Article

The Geography of Taste: Using Yelp to Study Urban Culture

Sohrab Rahimi [1,*], Sam Mottahedi [2] and Xi Liu [3]

[1] Pennsylvania State University; sur216@psu.edu
[2] Pennsylvania State University; s.mottahedi@psu.edu
[2] Pennsylvania State University; xiliu@psu.edu
* Correspondence: sur216@psu.edu; Tel.: +1-781-2965152

Abstract: This study aims to put forth a new method to study the socio-spatial boundaries by using georeferenced community-authored reviews for restaurants. In this study, we show that food choice, drink choice, and restaurant ambience can be good indicators of socio-economic status of the ambient population in different neighborhoods. To this end, we use Yelp user reviews to distinguish different neighborhoods in terms of their food purchases and identify resultant boundaries in 10 North American metropolitan areas. This data-set includes restaurant reviews as well as a limited number of user check-ins and rating in those cities.

We use Natural Language Processing (NLP) techniques to select a set of potential features pertaining to food, drink and ambience from Yelp user comments for each geolocated restaurant. We then select those features which determine one's choice of restaurant and the rating that he/she provides for that restaurant. After identifying these features, we identify neighborhoods where similar taste is practiced. We show that neighborhoods identified through our method show statistically significant differences based on demographic factors such as income, racial composition, and education. We suggest that this method helps urban planners to understand the social dynamics of contemporary cities in absence of information on service-oriented cultural characteristics of urban communities.

Keywords: Volunteered Geographic Information (VGI), Yelp, Natural Language Processing (NLP), Machine Learning, Cultural Boundaries, Consumption Behavior, Urban Computation, GIS, Word2Vec

1. Introduction

Socio-economic polarization is a defining characteristic of cities in the global economy [1], [2]. In global markets where economic regulations are minimized, social polarization is an inevitable consequence given the relatively small proportion of the population involved in this growing affluence [3]. In case of the U.S., this social polarization is also ethnic/racial as the prosperous economy in the U.S. was accompanied by massive immigration waves from other countries adding more dimensions to the long-lasting Black and white dichotomy. Not surprisingly, immigrants targeted large cities where most industries were located at and this, in part, led to more diversity in urban population. The multitude of cultural/ethnic groups led to cultural polarization and fragmentation of these global cities where every ethnic group occupied a piece of land [4]. Therefore, the American metropolis is plagued by both cultural and economic polarization [3].

During the past four decades, the debate over the definition and qualities of urban communities in developed countries grew significantly. Overall, scholars have different opinions regarding the strength of communities. Some believe that the notion of community is lost, some believe it has not

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 376; doi:10.3390/ijgi7090376

changed significantly and other say that it's been liberated from their constraints [5]. However, many of the recent studies have shown that the liberated hypothesis is more representative of the state of modern communities[5]–[7]. These studies assert that telecommunication and mobility has encouraged dispersed networks of friendship, kinship or communities of interest. Under this condition, the individual's network is a personal choice that she is free to choose from. Even though telecommunication has facilitated broad networks over space, the spatial segregation instigates sharp borders between communities in American cities. Emphasis on diversity and seeing the city as a melting pot, which is championed by postmodern thinking, has not addressed the gaps between ethnic and economic groups [8].

Many studies have attempted to fathom the socio-spatial complexities that emerged in post-war American cities. Most of classic studies of this kind were based on the Census data [9], [10]. Although the U.S. Census data provide valuable information about cities, these data hardly inform us about lifestyles, consumption behavior, cultural factors, and space-use patterns. The past two decades have seen a rapid advancement in the field of urban and social studies partly due to emergence of new crowd-sourced data sources and computation techniques [11]. The new data sources have enabled the researchers to go beyond basic demographics such as race and income and delve into a multitude of socio-spatial phenomena in modern cities. This study aims to contribute to this line of studies by proposing taste as an indicator of social status which integrates different facets of culture, economy and social networks of urban inhabitants. We argue that using businesses as sensors can provide new insight into the intricate social structure of the American metropolis. More specifically, this research aims to answer the following questions:

- To what extent is taste a good indicator of socio-economic status of communities in American cities?
- By utilizing the concept of taste, can we use restaurant-as-sensor instead of citizen-as-sensor [12] to examine the socio-economic dynamics of neighborhoods without having the User IDs? This issue is especially important to us since business data is far more accessible and plentiful than individual-level data [13].
- Are American cities comprised of regions with different dominant taste cultures [4]? Are different regions in every city similar to regions from other cities [2]?

2. Materials and Methods

2.1 Literature review

2.1.1 Previous attempts to define socio-spatial boundaries

Recently, many studies have addressed these problems by using heterogeneous data sources that are updated frequently and exist at the scale of buildings or individuals [11]. Some investigate the communities on a large scale. For example, one study used vehicle GPS traces in Pisa, Italy to build a network and used community detection algorithms (i.e. Infomap) to identify non-overlapping communities of people at the county and municipality scale [14]. A similar study was conducted on a larger scale in Great Britain using telecommunications data [15]. Recently, detecting communities on urban scale has been more popular. For example, one study uses human mobility between different regions and the Points of Interest (POI) data to find the dominant functions of each urban

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

region using topic-based inference model [16]. Using this model, this research identifies nine functional regions using clustering techniques.

Most often, urban studies that use crowd-sourced data to study the socio-spatial structure of cities incorporate Location-Based Social Network (LBSN) techniques, that is, a network consisting of people in a social structure who share location-embedded information [11] . Much of research in this area uses social media data which includes the geographical location as well as their tagged images, videos, and texts. Common examples of data used for LBSNs include GPS trajectories of taxis, Twitter, Call Data Records (CDRs), Flickr geo-tagged photos, and Foursquare check-in data. Georeferenced crowd-sourced data such as tweets, photos, and check-ins can help understand people's lifestyles (e.g. likes and dislikes) [17], [18], cities' socio-spatial structure [19], neighborhood functions and characteristics [20] and behavioral patterns [21] in cities.

One of the common techniques for studying urban structure is identifying similarities between users in terms of their use of urban spaces [19], [22]–[25]. For example, among the most well-known studies of this kind is the Livehoods project, which uses check-in data to identify the zones where their establishments (e.g. restaurants and bars) share similar clientele [19]. This study uses 18 million check-ins collected from Foursquare, a location-sharing service where users share their location by checking in via their smart phones. By using clustering techniques, this study identifies clients with similar points of interest (POI). In another study, the authors studied the semantics of different locations by analyzing different categories of POIs in many neighborhoods [25].

Although the state of the art techniques used in these studies have dramatically improved our understanding of cities, they still have some limitations. First, accessing data that include individuals' behavior is often hard and these data are not freely available to the public. For example, companies which maintain a great inventory of georeferenced social networks do not share such information due to privacy issues. Second, the data is not often representative of the entire population. For example, not everyone has a Foursquare account and not all those account owners use Foursquare every time they visit a place [19]. Third, these studies only address one aspect of an individual's life, for example, Foursquare only covers check-in data and points of interest (POIs), and taxi data cover some travel patterns. While these data-sets have proven helpful, multiple data sources need to be fused to provide an understanding of urban lifestyle.

2.1.2 Taste as an indicator of urban culture

The social construct of taste is a well-studied topic especially in the age of Internet where individual preferences are available to information-based companies (e.g. Amazon, Facebook, Spotify). In fact, many recommender systems (i.e. algorithms made for recommending products to users) are designed under the same assumption that people of same social groups are likely to consume similar products [26], [27]. The underlying mechanism of the relationship between social groups and taste was discussed by Pierre Bourdieu in his well-recognized book Distinction: A Social Critique of the Judgment of Taste [28].

In Bourdieu's view, both cultural and economic capital are the most important forms of capital. Economic capital has to do with individuals' access to economic resources while cultural capital is a collection of non-material traits in a person, such as knowledge and skills, attitudes, philosophical

 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 376; doi:10.3390/ijgi7090376

views, use of vocabulary, and language skills. Bourdieu believes that taste is the means of identifying class distinction. He argues that these differences are most obvious in the routine everyday choices in taste of food, furniture, and clothing as they are representative of the pure taste. For example, he argues, children of a lower social status like plentiful and good meals while those of higher status go for original and exotic. These choices, according to Bourdieu, become intrinsic to one's personality and thereon he/she rejects the tastes of other groups. Bourdieu argues that high-taste is characterized by how far it is from pure necessities. The upper classes in this regard use taste as the ideal weapon in strategies of distinction [28].

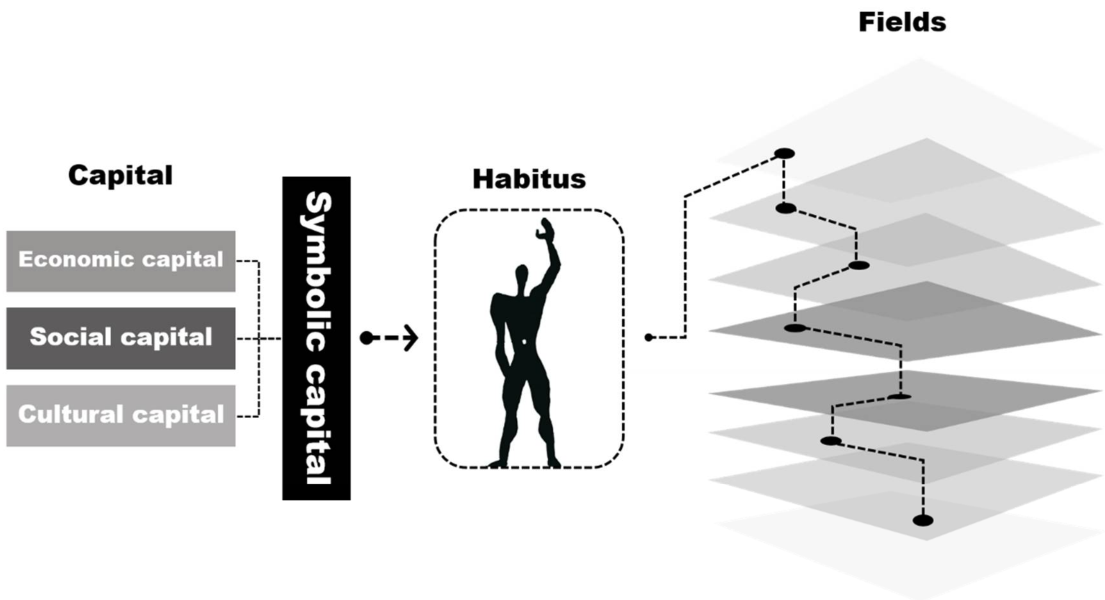

Figure 1. Bourdieu's theory of distinction. Fields refer to different sub-spaces of society such as family groups and work groups. Individuals' role in these fields is influenced by her symbolic capital.

