A Large-Scale Study of Cultural Differences Using Urban Data About Eating and Drinking Preferences

Thiago H Silva[a,*], Pedro O S Vaz de Melo[b], Jussara M Almeida[b], Mirco Musolesi[c], Antonio A F Loureiro[b]

[a]*Department of Informatics, Universidade Tecnológica Federal do Paraná, Curitiba, Brazil*
[b]*Computer Science Department, Universidade Federal de Minas Gerais, Belo Hozizonte, MG, Brazil*
[c]*Department of Geography, University College London, London, United Kingdon*

Abstract

Traditional ways to study urban social behavior, e.g. surveys, are costly and do not scale. Recently, some studies have been showing new ways of obtaining data through location-based social networks (LBSNs), such as Foursquare, which could revolutionize the study of urban social behavior. We use Foursquare check-ins to represent user preferences regarding eating and drinking habits. Considering datasets differing in terms of volume of data and observation window size, our results indicate that spatio-temporal eating and drinking habits of users voluntarily expressed in LBSNs has the potential to explain cultural habits of the users. From this, we propose a methodology to identify cultural boundaries and similarities across populations at different scales, e.g., countries, cities, or neighborhoods. This methodology is extensively evaluated in several aspects. For instance, by proposing some variations of it disregarding some of the considered dimensions, as well as analyzing the results using datasets from different periods and window of observation. The results indicate that our proposed methodology is a promising approach for automatic cultural habits separation, which could enable new urban services.

Keywords: Location-based social network, large scale assessment, urban data mining, cross-cultural study, Foursquare

*Corresponding author. Universidade Tecnológica Federal do Paraná. Av. 7 de setembro 3165 - 80230-901. DAINF. Curitiba, Brazil.
Email address: thiagoh@utfpr.edu.br (Thiago H Silva)

1. Introduction

The analysis of cultural differences among people and regions is usually performed using surveys based on questionnaires filled during face-to-face interviews [1], such as the Eurobarometer dataset [2]. Through these questionnaires, individual preferences, such as the taste for coffee and fast food, can be mapped into multidimensional vectors representing (and characterizing) each interviewee. From these vectors, it is possible, for instance, to quantify how similar or different two individuals are.

Although survey data are broadly used in the analysis of cultures, there are some severe constraints in its use, which are well known to researchers. First, surveys are costly and do not scale up. That is, it is hard to obtain data of millions, or even thousands of people. Second, they provide static information, i.e., they reflect the preferences of users at a specific point in time.

Recently some studies have revealed a new way of collecting data via location-based social networks (LBSNs) that could revolutionize the study of urban social behavior [3, 4, 5, 6]. Specifically, in this present study, we propose the use of publicly available data from Foursquare to map the individual preferences of users. This is interesting because a check-in on an LBSN expresses the preference of a user for a certain kind of place. In addition, LBSNs are accessible almost everywhere and by anyone, helping to solve the scalability issue and allowing data in various regions of the world to be collected, despite some possible limitation, as discussed on Section 7.

The study of the influence of cultural differences in human behavior is a challenging topic. Culture is a concept so complex and interesting that no single definition can capture it. Among the various aspects that define the culture of a society include its arts, religious beliefs, literature, and manners. Moreover, as Counihan [7], and Cochrane and Bal [8] pointed out, eating and drinking habits are also fundamental elements in a culture and may significantly mark social differences, boundaries, bonds, and contradictions. Since eating and drinking habits have such importance for a culture, we here address the topic of investigating and analyzing life and idiosyncrasies of different societies through them.

The identification of cultural boundaries could enable new/smarter urban services

and applications. Since culture is an important aspect for economic reasons [6], our methodology is valuable for companies that have businesses in one country and want to verify the compatibility of preferences across different markets. Another application that could rely on our methodology is a place recommendation system, which is useful for visitors and residents of a city. Foursquare estimates that only 10% to 15% of searches on Foursquare are for specific places [9]. Much more often users are searching within broader categories, such as "sushi" [9]. Based on this information, systems like Foursquare and other location-based search engines, as the one proposed in [10], could benefit from the introduction of new criteria and mechanisms in their recommendation systems that consider cultural differences between areas. For instance, a person who enjoyed a specific area of Tokyo could receive a recommendation of a similar area when visiting Chicago.

Based on this, the contributions of this work can be summarized as follows:

- We study properties of user's food and drink preferences worldwide, understanding how they change according to the time of day and geographic locations. For that, we use two Foursquare datasets differing in terms of volume of data and observation window, one of them covers a worldwide event: 2014 FIFA World Cup. We observe that the properties extracted from these different datasets are very similar to each other.

- We propose a new methodology for identifying cultural boundaries and similarities between societies, considering food and drink preferences. For this, we use Foursquare check-ins to represent a user's preferences regarding what he/she eats and drinks locally, for example, in a particular city. This proposed methodology can be used to identify similar cultures in regions of different sizes, such as countries, cities, or even regions inside cities;

- We perform an extensive evaluation of our proposed methodology in several aspects. First, using different Foursquare datasets in our methodology, we verify that the results of cultural separation are similar. We compare our results using a cultural map of the world based on the World Values Surveys (WVS), which

uses data from traditional surveys, and the similarities are striking. Then, we check the impact of observation window size on the results. In this analysis, big changes in the results were not observed exploring an observation window larger than one week. Finally, we evaluate two additional variations in our proposed approach for identifying cultural boundaries and similarities. The results of cultural separation obtained using these variations are inferior compared to those obtained by the original approach.

The rest of the work is organized as follows. Section 2 presents the related work. Section 3 describes our datasets and the core of our methodology for extracting cultural preferences from LBSNs. Section 4 investigates the cultural similarities between individuals and shows that food and drink check-ins outperform check-ins given in all types of places in this case. Section 5 investigates spatio-temporal properties of users' food and drink preferences considering different datasets. Using this knowledge, Section 5.3 proposes a methodology to identify similar areas around the planet according to their cultural aspects. Section 5.3 compares the results with survey data and also studies the impact of observation window size in the results. Section 6 proposes and evaluates two variations of our approach to identify culturally similar areas. Section 7 presents a final discussion and possible limitations of our results. Finally, Section 8 concludes the work.

2. Related Work

2.1. Social Media, The City, and Urban User Behavior

Several studies have focused on the spatial properties of data shared in location-based services such as Foursquare [11, 12, 13]. However, those prior efforts aimed mostly at investigating user mobility patterns or social network properties and their implications.

More recently, researchers have started looking at user activity as another data source that can be leveraged for studying social interactions. For example, Kershaw et al. [14] looked into the use of social media to monitor the rate of alcohol consumption. Venerandi et al. [15] proposed to use user-generated content to mine ur-

ban deprivation, data that otherwise is costly to be obtained. Cranshaw et al. [4] presented a model to extract distinct regions of a city according to current collective activity patterns. Similarly, Noulas et al. [5] proposed an approach to classify areas of a city by using all venues' categories of Foursquare. Besides these studies, we can also cite many other studies that present new insights about city dynamics such as, for example, their key characteristics and the behavior of their citizens. Zambaldi et al. [16] studied how to automatically identify appealing city pictures. Silva et al. [17] explored the transitions of users in the city as a way to differentiate them. There are several other topics of study in this direction, which includes event detection/study [18, 19, 20, 21, 22, 23, 24, 25, 26], gender studies [27, 28, 29, 30, 31], and food and dietary patterns [32, 33, 34, 35, 36, 37, 38, 39].

Particularly related to the latter topic, Wagner et al. [32] showed that dietary patterns observed in an online recipes system reflect well-known habits of the studies countries, a similar study is performed by West et al [39]. Abbar et al. [34] proposed a method to extract nutritional information in Twitter messages. Mejova et al. [36] used the Instagram to identify obesity patterns. Sharma and De Choudhury [38] and De Choudhury et al. [37] also used Instagram posts, but to understand food choices and nutritional characteristics. In this present study, we also explore user food, and also drink, consumption shared in social media. However, instead of studying linguistic characteristics around it, we study the potential of considering it as a way to perform cross-cultural studies (i.e., the study of cultural differences).

2.2. Social Media and Culture

Specifically about cultural differences studies based on social media, some recent work have shown how the use of Web systems vary across countries. For example, Hochman et al. [40] investigated color preferences in pictures shared through Instagram, showing considerable differences in the preferences across countries with distinct cultures. Garcia-Gavilanes et al. [6] and Poblete et al. [41] studied variations of Twitter usage across countries. In particular, Garcia-Gavilanes et al. showed that the culture of a country is associated with the way people use Twitter. In a more recent study, Garcia-Gavilanes et al. [42] perform a study of international Twitter commu-

nication which combines cultural information with geographic, economic, and social features. Large-scale microposts of Twitter are also studied by Gonalves et al. [43]. The authors showed that the considered data is able to reproduce the geospatial adoption of languages for a wide range of resolution scale, being able, for instance, to identify cultural diversity. Reinecke et al. [44] analyzed the use o Doodle[1] around the world and found that culture influences how we schedule events online. Laufer et al. [45] used data from Wikipedia to show particularities in the description of and the interest in different food cultures, and also propose an approach to mine cultural relations in this direction. State et al. [46] considered email and Twitter communications to revisit Samuel Huntingtons theory of changing international alignments [47]. Park et al. [48] explore tweets to study cross-cultural differences in people's use of emoticons as nonverbal cues.

We also have previously performed a study about cultural differences. Differing from all mentioned studies, in [3] we proposed a new methodology for identifying cultural boundaries and similarities across populations, considering eating and drinking patterns (e.g., what kind of food/drink people prefer as well as when they often have their meals). We evaluated this methodology considering a dataset spanning one week of data. The present study greatly builds upon our previous work [3] in three directions. First, we use a new dataset, larger and referring to more recent period, to study spatio-temporal properties of food and drink preferences, observing that the properties between the older and the newer dataset are very similar. We also extensively validate our proposed methodology using this new dataset. We show, for example, that the results of cultural habits agree between the considered datasets. Besides that, we evaluate two additional variations in the proposed approach to the study of cultural differences. The results indicate that these variations are inferior to the original approach. Finally, we evaluate the impact of observation window size in results.

It is worth mentioning that related studies of cultural differences considering social media do not constitute a new research area. Indeed, this type of study has been carried out by researchers working in the social sciences, particularly in cultural an-

[1] http://doodle.com.

thropology and psychology [49]. Despite globalization and many other technological revolutions [50], group formation might lead to the emergence of cultural boundaries that exist for millennia across populations [51]. Axelrod [52] proposed a model to explain the formation and persistence of these cultural boundaries, which are basically a consequence of two key phenomena: social influence [53] and homophily [54]. Homophily dictates that only culturally similar individuals are likely to interact and social influence makes individuals more similar as they interact. In a long run, these two phenomena lead to very culturally distinct groups of individuals, delimited by the so-called *cultural boundaries*.

World Values Surveys (WVS)[2] is a global research project that explores peoples beliefs and values, what social and political impact they have, and how they change over time. It is performed by a group social scientists worldwide who, since 1981, have conducted representative national surveys in several countries in different continents. Using data from the WVS, Ronald Inglehart and Christian Welzel asserts that there are two major dimensions of cross cultural variation in the world: Traditional values versus Secular-rational values and Survival values versus Self-expression values [55].

