To begin with, I graphically compare the spatial distribution of inhabitants to the distribution of restaurants across the grid cells in figure 5, where each rectangle represents a 1 x 1 km grid cell. As already seen in figure 2, there is a strong spatial concentration of restaurants in the centre of Berlin. Note that the distribution of restaurants is displayed on a more granular level in figure 2. There are 597 grid cells with at least one restaurant of which the average grid cell is populated by 16.4 restaurants and the highest restaurant density is given by 252 restaurants within a grid cell. The spatial distribution of inhabitants also reveals a concentration around the city centre. However, it is way more blurred out in comparison to the restaurant distribution, i.e. even in the outskirts of Berlin there are regions that are densely inhabited. Figure 6 a) then removes the spatial information and solely relates the number of inhabitants to the (log) number of restaurants within a grid cell. Table 5 panel a) (left side) provides information on the displayed regression line. Roughly 62% of the variation in the log of the number of restaurants can be explained by the number of inhabitants, when including grid cells where not a single restaurant is listed on Yelp. Based on the estimated regression line, an increase in the number of inhabitants by 1,000 coincides on average with an increase in the number of restaurants by roughly 25.7%. When restricting the sample to grid cells with at least one listed restaurant the R^2 decreases unsurprisingly to 0.54 and also the slope coefficient declines slightly (table 5 panel a) right side). To conclude, a very high number of inhabitants ($\geq$ 15,000) seems to be an almost certain sign for a large number of restaurants, whereas for less populated grid cells the relationship is more noisy. This could reflect partly the mentioned separation between consumption and housing areas.

Panel b) and c) of table 5 present the results of the simple regression models that relate the population share of an ethnicity to the share of matching restaurants. Figures 6 b) and c) display the respective scatter plots for the sample with at least 30 restaurants within a grid cell. There is a highly significant positive association between the population share with an ethnic background from Islamic countries and the share of matching restaurants,

which declines slightly when the sample size is increased. E.g., for the ≥ 30 sample (83 grid cells) an increase in the population share by one percentage point is associated with an increase in the share of matching restaurants by 0.69 percentage points. The worry that a smaller minimum number of restaurants per grid cell would result in noisier estimates is somewhat validated by the reduction in the goodness of fit. On the other hand, the sample size can be more than doubled when using the ≥ 10 instead of the ≥ 30 threshold.

In contrast, the results for the population share with an ethnic background from (South-) East Asia is only weakly significant for the ≥ 30 sample and becomes completely insignificant for the other samples.

This raises of course the question, how this disparity of findings can be rationalized. There might be two reasons: First, the variability in the population share is much smaller for the (South-) East Asian ethnicity (standard deviation: 0.95) than for the Islamic countries ethnicity (standard deviation: 3.94). This leads to less precise estimates for the coefficients. Second, there might be also a theoretical explanation. It appears as if Asian food is more often sold by chain restaurants that are more likely to operate in non-matching areas, since those restaurants are not linked to a single restaurant owner and its ethnicity.

Ultimately, Panel d) of table 5 and figure 6 d) display the relationship between the unemployment rate and the average price category of restaurants listed on Yelp. It turns out that there is a strongly significant negative association between these two variables, which is almost unaffected by the chosen sample. A one percentage point increase in the unemployment rate is related to a decrease of the average restaurant price estimate by 0.06 units in the ≥ 30 sample, where the minimum and maximum of the observed average price categories are 1.25 and 2.19 respectively. It seems stunning that there are (almost) no clear outliers in the scatter plot presented in figure 6 d) and that the explanatory power of the model is rather high for a single explanatory variable. The same holds true for the two other samples.

Unfortunately, this could be a partially artificial result. As already mentioned in the previous section, the price estimate provided by Yelp is not a part of the reviewing process, but is calculated internally in an unknown way. In the same time, the RWI-GEO-GRID dataset is aggregated partially based on data that is provided by companies to microm and is most likely also provided by microm to companies. It could well be that Yelp actually uses some input data in their price estimate algorithm that is also used to aggregate the RWI-GEO-GRID dataset. This would obviously invalidate this correlation completely.

