Use and Validation of Location-Based Services in Urban Research: An Example with Dutch Restaurants*

Daniel Arribas-Bel [a,b] (D.Arribas-Bel@liverpool.ac.uk, @darribas)
and
Jessie Bakens [c] (D.Arribas-Bel@liverpool.ac.uk)

[a] *Department of Geography and Planning, University of Liverpool, United Kingdom*
[b] *Geographic Data Science Lab, University of Liverpool, United Kingdom*
[c] *Research Centre for Education and the Labour Market, Masstricht University, The Netherlands*

Abstract

This paper focuses on the use of big data for urban geography research. We collect data from the location-based service Foursquare in the Netherlands and employ it to obtain a rich catalogue of restaurant locations and other urban amenities, as well as a measure of their popularity among users. Because the Foursquare data can be combined with traditional sources of socio-economic data obtained from Statistics Netherlands, we can quantify, document, and characterize some of the biases inherent to these new sources of data in the context of urban applications. A detailed analysis is given as to when this type of big data is useful and when it is misleading. Although the users of Foursquare are not representative of the whole population, we argue that this inherent bias can be exploited for research about the attractiveness of urban landscapes and consumer amenities in addition to the more traditional data on urban amenities.

Keywords: Big-data, Foursquare, Urban Amenities
JEL-classification: O18, R00

*The authors would like to specially thank participants of the VU-Eureka seminar, NARSC'14 meetings, TI workshop, ADETRE, and Gabriel Ahlfeldt for very useful discussions and comments on previous versions of the paper. Any errors remain the sole responsibility of the authors.

Use and Validation of Location-Based Services in Urban Research: An Example with Dutch Restaurants

Abstract

This paper focuses on the use of big data for urban geography research. We collect data from the location-based service Foursquare in the Netherlands and employ it to obtain a rich catalogue of restaurant locations and other urban amenities, as well as a measure of their popularity among users. Because the Foursquare data can be combined with traditional sources of socio-economic data obtained from Statistics Netherlands, we can quantify, document, and characterize some of the biases inherent to these new sources of data in the context of urban applications. A detailed analysis is given as to when this type of big data is useful and when it is misleading. Although the users of Foursquare are not representative of the whole population, we argue that this inherent bias can be exploited for research about the attractiveness of urban landscapes and consumer amenities in addition to the more traditional data on urban amenities.

Keywords: Big-data, Foursquare, Urban Amenities
JEL-classification: O18, R00

1 Introduction and Motivation

During the last few decades, the world has witnessed an explosion in the amount of data generated, much of which can be located somewhere on the Earth surface. In addition to the increase in computing power, location technology such as the global positioning system (GPS) has also undergone dramatic improvements and sharp drops in cost. The combination of these two trends is producing a vast amount of geo-referenced data, presenting many opportunities for research in the social and urban realms (Arribas-Bel, 2014). Powered by these technological advances (Goodchild, 2007), is the phenomenon called 'location-based services' (LBS). These are online applications that allow users to broadcast their location in real-time in what has come to be known as checkin's. Because traces of human behavior (in the form of checkin's) are stored and structured, they offer unprecedented opportunities to study questions urban researchers are interested in but have traditionally had no available data to work with. Additionally, many of these new data sources are available at a much finer spatial resolution than traditional ones, posing a great advantage when it comes to considering issues related to scale and the modifiable areal unit problem (MAUP).

Many references have argued that social science is at the dawn of a new era due to the increasing availability of these new, often geo-referenced, datasets (e.g. Lazer et al., 2009; Miller, 2010). Miller (2010) reviews the rise of new geo-referenced sources of data in the context of regional science and discusses how these can be mined to extract new knowledge. Novel datasets have also started to make appearances in scientific applications. As an example of use of LBSs data, Cranshaw et al. (2012) use checkin's to re-draw neighbourhood boundaries in several American cities, while Cheng et al. (2011); Noulas et al. (2012a,b); Rashidi et al. (2017) uses similar digital traces to deduce global patterns of human behavior and urban mobility across space.[1] Davis et al. (2017) use restaurant reviews given by users of Yelp to measure racial and ethnic segregation in New York City.

Much of the academic discussion and skepticism on LBSs is centered around their biases and coverage (or lack thereof) in representing the underlying general population. Despite the interest on this issue, the literature aiming to quantify it is, with exceptions (e.g. Hecht and Stephens, 2014), very scarce.

In this paper we use data from one of the leading companies in the LBS industry, Foursquare, and explore the geographical bias that arises when these new types of datasets are used for research in urban domains. Specifically, we compare the geographical map of restaurants present in the Foursquare data with the geographical restaurant data gathered by the official Statistics Bureau of the Netherlands.

In addition, we provide an illustration that demonstrates how these data may be used for empirical research in the field of urban and regional economics. Cities are thought to be attractive for consumers because they offer many amenities. Using checkins as a statement of a positive experience and a vote for popularity[2] (see, for example, Ludford et al., 2007, Lindqvist et al., 2011 or Frith, 2014) offers the possibility to explore the determinants of attractive areas in a city, as not all areas of attractive cities are typically considered attractive (rather, it is usually only a few of them). This approach has the advantage that it relies on revealed preferences. Datasets of this kind are rarely available and researchers usually have to resort to indirect channels of popularity, like the use of hedonic housing price models (Rosen, 1974; Roback, 1982; see for example Glaeser et al., 2001; Glaeser and Gottlieb, 2006; Dalmazzo and de Blasio, 2011; Roback, 1982). Hedonic models estimate the consumer's willingness to pay to *live* close to amenities, not to *consume* them. This is of interest in many regards, but also misses a wide range of urban amenities the population might consider attractive and derive utility from, but would not necessarily want to live close to (e.g. areas with high density of restaurants, bars, etc. See for example Talen and Anselin, 1988).

Our results show that new sources of location-based data becoming available as a byproduct of technological advances can be used to perform urban economic and geographic research that is relevant not only for academics but also for policy makers and other urban stake-holders. We explicitly measure, document and characterize some of the inherent biases in these data, providing a better guide as to when information from these sources can be useful or rather misleading, and we demonstrate a potential application of the data that overcomes some of the disadvantages inherit in using more traditional data.