Many studies followed Bourdieu's theory of distinction to determine how demographic factors were correlated with taste. For example, some studies showed that generally, people of higher economic status read more literature and quality papers [29] and have different taste in art [30]. More recent studies on Facebook and MySpace data-sets, argue that people with similar social networks share similar tastes of music, movies, TV shows and books [18], [31]. In all these studies, taste is seen as a means of distinction between different groups of people, which further supports Bourdieu's argument. According to Bourdieu, individuals may play different roles in different fields of a society (i.e. sub-spaces of society such as friend groups and institutions). The quality of these roles relies heavily on an individual's symbolic capital, which Bourdieu defines as a combination of social, economic and cultural capital. As discussed earlier, Bourdieu believes that taste best reflects the symbolic capital, which is the main reason of distinction in societies (Figure 1).

2.1.3 How can information about restaurants help us understand the socio-economic and cultural structure of cities?

Businesses are an effective type of sensors that can reflect what is accessible and offered to a neighborhood. Theoretically, it is not surprising to expect geographically concentrated clusters of similar tastes between individuals in American metropolitan areas: first, as discussed earlier, these

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

cities are characterized by highly fragmented social fabric with segregated communities of different taste, culture, ethnicity and economic status. Second, their economies are global and products of all types belonging to all different cultures and nationalities are offered in the marketplace and therefore the consumer is offered a variety of goods from which she can choose [32]. Third, in case of the U.S., the rise of individualism and diversity along with the economic growth of the post-war period has generated a dominant landscape in cities known as consumption spaces. These spaces gradually took the place of production spaces such as factories after the era of industrialization [33]–[36]. The emergence of these spaces is a result of the increasing impact of consumerism, pushing the individuals towards consuming goods and certain types of services [4], [37]–[40].

Restaurants are one of the most common and frequently used consumption spaces. In the U.S., restaurant expenditures exceed spending in higher education, computers, books, magazines, newspapers, movies and recorded music [41]. Data on consumption behaviors in restaurants is available in different social media venues such as Yelp. Yelp is a web-based application which maintains crowd-sourced reviews of local businesses (i.e. mostly restaurants, coffee shops and bars). Yelp users have generated nearly 127 million reviews for different businesses across the world [42]. Here, we used Yelp data to investigate the urban culture in different cities through the concept of taste.

2.2 Data

Two sets of data were used in this research:

- Data provided by Yelp [43] which includes 11 cities, 8 of which are in North America (i.e. Cleveland, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Toronto, and Montreal). This data includes 4.1M reviews by 1M users for 144K businesses as well as 1.1M business attributes (e.g., hours, parking availability and ambience). For the case of Montreal, of 86,054 reviews 11,284 were in French, as identified through langdetect 1.0.7 package in Python [44]. Since English reviews may not equally represent all Montreal neighborhoods, demographics, and resident population, we considered Montreal as an outlier and removed it from our analysis. For this study we were only interested in restaurants in English-speaking North American metropolitan areas therefore, we filtered out Montreal and Urbana-Champaign (a small city) as well as points that fell out of the metropolitan boundaries. Also, we only used businesses tagged as restaurants. This process resulted in 2,186,054 reviews for 34,231 restaurants. This data includes the following fields: Business ID, User ID, Reviews, Business Name, Star Rating, Address, City, State, Zip code, Business Category, Review Count, Longitude, Latitude. The geographic coordinates represent the location of businesses.
- As we discussed in the introduction section, we intended to see if we can characterize the socio-economic status of urban communities without having information about users. This is very important, because although it is possible to scrape data from different websites such as Yelp, the user IDs are often not provided in the interface and cannot be scraped easily. In other words, extracting information from businesses from the web is often easier than finding individual-level data. To investigate the extent to which business-level data scraped

from the web and stripped from user IDs can inform us about neighborhoods, we scraped restaurant reviews and attributes for Boston, Washington D.C., Detroit and Philadelphia metropolitan areas. All these cities are characterized by high segregation as well as ethnic and cultural diversity. This data includes 509,319 reviews for 120,801 restaurants. Using the earlier data-set, we expect to be able to study the communities in this data-set where the user IDs are absent. Also, the four cities are important metropolitan areas and studying the socio-spatial dynamics of these cities can be useful per se.

Table 1. Number of reviews and restaurants for the 10 cities

City	Number of Restaurants	Number of Reviews	Reviewers' User-ID	Number of Reviewers
Boston	44,597	172,401	Not available	Not available
Charlotte	2,780	139188	Available	39,813
Cleveland	3,996	139,824	Available	21,939
DC	8,206	40,420	Not Available	Not available
Detroit	35,823	81,301	Not available	Not available
Las Vegas	6,312	826,358	Available	275,012
Philadelphia	29,045	91,660	Not available	Not available
Phoenix	9,692	731,744	Available	97,476
Pittsburgh	3,130	124,170	Available	33,268
Toronto	11,451	357,940	Available	58,355

2.3 Methodology

In using the Yelp data-set, our assumption is that when a person talks about a food or drink in her comment, she has purchased or at least considered that food or drink and therefore, it can be used as an indicator of one's choice of food or drink. In the following sections we explain our methods for this research. Figure 2 summarizes our work-flow.

2.3.1 Feature generation

In this study we use the text provided by Yelp reviewers when they post restaurant reviews on Yelp.com. We use a bag-of-words model to define features for every restaurant. In this model the existence of a word, regardless of the way its embedded in the comment, is considered. A bag-of-words model is suitable for our case, as we are only interested in the frequency of these words and not the way they're used in the sentence. According to Bourdieu's theory of distinction, food, drink, and interior decoration are among the best indicators of taste reflecting one's everyday choice [28]. We are, therefore, interested in three categories of features: foods and drinks (e.g. pizza, martini), adjectives used to describe foods (e.g. fried, steamed), and adjectives described for ambience (e.g. rustic, minimalist). We assumed that ambience is an equivalent of decoration. Ambience are among those concepts that are frequently discussed in Yelp reviews along with food, price, and service [45] and provide an overview of the restaurants atmosphere and decorative features such as classy,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

intimate, romantic, hipster and so forth. In choosing features we avoided selecting words that have multiple connotations or are too general (e.g. nice, green).

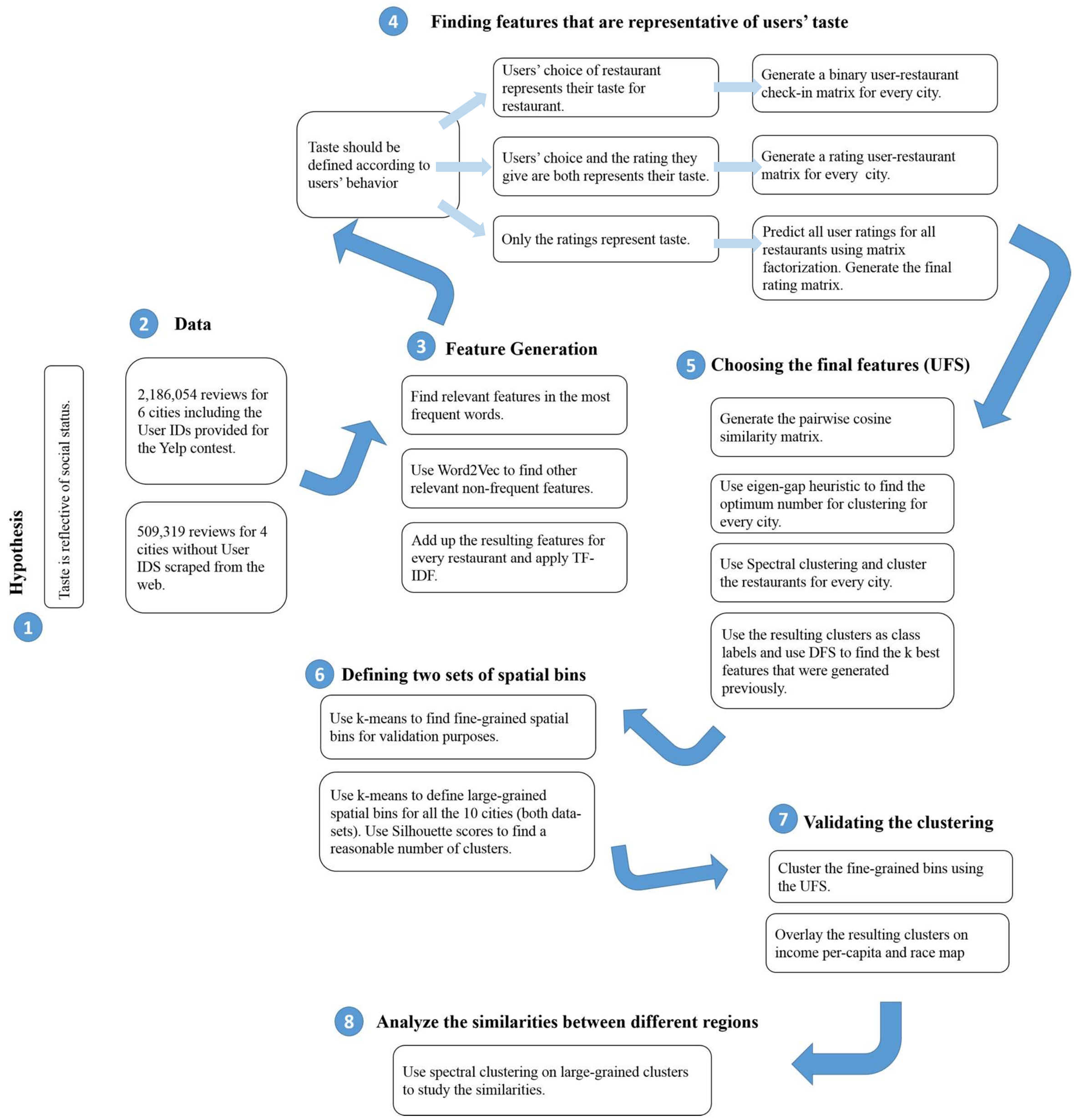

Figure 2. Research work-flow

In order to select relevant features from reviews, we used a four-step process:

1. first, we used English stop-words to remove commonly-used words [46] and then, chose features among the top 1000 frequent words. Fourty-five features of the three categories (i.e. foods and drinks, food adjectives, and ambience adjectives) were selected at this step (Appendix A).