Traditional values emphasize the importance of religion, parent-child ties, deference to authority and traditional family values. Secular-rational values have the opposite preferences to the traditional values. Survival values place emphasis on economic and physical security. Self-expression values give high priority to environmental protection, growing tolerance of foreigners, gays and lesbians and gender equality, and rising demands for participation in decision-making in economic and political life. Using the dimensions traditional versus secular-rational values and survival versus self-expression values, Inglehart and Welzel produce a cultural map of the world. In that map, countries can be divided into nine clusters: the English-speaking, Latin America, Catholic Europe, Protestant Europe, African, Islamic, South Asian, Orthodox and Confucian ones [55].

[2] http://www.worldvaluessurvey.org.

3. Extracting Cultural Preferences

In this section, we present our dataset and our methodology for extracting cultural preferences from LBSNs.

3.1. Mapping User Preferences

In order to overcome the aforementioned constraints regarding the acquisition of data to the analysis of cultural differences, we propose the use of publicly available data from LBSNs to map individual preferences. LBSNs can be accessed everywhere by anyone who has an Internet connection, solving the scalability problem and allowing data from (potentially) the entire world to be collected [56]. Moreover, these systems are dynamic, being able to capture the behavioral changes of their users when they occur, which solves the second mentioned constraint. However, data from such systems can be used if and only if they meet the requirements:

- **[R1]** It is possible to associate a user to its location;
- **[R2]** It is possible to extract a finite set of preferences from the data that is generated by the system;
- **[R3]** It is possible to map users' actions in the system into the preferences defined in **[R2]**.

Considering that these requirements are met, a dataset containing individual activities of N users of an LBSN can be used to map preferences as follows. First, associate each user n_i with a location l_i, which may be a country, a city or even a region within a city. Then, define a set of m individual preferences (or features) $f_1, f_2, \ldots, f_m$ that can be extracted from the dataset, which may represent the taste for the most varied things, such as Japanese food or a certain football team. Finally, the activities of each individual n_i should be mapped into an m-dimensional vector of preferences $F_i = (f_{1^i}, f_{2^i}, \ldots, f_{m^i})$ that characterizes the person's tastes, the same type of vector that is usually created from survey data [1].

Since the preference vector F_i is generated from self-reported temporal data of an individual n_i, we may populate and modify it in various ways. For instance, we can

use a binary representation, where $f_{k^i} = \{0|1\}$ represents whether user n_i has or not preference f_k (e.g., whether a person likes/dislikes a certain type of food), respectively. Alternatively, we may consider the intensity at which a user likes a feature, inferred from the number of times the corresponding preference is reported in the person's data, i.e., $f_{k^i} = [0; \infty)$. In this paper, we adopt a binary representation because we believe this is a good approximation. The evaluation of other approaches is out of the scope of this present study. Finally, one can group individuals by their geographic regions and sum up their preference vectors to characterize their regions. We adopt this approach in Section 5 to build preference vectors for regions (instead of individuals).

3.2. Data Description

In this work, the datasets used to infer user preferences were collected from one of the currently most popular Location Based Social Networks, namely Foursquare. We collect these data from Twitter[3], since Foursquare check-ins are not publicly available. To that end, we use the Twitter Streaming API[4], obtaining all public Foursquare check-ins returned by this API. Further details about this process can be found in [57]. We collect data representing two periods of time, resulting in two datasets: dataset 1 *(D1)* and dataset 2 *(D2)*. For each dataset tweets containing check-ins are gathered, each one providing a URL to the Foursquare website where information about the venue, in particular, its geographic location and category, was acquired.

In our datasets, each check-in consists of the latitude, longitude, identifier, and category of the venue as well as the time when the check-in was done. The current version of the Foursquare API groups venues into ten categories: Arts & Entertainment; College & University; Professional & Other Places; Residences; Outdoors & Recreation; Shops & Services; Nightlife Spots; Food; Travel & Transport; and Event. Each category, in turn, has subcategories. Table 1 presents all categories and some example of subcategories.

Since we are primarily interested in food and drink habits, we manually group

[3] http://www.twitter.com.

[4] https://dev.twitter.com/streaming/overview.

Table 1: Foursquare categories.

Name	Subcategories examples
Arts & Entertainment	Comedy Club, Movie Theater, Casino
College & University	College Lab, Fraternity House, Student Center
Residences	Home, Residential Building, Trailer Park
Professional & Other Places	Factory, Laboratory, Art Studio
Outdoors & Recreation	Baseball Field, Surf Spot, Park
Nightlife Spots	Bar, Rock Club, Nightclub, Strip Club
Shop & Service	Shoe Store, Nail Salon, Bike Shop
Food	Chinese Restaurant, Bakery, Pizza Place
Travel & Transport	Airport, Hotel, Pier
Event	Christmas Market, Festival, Parade

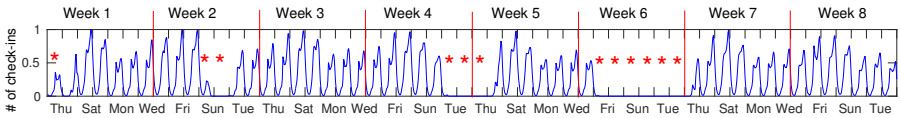


Figure 1: Number of data from Drink, Fast Food and Slow Food classes of D2 throughout the days.

relevant subcategories of the Food and Nightlife Spots categories into three classes: Drink, Fast Food, and Slow Food places. We exclude some subcategories that are not related to these three classes (e.g. Rock Club and Concert Hall) and moving some subcategories (e.g. Coffee Shop and Tea Room) from the Food category to the Drink class. Besides that, we also disregard the category Restaurant, because it is a sort of meta category that could fit in any of the two classes of food. The Drink class has 21 subcategories (e.g., brewery, karaoke bar, and pub), whereas the Fast Food class has 27 subcategories (e.g., bakery, burger joint, and wings joint) and the Slow Food class has 53 subcategories, including Chinese restaurant, Steakhouse, and Greek restaurant.

After this manual classification process, Table 2 summarizes the resulting data, as well as other information about the datasets. As we can see, D1 spans a single week of April 2012. The other dataset, D2, spans a much longer period of a more recent year: 2014. Having this two datasets is particularly interesting because it allows us to study our methodology to capture cultural dynamics in different observation windows.

Table 2: Information about our datasets.

	Dataset 1	Dataset 2	Dataset 3
		Period	
	One week May 2012	Apr/24/2014 to Jun/18/2014	May/8-14 and Jun/5-18 (2014)
		Number of check-ins	
Drink	279,650	1,170,084	540,587
Fast Food	410,592	2,234,502	1,053,530
Slow Food	394,042	1,265,473	586,308
		Number of unique users	
Drink	162,891	426,377	269,505
Fast Food	230,846	596,873	415,327
Slow Food	231,651	458,661	294,437
		Number of unique venues	
Drink	106,152	192,800	125,545
Fast Food	193,541	361,364	231,787
Slow Food	198,565	362,846	227,678

Figure 1 shows the number of data from Drink, Fast Food and Slow Food classes of D2 throughout the days. In this figure, stars represent days that our collection faced some issues, possibly not capturing all shared data on that day. For this reason, we decided to create a new dataset, dataset **D3**, which is a subset of D2 containing only weeks without days with collection issues that are: weeks 3, 7, and 8. It is also interesting to have this dataset because it covers partially a worldwide event: 2014 FIFA World Cup (week 8).

In order to study the similarity of our three datasets, we correlate the number of check-ins given in each of their subcategories (for Drink, Fast Food, and Slow Food classes), using Spearman correlation. Table 3 summarizes the results. As we can see, datasets D2 and D3 are very correlated to each other. The correlation of D2 and D3 with D1 are also very high for the Drink and Slow Food classes, and the correlation with the Fast Food class is fairly high. Despite the suggestion that D2 reflects correctly users' behavior, in the rest of the paper, we disregard D2 in most analysis, in order to prevent any misleading results. We use D2 only in a specific analysis related to observation windows size, shown in Section 5.6, where the incompleteness of this dataset is an interesting feature in the evaluation.

Table 3: Spearman rank correlation of the number of check-ins given in each subcategory for datasets D1, D2, and D3.

Drink class	
Datasets used	ρ (p-value)
D2, D3	0.99 (0)
D3, D1	0.94 (5.4e-07)
Fast Food class	
Datasets used	ρ (p-value)
D2, D3	0.99 (0)
D3, D1	0.8 (1.2e-05)
Slow Food class	
Datasets used	ρ (p-value)
D2, D3	0.99 (0)
D3, D1	0.96 (0)

To provide an idea of the size of the user population LBSNs can reach, consider the World Values Survey[5] project. That study is perhaps the most comprehensive investigation of political and sociocultural change worldwide, which was conducted from 1981 to 2008 in 87 societies, with about 256,000 interviews. Observe that D1 (the smallest dataset) has a population of users of the same order of magnitude of the number of interviews performed in that project in almost three decades.

3.3. Mapping Foursquare Data into User Preferences

Several characteristics of human beings are not directly observable, such as personality traits. Thus, we rely on face-to-face interactions or online signals to discover the presence of those hidden qualities [58]. In this direction, an LBSN check-in can be considered as a signal because it is a perceivable feature/action that expresses the preference of a user for a certain type of place. With that in mind, we use Foursquare check-ins to represent user preferences regarding food and drink places. Specifically, we use the three main classes defined in Section 3.2, namely, *Drink*, *Fast Food*, and *Slow Food*.

[5] http://www.worldvaluessurvey.org.

Table 4: The two most popular subcategories of places for all classes for D1 and D3.

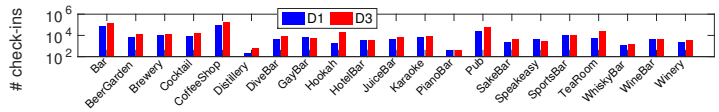
Drink			
Rank position	**Subcategory**	**# of check-ins in D1**	**# of check-ins in D3**
1	Coffee Shop	86,310	182,123
2	Bar	81,124	151,337
Fast Food			
Rank position	**Subcategory**	**# of check-ins in D1**	**# of check-ins in D3**
1	Café	91,303	473,512
2	Fast Food Restaurant	56,648	108,851
Slow Food			
Rank position	**Subcategory**	**# of check-ins in D1**	**# of check-ins in D3**
1	American Restaurant	47,373	48,071
2	Mexican Restaurant	28,712	-
2	Seafood Restaurant	-	47,777

Studying the popularity of different places according to people's preferences worldwide, we note that Coffee Shop and Bar are the two most popular subcategories of Drink places for D1 and D3. Table 4 summarizes these information for all classes. The two most popular Fast Food subcategories are Café[6] and Fast Food Restaurant for D1 and D3.