The remainder of the paper is structured as follows. Section 2 discusses in detail the dataset we use and compares it to conventional sources to explore the presence of geographical bias in the dataset. Section 4 describes our empirical strategy to explore the relationship between popular amenities and the local spatial composition. Section 5 presents the main results of this exercise. Finally, Section 6 concludes the paper.

2 Location-Based Services

Foursquare was created in March 2009 and, by the end of 2013, there were more than 45 million users who had checked in more than 5 billion times (Foursquare Inc., 2013). Its core mechanism is as follows: users sign in to an online site that allows them to post their location from an app that transmits the data into a central database that third-party developers can partially query through an Application Programming Interface (API).

This is the instrument we use to collect our dataset, as it allows us to obtain places (or venues, as they are called) in the surroundings of a specific location. Our approach to obtain a comprehensive list of Foursquare venues while remaining feasible and within Foursquare limits can be summarised as follows. During the month of June in 2013, we queried the Foursquare database from a grid of points equally spaced at 50 metres from each other in built-up areas of The Netherlands, and at 500 metres for the rest of the country. This is because the Foursquare API requires a single point to return nearby venues, and has a limit of data to be returned for every query, the rationale behind this was to be able to cover the entire extent of the country with a degree of resolution appropriate to capture all existing venues without querying more than needed in sparse regions. This returned information available from every venue in the vicinity of the pair of coordinates. Figure 1 displays the example of the famous Museum Square in Amsterdam, with the original grid points in black, used to query the API, as well as those returned in red, representing the location of venues.

After removing duplicates due to overlap of queries in parts of the country, we were left with around 800,000 unique places in The Netherlands for which we have access to the name, location, and category in the Foursquare classification, time on Foursquare, the total count of checkin's and the number of unique users who had checked in at the venue. In order to obtain a more consistent and complete set of checkin's count, we reran the query on the specific venues (not on the point grid) in August 2013, and use this dataset throughout the analysis. It is important to note the only information we have on the temporal dimension of this dataset is how long a given venue has been on Foursquare. Although there is a potentially very fruitful avenue of research in considering the different dynamics associated with the geography of checkin's, the current setup does not allow to explore it and hence we treat it as a cross-section.

Much of our in-detail analysis focuses on restaurants since they are non-tradable consumer goods and more closely mirror local demand and consumption (Waldfogel, 2008). This is crucial for our analysis, as it allows to identify preferences for local amenities. Our categorisation of restaurants relies on the labeling developed by Foursquare users and included in the collected dataset. This scheme, however, is very detailed (+80 labels for the category "Food") and results in most of them with very few restaurants. For the regression analysis, we simplify this detailed approach into two alternative aggregations: one that groups categories of restaurants based on the region of the cuisine, and one that splits restaurants in a similar way to how official data allow us to split residential population (i.e. Caribbean, Moroccan, Turkish, and other). Table 1 describes the number of restaurants in each category, and presents descriptive statistics of restaurants at the two aggregation schemes used: 500 meter buffers around each restaurant location; and the Buurt, the official enumeration unit adopted by the Dutch bureau of statistics (CBS).[3] A complete table with all the original categories is available in the appendix.

[Table 1 about here]

3 Coverage and bias of Foursquare dataset

There are legitimate concerns about the representativity of data extracted from LBS. A survey by the Pew Research Center (Zickuhr, 2013) on the use of these services in the US found that about 75% of adult smartphone owners use some form of LBS; this amounts to 45% of all adults in that country. More specifically, 12% of smartphone users check in using one of this services, which translates into 7% of all adults. In terms of the demographics of these users, the same report found few differences in the adoption of this practice among several population groups, with no statistical difference by gender or educational attainment. However, it did identify a significant leaning towards younger, suburban, and hispanic segments of the population. Income had mixed results with no clear trend.

Precisely establishing and documenting this bias requires accessing individual information and comparing it with overall population equivalents. As this degree of detail is not available to us, we explore this question by comparing the set of restaurants cataloged by Foursquare users with two official sources of statistical data. If the segment of the population who engage in LBS displays distinctive preferences from the overall population, some restaurants should not appear in the crowd-sourced list.

At this point, it should be noted this user-bias may, in some cases, represent an advantage for research. If this subgroup is composed by younger, better skilled, and more productive individuals, uncovering some of the patterns and stylized facts behind their urban consumption preferences should be of interest not only for researchers trying to understand what makes cities more attractive but also for policy makers and other stake holders that aim to attract this group (Florida, 2002, Moretti, 2004). The fact that users broadcast their location at particular venues and times can be interpreted as a special interest in their amenity choices. This results in a more informed picture of the amenity landscape in the city that allows them to more easily identify, and even determine, trending places before it becomes apparent to other residents. In addition, their behavior potentially influences others because each checkin is received by both their social network and, indirectly, other Foursquare users as a *recommendation* for a specific location.

In this section, we thus benchmark the collected data against two official sources to obtain the locations where the mismatches can potentially be more pronounced and hence help in a better understanding of the representativeness of Foursquare data.

Finding the right dataset to compare with is necessarily an imperfect endeavor. In fact, the very reason why we believe LBS data make a good research resource makes it a complicated one: there are no official sources to directly measure amenity popularity. This means we are not able to compare checkin activity. However, a tightly connected indicator is the mere presence of venues in the Foursquare database: locations so unpopular so as to not attract a single checkin will not be recorded. Since it is possible to obtain (somewhat crude) restaurant counts from official sources for the entire Netherlands, we compare the distribution of venues in the country, as portrayed by Foursquare, with building data from the cadaster and data from Statistics Netherlands. We employ different techniques that can best exploit the information provided by these sources. As the results will show, the final picture obtained is rather consistent.