2. Although frequent features can provide much information for restaurants, we expect to get more specific words from the comments. For example, different types of fish (e.g. haddock, tilapia) or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

different adjectives used to describe an ambience (e.g. divey, hipster) are not among frequent words. To address this problem, we used the Word2Vec model. This open-source model was developed by Google in 2013 which transforms words in a document to high-dimensional spatial vectors by using a Neural Network Language Model (NNLM) [47], [48]. Given N user comments and the n-th word in the comment $\mathbf{w}_n$, and the window size of the context centered on the n-th word as C, the maximum likelihood function of the NNLM model will be as follows:

$$I(\theta)=\frac{1}{N}\sum_{i=1}^{N} log\ p(\mathbf{w}_n|\mathbf{w}_{n-c}^{n+c}) \quad (1)$$

Where $\mathbf{w}_{n-c}^{n+c}$ represents a set of words at the center of which is $\mathbf{w}_n$ with context sampling window size of c. Word2Vec suggests two mathematical frameworks for solving Equation (1) i.e. Continuous Bag-of-Words (CBOW) and Skip-Gram. In summary, Skip-Gram uses stochastic processes to sample from the words whereas CBOW offers a continuous input and training mechanism. In this study, we use CBOW to train the model as some studies suggest it has a better performance at characterizing the words [**Error! Reference source not found.**]. We trained our Yelp corpus with this model and every word was turned into a 100-dimensional vector. As an example, Table (2) shows the closest words to the word classy. It is noteworthy that the model does not necessarily return synonyms of classy but rather, it considers the way word classy is used in a sentence and therefore, it returns all adjectives that are used to describe an ambience. The 45 words chosen in the last step were given as input to this model to find the 20 closest words in cosine distance. However, not all these 20 words were relevant to food, drink, or ambience. Accordingly, we went through all the 900 words (i.e. 45*20) and selected related words subjectively. It is important to note that Word2Vec model significantly simplified the filtering process and instead of going through all the words in the corpus, we just went through the Word2Vec outputs that is 900 words total. At the end of this step, a total of 454 features were selected.

Table 2. Top 10 most similar words to *classy*

Word2Vec output	Similarity to classy
swank	0.87688
trendy	0.86152
chic	0.85917
posh	0.84972
elegant	0.84592
stylish	0.84019
cozy	0.83344
modern	0.80526
contemporary	0.78569
homey	0.77934

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

3. We binarized the number of words selected from the last step in each comment (1 word exist 0 otherwise) and aggregated them for every restaurant. Given that these words are not equally common we use Term Frequency-Inverse Document Frequency model (TF-IDF) to weight these features:

$$idf(t,D)=log\ \frac{N}{1+|\{d\in D:t\in d\}|} \qquad (2)$$

Where N is the total number of restaurants in the corpus and $|\{d\in D:t\in d\}|$ is the number of times that term t appears in the restaurant d. We can then multiply IDF by the Term Frequency (TF) that we previously generated. After this step, for every restaurant, we will have 454 features that are properly weighted.

4. The features generated in the previous steps can sometimes fall into categories which can be even more important than the individual features themselves. For example, specific fish types (e.g. salmon) might be important but less informative than the combination of all types of fish. This information tells us that seafood is popular in a certain area. Appendix B indicates the groups of features that we combined in order to generate new features. By including these new features, a total of 477 potentially-unnecessary features remain (e.g. does the word "water" really explain anything about a community's taste?). In the next step, we explain our methodology for reducing the dimensionality and choosing the most important features.

2.3.2 User's taste and the curse of dimensionality

In the feature generation process, we took an inclusive approach and considered all features that could possibly represent user taste. Considering all these features for clustering is problematic due to high dimensionality. It is also unclear whether these features represent people's taste. In other words, we are interested in a subset of features that distinguishes between different groups of users in terms of their practiced taste. For example, the word water may be used equally in all restaurants. In this case, considering water not only doesn't add any additional information about different neighborhoods but also increases the dimensionality. Therefore, it is important to only select those features that have to do with people's taste.

Recall that the data-set provided by Yelp includes User IDs as well as user-generated ratings for rated restaurants. This data can assist us to select a subset of the 477 features that actually has to do with users' taste of food, drink, and decoration. Therefore, we examined three scenarios to select the best features related to taste:

1. **Users' choice of restaurant represents their taste for food, drink and decoration:** Under this assumption, a person's taste is only reflected in the type of restaurants she chooses to visit. Therefore, if we find clusters of restaurants that have been visited by similar users, we should be able to find distinguishing features between these clusters. To this end, we first create a matrix for every city showing whether a user has visited a restaurant (1) or not (0). We generate this matrix for each city separately to reflect how a user living in one city is more likely to go

 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

to restaurants in the same city. By separating the cities, the effect of geography is minimized and we can draw our focus on the effects of restaurant attributes on users' choice of restaurant. Of all the 525,863 Yelp reviewers, 311,866 reviewers have provided only one review. We removed users with only 1 review since first, these reviews are more likely to be biased and have extremely high or low ratings and we will use the ratings in the next steps. Second, excluding these would reduce the computational costs and also increase the accuracy of our clustering, which we will explain in the next steps. From these matrices, we generated a pairwise similarity matrix using cosine distance:

$$\cos(\mathbf{A},\mathbf{B}) = \frac{\mathbf{AB}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{A}_i^2}\sqrt{\sum_{i=1}^{n} \mathbf{B}_i^2}} \qquad (3)$$

Where restaurant A and restaurant B are n dimensional vectors with n being the number of Yelp users in each city. Every element of A and B is 1 if a given user has reviewed that restaurant and 0 otherwise.

In the next step, we used spectral clustering [49] to first find the restaurants with similar clientele. This method constructs a graph from the similarity matrix, where the data points (i.e. restaurants) are the nodes and the similarity between them are presented as weighted edges. The algorithm finds partitions of the similarity matrix by detecting low-weight edges. More specifically, this algorithm first performs a dimensionality reduction and then applies a k-means clustering [50] on the low-dimensional embedding. To reduce the dimension, the algorithm first generates a Graph Laplacian L [51]:

$$L = 1 - D^{-1}W \qquad (4)$$

Where D is the degree matrix with diagonal terms $d_i = \sum_{j=1}^{n} W_{ij}$, and W is the adjacency weight matrix of an undirected graph. The Laplacian matrix L, in fact, is used to calculate the eigenvalues for the matrix. The k-means clustering will then be applied to these eigenvalues, which represent an image of the similarity matrix in a lower-dimension space. Since the k-means is applied to a reasonably lower dimension, the resulting clusters are expected to be more distinguishable and informative. To ensure an optimal number of clusters, we use eigen-gap heuristic method [49] to find the largest difference between two consecutive eigenvalues of the Laplacian matrix and set the number of clusters equal to the rank of the eigenvalues. The check-in row in figure (4) shows the resulting eigen-gaps for different number of clusters. As we can see, for Pittsburgh for example, 2 is the best number of clusters for the check-in matrix.

We then select the k best features (from those 477 features) that affect the membership status of a restaurant in one of those previously defined clusters. In other words, we discover which subset of the 477 features actually distinguishes between the clusters using a Deep Feature Selection (DFS) model [52] to select features at the input level of the deep network. The DFS

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

model used in this study has the following network structure $\{477 \rightarrow 477 \rightarrow 256 \rightarrow 64 \rightarrow 16\}$ with a softmax output layer. The first one-to-one linear layer w, between the input layer and the first hidden layer with linear activation function is regularized using an elastic-net [53]. The resulting sparse one-to-one layer weights w only selects those features corresponding to none-zero terms in w. The model parameters are learned by minimizing this equation (5).

$$\min_{\theta} f(\theta) = l(\theta) + \lambda_1(\frac{1-\lambda_2}{2}\|\mathbf{w}\|_2^2 + \lambda_2\|\mathbf{w}\|_1) \\ +\alpha_1(\frac{1-\alpha_2}{2}\sum_{k=1}^{K+1}\|\mathbf{W}^{(k)}\|_F^2 + \alpha_2\sum_{k=1}^{K+1}\|\mathbf{W}^{(k)}\|_1) \qquad (5)$$

where $l(\theta)$ is the log-likelihood of the data, the matrix $\mathbf{W}^{(k)}$ is the kth hidden layer weights and $\lambda_{1,2} \in [0,1]$ is the parameter that controls the sparsity of w and the term $\alpha_{1,2}$ is another elastic-net like term that reduces the model complexity and increases the speed of optimization.

To find the best subset of features, we tuned hyper-parameters $\alpha_{1,2}$ and $\lambda_{1,2}$ corresponding to the sparsest model with the highest prediction accuracy measured using F_1 score which is a weighted harmonic mean of the precision and recall metrics described below:

$$recall = \frac{TP}{TP+FN} \quad and \quad precision = \frac{TP}{TP+FP} \qquad (6)$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \qquad (7)$$

Where TP, FP and FN stand for true positive, false positive and false negative respectively [54]. Since the data for each city is moderately small, 10-fold cross-validation was performed to prevent over-fitting to the training data set.

2. **Users' choice and the rating they provide both affect their taste for restaurant:** The only difference between this hypothesis and the first one is that the rating that one provides for a restaurant acts as a weight to the check-in matrix from the last hypothesis. Accordingly, in this hypothesis, not all restaurants visited by the user are equally important, but rather, we assume those that the user rates higher are more important in determining one's taste.

3. **Only the users' ratings determine their taste:** In the second assumption we assumed that taste is reflected in the way people rate a restaurant. The only difference here from the last assumption is that we try to see what would happen if every user rated every restaurant. Under this assumption, however, a problem arises: the rating matrix is sparse and many ratings for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

many restaurants are missing. Using the original rating matrix cannot help us identify how would every user like every restaurant. Therefore, we will need to predict the ratings by using matrix factorization method [55]. The fundamental assumption of this method is that there are d latent features in restaurants that affect the users' ratings. The advantage of this method is that without having to know what those d features are; we can predict how users might rate restaurants which they have not yet reviewed. We use Singular Value Decomposition (SVD) method to factorize the rating matrix [56]. To find the best number for d, we used 10-fold cross validation. The results indicate that there are approximately 20 latent features (d=20) that affect one's rating for a restaurant. The Mean Square Error (MSE) decreases significantly up to d=20 and gradually increases afterwards due to being over-fit (figure 3). After predicting the rating matrix with 20 latent features for every city, we repeat the steps described in the last two hypotheses. In all three hypotheses above, we selected the number of clusters with the largest eigen-gaps (figure 4) for every city. Table 3 shows the final number of clusters selected for different matrices and different cities.