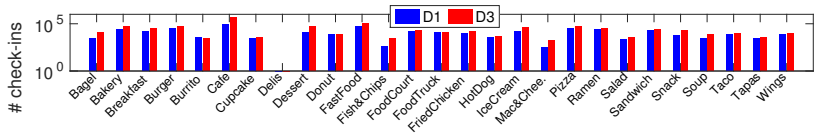
Finally, American Restaurant is the most popular subcategory for D1 and D3. The second most popular subcategory in Slow Food class for D1 is and Mexican Restaurant, and for D3 is Seafood Restaurant. Mexican Restaurant is still also very popular in D3 (among the top positions). This small change in the popularity might be related to a season that is only covered by the new dataset: summer. Figures 2a, 2b, and 2c show the number of check-ins at each subcategory of the Drink, Fast Food, and Slow Food classes, respectively, so we can have a general idea about the popularity of user preferences for different food and drink related places. These figures show the popularity of different places according to people's preferences worldwide.

In these datasets, a user is represented by a vector of $m = 101$ features correspond-

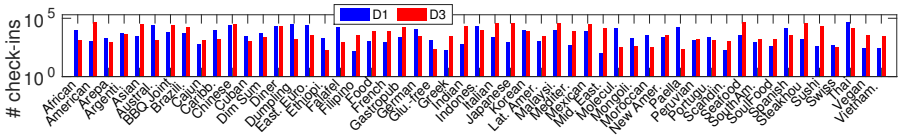
[6]Like in many European countries, this term is referred to a restaurant primarily serving coffee as well as pastries.



(a) Drink



(b) Fast Food



(c) Slow Food

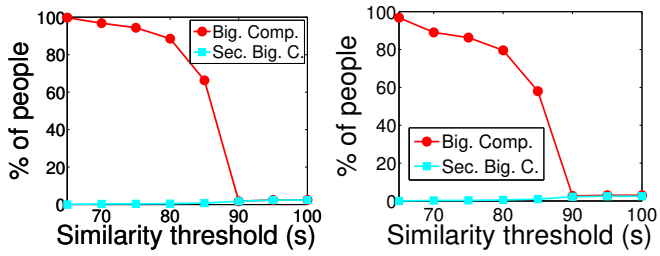
Figure 2: Number of check-ins at all subcategories of the three analyzed classes for D1 and D3. The names of some places are abbreviated but the semantics of the names is preserved.

ing to the 101 subcategories that comprise the three classes we have defined. A feature $f_i \in F = \{f_1, f_2, \ldots, f_{101}\}$ is equal to 1 if a user made at least one check-in at f_i, and 0 otherwise. In this way, a feature vector represents the positive and negative preferences of a user for fast food, slow food, and drink subcategories. With that, a finite set of preferences is extracted (requirement **[R2]**, see definition in Section 3.1) and users' actions are mapped into this set (requirement **[R3]**). To associate a user with a location (requirement **[R1]**), we analyzed the GPS coordinates of all check-ins performed by the user. If all check-ins performed are from the same country, according to the free reverse geocoding API offered by Yahoo[7], we assume that the user taken into consideration is from that country. Otherwise, we do not consider the user in our analysis. In this way, we minimize the wrong association of a user with a country. Following this procedure, approximately 1% of the users were disregarded from our analysis.

[7] http://developer.yahoo.com.

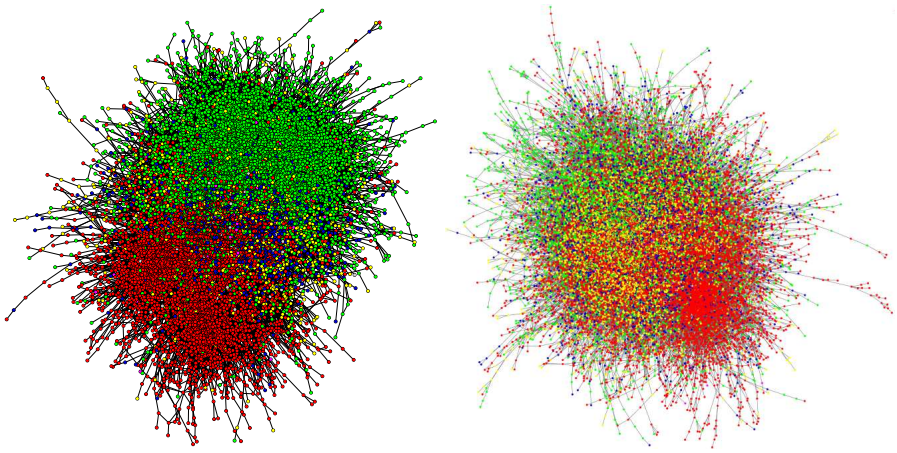
4. Cultural Analysis of Individuals

In this section, we use the map of preferences presented in Section 3.3 to analyze
305 the individual preferences of users, showing, among other results, that food and drink preferences are good indicators of cultural similarities.



(a) (b)

Figure 3: Characterization of similarity networks. (a) % of users in the 2 largest component of G_s^1. (b) % of users in the 2 largest component of G_s^2



(a) Largest component of G_s^1 (b) largest component of G_s^2

Figure 4: [Better in color] Similarity network for the 0.65-network. Node colors: Africa (Pink), Asia (Red), Central and South America (Yellow), Europe (Blue), North America (Green), Oceania (Grey).

In order to assess the cultural similarities among users, we construct a similarity network $G_s = (V_s, E_s)$, where s is a similarity threshold used to build the network, vertices V_s represent the set of users, and an edge (v_i, v_j) exists in E_s if users v_i and

v_j have a similarity value above s. The similarity value $s_{i,j}$ between two users v_i and v_j is the Jaccard index (JI) between their preference vectors[8] multiplied by 100. In this way, $s_{i,j}$ varies from 0 to 100 and measures the percentage of preferences shared by the users v_i and v_j. For example, considering a similarity threshold $s = 65$ (or 65%-network[9]), there is an edge between vertices v_1 and v_2 if the corresponding users have, at least, 65% of preferences in common. We have built two similarities networks, G_s^1 and G_s^2, taking into account dataset D1. The network G_s^1 considers only food and drink preferences, i.e., only check-ins at food and drink places. On the other hand, G_s^2 consider all preferences, i.e., all Foursquare subcategories, 435 in our dataset, including food and drink venues. To build both networks we consider only users who performed at least 7 check-ins in the dataset (i.e., at least one check-in per day on average). In total, 28,038 users were considered in G_s^1 and 194,902 in G_s^2. Moreover, isolated nodes were disregarded. We here consider the following values of $s \in \{65, 70, 75, 80, 85, 90, 95, 100\}$. Note that G_s^1 and G_s^2 are undirected unweighted and symmetric graphs.

We first analyze relevant properties of G_s^1 and G_s^2. Figures 3a and 3b show the percentage of vertices (i.e., users) in the two largest components of the network G_s^1 and G_s^2, respectively, for various values of s. These figures show that the largest component of the 65%-network practically contains all nodes, reason why we present results starting from this network. The percentage of users in the largest component slowly decreases as the similarity threshold increases, until s reaches 85. For larger values of s, the number of users in the largest component drops sharply, becoming comparable to the size of the second largest component. This is explained by observing networks built using large values for s, such as the 100%-network, where every component is composed of very similar users. Since users with very similar preferences are rare, the largest components tend not to have very large differences in size.

Next, we visualize the 0.65-network for the largest components of G_s^1 and G_s^2 in Figures 4a and 4b, respectively. In these figures, the color of a node indicates that a user is from a specific region of the world: Africa (Pink), Asia (Red), Central and South

[8]The Jaccard index of sets A and B is computed as $\frac{A \cap B}{A \cup B}$.

[9]Network created with a threshold s is referred to as s-network.

America (Yellow), Europe (Blue), North America (Green), Oceania (Grey). This figure suggests that vertices with the same color are grouped together more in the network G_s^1 than G_s^2. This indicates that in G_s^1 there is a significantly larger number of edges between users from geographic areas close to each than in G_s^2. Observe in the Figure 4a the clearer separation between Western and Eastern countries and the centrality of Europe, representing the cultural separation among those regions, a fact that is not clearly observed in Figure 4b.

In order to further investigate this insight, we calculate the assortativity of these networks shown in Figure 4. Assortativity measures the similarity of connections in the network with respect to a given attribute and varies from -1 to +1 [59]. In an *assortative network* (with positive assortativity), vertices with similar values of the given attribute (e.g., same country) tend to connect with (be similar to) each other, whereas in a *disassortative network* (with negative assortativity), the opposite happens. We calculated the assortativity analysis with respect to the geographic attributes: i-Western-Eastern parts of the world (i.e., west-east hemispheres); ii-Continent; iii-and Country. The assortativity analysis for the networks formed considering $s = 0.65$ are 0.31 and 0.12 for the attribute Country, 0.48 and 0.14 for Continent, and 0.38 and 0.19 for Western/Eastern division, for G_s^1 and G_s^2, respectively. The results imply that all these similarity networks are assortative. However, the assortativity values of the geographic attributes for G_s^1 are higher compared to those obtained for G_s^2. This suggests that a similarity network considering only food and drink preferences might provide better insights in the study of cultural differences. For this reason, in this study, we dedicate our efforts to investigate only features extracted from food and drink check-ins.

5. Extraction of Cultural Signatures

Given the results discussed in Section 4, we hypothesize that it is possible to define cultural signatures of different areas around the planet. In this section, we show how to extract features from Foursquare data that are able to describe regions from their cultural elements. In particular, we investigate two properties of food and drink pref-

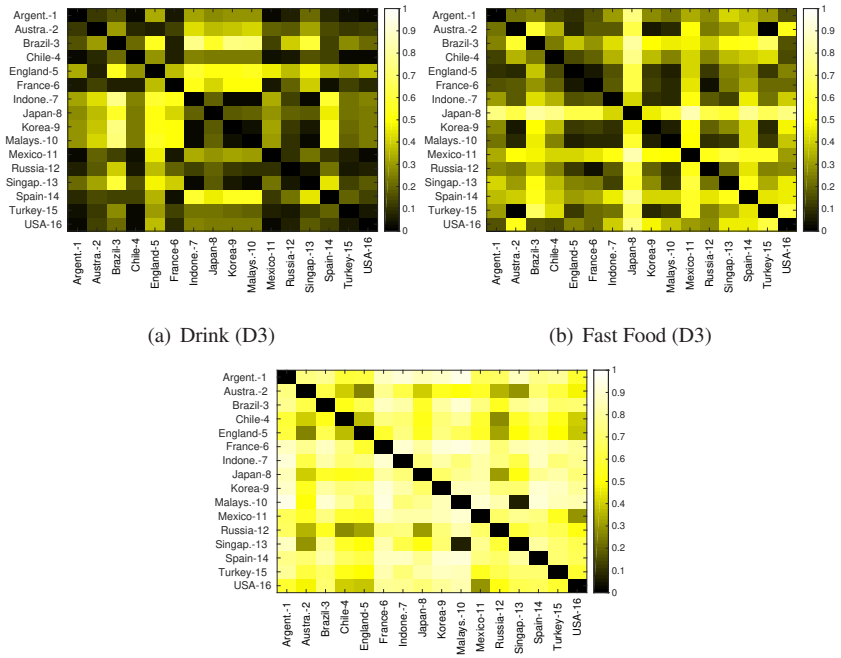
erences: their geographic (Section 5.1) and temporal (Section 5.2) aspects. All these analyses were performed considering all datasets studied in this work. However, in order to increase readability of the study, we are not presenting all results. We show only results for dataset D3 because it represents a period without interruption in the data collection process.