5 Conclusion

In recent years, crowdsourced data from online platforms has become increasingly popular in economic research. One example is Yelp, a reviewing platform for local businesses around which a small literature has emerged that focuses on the US.

The first goal of this paper was to investigate whether Yelp data can also be used to conduct research concerning other countries like Germany. For this purpose I compiled a large dataset of restaurants in Berlin listed on Yelp containing information like precise location, number of reviews, restaurant categories, and prices. To assess the data quality I first looked at Google Trends scores, which can be seen as a proxy for the popularity of the platform across time and between geographical regions. This then allows to draw indirect conclusions about data quality. It turned out that Yelp is relatively unpopular in Germany compared to the US and that it experienced its highest popularity between the years 2013 and 2017, while facing a loss of meaning since then. Additionally, I drew a random sample of 100 restaurants from the original dataset to estimate how many of the listed restaurants actually do exist (up to today). Unfortunately, only three out of four restaurants could be verified as existent, what reinforces the notion that the data may not be up to date, implying both over- and undercoverage issues. In order to mitigate at least the overcoverage issue, I employed a logistic regression model that uses the number of reviews as a predictor for the existence of a restaurant. It's feasible to filter 66% of the non-existent restaurants by this procedure (in the subsample), while misclassifying 12% of the truly existing restaurants and reducing sample size to roughly 67% when applying the decision rule to the full sample.

The second objective of this paper was then, to shed light on the question whether local restaurants match socioeconomic features of neighbourhoods like ethnic composition and purchasing power. For this purpose I linked the Yelp data to the RWI-GEO-GRID dataset, which contains socioeconomic spatial information for Germany on a 1 x 1 km grid level, by aggregating restaurant features on the same spatial level. There is significant correlational evidence that a larger population share with an ethnic background from Islamic countries coincides with a higher density of matching restaurants. On the other hand, there is no robust analogous association for areas with a relatively high share of inhabitants that have a (South-) East Asian ethnic background. This inconsistency may be explained by the notion that there are more chain restaurants that sell the matching dishes of the latter than of the former population group. Finally, I found a robust statistical significant negative association between the unemployment rate (as a proxy for purchasing power) and the average restaurant price level. This result may be however fishy, since it's possible that the data generating processes of the RWI-GEO-GRID dataset and the price information overlap.

Given the correlational nature of my analysis, a natural extension would be to extent it to a multivariate setting that controls for observable characteristics. In particular, one could try to account for chain restaurants in the analysis which arguably distort associations. Due to the missing time dimension in my Yelp data (and other obvious sources of exogenous variation), claims for causality will always be somewhat shady. This is a clear limitation compared to the papers that use Yelp data from the US, where opening dates are either directly available or can be reasonably imputed by the date

of the first (and the last) review. However, for the given research question it is not absolutely necessary to identify precise causal effects. This potentially opens the door to use algorithms like Random Forests (Breiman, 2001) from the Statistical Learning paradigm to make predictions. Another potential extension would be to choose a more sophisticated way of data aggregation, to better account for neighbourhood structure.

References

Michael Anderson and Jeremy Magruder (2012): Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989.

Philipp Breidenbach and Lea Eilers (2018): RWI-GEO-GRID: Socio-economic data on grid level. *Jahrbücher für Nationalökonomie und Statistik*, 238(6):609–616.

Leo Breiman (2001): Random forests. *Machine Learning*, 45:5–32.

Donald R. Davis, Jonathan I. Dingel, Joan Monras, and Eduardo Morales (2019): How segregated is urban consumption? *Journal of Political Economy*, 127(4):1684–1738.

Edward L. Glaeser, Hyunjin Kim, and Michael Luca (2017): Nowcasting the local economy: Using Yelp data to measure economic activity. Working Paper 24010, National Bureau of Economic Research.

Edward L. Glaeser, Hyunjin Kim, and Michael Luca (2018a): Nowcasting gentrification: Using Yelp data to quantify neighborhood change. *AEA Papers and Proceedings*, 108: 77–82.

Edward L. Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik (2018b): Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137.

Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi (2013): Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1443–1448.

Chun Kuang (2017): Does quality matter in local consumption amenities? An empirical investigation with Yelp. *Journal of Urban Economics*, 100(C):1–18.