[Figure 1 about here]

The national cadaster (BAG) provides access to the location and a few attributes of every single building in the country and the individual units that compose them. One of these characteristics provided is a rough indicator of the function they fulfill. We extract building units devoted to social gathering.[4] Although arguably imperfect, this is the closest match to our restaurant venues. Since the spatial location of these units is given at the point coordinate level, the best way to take advantage of this resolution is a method that does not require any aggregation (e.g. unit count at the neighbourhood level), such as kernel density estimation (KDE, see for example Rushton and Tiwari, 2009). This is a standard technique in point pattern analysis that essentially computes a probability surface for the location of points. In this context, KDE is a tool to help us compare the spatial distribution of the two sets of points.

Figure 2 (a) and (c) display the individual distributions of Foursquare restaurants and the BAG social gathering venues, respectively, using a color map based on the same range of values so it is directly comparable. Both maps capture the same general spatial arrangement. The Randstad area to the west appears clearly more populated, while the north of the country is rather empty. However, it also becomes clear both maps are not exactly equal. Upon visual inspection, it is possible to tell the area around Amsterdam is darker in the Foursquare map, while some parts of the north have a higher probability when one uses BAG data. To obtain a more clear comparison, Figure 2 (b) represents the difference between the two maps: BAG estimated probabilities are subtracted from Foursquare ones and plotted in a scale from -1 (for pixels where Foursquare probability is zero and that for BAG is 1) to 1 (the opposite case), effectively providing an indicator of Foursquare over-representation. The single area with the largest gap between both sources is clearly around Amsterdam. In this region, we find a notably higher density of Foursquare venues as compared to the underlying BAG benchmark. Differences in the rest of the country are much milder, with the center slightly leaning towards over-representation and the opposite holding true for the north and the bottom south.

The second comparison uses data from Statistics Netherlands (CBS). CBS provides accessibility measures for several types of urban amenities at the neighbourhood level, albeit in a slightly different way than simple counts: for each area, it is possible to obtain the average number of locations within three kilometers by road of all residents in an area. We use this neighbourhood index of accessibility to restaurants and set up a sensible comparison with our dataset, aggregated at the same spatial level[5]. Using neighbourhood counts of Foursquare venues, we estimate the following regression:

$$cbs_b = \alpha + \beta \, 4sq_b + \gamma \sum_k w_{bk} \, 4sq_k + u_b, \tag{1}$$

where cbs_b represents the CBS accessibility measure for neighbourhood b, $4sq_b$ the number of Foursquare restaurants in the same neighbourhood, and w_{bk} is the bk-th element of a matrix of spatial weights where $w_{bk} = 1$ if the centroid of neighbourhood j is closer than three kilometers to that of b, zero otherwise and $w_{bb} = 0$. In other words, this equation is predicting the Statistics Netherlands accessibility index with the combination of Foursquare locations in a given neighbourhood, plus those in 'roughly'[6] a three kilometer buffer.

The estimates we obtain for Equation 1 are $\beta = 0.54$ and $\gamma = 0.22$. Both are significant at the 0.1% level. They are also smaller than 1, pointing to, on average, a larger number of venues in the Foursquare dataset than in the CBS one. The critical aspect that we are interested in for this paper is the proportion of variation in the latter source that we are able to explain with the former. A close match between the two will point to a good coverage of the Foursquare data, at least as measured by the official statistics to which we have access. Even bearing in mind the mismatch produced by the differences in the exact definition of the variables, Foursquare data are able to explain more than 90% of the variation ($R^2 = 0.91$) in the Statistics Netherlands variable for restaurants.

By examining the error of the model, it is possible to characterize the potential bias in an alternative but complementary fashion to the one in Figure 2 (b) using KDE. The mismatch can be due to the methodological differences in calculating the variables outlined above, or to a true lack of proper coverage in the Foursquare dataset. Equally, poor alignment of the two variables may very well vary over space. To assess this situation, we explore the residuals of Equation 1.

The bottom panel of Table 2 shows the estimates of regressing $|u_b|$ on population, area and coordinates of the neighbourhoods. Using the absolute value of the residual, rather than the value with its original sign, allows us to model the degree of error in the prediction, irrespective of whether Foursquare data are over or underpredicting the CBS count. A complimentary exercise in this context would be to also model the sign of the error, to study potential differences between under- and overrepresentation of Foursquare venues in relation to CBS. We recognise the utility of this approach but, due to space constraints, do not tackle it in this context. Results are significant at the 0.1% level, although the explanatory power is substantially lower than in the previous model ($R^2 = 0.09$). Denser areas close to the west of the country, such as the dark blue Randstad area of the KDE map, tend to see higher disparity with the predicted values; equally, the higher the latitude of a neighbourhood, the more error in our model.

In sum, based on different but coinciding comparisons of Foursqaure data with two official sources, we can say checkins are biased with a marked geographical and socio-demographic dimension. Foursquare tends to be over-represented in the center of the country, where urban density, as well as a series of variables the literature has known to be associated with them (e.g. productivity, skills, income), is concentrated. Equally, it is under-represented in more remote, rural areas. These findings are in line with previous research on location-based services that characterizes them as an urban (Hecht and Stephens, 2014) phenomenon particularly focused on a younger, more educated, layer of the population (Zickuhr, 2013).

We believe Foursquare restaurants (and information about their popularity) can offer interesting insights of relevance for a series of debates. The group it represents, which we may call the *trend-setters*, are of interest in themselves as there is an interest in local and regional policy to attract them. We also argue that their more visible position, on- and offline, may exert an effect on other sectors of the population – setting trends – and hence it is also of interest to understand their behaviour as an anticipation of wider preferences.

4 An empirical exercise: diversity & popularity

To demonstrate some of the advantages of LBS data, we perform an exploration of the determinants of restaurant popularity. In doing so, we are able to further explore what makes certain areas of a city attractive. Restaurant differentiation is used more often as a measure of local product differentiation (Waldfogel, 2008, Mazzolari and Neumark, 2012, or Schiff, 2015) but research focusing on the consumer utility from product heterogeneity is however scarce and can be explored with a dataset that is obtained from a LBS such as Foursquare.