5.1. Spatial Correlations

Here our goal is to define a set of features that are able to characterize the cultural preferences of a given geographic area on the planet, such as a country, a city or a neighborhood. Thus, for a given delimited area a (e.g., the city of Chicago), we sum up the values of the features in the preference vectors of the users who checked in at venues of that area. In other words, we count the number of check-ins $C^a = c_1^a, c_2^a, \ldots, c_{101}^a$ performed in venues of each of the 101 subcategories $s_1, s_2, \ldots, s_{101}$ of the Fast Food, Slow Food and Drink classes (Section 3.2) that are located within the perimeter of area a. Next, we represent each area a by a vector of 101 features $F^a = f_1^a, f_2^a, \ldots, f_{101}^a$, where each feature f_i^a is equal to $c_i^a / \max(C^a)$. That is, we normalize the number of check-ins at each subcategory by the maximum number of check-ins performed in a single subcategory in the area a ($\max(C^a)$). Thus, each area a is represented by a feature vector F^a containing values from 0 to 1, indicating the preferences of people who visited that area, i.e., the profile of preferences for that area. From now on, we use F^a_{drink}, F^a_{sfood} and F^a_{ffood} to refer, respectively, to the subset of features that correspond to subcategories belonging to the Drink, Slow Food and Fast Food classes in the area a.

In order to verify if two areas a and b are culturally similar, we compute the cosine similarity between the two feature vectors F^a and F^b of those areas. We compute the similarity considering all features (F^a and F^b) as well as a subset of them (e.g., F^a_{drink} and F^b_{drink}). In particular, Figure 5 shows the similarity of preferences between 16 different popular countries for the Drink, Fast Food, and Slow classes; the darker the color, the stronger the similarity. The same similarity computed for city level areas (27 cities around the world) are shown in Figure 6.

The results considering dataset D1 and D3 for all classes and countries or cities are



(a) Drink (D3)

(b) Fast Food (D3)

(c) Slow Food (D3)

Figure 5: Similarity of preferences between countries considering dataset D3.

very similar[10]. Analyzing the results for the Drink class for dataset D3 (Figure 5a), we find countries with very strong similarities, such as Argentina and Chile, as well as countries with low similarities, such as Brazil and Indonesia. Moreover, although regions close geographically tend to have stronger similarities, this is not always the case. For example, the similarity between Argentina and France is stronger than the similarity between England and France, which are geographically closer. The similarity of Argentina and France is slightly stronger when considering data from dataset D3 compared with D1. This particular example helps to illustrate that some small changes may, not surprisingly, happen because a larger dataset might enable a clearer image of the analyzed users' preferences. For example, some activities may be more common during the summer, period that is only covered by the new dataset.

[10]The results for D2 are also very similar.

Similarly, Figure 6[11] shows that cities in the same country tend to have very similar drinking habits in most cases, but there are exceptions. Manaus (Brazil), for example, has weak similarity with other cities in Brazil. We conjecture that this result might be associated with the fact that this city is located in the North region of Brazil, which is known for having a strong cultural diversity compared to other parts of the country.

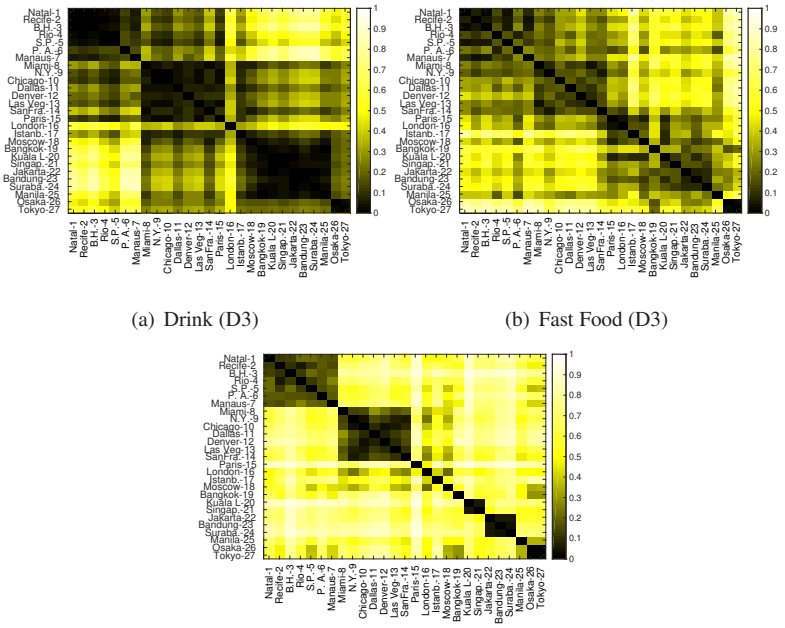
Turning our attention to food practices, we observe in Figures 5b and 6b a strong similarity of fast food habits between several localities at both country and city levels. This is not observed in the same intensity for the Slow Food class (Figures 5c and 6c). The explanation might be related to the diffusion of fast food places worldwide [60].

The Slow Food class presents the highest distinction, or smaller similarity, across most of the countries and cities. This is expected since Slow Food venues usually are representative of the local cuisine. Note, for instance, that cities from Brazil and USA have highly similar drinking and fast food habits, but almost no similarity in slow food habits.

Finally, we turn our attention to the cultural habits within city boundaries. It is known that, in many cities, there is a strong cultural diversity across different neighborhoods [4], reflecting distinct activities typically performed in these areas. To analyze these local cultures, we focus on three populous cities, namely London, New York, and Tokyo. We divide each city's geographic area using a grid structure. Next, we select the most popular cells in the grid of each city, according to the number of check-ins, and label them with a number, as shown in Figure 7. We then compute the similarity between the selected cells. Note that we here assume a grid with regular (rectangular) cells to show the potential of the proposed analysis. However, our approach can be applied to any other segmentation of the city areas (e.g., by city districts).

Figure 8 shows the similarities for pairs of cells within the same city and from different cities for dataset D3. Note that, for the Drink class (Figures 8a), different areas within the same city tend to have very strong similarities. There are also areas from different cities with strong similarities (e.g., areas NY-7 and TKO-1). For Fast Food places (Figures 8b), the similarities between areas within the same city are much

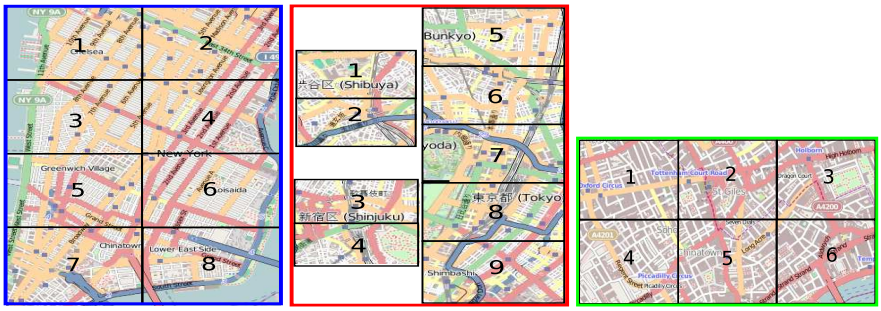
[11] The ratio of check-ins per inhabitant is similar among all the cities taken into consideration.



(a) Drink (D3) (b) Fast Food (D3)

(c) Slow Food (D3)

Figure 6: Correlation of preferences between cities considering dataset D3.

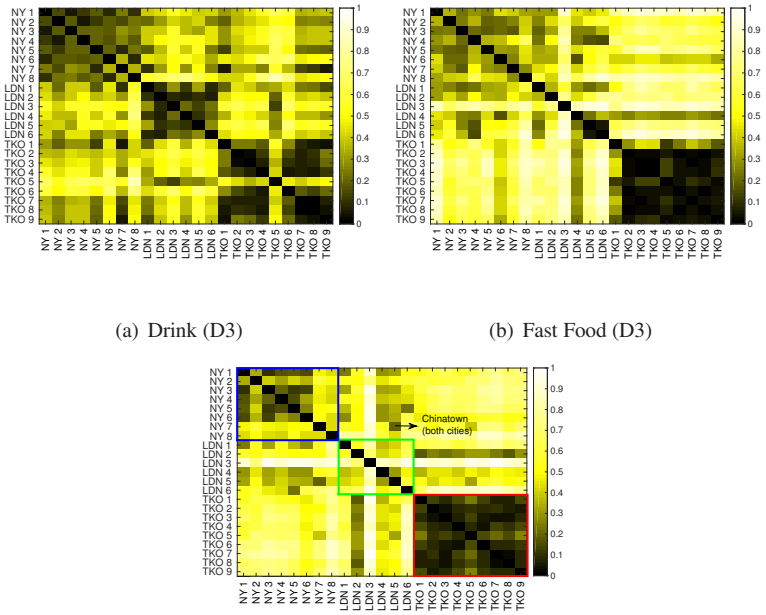


(a) NY (b) Tokyo (c) London

Figure 7: [Better in color] Areas of cities taken into consideration: London/England; New York/USA; and Tokyo/Japan.

stronger for Tokyo, although the similarities between New York and London areas are fairly moderate. In contrast, there are areas very distinct in terms of cosine similarity.

Finally, for the Slow Food class, Tokyo areas, once again, are very strongly similar among themselves. In comparison with the Fast Food class, there is a more clear



(a) Drink (D3)

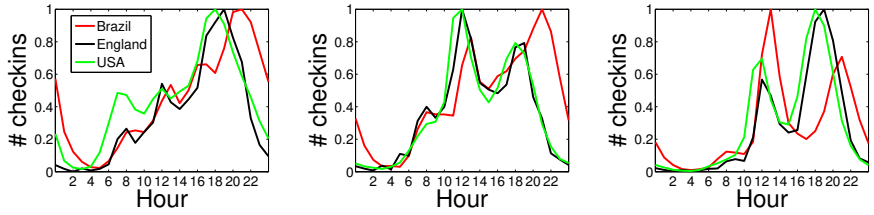
(b) Fast Food (D3)

(c) Slow Food (D3)

Figure 8: Correlation of preferences between regions considering dataset D3.

distinction (weaker similarity) between London and New York areas as well as among distinct areas in London. This last observation is probably due to a specific characteristic of London, that has neighborhoods with a strong presence of a cuisine of a particular region of the globe. Observe also that two specific areas of New York, namely NY-7 and NY-8, are particularly not similar with the others from this city. This is probably related to the location of Chinatown in those areas (mainly NY-7). Indeed, this particular area (NY-7) has a strong similarity with a particular area of London, LND-5, where Chinatown/London is located.

All messages discussed in this section are valid for D1 and D3. Some small differences in the cosine similarity calculated between certain areas were observed, however, these differences are not significant for the vast majority of cases, i.e., the results are very similar.