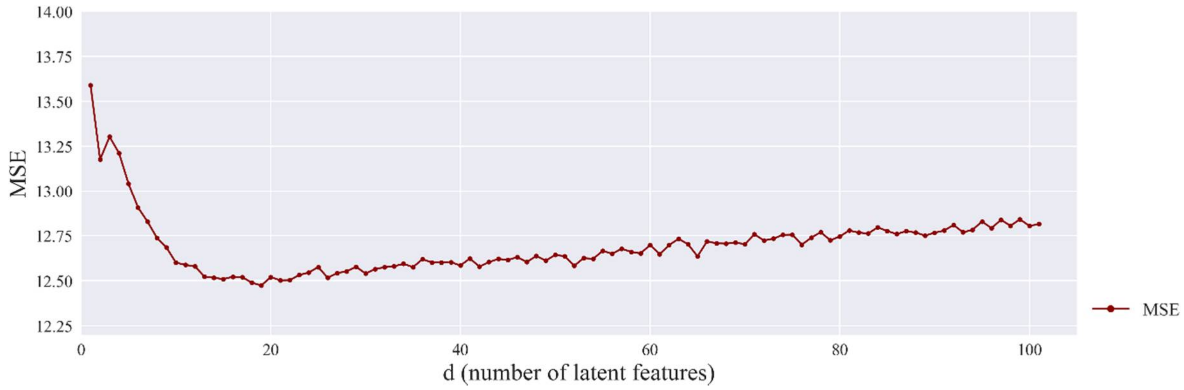

Figure 3. 10-fold cross-validation results for rating predictions.

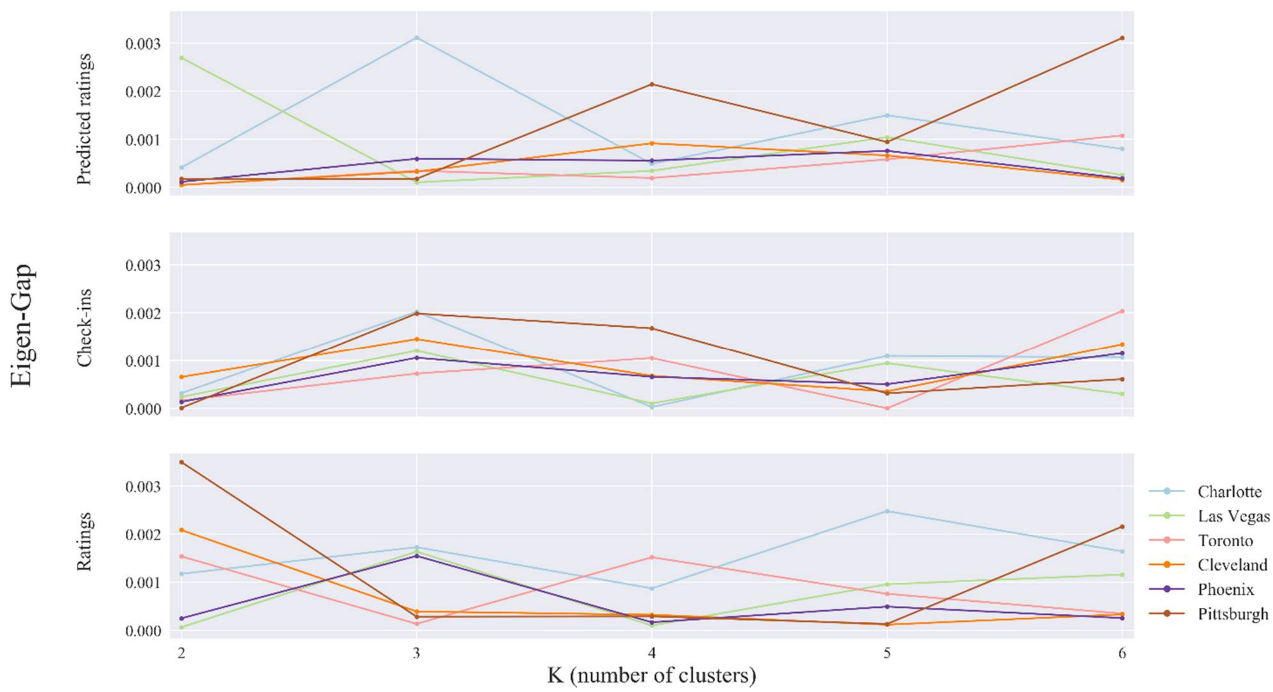

Figure 4. Eigen-gaps for different number of clusters and different

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

Table 3. Selected number of clusters for different matrices and cities

City	Predicted Matrix	Rating Matrix	Check-in Matrix
Charlotte	3	5	3
Cleveland	4	2	3
Las Vegas	2	3	3
Phoenix	5	3	3
Pittsburgh	6	2	3
Toronto	6	2	6

2.3.3 Defining the spatial bins

The features generated from the previous steps reflect Yelp reviewers' preferences in different urban areas. We next aggregate restaurant features on some spatial units which represent the urban fabric to ensure that nearby restaurants will belong to the same spatial bin. Aggregating restaurants on geographic units will enable us to minimize the impact of outliers and noise. It also enables us to get an overall sense of taste preference given all different types of restaurants in a region. Since our sensors are restaurants, we define these geographical units based on their density and configuration and avoid using administrative boundaries e.g. block groups. Two sets of spatial bins are required to answer our research questions:

1. **Large-grained spatial bins:** These spatial bins enable us to compare different parts of cities together as to see how different cities interact in terms of food, drink, and decoration related attributes. The existing administrative boundaries are too small for this purpose. For example, we are looking at dividing up Washington DC to 3-6 parts and conventional administrative boundaries are too fine-grained for this purpose. Also, we intend to have reasonable spatial bins that are actually representative of the city form. The number of these bins are actually a matter of preference, however, for visualization and simplification purposes we choose large-grained clusters. Accordingly, we use k-means clustering on the restaurants' geographic coordinates to find reasonable spatial clusters. To find the best number of clusters for each city, we use the silhouette scores [57] for different number of clusters for every city. Silhouette score measures the extent of tightness and separation for each cluster. In other words, it specifies which objects are within their clusters and which ones are somewhere in between:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)} \qquad (8)$$

where $a(i)$ is the average dissimilarity of datum i with all other data points and $b(i)$ is the lowest average dissimilarity of i to any other cluster. We then average $s(i)$ over all data points, a measure that we used for goodness of clustering. Silhouette score ranges from -1 to 1, where 1 means that the clustering configuration is appropriate. Figure 5 shows the Silhouette scores when we divide each city to less than 30 clusters. At this point, we make a compromise between the number of restaurants in every city, area of the city as well as the Silhouette score.

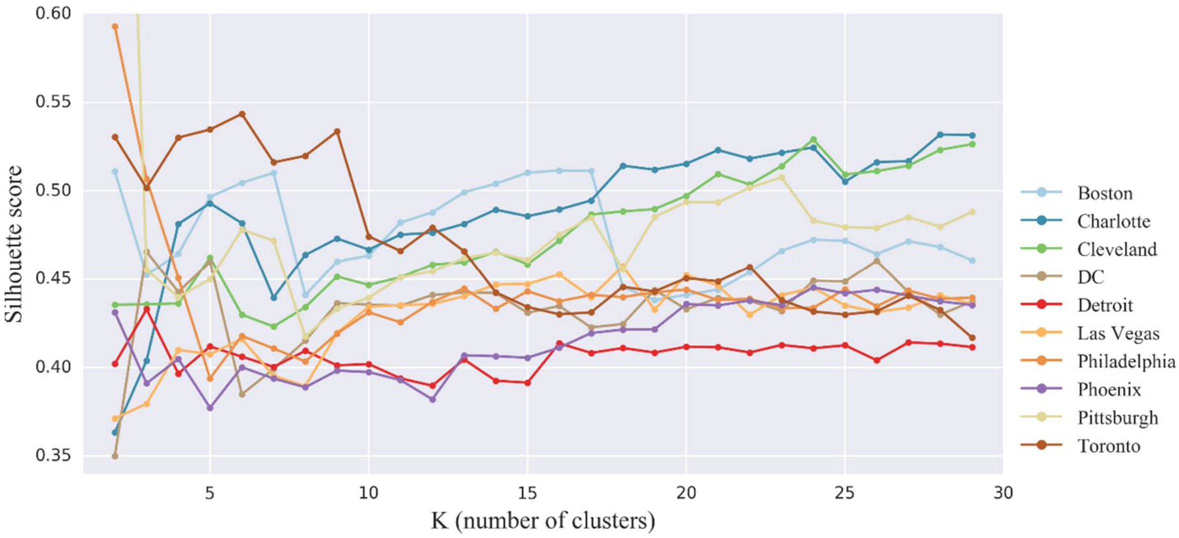

Figure 5. Silhouette scores for different Ks for different cities (large-grained spatial bins)

2. Small-grained spatial bins: To validate the results and compare it with other demographic datasets, more fine-grained spatial bins are needed. Administrative boundaries are not helpful in this case either since these boundaries do not consider the formality of the built environment. For example, restaurants located on the Woodward Ave and East 9 Mile Rd cross section in Detroit, MI have been divided between four Census tracts, whereas they are all located near the same cross section and are very close to one another. Another problem with the administrative boundaries is that their sizes are not consistent with the distribution of the restaurants. For example, as we move to the suburbs of Detroit we can see tracts which contain one or two restaurants in them. Accordingly, same as the last step, we use k-means clustering and Silhouette scores to define these spatial bins. This method enables us to consider for the distribution of restaurants while defining the spatial bins. Figure 6 shows the Silhouette scores for the four cities. As we can see, for all these cities the Silhouette score improves as we increase the number of clusters. At this point, Silhouette scores are not useful for our purposes as they do not suggest any optimum number of clusters. Therefore, we base our decision on the number of restaurants and city area. Given the number of restaurants we have for every city (table 1), we expect about 200 clusters for Washington D.C., 500 for Detroit and Philadelphia and 600 for Boston. It is important to note that there are more census tracts in these areas than the number of clusters that we determined. For example, Detroit metropolitan area has 909 census tracts however, as discussed earlier, due to the uneven spatial distribution of restaurants, our spatial bins are larger than census tracts in the suburban areas with low number of restaurants, but smaller than block-groups in the city centers. It is important to note that the size and number of these spatial bins can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

change depending on one's research question as well as spatial resolution of the original data-set (i.e. Yelp in this case).