We run a set of regressions at the level of the individual restaurant. In this case, we benefit from the availability of high spatial resolution and exploit variation within 500 meter of the immediate surroundings of the restaurant. This is necessarily an arbitrary threshold that could give rise to issues of scale and modifiable areal unit problem (MAUP). However, in additional regressions, we modified the radious finding largely the same results. We can analyze issues such as the correlation between co-location of restaurants of similar characteristics and bigger traction, but also general spatial economic processes such as agglomeration externalities. A diversity bonus, the so-called Jacobs externalities, could be interpreted in terms of the benefit of clustering of different industries (types of restaurants in this case). Equally, a concentration effect, consistent with Marshall-Arrow-Romer (MAR) type of externalities, might lend support for the localization argument, in which it is only industries in the same sector that benefit from being close to each other because consumers have lower search costs, or they can benefit from shared (labor) inputs.

At the level of individual restaurants, we can measure the concentration (C_i) for restaurant i, defined as the proportion of all the restaurants in its surroundings that belong to the same ethnic group g_i:

$$C_i = \frac{\sum_{j=1}^{N} w_{ij} \times K(g_i)_j,}{\sum_{j=1}^{N} w_{ij}} \tag{2}$$

where N is the total number of restaurants; K is an indicator function that returns 1 if restaurant j is in the same group g_i and 0 otherwise; and w_{ij} the ij-th element of a spatial weights matrix that assigns 1 if i and j are neighbors (including $i = j$, contrary to common practice in the use of these matrices) and 0 otherwise. We use distance band weights that consider every observation within a 500m radious a neighbor. Only the satisfaction of both constraints, spatial and ethnic, results in restaurant j increasing the local concentration index C_i for location i.

Additional CBS data is used to incorporate socio-economic measures and residential population composition at the neighbourhood (buurt) level.[7] Neighbourhoods are the most spatially detailed unit for which this kind of data are available publicly, providing the best possible match between Foursquare venues and characteristics of their surroundings. We include population breakdown by country of origin, following the main groups established by the statistical agency (i.e. Moroccan, Turkish, Caribbean, and other).

We consider residential and restaurant diversity in a given part of the city. These variants of diversity aim to capture two of the main factors we hypothesize influence the popularity of an area: the range of available alternatives (restaurants) and its 'ethnic profile' (residents). The groups we use to categorize residents and restaurants are based on nationality and the ethnic categories developed in Table 1 (and fully available in Table A in the appendix), respectively. Similarly, we use the best possible unit of analysis to measure diversity. This means that restaurant diversity, built from individual data, can be calculated using the 500m buffer area; while residential diversity is calculated at the buurt level. We use a common fractionalization index to calculate diversity (Mauro, 1995). The index represents the probability that two observations randomly selected from neighbourhood b belong to different groups. It is bounded $0 < \text{frac}_b < 1 - 1/G$, where G is the total number of groups, so the closer to one the more diverse an area.

The data presented above is combined in an equation that describes popularity of a restaurant as a (linear) function of several of its own characteristics as well as additional features of its location:

$$log(ch_i) = F_i'\alpha_v + S_l'\beta_v + \gamma_v\text{frac}_l + \delta_v C_l + B_b'\kappa_v + \lambda_v\text{frac}_b + \mu_{m-v} + \eta_v\left(F_i \times B_b\right) + u_i, \quad (3)$$

where ch_i is the total volume of checkin's in restaurant venue i, transformed to its log so interpretation of the parameters can be in terms of percentage changes, F_i is a vector of venue characteristics, S_l is a vector with counts of other Foursquare venues within 500 meter surrounding i, frac_l and C_l are the factionalization index of restaurants and the concentration of restaurants, respectively, within 500 meter surrounding i, B_b is a vector of neighbourhood variables for area b where i is located, frac_b is the fractionalization index of the population in the neighbourhood b in which i is located, μ_{m-v} are municipality or area fixed effects, u_i is the error term, and the subscript v in each of the parameters implies the estimates relate to the venue level regression.

In a second set of regressions we include an interaction term of the restaurant's ethnic category and the proportion of population with that ethnic background in the neighbourhood to test whether restaurants of a specific cultural background will benefit[8] from being in areas with a higher population with that specific background (e.g. Turkish restaurants will be more popular in neighbourhoods with more Turkish resident population) because of the presumably more demanding and experienced set of customers. The equation we present is non-spatial in the sense that we assume independence between observations irrespective of where they are located. We recognise that, given the nature of our dataset, it is possible that spatial autocorrelation is present. However, to keep the analysis simple and illustrative, we do not show spatial estimates.[9]

5 Results

The analysis of the neighbourhood components that relate to the popularity of a restaurant allows us to control for many unobserved characteristics of neighbourhoods, such as other amenities or aesthetic views and buildings, and unobserved characteristics of restaurants that could drive the above found results. The regression results for the individual restaurant checkin's are given in Table 3. A municipality fixed effect is included in each regression. The first regression additionally includes fixed effects for the region from which the cuisine originates, as designated in Table 1. This controls for any variation due to unobserved characteristics shared by al restaurants in a given category. For example, if European restaurants tend to be higher-end than their Asian counterparts and that translates in their checkin volume, or tend to attract more checkin's because they are more likely to be on Foursquare, this will be captured by the cuisine/region fixed effect. In the second regression we do not include a region category fixed effect, but look at the ethnic origin of the cuisine of restaurants in relation to local presence of the population from the same ethnic origin.

The positive coefficient of the average taxation value of the dwellings surrounding the restaurant –a proxy for house prices– suggests that restaurants are more popular, in the sense of attractive and preferred, in expensive areas. Or popular areas, of which the number of restaurant check-ins is a proxy, become more expensive, highlighting the role of neighbourhood characteristics on restaurant popularity. If local housing prices signal local income levels, and local population preferences mirror the local availability of non-tradable consumer goods, this is an indication that more expensive restaurants tend to be more popular. This results is what is generally to be expected based on urban economics literature. The negative coefficient on the fast food dummy points into the same direction. For individual restaurant checkin's, the local population composition plays a small negative role. Restaurants in areas with higher shares of Moroccans and Turks tend to be, on average, less popular. Population diversity plays no role for the number of checkin's, so the sheer presence of immigrants does not make restaurants more attractive.