(a) Drink (D3) (b) Fast Food (D3) (c) Slow Food (D3)

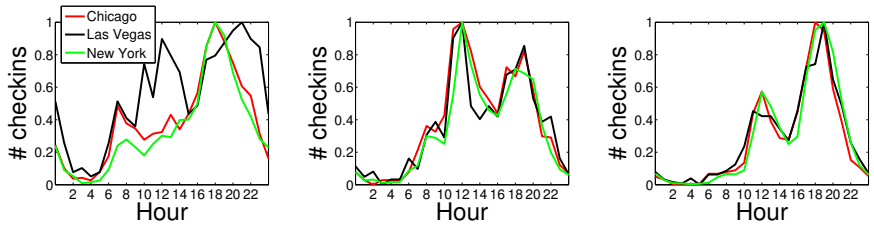
Figure 9: # of check-ins throughout the hours of the day in different countries during weekdays. Results for dataset D3.

5.2. Temporal Analysis

We now turn our attention to the temporal and circadian aspects of cultural habits. The time instants when check-ins are performed in food and drink places may also provide valuable insights into the cultural aspects of a particular region. For example, in a particular area, one may like to drink beer during the weekends but not during the weekdays.

To that end, we first count the number of check-ins per hour during the whole week covered by our dataset in venues of each class (Drink, Fast Food, and Slow Food) for different regions. Next, we group days into weekdays (Monday to Friday[12]) and weekends (Saturday and Sunday), summing up the check-ins performed on the same hour of the day in each group and for each region. We then normalize this number by the maximum value found in any hour for the specific region, so that we can compare the patterns obtained in different regions. For illustration purposes, we show the results for three countries (Brazil, USA, and England) and for three American cities (Chicago, Las Vegas, and New York). Results for each class are shown separately for weekdays (Figure 9 for countries and Figure 10 for cities) and weekends (Figure 11 for countries and Figure 12 for cities).

[12]One could argue that Friday evening is part of the weekend. We evaluated the impact in the time series considering this possibility and we did not observe significant changes. In addition, we also evaluated the impact on the results provided by our methodology for cultural boundaries identification, explained in Section 5.3, and we did not observe alterations.



(a) Drink (D3) (b) Fast Food (D3) (c) Slow Food (D3)

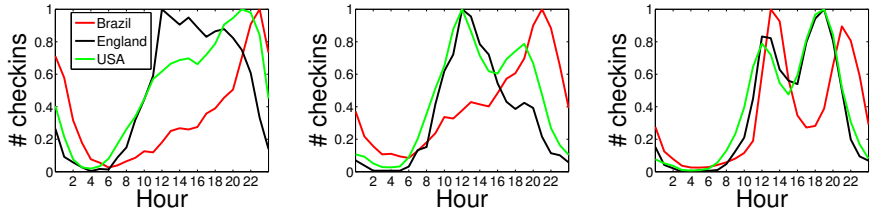
Figure 10: # of check-ins throughout the hours of the day in different American cities during weekdays. Results for dataset D3.

As observed for users' preferences of type of places discussed above, the temporal correlation when users perform eating and drinking activities, in all datasets and in all cases studied, are very similar.

Discussing first weekday patterns, Figure 9 shows that American and English people have similar peaks of activities, despite differences in their preferences for different categories of places, as previously shown (Figure 5). In contrast, Brazilians tend to have significantly different temporal patterns, particularly in terms of activities in Slow Food places (Figure 9c): whereas Americans and English people tend to have their main meal at dinner time, Brazilians have it at lunch time. Observe also that Brazilians have their meals later, compared to Americans and English people.

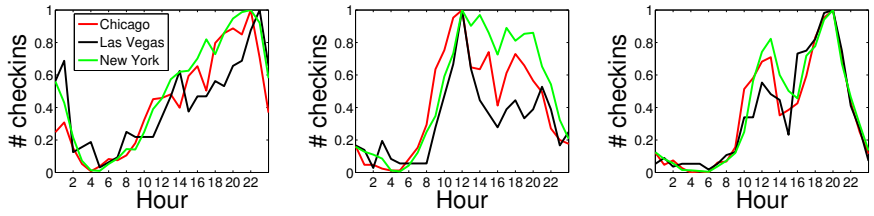
Concerning the times when people go to drink venues, it is possible to note similarities among most of the cities from the same country, but also some different patterns. For example, most of the analyzed cities from USA exhibit a weekday pattern similar to New York and Chicago, shown in Figures 10a, with two distinct peaks around breakfast, lunch, and happy hour time (around 6 $p.m.$). This behavior is consistent with the general pattern observed for the country, shown in Figure 9a. However, Las Vegas is one exception, since there is an intense activity during the dawn, besides many other peaks of activities throughout the day that do not occur in other cities. Despite a more smooth curve pattern compared to the result using D1, the pattern for this city still very distinct from other cities.

Turning our attention to eating habits on weekdays, Figure 10 shows that most cities in the United States present activity patterns very similar to the general pattern



(a) Drink (D3) (b) Fast Food (D3) (c) Slow Food (D3)

Figure 11: # of check-ins throughout the hours of the day in different countries during weekends.



(a) Drink (D3) (b) Fast Food (D3) (c) Slow Food (D3)

Figure 12: # of check-ins throughout the hours of the day in different American cities during weekends.

identified for the country, both in terms of Slow and Fast Food places. Note that Las Vegas in this aspect, differently from drink habits, follows the general pattern of the country, as we can see in Figure 10c. This result is not observed using D1, where Las Vegas presents distinct trends. This suggests that cities with inherent idiosyncrasies, such as Las Vegas, might need a larger time observation window to minimize the impact outlier behavior in the results.

We also note relevant similarities and differences in eating habits of people from cities in different countries. For instance, comparing Figure 10c with similar graphs produced for different Brazilian cities, we find that the curves for Slow Food places are quite different, reflecting distinct habits for each country, as discussed previously. This was also observed for photo sharing pattern [61].

The curves for weekends have very distinct peaks of activities from those of weekdays, both at the country and city levels. Again, this is the case for all studied datasets. For instance, as shown in Figure 11a, English people have a very distinct drinking pattern from Americans on weekends. The pattern representing Slow Food activities is

the only one that is more similar to the pattern observed for weekdays. This could be explained because such places (often restaurants) have well-defined opening hours, serving meals around lunch and dinner times only, which coincide with the times of check-in peaks (Figures 11c, 11d, 12c, and 12d). For this reason, two clear peaks are expected to represent those classical periods. The curve representing Brazil, in this case, presents a bigger number of activity around dinner times compared to weekdays. This pattern could be explained by the fact that during weekends Brazilians tend to go more in restaurants at dinner time.

Specifically about the pattern observed at city level during weekends, we note that there is no clear (dominant) temporal check-in pattern for Fast Food places, as observed for weekdays when considering different cities of a country (Figure 12b). However, we do note that most activities happen after noon, which was expected. In contrast, there is a dominant pattern for check-ins at Slow Food places during weekends (Figures 12c and 12d), and it is similar to the one observed on weekdays. This result could be also explained by the classical periods of restaurants mentioned previously.

5.3. Discussion

We analyzed temporal and spatial patterns of check-ins at different types of places for datasets that span different periods of time, and the results are very similar for all of them.

In addition to this analysis, we also compute Shannon's entropy [62] of preferences for each venue subcategory among all considered areas. The goal is to analyze whether the check-ins at specific subcategories are more concentrated at specific areas (low entropy) or not (high entropy).

We compute the entropy for subcategories of each class (Drink, Fast Food and Slow Food) at country and city levels. Appendix A presents the results for all the subcategories of the Drink, Fast Food, and Slow Food classes, respectively. For all cases, the minimum entropy value is 0 and the maximum one is 4.76 for cities and 4 for countries. Sake bar is one example with low entropy, the value is 1.14 and 1.63 for countries and cities, respectively. This indicates that this subcategory is popular in very few countries and cities. Surely Japan contributes considerably to this result.

Looking at fast food related places, one of the highest entropy was found for the cafe subcategory (3.96 for countries and 4.62 for cities). This is not a surprise because it is a type of place very common worldwide. Considering the Slow Food class, some of the high entropy values reflect the widespread popularization of various cuisines. For example, a check-in at an Italian restaurant does not necessarily mean that it represents a behavior of an Italian since it is a very international type of restaurant, confirmed by the high entropy, almost reaching the maximum value for countries and for cities. Note, however, that if the check-in at an Italian restaurant is made at lunch time it could be more likely to represent a Brazilian behavior than American, since Brazilians have their main meal at lunch time, as presented in Section 5.2. Time plays an important role in this case.

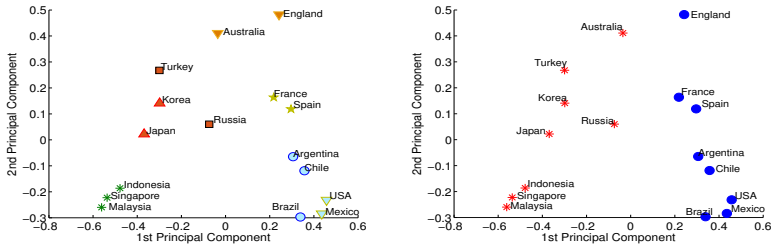
All these observations increase our confidence that spatio-temporal similarities of check-ins are good cultural signatures of regions, as used in the technique described in the next section.

sectionIdentifying Cultural Boundaries

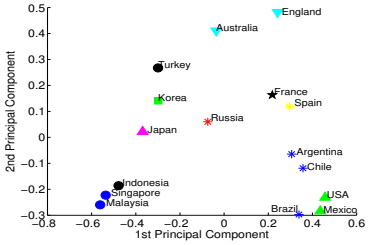
5.4. Clustering Geographical Areas

In this section, we use the cultural signatures of areas described above to identify similar areas around the planet according to their cultural aspects, delineating their so-called "cultural boundaries". To that end, we first represent each area a by a high dimensional preference vector composed of 808 features, namely the normalized number of check-ins at each of the 101 subcategories in four[13] disjoint periods of the day (0 $a.m.$ to 5:59 $a.m.$ (dawn), 6 $a.m.$ to 11:59 $a.m.$ (morning), 12 $a.m.$ to 5:59 $p.m.$ (afternoon), and 6 $p.m.$ to 11:59 $p.m.$ (night)), on weekdays and on the weekends. We then apply the Principal Component Analysis (PCA) [63, 64] technique to these vectors to

[13] We tested the results for countries considering 2 divisions (0 $a.m.$ to 11:59 $a.m.$ and 12 $a.m.$ to 11:59 $p.m.$) and 8 divisions (periods of 3 hours starting from 0 $a.m.$). The results with 2 divisions were inferior related to the one obtained with 4 divisions. The results considering 8 divisions was the same as considering 4 divisions, so we considered 4 divisions because it leads to fewer dimensions. Note that other periods could be considered, but their evaluation is left for future work.



(a) $k = 7$ (b) $k = 2$



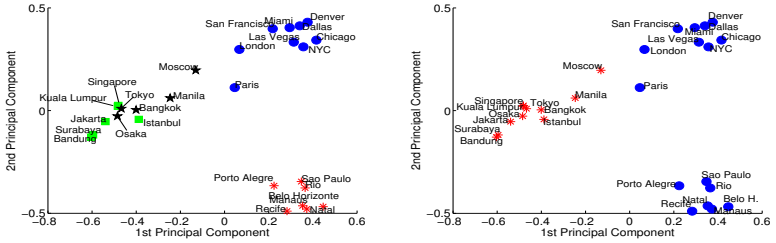
(c) $k = 10$

Figure 13: Clustering results for countries considering dataset D3.