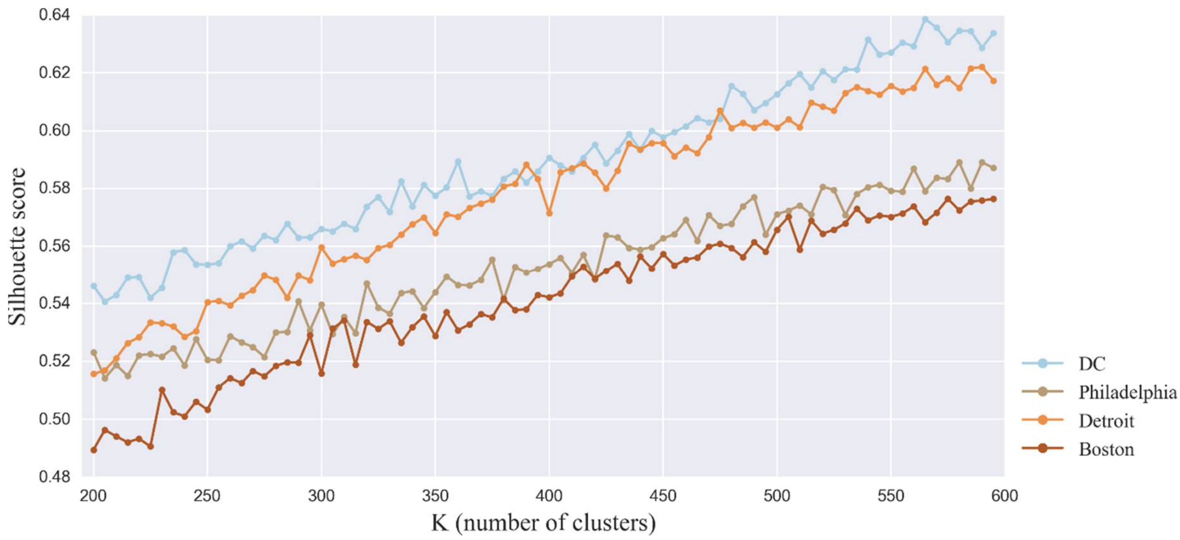

Figure6. Silhouette scores for different Ks for different cities (small-grained spatial bins)

We use the small-grained and large-grained spatial bins defined in the last step in two different ways. The small-grained clusters are for validation purposes. Our purpose is to see if we can find any clear spatial pattern by clustering these fine-grained clusters. Using small bins enables us to assess the accuracy of this method and compare it with other high-resolution data sources. We will first average the selected set of features from the last step on these spatial bins, scale the features using min-max scaling for every bin, and then calculate the pairwise cosine similarity between the fine-grained bins separately for every city which we didn't have information about user IDs (i.e. Philadelphia, DC, Detroit, Boston), using formula 3. To calculate the similarities, we will use principal components instead of the actual features, to further reduce the dimension and improve the clustering results. For every resulting matrix, we will use spectral clustering method [58] as described in section 3.2. We will then overlay the resulting clusters on the block-group level map of 2017 income per-capita provided by Tableau 10.0 software for those four cities. At this point, we expect to see a geographic pattern in our clustering as well as a reasonable alignment between the clustering results and the block-group income per-capita layer.

After validating, we can use the selected set of features from the last steps to study the interactions between different regions in cities. It is important to note that this capacity is the advantage of this set of features over using user ID data since, at this point, this feature set only relies on the aggregated comments for every restaurant and not the users' check-ins and ratings. To this end, we will average these features on the large grained clusters, calculate the pairwise similarities and cluster, same as the last step. Due to the extreme cultural, economic, and racial divisions in the American metropolis [4][59] we expect to see different clusters in every city and due to the global nature of these cities [60] we expect some regions from some cities to be similar to other regions in another cities.

3. Results

doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 376; doi:10.3390/ijgi7090376

3.1. Selected features

We took the steps described in section (2.3.2), to reduce the dimension of the data set and only focus on those features that are actually representative of users' choice of food, drink and ambience. Figure 7 shows the resulting F1 scores for the three hypotheses (i.e. check-ins, ratings, and predicted ratings) and the 6 cities where the user IDs were available (table 1). For every city we selected a taste scenario that returned the highest F1 score. The resulting features along with the scenarios that returned the highest F1 scores, as well as the F1 scores are presented for every city in table 4. By considering all these features, we will have a total of 105 features which we call the *Universal Feature Set* (UFS). We can now use the UFS to study those cities where the user data is not available. The underlying assumption here is that the 6 cities that we have based the UFS on, are diverse enough that cover the types of food, drink and ambience that one expects to find in the four other cities where the user IDs are not available.

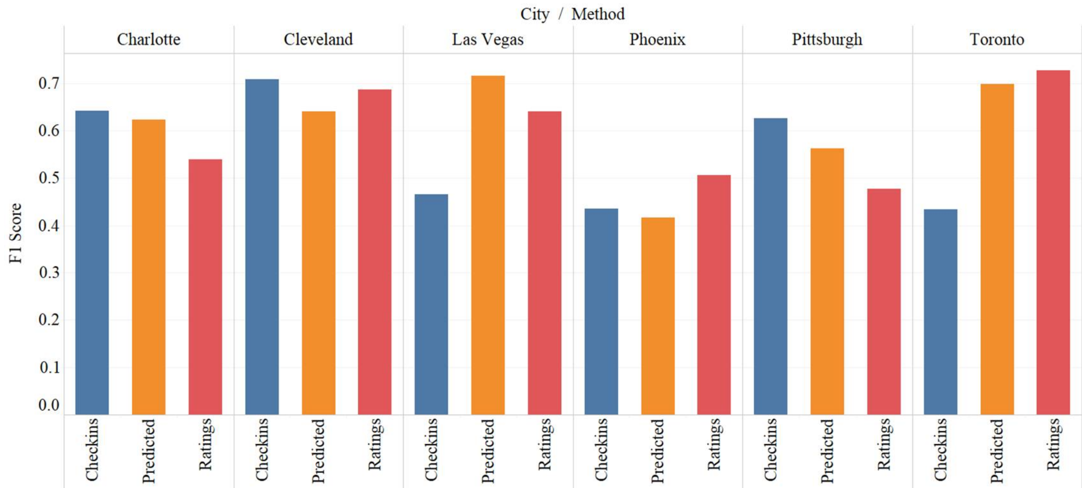

Figure 7. F1 scores resulting from classification for different cities

Table 4. Selected features for different cities

City	Best Method	F1 Score	Selected Features
Charlotte	Check-ins	0.64244	salty, vegetarian, creamy, hipster, divey, dessert, calamari, asparagus, vodka
Cleveland	Check-ins	0.70865	sweet, spicy, hipster, tomato, lime, meat_types, vegie_types, herb_types
Las Vegas	Predicted Ratings	0.71563	braised, seared, salty, creamy, intimate, classy, modern, casual, upscale, elegant, rice, soup, wine, crab, salmon, lobster, lamb, dessert, duck, cocktail, calamari, martini, ranch, steak_types, vegie_types, herb_types, hardliq_types, sofliq_types, sweet_types, asian_types, seafood_types, pos_ambience, neg_ambience, style_types
Phoenix	Ratings	0.50608	spicy, upscale, wine, pos_ambience
Pittsburgh	Check-ins	0.62651	crispy, vegetarian, hipster, romantic, rice, noodle, curry, sausage, cocktail, tofu, coleslaw, wing,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

			cheesesteak, lettuce, provolone, ranch, fast_food, dressing_types, pos_ambience, style_types
Toronto	Ratings	0.72686	fried, Chinese, salty, Asian, Japanese, steamed, oily, hipster, rice, beer, soup, pork, shrimp, wine, tea, noodle, seafood, cocktail, sashimi, soy, squid, milk, sesame, Fanta, meat_types, softliq_types, Asian_types, soda_types, seafood_types, ethnic_food

3.2 Clustering results

Results derived from clustering the small-grained spatial bins with the selected set of features reveal clear geographic patterns which correspond with block-level per-capita income for the four cities where the user IDs were absent (figure 8). We set the number of clusters on two (k=2) for the ease of comparison.

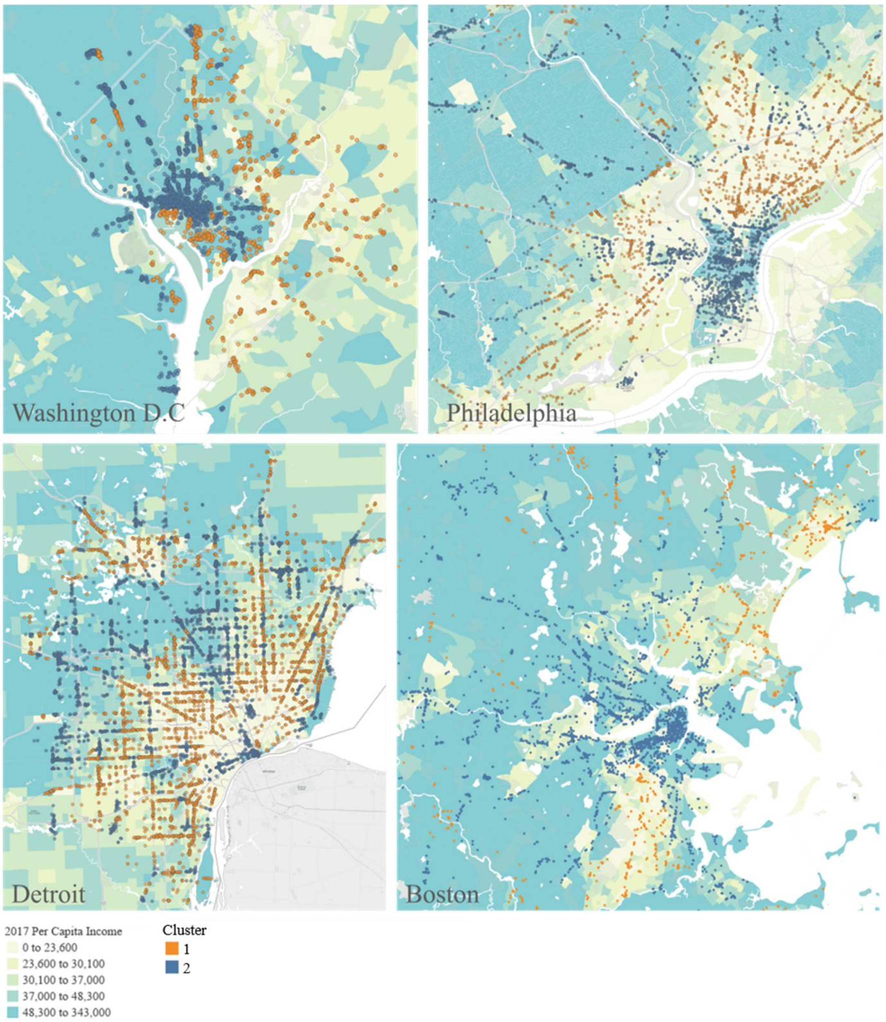

Figure 8. Clustering results overlaid on per-capita income map for four cities. As we can see the two clusters clearly correspond with income per-capita map from Census

The difference between the type of tastes practiced between the two clusters is shown in figure 9. This figure shows the top 30 features with highest average difference between the two clusters. As we can see, features such as seafood, salad, ethnic foods, vegetables, fruits, and Asian food types

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

show higher values in high-income communities whereas the low-income cluster shows higher consumption of fast food.