[Table 3 about here]

Our results suggest restaurants are more popular in areas that offer both greater diversity of choice *and* a higher concentration of their own type. The effect for the latter is smaller than the former. A standard deviation increase in the diversity of the number of surrounding restaurants is associated with 5% increase in the volume of checkin's into a given restaurant. At the same time, if the concentration of restaurants of the same ethnic background in the surrounding of a restaurant increases by a standard deviation, the volume of checkin's into that restaurant goes up by 2.5%, all other things equal. These results again support the findings of Glaeser et al. (2001) and Fujita et al. (1999) that consumers value diversity of products. Our results may be considered as indirect evidence of the mechanism described in Ottaviano and Peri (2006), who state that product diversity brought along by immigrants, like restaurants, can increase consumer utility. They can also be interpreted along the lines of Glaeser et al. (1992)'s agglomeration externalities: diversity leads to positive agglomeration externalities (Jacobs) while concentration of the same type of restaurants translates into externalities in production (MAR), or externalities from competition between them (Porter). Both MAR and Porter-type processes lead to positive externalities of concentration of the same firms but we cannot distinguish between these effects. However, the negative coefficient of the total number of restaurants in the local surrounding seems to underwrite the hypothesis that restaurants compete for consumers' time when they are located nearby. The overall effects are not very large in size but are significant and non-negligible, showing that the local composition of supply and its spatial arrangement has an influence on a restaurant's popularity.

Results in the second column of Table 3 closely resemble those in the first one, but allow us to obtain an additional insight into whether restaurants with specific ethnic backgrounds are more popular if located in an area with higher population of such background. The literature has identified several mechanisms that can be at work in this context. On the demand side, a more rigorous and experienced consumer can lead to more competition that pushes quality up. For example, a neighbourhood with a larger proportion of Turks could have better Turkish food because residents are not content with mediocre options and hence restaurants need to improve if they want to survive. Adopting a supply perspective provides an alternative mechanism (Mazzolari and Neumark, 2012). If restaurant owners tend to live nearby their businesses and populations with a given background are expected to have better skills at cooking their traditional cuisines, then a similar outcome would arise. A classical example in this line would be China towns, where a larger share of Chinese residents leads to finer quality (and range of choices) in Chinese food.

All other things equal, a Caribbean restaurant receives 35% less checkin's than other restaurants, and that percentage goes up to 44% for a Moroccan one.[10] Interestingly, the contrary seems to be true for Turkish restaurants: venues of this kind located in areas with higher Turkish population experience a checkin premium, albeit rather small, as evidenced by the significant interaction term.

6 Conclusion

This paper contributes to an emerging literature building evidence about the relevance of data derived from location-based services like Foursquare, and how these can be linked with conventional sources to study consumer behavior or local consumption patterns. Although the data we use is most likely not representative of the entire population, as users are not randomly selected, the type of information obtained is valuable for researchers, urban planners and municipalities. This is of particular relevance given the increasing importance of local consumer amenities and their role in city development and policy making.

The research in this paper shows that data from LBSs can be used for urban and regional economic analysis of, for example, horizontal product differentiation of restaurants. Insights into local amenity composition based on quality or prices, and the use of other consumer products and local relative consumer good composition would be valuable extensions for future research. Our analysis of the individual restaurants already deals with the drawback of predetermined administrative areas by composing local areas based on a fixed radius. However, part of the analysis of popular areas should take into account that an area is not necessarily a predefined administrative region but, especially when looking at dynamic phenomena such as popularity, should be endogenously determined. As this is an exercise worth the entire focus of a project, we warrant this for future research.

Notes

[1]In computer engineering LBSs data is used to uncover patterns of mobility and use of urban landscapes. See, for example, also Karamshuk et al. (2013); Hasan and Ukkusuri (2014, 2015).

[2]Lindqvist et al. (2011) find that one of the main reasons why people check in is to use "the history of places you go as a form of presentation of self". One of the main reasons people do not check in at some places is if the venue would make them feel embarrassed.

[3]The average population of a buurt is typically around a couple of thousands. However, in very urban areas the number is much higher and in rural areas much lower.

[4]Those under the category `bijeenkomstfunctie`.

[5]In particular, the CBS variable employed is `AV3_RESTAU`. It is important to note that, the way the variable is constructed, it is not possible for us to exactly replicate it with our Foursquare data, as that requires to know the exact location of every individual within a neighbourhood to calculate the number of restaurants within three kilometres and to then average those across areas. Our approach, although introducing some level of noise, attempts to closely mirror the variation in the original variable.

[6]This is indeed less accurate than the CBS measure because we take the geographical centroid, without using any population weighting scheme and either include the entire neighbourhood, or discard it, while CBS is effectively including only that part of adjacent neighbourhoods exactly within three kilometers.

[7]The geographical area (`'buurt'`) is used with the data from 2010. Statistics Netherlands files are available from: http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2011/default.htm

[8]As already mentioned, at this stage, we cannot distinguish a causal link from sorting or a 'survival of the fittest' type of process. Hence, a positive and significant coefficient in this term could equally point to an effect of locating in neighbourhoods with higher populations of the same cultural background, or simply signaling that those located there hold unobserved characteristics that make them more popular.

[9]In ancillary regressions, we included a spatial lag and a spatial HAC correction in the estimation of the variance-covariance matrix but found largely comparable results. Results are available from the authors.

[10]Calculated as $100 \times \left(\exp^{-0.43} - 1\right)$. See Halvorsen and Palmquist (1980) for the interpretation of dummy variables if the dependent variable is log-transformed.

References

Arribas-Bel, D.: 2014, Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities, *Applied Geography* **49**, 45–43.

Cheng, Z., Caverlee, J., Lee, K. and Sui, D. Z.: 2011, Exploring Millions of Footprints in Location Sharing Services, *Proceeding of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Barcelona.