Michael Luca (2011): Reviews, reputation, and revenue: The case of Yelp.com. Harvard Business School Working Papers 12-016, Harvard Business School.

Michael Luca and Georgios Zervas (2016): Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, 62(12):3412–3427.

RWI and microm (2019): RWI-GEO-GRID: Socio-economic data on grid level - scientific use file (wave 8). version: 1. Technical report, RWI – Leibniz Institute for Economic Research.

Appendix A: Tables

Table 1: Descriptive statistics of selected variables

Variable	Scale	Obs.	Mean	Median	Std. Dev.	Min.	Max.
No. of reviews	metric	9817	17.37	4.00	37.67	0.00	1084.00
Price category	metric	5768	1.76	2.00	0.68	1.00	4.00
Dist. to main stat.	metric	9817	5.99	4.87	3.87	0.02	27.91
City district	nominal	9817					
Restaurant cat.	nominal	9817					

Table 2: Most frequent restaurant categories and distribution of restaurants between the city districts in the dataset

Restaurant category	Frequency	City district	Frequency
Café	2019	Mitte	1123
Italian	1280	Charlottenburg	983
German	1046	Kreuzberg	818
Pizza	751	Prenzlauer Berg	817
Fast Food	607	Schöneberg	667
Breakfast & Brunch	517	Friedrichshain	631
Kebab	489	Wilmersdorf	513
Vietnamese	469	Neukölln	464
Asian	436	Tiergarten	451
Burger	403	Wedding	365
Sushi	341	Reinickendorf	363
Snack	265	Steglitz	358
Turkish	253	Spandau	308
Bakery	250	Tempelhof	287
International	232	Köpenick	232
Indian	229	Treptow	207
Coffee Shop	227	Zehlendorf	191
Mediterranean	220	Lichtenberg	168
Chinese	219	Pankow	150
Bar	208	Reuterkiez	144
Currywurst	195	Bergmannkiez	116
Thai	195	Weißensee	113
Cocktail bar	181	Hohenschönhausen	94
Bistro	176	Marzahn	79
Greek	156	Hellersdorf	74
Beer garden	154	Schillerkiez	54
Steakhouse	144	Britz	47

Notes: The list of city districts is exhaustive, while the list of restaurant categories is truncated. Note that restaurants listed on Yelp can be assigned to more than one category.

Table 3: Logistic regression models to predict the existence of restaurants

	Model 1	Model 2	Model 3
Intercept	−0.20	0.11	−0.59
	(0.35)	(0.39)	(0.56)
No. of reviews	0.22***	0.33***	
	(0.08)	(0.11)	
𝟙(Price cat. avail.)		−1.32**	−15.98
		(0.66)	(1696.73)
𝟙(No. of reviews > 0)			2.32***
			(0.84)
𝟙(No. of reviews > 0)*𝟙(Price cat. avail.)			15.82
			(1696.73)
AIC	86.03	83.70	101.89
Accuracy	0.81	0.77	0.81
Sensitivity	0.93	0.95	0.93
Specificity	0.44	0.24	0.44
No. of obs.	100	100	100

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Accuracy: Fraction correctly predicted. Sensitivity: True positive rate. Specificity: True negative rate. Positive class: 'exists'. The accuracy of a non-informative classifier, i.e. a classifier that predicts for all observation 'exists', is 0.75. (Predicted) probability threshold to assign an observation to 'exists': 0.5.

Table 4: Descriptive statistics of 1 x 1 km grid cells

Variable	Obs.	Mean	Median	Std. Dev.	Min.	Max.
Number of inhabitants (in 1000)	880	4.17	2.36	4.79	0	26.12
Share ethnic background Islamic countries	933	5.16	4.11	3.94	0.00	28.25
Share ethnic background Asia	933	0.73	0.38	0.95	0.00	6.24
Unemployment rate	933	6.43	6.43	3.45	0.04	14.77
Number of restaurants	597	16.44	4.00	33.65	1	252
Share matching rest. Islamic countries (10)	178	10.78	9.50	8.33	0.00	37.71
Share matching rest. Islamic countries (20)	108	11.39	9.52	7.48	0.00	37.71
Share matching rest. Islamic countries (30)	83	10.97	9.38	7.36	0.00	37.71
Share matching rest. Asian countries (10)	178	16.09	16.58	8.48	0.00	50.00
Share matching rest. Asian countries (20)	108	17.05	17.36	6.97	0.00	33.33
Share matching rest. Asian countries (30)	83	17.28	17.50	6.60	3.23	33.33
Average price category (10)	178	1.72	1.71	0.28	1.11	3.25
Average price category (20)	108	1.72	1.73	0.22	1.17	2.19
Average price category (30)	83	1.72	1.72	0.22	1.25	2.19