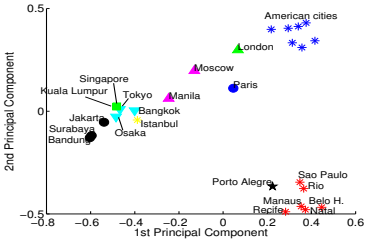
obtain their principal components[14]. PCA simplify the description of the original preferences in terms of the differences and similarities between them [63, 64]. Besides, the principal components enable visualizing the clustering results in the dimensions that explain most of the variation in the data (first and second principal components). Finally, we use the k-means algorithm, a widely used clustering technique, to group areas in the space defined by these principal components. We perform this analysis for areas defined at the country, city and neighborhood levels.

The score values for the first two principal components generated by the PCA for countries, cities, and regions are used in Figures 13, 14, and 15, respectively. Each color/symbol in those figures indicates a cluster obtained by k-means, which used the p first principal components that explain 100% of the variation in the data (p=15 for countries, p=26 for cities, and p=22 for regions).

[14] Alternative methods could be applied to reduce the dimensionality of these vectors. A comparison of these methods is out of the scope of the present work.

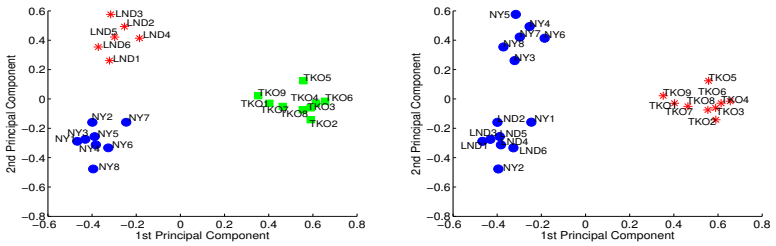


(a) $k = 4$ (b) $k = 2$

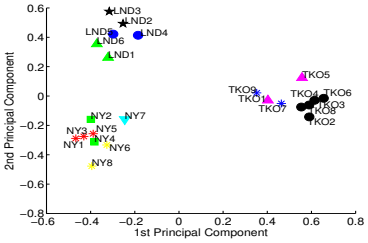


(c) $k = 10$

Figure 14: Clustering results for cities considering dataset D3.



(a) $k = 3$ (b) $k = 2$



(c) $k = 10$

Figure 15: Clustering results for regions considering dataset D3.

The k value in the k-means varied according to the characteristics of the considered areas. For countries, we set k=7 (same number of clusters used in [55]), result shown in Figure 13a. Following the same logic, we set k=4 for cities (Figure 14a), since we considered cities from 4 different continents/countries, and k=3 for regions inside a city (Figure 15a), because we considered 3 cities. Besides those values of k we computed the clusters for $k = 2$ and $k = 10$ for all considered areas, in order to evaluate the clustering result (figures labeled with "b" and "c" of Figures 13, 14, and 15). The parameters $k = 2$ and $k = 10$ are used to study relaxed and tight clusters, respectively. The evaluation of other strategies for cluster formation is possible, however, this is out of scope of the present study. We used the cosine similarity to compute the similarity between locations.

To help in the analysis of these results, we propose a cluster similarity score $c_{i,j}$ that represents how similar one set of clusters (i) is from another set of clusters (j). This score can be used, for instance, to evaluate how good is the matching between the clusters obtained using our new dataset with the old one. The result of c is a value up to 1. The closer to 1 the more similar are the compared clusters. Appendix B presents more details about how the score is calculated.

Table 5 shows the cluster similarity score between clusters found with D3 and D1 ($c_{D3,D1}$). The score is generated for countries, cities, and regions and for all k values considered. Analyzing first the clustering results for country level, the clusters for $k = 7$ and $k = 2$ are the same in all datasets ($c_{D3,D1} = 1$). For $k = 10$ we note a few differences, for example, the dissolution of the cluster formed by Turkey and Russia, to form a new cluster containing Indonesia and Turkey, and another one containing only Russia (this happened using D3). In general, these new clusters agree slightly more with our ground-truth than the result using D1, as we discuss in the next section (Section 5.5).

Besides that, it is possible to observe in Figure 13 that countries with close geographic proximity are not necessarily associated with the same cluster. For example, Australia and Indonesia are *not* likely to be in the same cluster. Although they are geographically neighboring countries, they are culturally very distinct. It is worth mentioning that if we had considered two disjoint periods of the day, instead of four, the

Table 5: Cluster similarity score between clusters found with D3 and D1. The score is generated for countries, cities, and regions and for all k values considered.

Countries	
Number of clusters (k)	$c_{D3,D1}$
$k = 7$	1
$k = 2$	1
$k = 10$	0.78
Cities	
Number of clusters (k)	$c_{D3,D1}$
$k = 4$	0.9
$k = 2$	1
$k = 10$	0.77
Regions	
Number of clusters (k)	$c_{D3,D1}$
$k = 3$	1
$k = 2$	1
$k = 10$	0.59

similarity score for D3 and $k = 7$ is $c = 0.25$. Whereas considering eight disjoint periods, in the same scenario, $c = 1.0$.

We now turn our attention to the clustering results for large cities. We observe using dataset D1 that cities are well clustered by the geographic regions where they are located: Asia, Brazil, Europe, and the USA. The European cluster, in this case, is formed by Paris, London, Moscow, and Istanbul. Figure 14a shows the clusters considering $k = 4$ for dataset D3. Considering this value of k the cluster similarity score with D1 is: $c_{D3,D1} = 0.9$. This outcome is due to very small changes in Asia and Europe clusters found using D1. Asia cluster was divided in two: one composed of Jakarta, Kuala Lumpur, Singapore, Surabaya and Bandung (cluster named Asia-1), and the other one composed by Tokyo, Osaka, Manila, and Bangkok (cluster named Asia-2). Europe cluster was also divided. London and Paris were associated with American cluster, while Istanbul was grouped with Asia-1 and Moscow with Asia-2. This result might represent more precisely cultural similarities between those cities. Istanbul was grouped with other cities, as itself, with a strong influence of Islam, which might impact in the cultural habits of the inhabitants. Eating and drinking practices of inhabitants of

London and Paris might be more similar to the Americans compared to other studied cities, a fact that helps to explain their grouping. Moscow is a particular city in Europe, and among other cities, it might be more similar to the group Asia-2. This suggests that a larger time window might offer better precision in cultural boundaries identification in some cases.

Considering $k = 2$, as we can see in Figure 14b, the results are exactly the same for all datasets ($c_{D3,D1} = 1$). For $k = 10$, the cluster similarity score is: $c_{D3,D1} = 0.77$. This result is also due small differences in the results. In the cluster found using D1 not all Brazilian and American cities were clustered together. This fact does not happen when using dataset D3 (see Figure 14c), where all cities from the United States were grouped together in one separated cluster. Besides that, the cluster found with D1 composed by Manila, Singapore, and Kuala Lumpur and the cluster formed by Istanbul and Moscow also suffered changes. According to the results with dataset D3, Istanbul is now a cluster by itself, and Manila and Moscow form another cluster. One thing that might help to explain this result is that the Christianity is the largest religion in both cities, which, as discussed before, might strongly influence cultural habits. The other clusters were unchanged.

Turning our attention to regions inside London, NY, and Tokyo, we observe in Figure 15a, which show results for $k = 3$, that all regions in the same city are grouped together, and this result is the same for all datasets. The clusters for $k = 2$ (Figure 15b), are also the same considering all studied datasets. Analyzing the results for $k = 10$, we can see a few changes in the clusters containing regions of New York and London, regions of Tokyo clusters remained the same for all datasets. One interesting results using D3 is that NY7 (where most part of Chinatown is located) represents a group by itself, instead of being grouped with NY8, as observed using D1. Chinatown in New York is very distinct from the other areas studied in that city.

In this analysis, we only see considerable changes when grouping a large number of clusters, i.e., $k = 10$. For these cases, we have an indication that a dataset with a larger observation window might offer a better cultural boundaries identification.

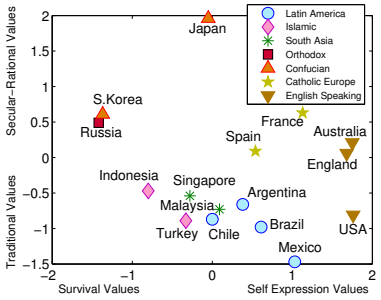


Figure 16: The cultural map of the World given by the World Values Survey [55].

5.5. Comparing with Survey Data

Similarly to us, Ronald Inglehart and Christian Welzel proposed a cultural map of the world based on the World Values Survey (WVS) data [55]. This map is shown in Figure 16 and contains only the countries we analyze in this study. The x-axis represents the dimension survival values versus self-expression values, and the y-axis represents the dimension traditional values versus secular-rational values, both explained in Section 2. Moreover, it offers a division of the world into clusters, similarly to what we have done in the previous section.

Comparing Figure 16 with Figure 13a, (figure with the same number of clusters of Figure 16), observe that the similarities are striking, with only two major differences. First, the "Islamic" cluster dissolved, with Turkey joining Russia and Indonesia joining Malaysia and Singapore. Second, USA and Mexico left the "English Speaking" and the "Latin America" clusters, respectively, and paired up to form a new one. Note, nevertheless, that these differences might not be surprising as these new boundaries.

We further investigate the differences between boundaries given by the WVS study and by our approach. In order to do so we first rank for a given country, all the other countries according to their cosine similarity towards it. We compute the similarity using the dimensions produced by the WVS data [55] and the dimensions computed by our approach. Then, we compute the Spearman's rank correlation coefficient ρ between these two ranks to see, for instance, if the most similar (and distinct) countries to England using the WVS data are ranked similarly when we use our approach.

Table 6 shows these results. We highlight in bold all the coefficients that are sta-

Table 6: The Spearman's rank correlation coefficient ρ (and its respective p-value) between the rank of similar countries generated from WVS and by our approach.

Country	Dataset D3	
-	ρ	p-value
Argentina	**0.56**	**0.027**
Australia	0.26	0.33
Brazil	**0.52**	**0.04**
Chile	0.09	0.74
England	**0.77**	**0.0009**
France	**0.73**	**0.002**
Indonesia	**0.61**	**0.015**
Japan	0.34	0.2
Korea	**0.69**	**0.004**
Malaysia	-0.19	0.48
Mexico	0.46	0.47
Russia	0.83	0.07
Singapore	**0.24**	**0.0001**
Spain	**0.8**	**0.0003**
Turkey	0.08	0.75
USA	**0.64**	**0.009**

tistically significant, i.e., with a $p\text{-}value < 0.05$. Observe that the correlation ρ is significant and positive for several countries. Considering also the strong similarity of clusters found using our approach compared with clusters using WVS, these results suggest that our approach, which is based solely on a small amount of participatory data, has a potential to reproduce/complement cultural studies performed using surveys, such as the one relying on the WVS, which is based on 4 years of survey data.