Cranshaw, J., Schwartz, R., Hong, J. and Sadeh, N.: 2012, The livehoods project: Utilizing social media to understand the dynamics of a city, *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM*, Vol. 12.

Dalmazzo, A. and de Blasio, G.: 2011, Amenities and skill-biased agglomeration effects: Some results on Italian cities, *Papers in Regional Science* **90**(3), 503–527.

Davis, D.R., Dingel, J.I., Monras, J. and Morales, E.: 2017, How Segregated is Uran Consumption?, *NBER Working Paper* **23822**.

Florida, R.: 2002, The economic geography of talent, *Annals of the Association of American Geographers* **92**(4), 743–755.

Foursquare Inc.: 2013, Ending the year on a great note (And with a huge thanks and happy holidays to our 45,000,000-strong community). Foursquare Blog (accessed on Jan. 22nd, 2014), http://blog.foursquare.com/post/70494343901/ending-the-year-on-a-great-note-and-with-a-huge-thanks.

Foursquare Inc.: 2015, Foursquare.com.
URL: *http://www.foursquare.com*

Frith, J.: 2014, Communicating through location: The understood meaning of the foursquare check-in, *Journal of Computer-Mediated Communication* .

Fujita, M., Krugman, P. and Venables, A. J.: 1999, *The Spatial Economy: cities, regions, and international trade*, The MIT Press, Cambridge.

Glaeser, E., Kolko, J. and Saiz, A.: 2001, Consumer city, *Journal of Economic Geography* **1**(1), 27–50.

Glaeser, E. L. and Gottlieb, J. D.: 2006, Urban Resurgence and the Consumer City, *Urban Studies* **43**(8), 1275–1299.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A. and Shleifer, A.: 1992, Growth in cities, *Journal of Political Economy* **100**(6), 1126–1152.

Goodchild, M.: 2007, Citizens as sensors: the world of volunteered geography, *GeoJournal* **69**(4), 211–221.

Halvorsen, R. and Palmquist, R.: 1980, The interpretation of dummy variables in semilogarithmic equations, *The American Economic Review* **70**(3), 474–475.

Hasan, S. and Ukkusuri, S.V.: 2014, Urban Activity Pattern Classification Topic Models from Online Geo-Location Data, *Transportation Research Part C: Emerging Technologies* **44**, 363–381.

Hasan, S. and Ukkusuri, S.V.: 2015, Location Contexts of User Check-Ins to Model Urban Geo Life-Style Patterns, *PLOS ONE* **10** (5).

Hecht, B. and Stephens, M.: 2014, A Tale of Cities: Urban Biases in Volunteered Geographic Information, *Proceedings of the International AAAI Conference on Weblogs and Social Media, ICWSM*, Menlo Park, CA: AAAI Press.

Jacobs, J.: 1961, *The death and life of great American cities*, Random House LLC.

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C.: 2013, Geo-Spotting: Mining Online Location-Based Services for Optimal Retail Store Placement, *ACM*, 793–801.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Alstyne, M. V.: 2009, Computational Social Science, *Science* **323**, 721–723.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J. and Zimmerman, J.: 2011, I'm the mayor of my house: Examining why people use foursquare - a social-driven location sharing application, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM, New York, NY, USA, pp. 2409–2418.

Ludford, P. J., Priedhorsky, R., Reily, K. and Terveen, L.: 2007, Capturing, sharing, and using local place information, *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 1235–1244.

Mauro, P.: 1995, Corruption and Growth, *The Quarterly Journal of Economics* **110**(3), 681–712.

Mazzolari, F. and Neumark, D.: 2012, Immigration and product diversity, *Journal of Population Economics* (25), 1107–1137.

Miller, H. J.: 2010, The data avalanche is here. Shouldn't we be digging?, *Journal of Regional Science* **50**(1), 181–201.

Moretti, E.: 2004, Chapter 51 human capital externalities in cities, *in* J. V. Henderson and J.-F. Thisse (eds), *Cities and Geography*, Vol. 4 of *Handbook of Regional and Urban Economics*, Elsevier, pp. 2243 – 2291.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C.: 2012, A Tale of Many Cities: Universal Patterns in Human Urban Mobility, *PLOS ONE* **7**(9).

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C.: 2012, A Random Walk around the City: New Venue Recommendation in Location-Based Networks, *Privacy, Security, Risk, and Trust (PASSAT)* 144–153.

Ottaviano, G. and Peri, G.: 2006, The economic value of cultural diversity: evidence from US cities, *Journal of Economic Geography* **6**(1), 9–44.

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S. and Waller, T.S.: 2017, Exploring the Capacity of Social Media Data for Modeling Travel Behaviour: Opportunities and Challengers, *Transportation Research Part C: Emerging Technologies* **75**, 197–211.

Roback, J.: 1982, Wages, Rents, and the Quality of Life, *The Journal of Political Economy* **90**(6), 1257–1278.

Rosen, S.: 1974, Hedonic Prices and implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy* **82**(1), 34–55.

Rushton, G. and Tiwari, C.: 2009, Spatial filtering/kernel density estimation, *in* R. Kitchin and N. Thrift (eds), *International Encyclopedia of Human Geography*, Elsevier, Oxford, pp. 359 – 364.