Notes: The numbers in parentheses indicate the minimum number of restaurants in a grid cell required to be included in the sample.

Table 5: Results of the simple regression models by minimum number of restaurants per grid cell

	a)		b)			c)			d)		
Dep. Var.	log No. of restaurants		Sh. of matching restaurants			Sh. of matching restaurants			Average price category		
Indep. Var.	No. of inhabitants		Pop. share Islamic background			Pop. share Asian background			Unemployment rate		
Min. Rest.			≥ 30	≥ 20	≥ 10	≥ 30	≥ 20	≥ 10	≥ 30	≥ 20	≥ 10
Intercept	-0.07^{**}	0.34^{***}	3.29^{*}	4.58^{***}	5.38^{***}	15.17^{***}	17.40^{***}	15.14^{***}	2.25^{***}	2.18^{***}	2.13^{***}
	(0.03)	(0.05)	(1.79)	(1.50)	(1.36)	(1.40)	(1.34)	(1.10)	(0.06)	(0.05)	(0.06)
Slope	0.23^{***}	0.20^{***}	0.69^{***}	0.66^{***}	0.58^{***}	1.25^{*}	-0.22	0.63	-0.06^{***}	-0.05^{***}	-0.05^{***}
	(0.006)	(0.006)	(0.19)	(0.17)	(0.15)	(0.68)	(0.74)	(0.59)	(0.007)	(0.006)	(0.006)
R^2	0.62	0.54	0.25	0.22	0.13	0.03	0.001	0.01	0.47	0.38	0.22
No. of Obs.	880	585	83	108	178	83	108	178	83	108	178

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Observational units: 1 x 1 km grid cells. Heteroskedasticity robust standard errors (White standard errors) in parentheses. 'Min. Rest.' is the minimum number of restaurants per grid cell required to be included in the sample. Sample construction of Panel a): All grid cells that are either missing completely in the RWI-GEO-GRID dataset (80) or where the number of inhabitants is missing (54) are excluded from the sample. Grid cells without restaurants, where the number of inhabitants is available are included in the first but not in the second model. The number of restaurants is increased by 1 for each grid cell to allow for the log transformation in the case of no restaurants. Panel b): Matching restaurant categories: Kebab, Falafel, Syrian, Afghan, Pakistani, Halal, Persian, Turkish, Moroccan, Lebanese, Middle Eastern, Arabic, Oriental. Panel c): Matching restaurant categories: Sushi, Wok, Ramen, Vietnamese, Korean, Thai, Chinese, Indonesian, Cambodian, Philippine, Taiwanese, Malaysian, Nepalese, Bengal, Japanese, Indian, Asian, Asian fusion cooking.