We would also like to point out possible reasons for the differences between our cultural map and the WVS map, as well as for the negative correlation seen in Table 6. First, the traits of each dataset are significantly different. While the WVS looked at several cultural dimensions, from religion to politics, from economics to lifestyle, we looked only at food and drink preferences. Second, the WVS data has a distance of 4 to 9 years to our data. During this time, significant cultural changes may have happened, given that the world is getting more connected every day.

Third, the most significant differences are related to multi-ethnic, multicultural, and multilingual countries, such as Malaysia. In these countries, it is probably hard to find culturally homogeneous samples of individuals, which might be the cause of the discrepancies seen between our results and those described in [55]. Fourth, this is, perhaps, also related to the fact that social media data might be biased towards a certain type of individuals, as discussed in Section 7, and in certain countries, this bias might be more evident. More data might help reduce this possible bias.

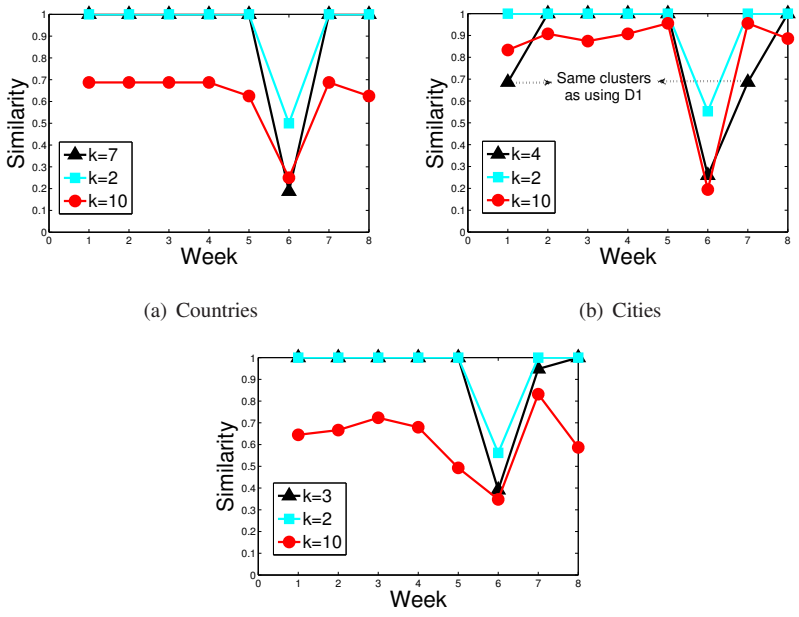
5.6. Impact of observation window size

We observe that the results obtained for D1 and D3 are very similar. With that, a natural question that emerges is: what is the impact of observation window size in the results?

Remember that D1 has one full week, D3 has three full weeks, which is a subset of D2, and D2 has eight weeks, but some of them probably do not represent all data that could be collected. In order to answer the posed question, we now investigate the impact in the results considering each week of D2 individually. This particular window size was chosen to agree with the size of D1. Figure 17 shows the cluster similarity score of clusters obtained using each individual week of dataset D2 (1 to 8) with clusters using D3, for countries (Figure 17a), cities (Figure 17b), and regions (Figure 17c). The results refer to all values of k considered in this work.

We found $c = 1$, in all figures, for most clusters identified for all values of k, except $k = 10$. Another thing in common in all results is the very low value of c for week 6. This is expected because this particular week has almost no data (6 days without data, representing a very short observation window).

Considering $k = 4$ we have two cases for cities that $c \neq 1$ (besides using week 6): using week 1 and 7. For both cases $c = 0.7$ and the clusters are equal ($c = 1$) to those using $D1$. This suggests that cultural differences observed using D1 are representative. Due to this short time window of observation, variations in the behavior of people under any atypical situation, e.g., bad weather condition, are more susceptible to be captured. This might be the case for all these mentioned weeks. For regions considering $k = 3$ the similarity score is $c = 1$ for all weeks, except week 6 (expected) and week 7



(a) Countries

(b) Cities

(c) Regions

Figure 17: Cluster similarity score of the clusters obtained for each individual week of D2 (1 to 8) with the clusters for D3.

($c = .95$).

Analyzing particularly the results for $k = 10$ for countries and cities, we observe that most clusters for all weeks are similar to each other, explaining the similar values of c, around 0.7 (for countries) and 0.9 (for cities). This value is considerably high, indicating that all clusters are similar with clusters found using D3.

Turning our attention to clusters found for regions considering $k = 10$, similarity scores lower than 1 were also observed, however, with a slightly bigger variation than for countries and cities. This result might be explained by the fact that regions, due its much smaller size, tend to be more susceptible to variation in the behavior of people that came to visit them, fact that might be attenuated using a dataset spanning larger time window. Another possible explanation might be due to the fact that $k = 10$, for our analysis, might represent too many clusters which could lead to arbitrary groupings in some cases.

Based on the analysis performed, we do not observe big changes using an observation window larger than one week. Recall that in this analysis we are comparing results from isolated weeks with those composed of three weeks. However, according to some insights discussed, we have an indication that exploring a dataset with a bigger observation window could enable more accurate results. It is also important to note that using a dataset spanning considerably less than one week, i.e. a small observation window to capture the routine of users, such as week 6, the results tend to degrade considerably. One week of data seems to be enough to capture at least the strongest cultural differences and, perhaps, exploring a dataset bigger (e.g., spanning a year or several years) the results may better capture cultural nuances and minimize the impact of atypical situations and any other type of bias in the data.

6. Variants of Cultural Boundary Inference Methodology

In this section, our goal is to assess whether the clustering methodology we are following, proposed here, is satisfactory. For that, we analyze two variations in the original approach.

6.1. Description of the Analysis

We disregard the time dimension in our evaluation, to propose two additional analysis (AA) for cultural boundaries identification.

- AA1: in this analysis, the vector of preferences considers only the types of venues (i.e., subcategories of places) presented in each city. For example, a city could be described by the subcategories $[Bar, PizzaPlace, AmericanRestaurant]$ and another one by $[SushiPlace, SakeBar]$. AA1 do not consider the popularity of subcategories, i.e. number of check-ins performed by the users;
- AA2: in this analysis, the vector of preferences considers the types of venues, as well as their popularity, i.e. we consider the normalized number of performed check-ins at each of the 101 subcategories.

With AA1 we try to answer the question: does the existence of certain types of venues in an area a are enough to explain cultural differences? AA2 help us to complement the former question, answering: is the popularity of those venues useful/essential in this task?

The rest of the methodology remains the same as we presented in Section 5.3. In sum, we now represent each area a by a preference vector as described in AA1 and AA2, disregarding the time dimension. We then apply the PCA technique to these vectors to obtain their principal components. Finally, we use the k-means algorithm to group areas in the space defined by these principal components. We perform this analysis for areas defined at the country, city and neighborhood levels. For this analysis, we consider only D3.

6.2. Evaluating AA1

First, we study the results obtained for AA1. The clusters found for countries considering $k = 7$, $k = 2$, and $k = 10$ do not agree with our ground-truth nor with common sense. There is always a cluster with the maximum number possible according to k. In other words, since we have 16 countries, when we set $k = 7$ we have a cluster of 10 countries and other 6 clusters containing one each. Those clusters are practically randomly selected just to satisfy the chosen k, resulting in clusters very different from WVS ones. In fact, the cluster similarity score between the clusters found considering $k = 7$ and the clusters found by WVS is: $c_{aa1,wvs} = 0.18$.

Since we tend to have many data representing one country, the unsatisfactory result for this approach is expected because it is very likely that we find all types of places (in our preference vector) for all countries. For this reason, the distances from each vector of preferences tend to be zero, making the quality of clustering very low.

We also calculated the cluster similarity score between all results obtained considering AA1 with the original methodology using dataset D3, generating then $c_{aa1,D3}$, as shown in Table 7. The score is obtained for all types of areas and k values considered. As we can see, the results for countries, to all values of k, obtained with AA1 are also very distinct from those obtained with the original methodology.

Figures 18a and 18b show the clustering results for cities ($k = 4$) and regions

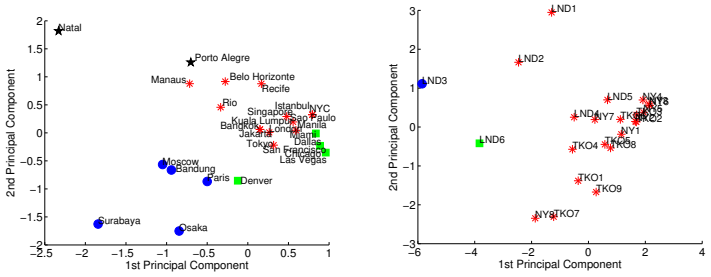
Table 7: Cluster similarity score between the clusters found using the original methodology (considering dataset D3) and those using AA1. The score is generated for all types of areas and k values considered.

Countries	
Number of clusters (k)	$c_{aa1,D3}$
$k = 7$	0.18
$k = 2$	0.5
$k = 10$	0.31
Cities	
Number of clusters (k)	$c_{aa1,D3}$
$k = 4$	0.39
$k = 2$	0.57
$k = 10$	0.32
Regions	
Number of clusters (k)	$c_{aa1,D3}$
$k = 3$	0.39
$k = 2$	0.56
$k = 10$	0.34

($k = 3$), respectively. As we can see, the results are much inferior compared to those obtained with the original approach. For instance, in Figure 18a we have a cluster composed by Moscow, Surabaya, Paris, and Osaka, cities very culturally distinct. If we study the results for cities in Table 7, we can see that the results obtained with AA1 are very distinct for those obtained by the original approach, which, as discussed previously, agrees more with WVS.

Studying the clusters for regions, the quality of the results are also not satisfactory. For example, looking at Figure 18b, we see that NY8 and NY5 represent two individual clusters and the rest of the areas form another one. This type of unsatisfactory results happens also for other values of k.

With that, we have enough evidence that AA1 do not capture appropriately cultural differences. Besides, this approach is likely to reduce quality when increasing the number of data for the considered areas, because it tends to make all the distances between vectors of preferences of areas very close to zero, as observed for countries.



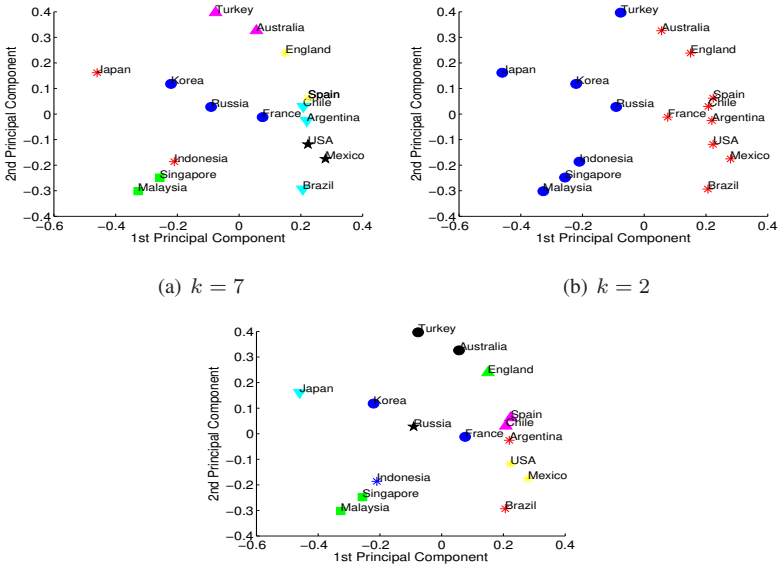
(a) $k = 4$, cities (b) $k = 3$, regions

Figure 18: Clustering results for cities and regions obtained for AA1, considering dataset D3.