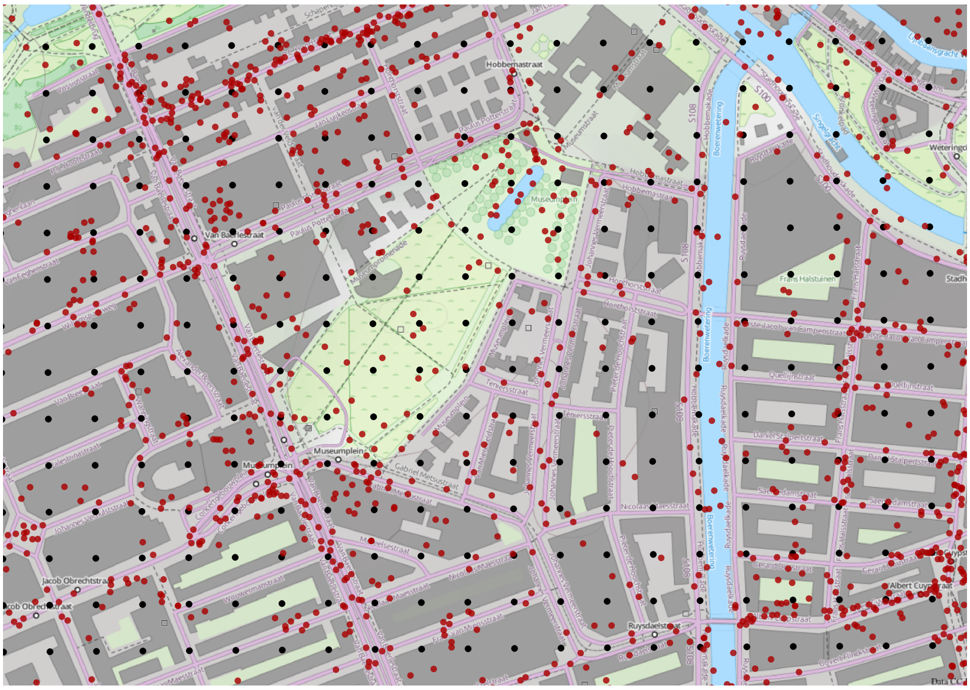
Schiff, N.: 2015, Cities and Product Variety: Evidence from Restaurants, *Journal of Economic Geography* **15**(6), 1085–1123.

Talen, E., and Anselin, L., 1998. Assessing spatial equity: an evaluation of measures of accessibility to public playgrounds. *Environment and planning A*, **30**(4), 595-613.

Waldfogel, J.: 2008, The median voter and the median consumer: Local *private* goods and population composition, *Journal of Urban Economics* (63), 567–582.

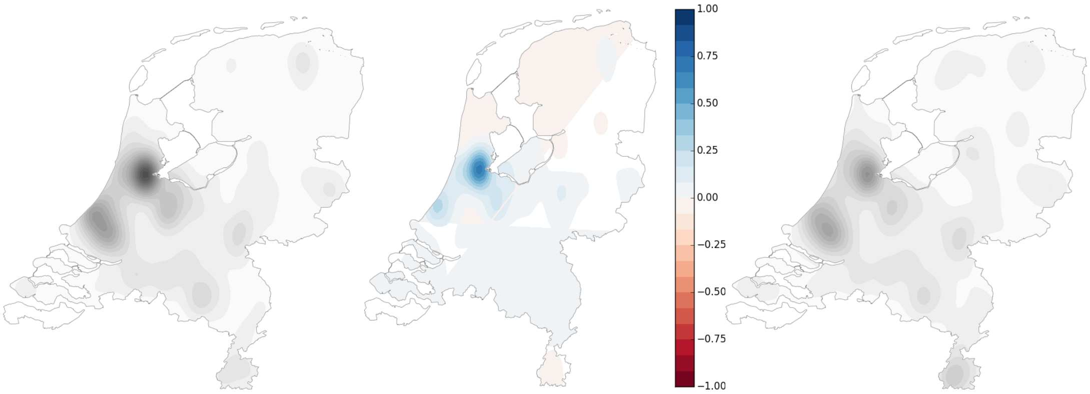
Zickuhr, K.: 2013, Location-based services, *Pew Internet and American Life Project* .

A Figures and Tables



NOTE: Black dots correspond with locations in the grid used to query the Foursquare database. Red dots represent Fourquare venues obtained as a response to the queries. Background data come from OpenStreetMap and are available under a CC-By-SA license.

Figure 1: Foursquare data collection. Amsterdam's Museum Square.



(a) Foursquare venue KDE (b) Density Difference (c) BAG KDE

Figure 2: Foursquare bias

Tables

	Count	Pct.	Max	Mean	Min	Std.
Total	50,847	100				
500m. Buffer			685	68.100	1	101.68
Buurt			547	9.817	1	21.083
Regional classification						
African	362	0.712	9	0.498	0	1.028
Asian	1,216	2.391	22	1.968	0	3.505
Central America	429	0.844	15	0.932	0	1.903
Central Asian	331	0.651	14	0.797	0	1.824
East Asia	3,693	7.263	74	5.180	0	9.279
East Europe	145	0.285	4	0.220	0	0.563
European	8,452	16.622	152	14.200	0	23.362
Middle East	1,557	3.062	27	2.647	0	4.238
North America	10,346	20.347	77	9.883	0	12.417
Other	23,755	46.719	296	30.380	0	44.775
Pacific	27	0.053	2	0.051	0	0.227
South America	534	1.050	35	1.349	0	3.963
Residential classification						
Caribean	207	0.481	7	0.328	0	0.771
Moroccan	131	0.304	6	0.258	0	0.655
Turkish	354	0.822	6	0.510	0	0.930
Other	42,374	98.393	488	0.032	0	3.429

Table 1: Descriptive statistics of restaurant classifications used in the analysis

	Restaurants
Constant	-4.29***
4sq venues	0.54***
W 4sq venues	0.22***
R^2	0.91
Constant	6.46***
Population	19.86***
Area	-21.58***
Easting	-0.35***
Northing	0.14***
R^2	0.09
N	11,151

Upper panel shows estimates from Eq. 1; the absolute value of its residuals is used as dependent variable in the bottom panel. Explanatory variables in the bottom regressions are rescaled to per 10,000 units to obtain more readable coefficients.