Appendix B: Figures

Figure 1: Exemplary search result from Yelp

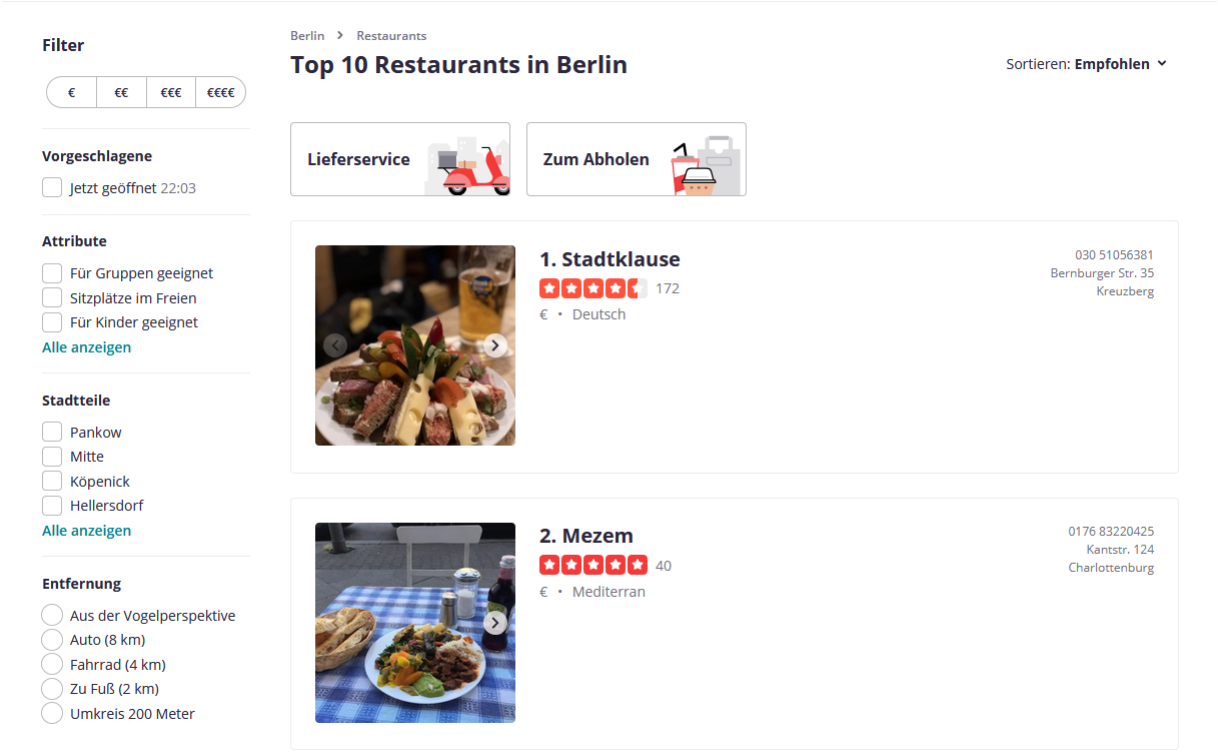
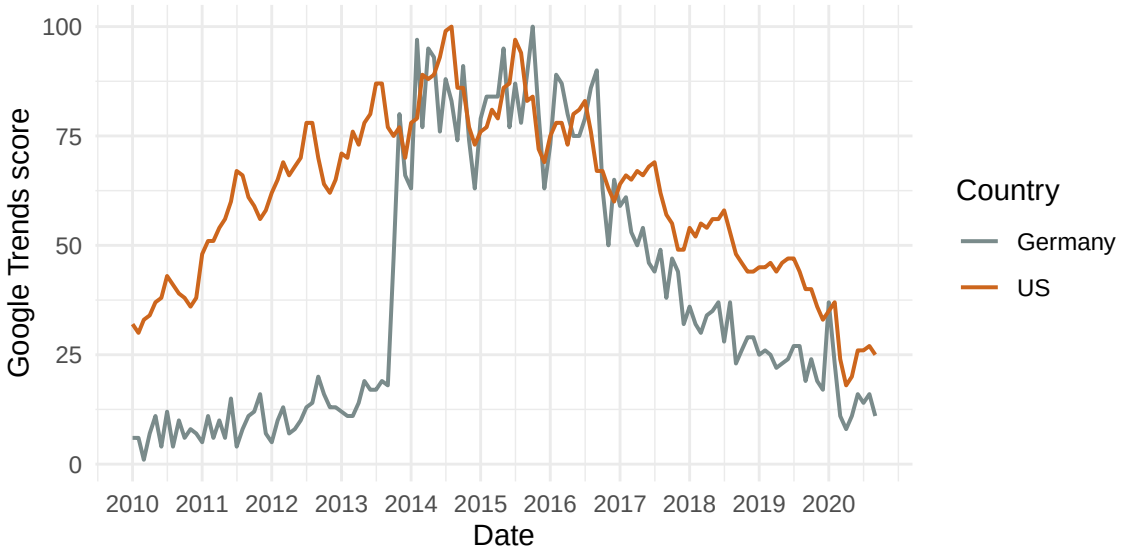