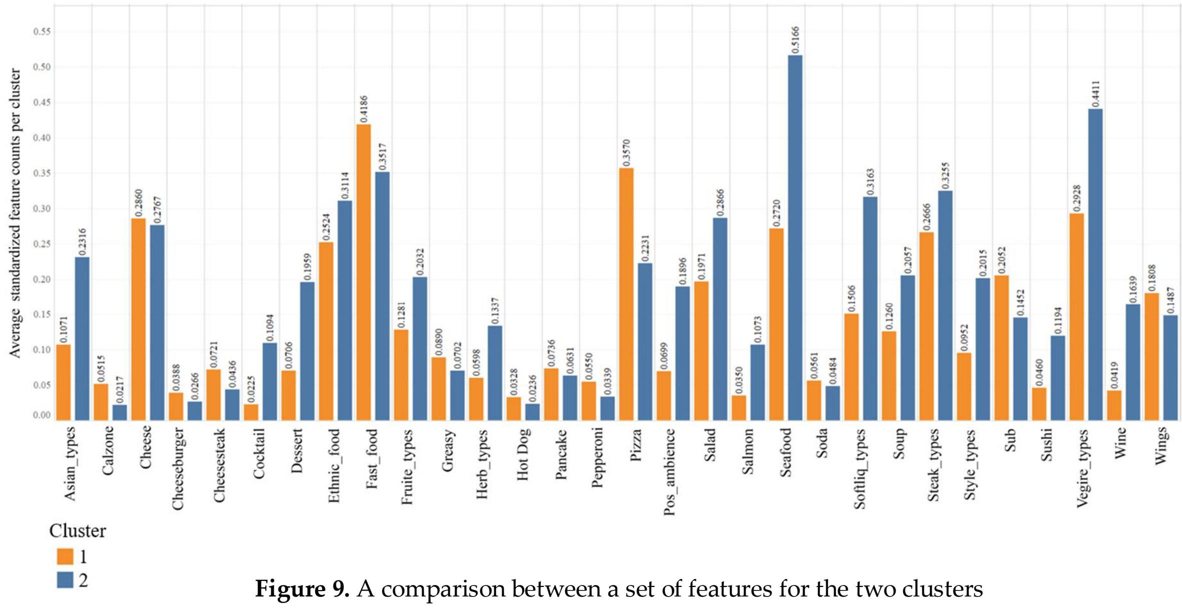

Figure 9. A comparison between a set of features for the two clusters

Table 5. Number of restaurants in the two clusters for different cities

Cluster	Boston, MA	Detroit, MI	Philadelphia, PA	Washington, D.C.
Cluster 1	16,827	17,226	13,849	2,780
Cluster 2	27,770	18,597	15,180	5,419

The fact that this spatial distribution has been derived from small spatial bins indicates the high accuracy of taste as indicator. These maps show that income can be an important factor in determining a communities' taste. To see empirically how our clusters, correspond with demographic factors, we considered racial composition, educational status, and annual household income at the block-group level for the four cities. Block-group level data is the highest spatial resolution available on Census for these demographics. The data was collected from the American Community service(ACS) website [61]. We defined educational ratio as the ratio of population that have a bachelor degree or higher, in each block-group. The racial composition was defined as the population ratio of Black/A.A., White, and Asian for different block-groups. The income variable is the annual household income in U.S. Dollars. All these demographic factors were estimates provided by the ACS for 2016. We spatially joined the restaurants to the block-groups and conducted t-tests to evaluate the extent to which our clustering results compare with these demographic factors. Table 6 provides a summary of the results. As we can see, the two clusters show significantly different demographic features in all four cities. Looking at all four cities together, we can see that education is the most different demographic factor between the two clusters. Considering the restaurants in all four cities, we can see that education and the Asian population ratio are the most distinctive factors with the highest T-statistics. As we consider each city individually, we can see that the order of importance for different demographic factors differs among different regions. For example, in Boston, the top

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

distinctive factors are education and Asian population ratio whereas in Washington D.C. the Black population ratio and annual household income have the highest T-statistics. It is important to note that all the four cities show clear spatial boundaries separating the two clusters. In other words, this method proves to be capable of identifying spatial segregation patterns that may have different demographic reasons in different regions (e.g. education level and Asian population in Boston, MA versus income and Black/A.A. population ratio in Washington D.C.).

Table 6. T-test results between the two clusters for demographic variables

City	Factor	Mean value in cluster 1	Mean value in cluster 2	T statistic (absolute value)	P value
Boston, MA					
	Educated population ratio	0.06	0.10	97.46	0.000
	Annual household income (USD)	66985.93	68655.57	5.47	0.000
	Black/A.A. population ratio	0.41	0.40	13.49	0.000
	White population ratio	0.53	0.56	32.97	0.000
	Asian population ratio	0.02	0.07	73.04	0.000
Detroit, MI					
	Educated population ratio	0.04	0.07	59.39	0.000
	Annual household income (USD)	50359.52	61600.40	40.69	0.000
	Black/A.A. population ratio	0.41	0.38	21.84	0.000
	White population ratio	0.55	0.60	35.75	0.000
	Asian population ratio	0.01	0.03	43.08	0.000
Philadelphia, PA					
	Educated population ratio	0.05	0.09	72.51	0.000
	Annual household income (USD)	55067.55	64436.73	25.42	0.000
	Black/A.A. population ratio	0.39	0.35	24.14	0.000
	White population ratio	0.55	0.62	41.79	0.000
	Asian population ratio	0.03	0.05	33.30	0.000
Washington, D.C.					
	Educated population ratio	0.11	0.15	25.48	0.000
	Annual household income (USD)	53222.42	80220.32	28.74	0.000
	Black/A.A. population ratio	0.36	0.22	41.42	0.000
	White population ratio	0.55	0.68	23.94	0.000
	Asian population ratio	0.02	0.04	15.19	0.000
All four cities combined					
	Educated population ratio	0.06	0.09	134.74	0.000
	Annual household income (USD)	57322.86	66673.53	51.06	0.000
	Black/A.A. population ratio	57322.86	0.37	29.46	0.000
	White population ratio	0.40	0.60	69.42	0.000
	Asian population ratio	0.54	0.05	91.96	0.000

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 376; doi:10.3390/ijgi7090376

Figure 10 illustrates the clustering result with 5 clusters for Boston, MA. In this case, as well, we can see clear geographic patterns. For example, we can see orange and green points are both clustered together around the low-income areas. By overlaying these clusters on the African American population, we can see that most of the green points are located in areas with high concentration of African American population. On the other hand, many purple points are located at areas with high income and high concentration of African Americans. This issue gets to the heart of Bourdieu's argument [28], that taste as an indicator of social status, is not merely a construct of economic capital, but rather it's derived from symbolic capital, which is in turn, a combination of social, cultural and economic capitals. Accordingly, using taste as an indicator of symbolic capital can shed light on different aspects of communities' lifestyles which may not be explained similarly with conventional demographic indicator (e.g. income, race) for different geographic and cultural contexts.

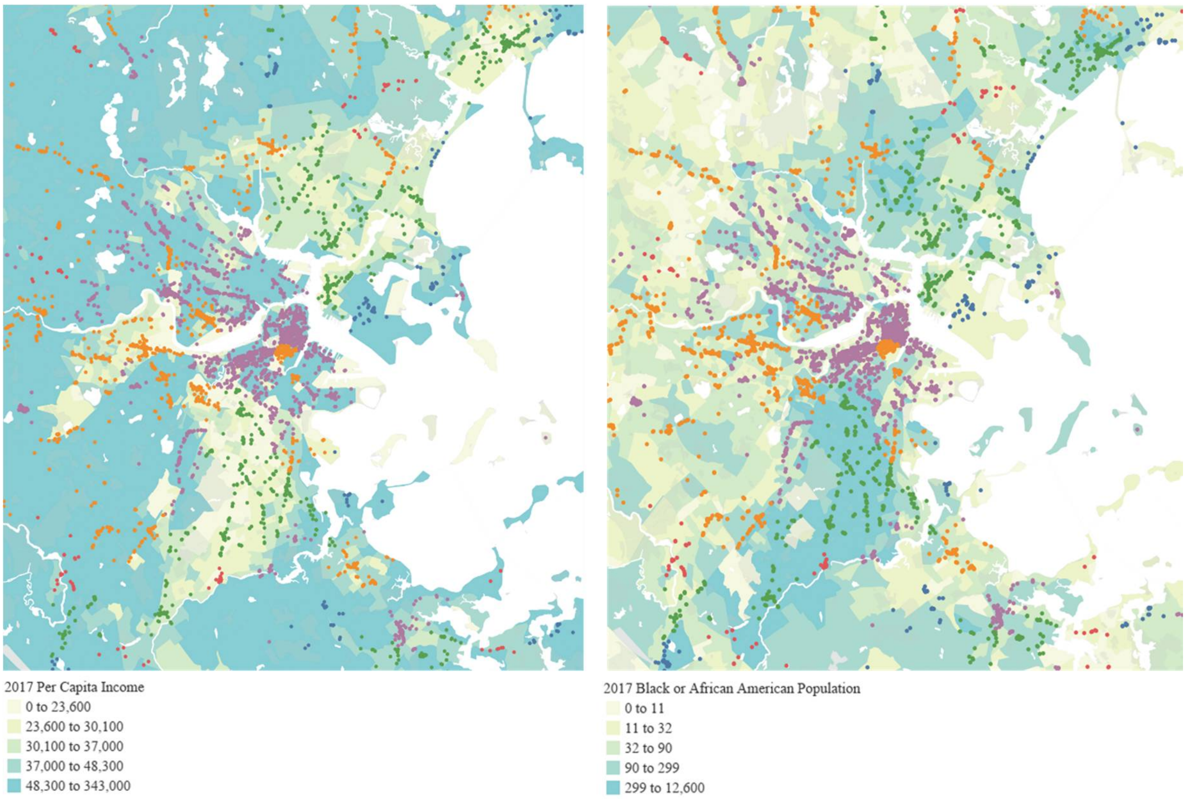

Figure 10. Clustering results with 5 clusters for Boston

Having set our new indicator, we can now use this indicator to study the socio-economic interactions between different regions in different cities. We use the large-grained spatial bins that we previously defined for all cities and choose 5 clusters for simplification purposes (figure 11). The results are consistent with our understanding of global cities. American cities are comprised of spatially separated cultural groups [4]. We can also see that the distribution of these cultural clusters is consistent with our knowledge of some cities. For example, we know that the racial and economic segregation pattern for Phoenix, Pittsburgh, and Washington D.C. approximately corresponds with our results. In some cases, the clusters do not necessarily match with racial and economic measures of those regions. For example, the north-eastern side of Phoenix is in the same cluster as downtown Cleveland while the two regions are demographically different. The earlier is dominantly white and high-income whereas the latter is a low-income mixed-race region. Another anomaly is Toronto

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

which seems to have all its regions in the same cluster colored in cyan. Clusters shown in cyan signify high-income multicultural areas with a variety of restaurant types and cultural groups. This issue might be due to the fact that Toronto does not suffer from extreme racial and economic segregation as is the case for American metropolitan areas [62].