Table 2: Comparison CBS-Foursquare

A Foursquare categories

	categorynm	categorycnt	categoryff
0	Other	Other	Non Fast Food
1	Afghan Restaurant	Central Asian	Non Fast Food
2	African Restaurant	African	Non Fast Food
3	American Restaurant	North America	Non Fast Food
4	Arepa Restaurant	South America	Non Fast Food
5	Argentinian Restaurant	South America	Non Fast Food
6	Asian Restaurant	Asian	Non Fast Food
7	Australian Restaurant	Pacific	Non Fast Food
8	BBQ Joint	North America	Non Fast Food
9	Bagel Shop	Other	fastFood
10	Bakery	Other	Non Fast Food
11	Brazilian Restaurant	South America	Non Fast Food
12	Breakfast Spot	Other	Non Fast Food

13	Brewery	Other	Non Fast Food
14	Burger Joint	North America	fastFood
15	Burrito Place	Central America	fastFood
16	Caf	European	Non Fast Food
17	Cajun / Creole Restaurant	South America	Non Fast Food
18	Caribbean Restaurant	Central America	Non Fast Food
19	Chinese Restaurant	East Asia	Non Fast Food
20	Coffee Shop	Other	Non Fast Food
21	Cuban Restaurant	Central America	Non Fast Food
22	Cupcake Shop	Other	Non Fast Food
23	Deli / Bodega	Other	Non Fast Food
24	Dessert Shop	Other	Non Fast Food
25	Dim Sum Restaurant	East Asia	Non Fast Food
26	Diner	North America	Non Fast Food
27	Distillery	Other	Non Fast Food
28	Donut Shop	North America	fastFood
29	Dumpling Restaurant	East Asia	Non Fast Food
30	Eastern European Restaurant	East Europe	Non Fast Food
31	Ethiopian Restaurant	African	Non Fast Food
32	Falafel Restaurant	Middle East	fastFood
33	Fast Food Restaurant	North America	fastFood
34	Filipino Restaurant	East Asia	Non Fast Food
35	Fish & Chips Shop	European	fastFood
36	Food Truck	North America	fastFood
37	French Restaurant	European	Non Fast Food
38	Fried Chicken Joint	North America	fastFood
39	Gastropub	Other	Non Fast Food
40	German Restaurant	European	Non Fast Food
41	Gluten	Other	Non Fast Food

42	Greek Restaurant	European	Non Fast Food
43	Hot Dog Joint	North America	fastFood
44	Ice Cream Shop	Other	Non Fast Food
45	Indian Restaurant	Central Asian	Non Fast Food
46	Indonesian Restaurant	East Asia	Non Fast Food
47	Italian Restaurant	European	Non Fast Food
48	Japanese Restaurant	East Asia	Non Fast Food
49	Juice Bar	Other	Non Fast Food
50	Korean Restaurant	East Asia	Non Fast Food
51	Latin American Restaurant	South America	Non Fast Food
52	Mac & Cheese Joint	North America	fastFood
53	Malaysian Restaurant	East Asia	Non Fast Food
54	Mediterranean Restaurant	Middle East	Non Fast Food
55	Mexican Restaurant	Central America	Non Fast Food
56	Middle Eastern Restaurant	Middle East	Non Fast Food
57	Molecular Gastronomy Restaurant	Other	Non Fast Food
58	Mongolian Restaurant	Central Asian	Non Fast Food
59	Moroccan Restaurant	African	Non Fast Food
60	New American Restaurant	North America	Non Fast Food
61	Peruvian Restaurant	South America	Non Fast Food
62	Pizza Place	Other	fastFood
63	Portuguese Restaurant	European	Non Fast Food
64	Ramen / Noodle House	East Asia	fastFood
65	Restaurant	Other	Non Fast Food
66	Salad Place	Other	Non Fast Food
67	Sandwich Place	Other	fastFood
68	Scandinavian Restaurant	European	Non Fast Food
69	Seafood Restaurant	Other	Non Fast Food
70	Snack Place	Other	fastFood

71	Soup Place	Other	fastFood
72	South American Restaurant	South America	Non Fast Food
73	Southern / Soul Food Restaurant	North America	Non Fast Food
74	Spanish Restaurant	European	Non Fast Food
75	Spanish Restaurant	European	Non Fast Food
76	Steakhouse	North America	Non Fast Food
77	Sushi Restaurant	East Asia	Non Fast Food
78	Swiss Restaurant	European	Non Fast Food
79	Taco Place	Central America	fastFood
80	Tapas Restaurant	European	fastFood
81	Tea Room	Other	Non Fast Food
82	Thai Restaurant	East Asia	Non Fast Food
83	Turkish Restaurant	Middle East	Non Fast Food
84	Vegetarian / Vegan Restaurant	Other	Non Fast Food
85	Vietnamese Restaurant	East Asia	Non Fast Food
86	Winery	Other	Non Fast Food
87	Wings Joint	North America	fastFood
88	Yogurt	Other	Non Fast Food

Table 4: Regional aggregation of Foursquare original categories

Table 3: Regression results individual restaurants

	Dependent variable: restaurant checkin's	
	(1)	(2)
Population size (× 10,000)	0.013	0.011
Population density	−0.068***	−0.069***
Share social rent houses	−0.062***	−0.06***
Share owner occupied houses	−0.1***	−0.102***
Average taxation value (× €100,000)	0.045***	0.043***
Share of Caribbeans	0.01	−0.007
Share of Moroccans	−0.024**	−0.03**
Share of Turks	−0.026**	−0.039***
Population fractionalization		−0.019
Caribbean restaurant (dummy)		−0.43***
Moroccan restaurant (dummy)		−0.582***
Turkish restaurant (dummy)		0.006
Restaurant time on Foursquare	1.035***	1.036***
Fast food restaurant (dummy)	−0.052***	−0.069***
Number of arts & entertainment venues	−0.079**	−0.086**
Number of restaurants	−0.348***	−0.355***
Number of nightlife spots	0.049	0.046
Number of outdoors & recreation spots	−0.0	−0.003
Number of shops & other services	0.186***	0.186***
Number of travel & transport locations	0.048*	0.046*
Total foursquare locations	0.267**	0.285***
Concentration index	0.025***	0.033***
Restaurant fractionalization	0.052***	0.054***
Carribean restaurant × Share of Caribbeans		0.018
Moroccan restaurant × Share of Moroccans		−0.08
Turkish restaurant × Share of Turks		0.091**
Municipality fixed effects	Yes	Yes
Restaurant ethnic origin fixed effect	Yes	No
Adjusted R^2	0.354	0.352
N	43066	43066

[a] * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All independent continuous variables are standardized. The dependent variable is in logs.