Figure 2: Spatial distribution of restaurants in Berlin listed on Yelp and kernel density estimate of the shortest distance to Berlin main station

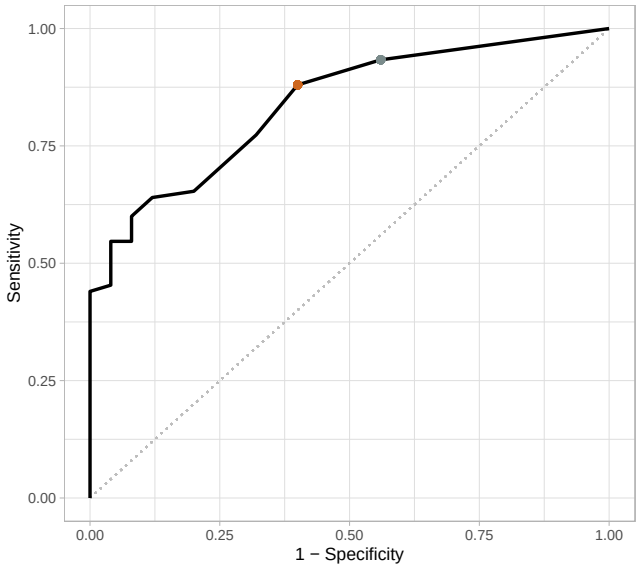
Notes: Left: The grey area displays the administrative area of Berlin and the orange point marks Berlin main station. Note that the colour scheme follows a log scale, where each rectangle represents the number of restaurants on Yelp corresponding to the colour displayed in the legend. Right: The estimate is based on a Gaussian kernel with a bandwidth of approximately 0.493. Displayed distances correspond to the shortest distance between two points (the restaurants and Berlin main station) on an ellipsoid approximation of planet earth. Distance measurement was conducted with the distGeo function of the geosphere package in R.

Figure 3: Time series of monthly Google Trends scores for the search term 'Yelp' in the time frame January 2010 until September 2020 by country



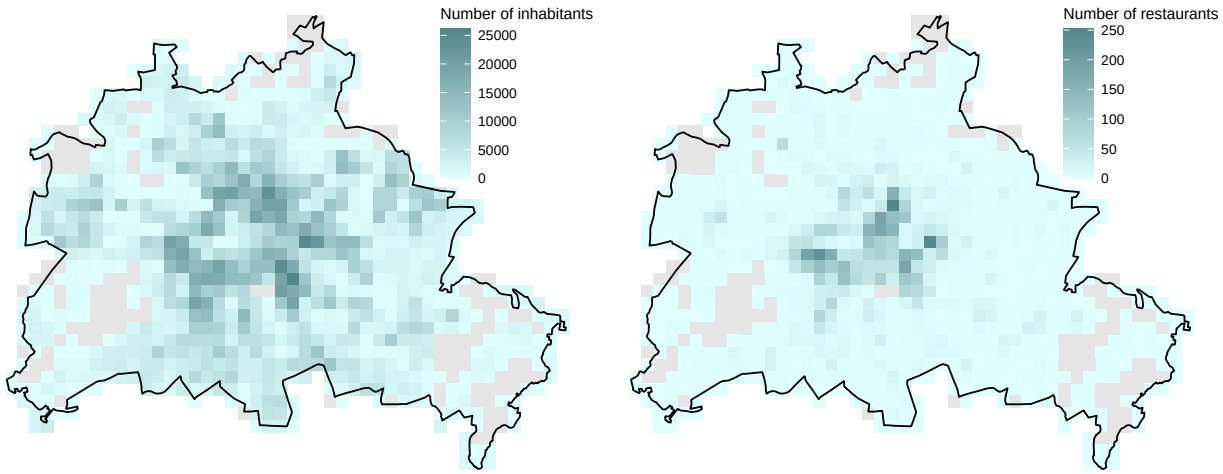
Notes: The Google Trends score quantifies the volume of a search term on Google within a geographic region by a value between 0 and 100. The respective score value of a month and region is calculated by building the ratio of the number of search queries in that month to the maximum number of search queries within a month in the displayed time frame. Hence, it's possible to compare scores for different points in time within a region, but not for a given month across regions.
Data sources: US, Germany.