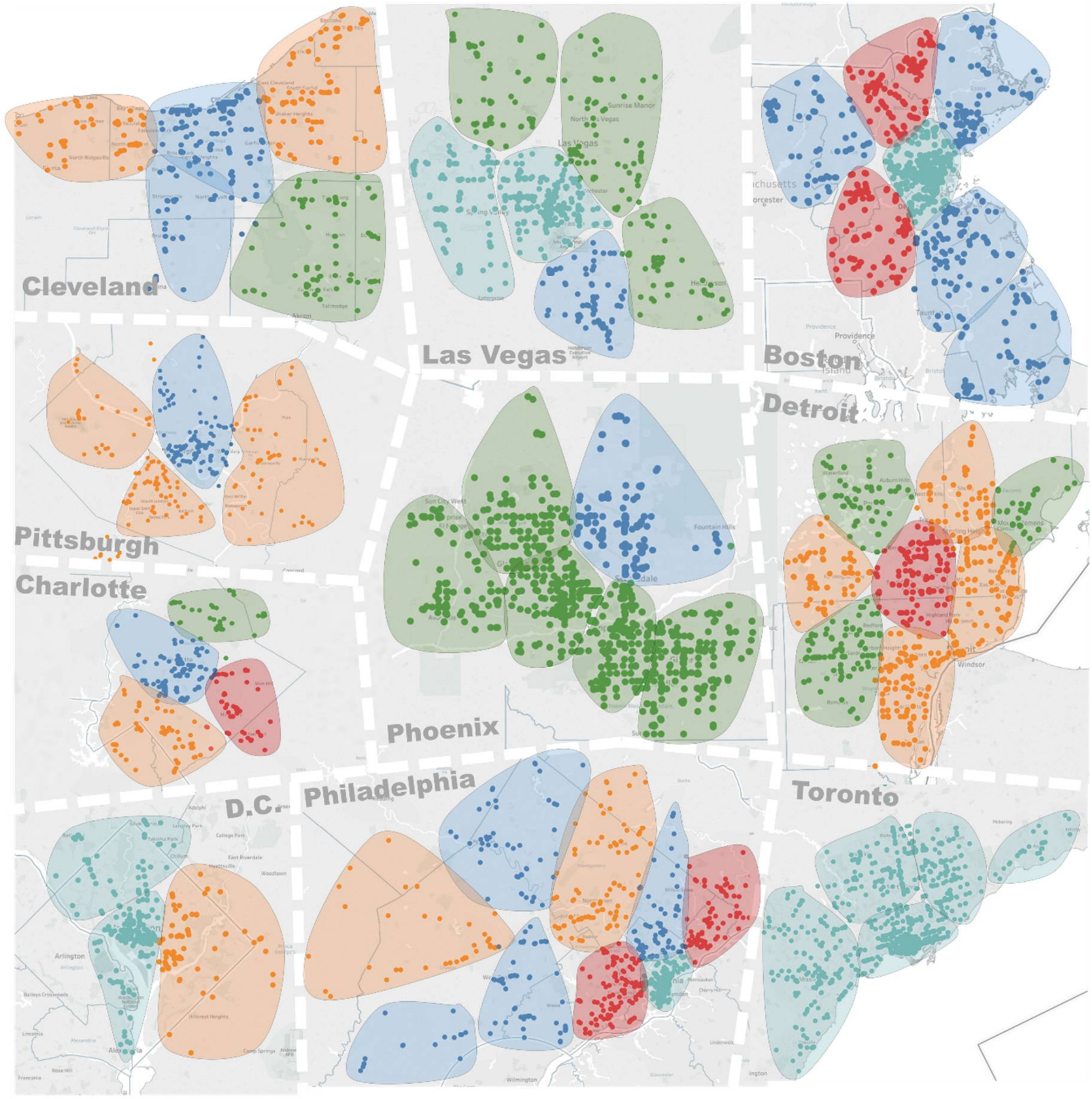

Figure 11. Cultural interactions between different cities. Similar colors across cities indicate similar tastes.

4. Conclusion

In this study we first used Google Word2Vec model to generate a total of 477 features pertaining to foods, drinks, food qualities, and the interior ambience of restaurants. We extracted these features for the 6 metropolitan areas where the User IDs of the reviewers were available (i.e. Toronto, Phoenix, Las Vegas, Pittsburgh, Cleveland, Charlotte). We then hypothesized three possible scenarios for defining taste and limiting the number of features to those that have to do with users' behaviors: first, taste may be seen as the combination of factors that encourages a Yelp user to visit a place. Second, the rating that the user gives to a place is also a factor in determining one's inclination towards a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

restaurant. Third, the predicted rating of every user for every restaurant should be used as a basis for an individual's taste. For every one of the matrices generated from these hypotheses, we used the eigen-gap heuristic method to find the best number of clusters. We then solved a classification problem by incorporating the Deep Feature Selection (DFS) model for every hypothesis to see what would be the best subset of those 477 features (i.e. returns the largest F1 score) if we used the clusters defined from the last step class labels. We repeated this process for every one of the 6 cities and for every city we chose the highest F1 score derived from the three hypotheses (Section 2.3.2). We named the union of these 6 sets of features (i.e. one set of features for every city) as the Universal Feature Set (UFS) which became the basis for clustering the cities where the User ID for reviews were absent.

By overlaying the clusters identified using the UFS for the four cities with absent user IDs on the 2017 block-group level income per-capita map, we showed that our definition of taste can be used as an indicator for studying the socio-economic structure for the four cities where we didn't have the user IDs. We found a clear alignment between areas of low-income and high-income and our clusters for all the four cities (figure 8). We also showed statistically that the two clusters are significantly different based on different demographic factors representing income, education and racial composition. We showed that education is the most distinctive factor between the two clusters once we consider all four cities combined. We also showed that the two clusters in different cities, while forming clear spatial boundaries, are different in terms of demographic differences between the two clusters. For example, We found that Education and Asian population ratio are the most distinctive factors in Boston while in case of Washington D.C., Black/A.A population ratio and Annual household income are the main distinctive factors.

Once we increased the number of clusters we still observed a geographic pattern (figure 10) which results from a combination of demographic factors such as race and income. This issue reflects the multifaceted nature of taste as argued by Bourdieu nature of taste [28]. We showed that this method also works well for more than 2 clusters, although the performance of this method depends highly on the quality of data and number of reviews. Lastly, we used UFS to study the inter-regional similarities for 10 North American cities. Our results showed that all the 9 American cities were comprised of regions that are less similar to one another and more similar to some regions in other cities. This observation is close to our understanding of the global cities as described in the literature [4]. In case of Toronto, all the spatial bins were in the same cluster which might be due to the fact that extremely disadvantaged neighborhoods for different racial groups do not exist compared to the U.S. metropolitan areas [62].

As discussed earlier, we do not expect to see a direct relationship between clusters derived from taste and racial and economic segregation patterns in all cultural and geographic contexts: First, commonly used foods and drink in a White community in one city might be quite popular in the African American communities in another. From a theoretical point of view, the taste index assists us to see cities regardless of their mere economic and racial composition, but rather the symbolic capital of the inhabitants which results from social, economic, and cultural capital, combined. Second, reviews provided by Yelp users in a region might not have necessarily been authored by the residents of that region. It is not surprising to see that a considerable number of reviews in downtown Cleveland, for example, have been authored by visitors who do not reside in that region. This issue can be seen as both a limitation or potential [63] . It is a limitation in a sense that restaurants-as-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

sensors, may fail to capture the cultural characteristics of the resident population in a neighborhood as these restaurants may target the visitors and not the resident population. On the other hand, it could be a potential since most of the information collected by different agencies such as Census are collected from residents while ignoring the ambient population. This issue has also been discussed by other studies [63] that argue about the mismatch between density of tweets and residents' population. The taste index, therefore, enables us to see the cultural preferences practiced by the ambient population who actually are the clientele of these restaurants. Using ambient population can help urban planners to gain a better understanding of the people who actually use urban spaces and design spaces accordingly [64].

Working with socially sensed data comes with many limitations. First, Yelp reviewers may be a biased sample of the population and therefore, the comments that they provide might not be reflective of the entire population's judgment for a restaurant. Second, our definition of taste was limited to the types of food, drinks and restaurants' ambience. Although this definition may reflect the characteristics of neighborhoods to some extent, additional data on people's lifestyle such as the interior decorations, grocery purchases, and types of movies they watch will provide a more accurate understanding of different neighborhoods. The extent of these limitations for different geographic contexts may affect the final results, significantly. In case of Phoenix for example, we can see that the final F1 score, according to table 4 was low (i.e. 0.50608) compared to other cities, which may be due to data bias or similar food tastes between different user groups. Despite all these limitations, our method uses community-authored comments scraped from the web at no cost with a reasonable spatial and temporal resolution. Given the variety and accessibility of business data [13], the information derived from this method can complement the conventional demographic data of the cities and provide a multifaceted understanding of cities which integrate economic, social and cultural components at once.

Appendix A. List of 45 features used as seed to Word2Vec model

Category	List of features
Food	chicken, pizza, ketchup, cheese, salad, hot dog, burger, bacon, burrito, mushroom, fish, wings, strawberry
Drink	coffee, tea, beer, soda, water, wine, cocktail, alcohol, smoothie
Food adjectives	Mexican, Italian, Chinese, sweet, fried, spicy, vegetarian, greasy, homemade, juicy, organic, stuffed, crispy
Ambiance	cozy, hipster, trendy, classy, modern, homey, intimate, romantic, upscale, divey

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 376; doi:10.3390/ijgi7090376

833 **Appendix B. List of features generated by aggregation**

New Feature	List of combined features
steak_types	meatloaf, Barclay, flank, wagyu, kalbi, tenderloin, striploin, bavette, rib, brisket, mignon, steak, ribeye
meat_types	chicken, meat, beef, pork, lamb, veal, duck, turkey, steak
sweets_types	yogurt, gelato, pudding, cupcake, biscuit, pie, tiramisu, crepe, custard, tart, sorbet, Nutella, cheesecake, cream, cannoli, muffin, donut, cookie, cake, shake
fast_food	pizza, hot fog, sandwich, burger, chips, pepperoni, max, finger, cheeseburger, cheesesteak, calzone, meatball, hoagie, poutine, blt, Rueben, wing
vegie_types	turnip, lettuce, celery, seaweed, parsley, scallion, eggplant, broccoli, zucchini, kale, cilantro, veggie, ceasar, cabbage, cucumber, basil, vegetable, mushroom, sprout, carrot, asparagus, bean, onion, tomato, coleslaw, avocado, spinach, artichoke
breakfast_types	bacon, sausage, egg, benedict, scramble, omelet, bagel, pancake, croissant, pretzel, syrup, waffle, roast
fruite_types	pineapple, peach, strawberry, raspberry, blueberry, coconut, apple, mango, banana, orange
nut_types	walnut, pecan, peanut, almond
herb_types	oregano, thyme, fennel, sumac, paprika, garnish, herb, radish, chive, dill, arugula, mint
dressing_types	ranch, ketchup, mayo, gravy, marinara, siracha
coffee_types	espresso, cappuccino, decaf, americano, mocha, latte
soda_types	Pepsi, Fanta, spirit, coke, soda
softliq_types	champagne, beer, wine, margarita, sangria, mimosa, cider
hardliq_types	tequila, whiskey, vodka, martini, bourbon, shot
ethnic_food	Thai, Chinese, Mexican, Italian, Asian, Indian, Japanese, Vietnamese, Hawaiian, Sicilian, Arabic, Middle Eastern, Korean, Taiwanese, Persian, Greek, Lebanese, Portuguesem Ethiopian, Spanish
latin_types	salsa, burrito, quesadilla, taco, carnitas, tamal, guac, tapa, enchilada, tortilla, fajita, carne, jalapeno, nacho, ceviche, empanada
Italian_types	pastrami, panini, lasagna, bruschetta, pasta, prosciutto, stromboli, vermicelli, risotto, spaghetti, pesto, chorizo, gnocchi
Asian_types	fusion, sesame, wonton, spring roll, omakas, sushi, aman, tofu, kimchi, nigiri, sashimi, mushi, noodle, teriyaki
Mideast_types	shawarma, flatbreadm pitam naan, hummus, falafel
pos_ambience	cozy, homey, classy, trendy, artsy, urbane, posh, swanky, upscale, festive, romantic, eclectic, elegant, chic, stylish
neg_ambience	casual, divey, kitschy, masculine
style_stypes	hipster, hippie, bohemian, rustic, modern, minimalistic, contemporary, retro, deco, quaint
material_types	wooden, hardwood, marble, concrete, mosaic, metal, steel, brick

834
835

References

[1] S. Musterd and W. Ostendorf, *Urban segregation and the welfare state: Inequality and exclusion in western cities*. Routledge, 2013.

[2] S. Sassen, *The global city*. 1991.

[3] B. Badcock, "Restructuring and spatial polarization in cities," *Prog. Hum. Geogr.*, vol. 21, pp. 251–262, 1997.

[4] M. Dear and S. Flusty, "Postmodern Urbanism," *Ann. Assoc. Am. Geogr.*, vol. 88, no. 1, pp. 50–72, 1998.

[5] B. Wellman, "The Community Question: The Intimate Networks of East Yorkers," *Am. J. Sociol.*, vol. 84, no. 5, pp. 1201–1231, 1979.

[6] G. C. Wenger, "A Comparison of Urban with Rural Support Networks: Liverpool and North Wales," *Ageing Soc.*, vol. 15, no. 01, pp. 59–81, 1995.

[7] K. Hampton and B. Wellman, "Neighboring in Netville: How the Internet supports community and social capital in a wired suburb," *City Community*, vol. 2, no. 4, pp. 277–311, 2003.

[8] A. Madanipour, G. Cars, and J. Allen, *Social exclusion in European cities: processes, experiences, and responses*, vol. 23. Psychology Press, 2000.

[9] W. E. Lyons and D. Lowery, "Citizen Responses to Dissatisfaction in Urban Communities: A Partial Test of a General Model," *J. Polit.*, vol. 51, no. 04, p. 841, 1989.

[10] I. Bracken and D. Martin, "The generation of spatial population distributions from census centroid data," *Environ. Plan. A*, vol. 21, pp. 537–543, 1989.

[11] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban Computing: Concepts, Methodologies, and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, 2014.

[12] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4. pp. 211–221, 2007.

[13] D. Arribas-Bel, "Accidental, open and everywhere: Emerging data sources for the understanding of cities," *Appl. Geogr.*, vol. 49, pp. 45–53, 2014.

[14] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti, "Discovering the Geographical Borders of Human Mobility," *KI Künstliche Intelligenz*, vol. 26, no. 3, pp. 253–260, 2012.

[15] C. Ratti *et al.*, "Redrawing the map of Great Britain from a network of human interactions," *PLoS One*, vol. 5, no. 12, 2010.

[16] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 186–194, 2012.

[17] N. Yuan, F. Zhang, D. Lian, and K. Zheng, "We know how you live: exploring the spectrum of urban lifestyles," *Proc. first …*, pp. 3--14, 2013.

[18] H. Liu, "Social network profiles as taste performances," *J. Comput. Commun.*, vol. 13, no. 1, pp. 252–275, 2007.

[19] J. Cranshaw, J. I. Hong, and N. Sadeh, "The Livehoods Project : Utilizing Social Media to Understand the Dynamics of a City," *Icwsm*, pp. 58–65, 2012.

[20] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 247–256.

[21] S. Wakamiya and R. Lee, "Crowd-sourced Urban Life Monitoring: Urban Area Characterization based Crowd Behavioral Patterns from Twitter Categories and Subject Descriptors," *Proc. 6th Int. Conf. Ubiquitous Inf. Manag. Commun.*, p. 26, 2012.

[22] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Y. Ma, "Mining user similarity based on location

history," *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, no. c, p. 34, 2008.

[23] C.-C. Hung, C. Hung, C.-W. Chang, C. Chang, W. Peng, and W.-C. Peng, "Mining Trajectory Profiles for Discovering User Communities," *LBSN '09 Proc. 2009 Int. Work. Locat. Based Soc. Networks*, pp. 1–8, 2009.

[24] Y. Zheng, X. Xie, and W. Ma, "GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory.," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–40, 2010.

[25] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Finding similar users using category-based location history," *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. - GIS '10*, no. 49, p. 442, 2010.

[26] J. He and W. W. Chu, *A Social Network-Based Recommender System (SNRS)*, vol. 12. 2010.

[27] P. Bonhard, C. Harries, J. McCarthy, and M. Sasse, "Accounting for taste: using profile similarity to improve recommender systems.," *CHI 06 Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 1057–1066, 2006.

[28] P. Bourdieu, *Distinction: A social critique of the judgment of taste*, vol. 1, no. 3. 1984.

[29] W. Knulst and G. Kraaykamp, "Trends in leisure reading: Forty years of research on reading in the Netherlands," *Poetics*, vol. 26, no. 1. pp. 21–41, 1998.

[30] K. van Eijck, "Social Differentiation in Musical Taste Patterns," *Soc. Forces*, vol. 79, no. 3, pp. 1163–1185, 2001.

[31] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, "Tastes, ties, and time: A new social network dataset using Facebook.com," *Soc. Networks*, vol. 30, no. 4, pp. 330–342, 2008.

[32] S. Zukin, "Urban Lifestyles: Diversity and Standardisation in Spaces of Consumption," *Urban Stud.*, vol. 35, no. 5–6, pp. 825–839, 1998.

[33] D. Harvey, *The Condition of Postmodernity*, vol. 67, no. 2. 1991.

[34] S. Lash and J. Urry, *Economies of Signs and Space*. 1994.

[35] S. Zukin, *The Cultures of Cities.*, vol. 25, no. 6. 1996.

[36] P. Mullins, K. Natalier, P. Smith, and B. Smeaton, *Cities and Consumption Spaces*, vol. 35, no. 1. 1999.

[37] R. Bocock, *Consumption*. 1993.

[38] M. Featherstone, *Consumer culture and postmodernism*. 1991.

[39] S. Clarke, "Review: The World of Consumption. by Ben Fine; Ellen Leopold Review," *Contemp. Sociol.*, vol. 23, no. 6, 1994.

[40] D. Miller, "Consumption Studies as the Transformation of Anthropology," in *Acknowledging Consumption*, 1995, pp. 272–301.

[41] E. Schlosser, *Fast food nation: The dark side of the all-American meal*. Houghton Mifflin Harcourt, 2012.

[42] "Yelp information." Yelp, 2017.

[43] Yelp Dataset Challenge, "Yelp Dataset Challenge," 2017. [Online]. Available: https://www.yelp.com/dataset_challenge. [Accessed: 15-Feb-2017].

[44] M. Danilak, "langdetect 1.0.7: Language detection library ported from Google's language-detection," 2014. [Online]. Available: https://github.com/Mimino666/langdetect. [Accessed: 15-Jan-2018].

[45] H. Sajnani and V. Saini, "Classifying Yelp reviews into relevant categories," 2008. [Online]. Available: http://www.ics.uci.edu/ vpsaini/. [Accessed: 15-Jan-2018].

[46] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.," 2009.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Advances in neural information processing systems. 2013," *Distrib. Represent. words phrases their Compos.*, pp. 3111–3119.

[48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 June 2018 doi:10.20944/preprints201806.0389.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 376; doi:10.3390/ijgi7090376

space," *arXiv Prepr. arXiv1301.3781*, 2013.

[49] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[50] J. a. Hartigan and M. a. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *J. R. Stat. Soc. C*, vol. 28, no. 1, pp. 100–108, 1979.

[51] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, 2008.

[52] Y. Li, C.-Y. Chen, and W. W. Wasserman, "Deep feature selection: theory and application to identify enhancers and promoters," *J. Comput. Biol.*, vol. 23, no. 5, pp. 322–336, 2016.

[53] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[54] D. M. W. POWERS, "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, 2011.

[55] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer (Long. Beach. Calif).*, vol. 42, no. 8, pp. 42–49, 2009.

[56] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems," *Fifth Int. Conf. Comput. Inf. Sci.*, pp. 27–28, 2002.

[57] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[58] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[59] H. El Nasser, "Census data show 'surprising' segregation," *USATODAY.com*, 2010. [Online]. Available: http://usatoday30.usatoday.com/news/nation/census/2010-12-14-segregation_N.htm. [Accessed: 11-Feb-2017].

[60] S. Sassen, "The Global City: New York," *London, Tokyo*, vol. 41, 1991.

[61] "C. B. A. C. S. Office, United states census bureau / american factfindeler," 2016.

[62] E. Fong and M. Gulia, "Differences in neighborhood qualities among racial and ethnic groups in Canada," *Sociol. Inq.*, vol. 69, no. 4, pp. 575–598, 1999.

[63] B. Jiang, D. Ma, J. Yin, and M. Sandberg, "Spatial distribution of city tweets and their densities," *Geogr. Anal.*, vol. 48, no. 3, pp. 337–351, 2016.

[64] J. Jacobs, "The Death and Life of Great American Cities," in *New York*, vol. 71, 1961, p. Alexander, C., Ishikawa, S., Silverstein, M. (19.