Figure 4: Receiver operating characteristic curve (ROC curve) of the logistic regression model (model 1)



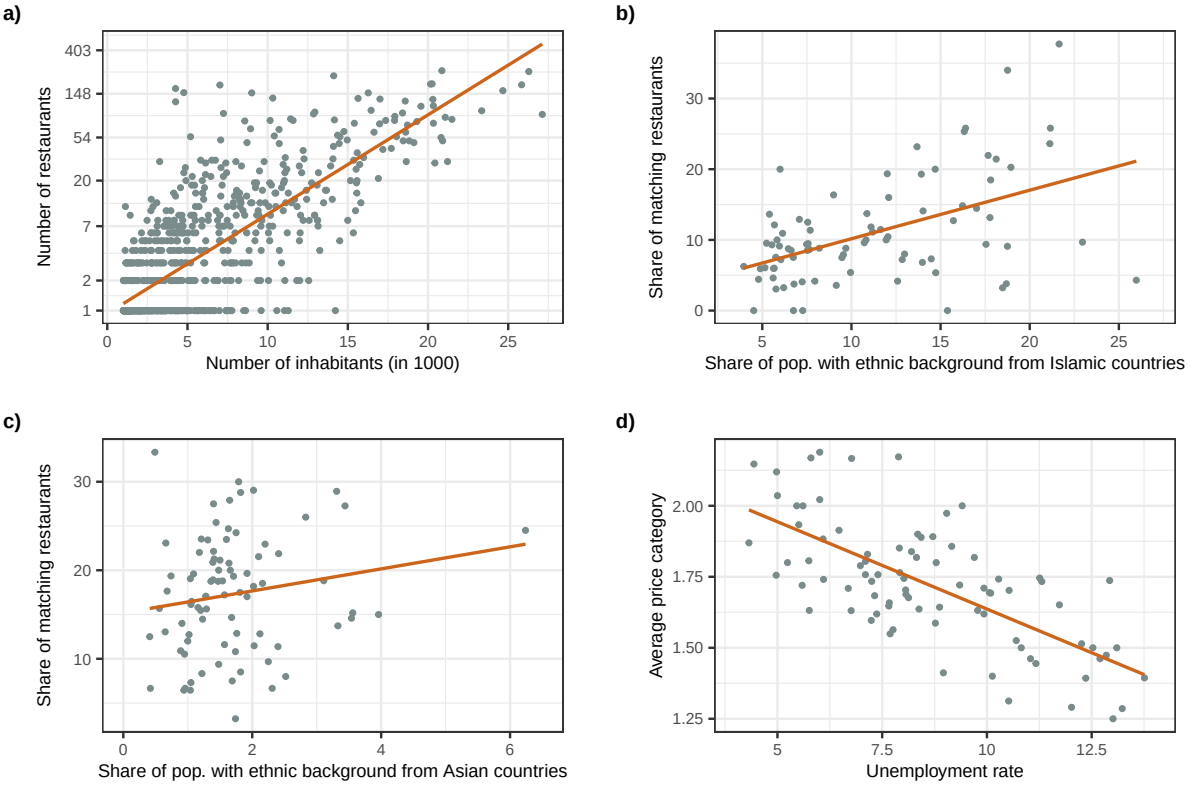
Notes: The solid black line represents the ROC curve of the fitted logistic regression model. It displays sensitivity and specificity for all feasible probability thresholds (between 0 and 1) applied to classify an observation as existent. The grey point marks the default threshold of 0.5 and the red point the optimised threshold of 0.562. Finally, the dotted grey line represents the expected performance of random classifiers.

Figure 5: Spatial distribution of inhabitants and restaurants across the grid cells



Notes: Missing grid cells are represented by grey rectangles. 880 grid cells are included, i.e. only grid cells which are missing in the RWI-GEO-GRID dataset are excluded. Grid cells that do not contain any restaurants are included.

Figure 6: Scatter plots of key variables



Notes: a) corresponds to the model with 880 observations from table 5. The y-axis follows a logarithmic scale. b), c) and d) display the models from table 5 which include only grid cells with at least 30 restaurants.