6.3. Evaluating AA2

We now turn our attention to the results obtained for AA2. Figures 19, 20, and 21 show the results for countries, cities, and regions, respectively. Some results are omitted because they are identical to those found using the original methodology. Studying results for countries, we observe they agree more with those found by Ronald Inglehart and Christian Welzel using WVS data than those obtained using AA1. However, they are less precise than results obtained using the original methodology. For example, inspecting 19a, we find results not very similar with those using WVS data. The cluster Turkey and Australia, for instance, is not identified using WVS data. The cluster similarity score between the clusters found considering $k = 7$ and the clusters found by WVS is: $c_{aa2,wvs} = 0.35$.

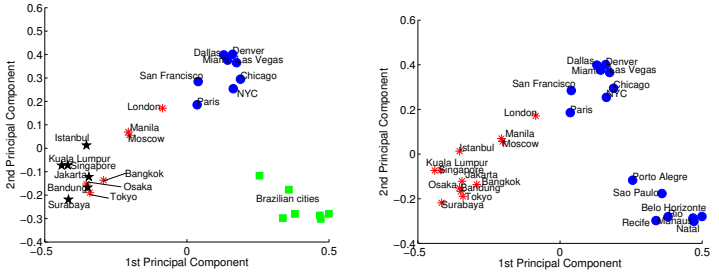
To further study this case, Table 8 shows values of c between the clusters found using the original methodology and those using AA2. As we can see, the scores for countries do not agree considerably with the original methodology for $k = 7$ and $k = 10$. This is not the case for cities and regions, cases where results with AA2 are much more similar with those obtained using the original methodology. Despite that, a very high similarity is not always obtained. Besides, using the original methodology we are more likely to get results that are expected according to common sense. For instance, using AA2 for $k = 4$ (Figure 20a), London was grouped with Bangkok, Tokyo, Manila, Moscow, and Osaka, fact not observed using the original methodology. In addition, the original methodology tends to cluster regions inside the same city more



(a) $k = 7$ (b) $k = 2$

(c) $k = 10$

Figure 19: Clustering results for countries obtained for AA2, considering dataset D3.



(a) $k = 4$ (b) $k = 2$

Figure 20: Clustering results for cities obtained for AA2, considering dataset D3.

than with AA2 (see results for $k = 10$ in Figures 15c and 21). Not necessarily neighbor regions are more similar, for example, the case of Chinatown shown in Figures 8c and 8d, however, in most of the cases this might happen (especially in the way we divided regions: using a grid). This is another indication the original methodology separates better culturally distinct areas.

Table 8: Cluster similarity score between the clusters found using the original methodology (considering dataset D3) and those using AA2. The score is generated for all types of areas and k values considered.

Countries	
Number of clusters (k)	$c_{na2,D3}$
$k = 7$	0.4
$k = 2$	0.92
$k = 10$	0.59
Cities	
Number of clusters (k)	$c_{na2,D3}$
$k = 4$	0.95
$k = 2$	0.96
$k = 10$	1
Regions	
Number of clusters (k)	$c_{na2,D3}$
$k = 3$	1
$k = 2$	1
$k = 10$	0.88

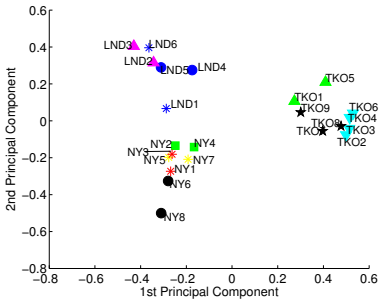


Figure 21: Clustering results for regions obtained for AA2 for $k = 10$, considering dataset D3.

6.4. Further Possibilities

It is important to emphasize that the comparison performed here with the original methodology and the approaches AA1 and AA2 was regarding cultural boundaries identification. However, we have to keep in mind that dimension reduction, in the same direction used in the approaches AA1 and AA2, could be useful to obtain other types of information about the considered areas.

In order to let more clear the usefulness of dimension reduction, when we analyze a

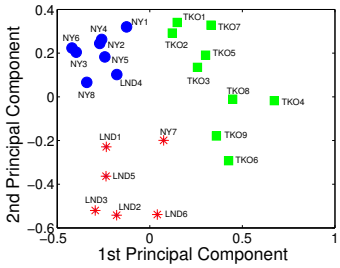


Figure 22: Clustering results for regions obtained with the original approach for dataset D1 ($k = 3$), considering only the drink class.

subset of features, for example, drinking habits during weekends in all regions of London, NY, and Tokyo, result shown in Figure 22, we find that some regions of London and NY are clustered together. This is corroborated by the results shown in Section 5: for certain categories, there are regions from different cities that are very similar and, thus, end up clustered together.

One might also create groups of subcategories according to a certain semantic of interest. For instance, aggregating all subcategories related to Bar, e.g. Sake Bar, Pub, Karaoke Bar, among others. Despite the probable cost of losing implicit cultural aspects in this step, this aggregation could still be useful to find similar areas.

All these possibilities are examples that could be useful in an application to suggest areas for users, for instance, to perform drinking activities.

7. Final Discussion and Limitations

In our analysis, we observe that using an observation window, perhaps, not enough to capture the routine of users, such as week 6 of dataset D2, which do not capture routines performed during weekdays and weekends, the results of cultural separation were not satisfactory. According to our analysis, using at least one week of data, from Monday to Sunday, the results do not change significantly for different scenarios and datasets (covering different years). This might mean that one week of data is enough to capture strong cultural differences between the analyzed places. However, this does not mean that more data does not provide a more precise result. Probably, exploring

a bigger dataset, e.g., spanning several months, may help to attenuate changes in the behavior due to atypical situations, such as weather conditions, which may induce users to leave their routines.

Particularly related to this aspect, our study might have some limitations. The results observed might be representative only of the period analyzed because we study different datasets around the same time of the year, May-June, despite some years of difference in the collection. Further investigation considering a dataset covering a different period of the year is important to be performed in a future study.

Our study might also face other possible limitation, for instance, regarding data collection. It may reflect the behavior of a fraction of consumers. Our collection was based on data shared by users of Foursquare on Twitter. Therefore, there could be biases relating to the fact that the users of such application are not necessarily representative of all population of a particular region. They are likely to be young, owners of smartphones, and urban dwellers. Consequently, urban areas with older and poorer populations might provide fewer data and be underrepresented in whatever analysis is made. Besides, users may not share data concerning all of their destinations, considering the info will be made public on Twitter. Thus, our dataset might offer a partial view of users preferences and habits, which needs to be taken with care.

In addition, our methodology assumes that the data shared by users are correct. Twitter is powerful a tool that opens opportunities for new forms of spam [65, 66]. Data quality, one of the challenges discussed in [57], under this circumstances becomes even more serious, because the production of false data might be possible. So far we are not aware of any significant production of false data in the systems we analyze, however this could potentially compromise the methodology. Exploring a dataset spanning a bigger observation window might help to minimize these problems in case they happen.

In spite of all that, our study provides solid aggregate information that, as shown, can capture important cultural habits of users.

8. Conclusions

Considering datasets from Foursquare differing in terms of volume of data and observation window size, we identify particular individual preferences, such as the taste for a certain type of food or drink, as well as temporal habits, such as the time and day of the week when an individual goes to a restaurant or a bar. This enabled the proposition of a new methodology to identify cultural boundaries and similarities across populations at different scales using LBSN data. We extensively evaluated our proposed methodology from different aspects, for instance, disregarding some of the considered dimensions, as well as analyzing the results using datasets from different periods and observation window. The results indicate that our proposed methodology is a promising approach to capture cultural boundaries and similarities across populations in a faster and cheaper way than traditional methods. From this, our study could enable several new urban services and applications.

Acknowlegments

This work was supported by the INCT-Web (MCT/CNPq grant 57.3871/2008-6), and by the authors individual grants and scholarships from CNPq, CAPES, FAPEMIG, Fundação Araucária and EPSRC Grant "The Uncertainty of Identity" (EP/J005266/1).

References

[1] L. Valori, F. Picciolo, A. Allansdottir, D. Garlaschelli, Reconciling long-term cultural diversity and short-term collective social behavior, Proceedings of Nat. Acad. of Sci. 109 (4) (2012) 1068–1073.

[2] H. Schmitt, E. Scholz, I. Leim, M. Moschner, The Mannheim Eurobarometer Trendfile 1970-2002. Data Set Edition 2.00: Appendix, Zentralarchiv fur Empirische Sozial., 2005.

[3] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, M. Musolesi, A. A. F. Loureiro, You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare, in: Proceedings of ICWSM'14, Ann Arbor, MI, USA, 2014.

[4] J. Cranshaw, R. Schwartz, J. I. Hong, N. Sadeh, The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City, in: Proceedings of ICWSM'12, Dublin, Ireland, 2012.

[5] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks, in: Proceedings of ICWSM'11, AAAI, Barcelona, Spain, 2011.

[6] R. Garcia-Gavilanes, D. Quercia, A. Jaimes, Cultural dimensions in twitter: Time, individualism and power, in: Proceedings of ICWSM'13, Boston, USA, 2013.

[7] C. Carole, Food And Culture: A Reader, 2nd Edition, Routledge, 1997.

[8] R. Cochrane, S. Bal, The drinking habits of sikh, hindu, muslim and white men in the west midlands: a community survey, British Journal of Addiction 85 (6) (1990) 759–769. doi:10.1111/j.1360-0443.1990.tb01688.x.

[9] C. Chaey, Foursquare explore is now a search tool anyone can use, no check-ins required, http://goo.gl/MQ22DU (October 2012).

[10] P. Shankar, Y.-W. Huang, P. Castro, B. Nath, L. Iftode, Crowds replace experts: Building better location-based services using mobile social network interactions, in: Int. Conf. on Perv. Comp. and Comm. (Percom'12), Lugano, Switzerland, 2012, pp. 20–29.

[11] S. Scellato, A. Noulas, R. Lambiotte, C. Mascolo, Socio-spatial Properties of Online Location-based Social Networks, in: Proceedings of ICWSM'11, Barcelona, Spain, 2011.

[12] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of KDD'11, ACM, San Diego, California, USA, 2011, pp. 1082–1090.

[13] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An Empirical Study of Geographic User Activity Patterns in Foursquare, in: Proceedings of ICWSM'11, Barcelona, Spain, 2011.

[14] D. Kershaw, M. Rowe, P. Stacey, Towards tracking and analysing regional alcohol consumption patterns in the uk through the use of social media, in: Proceedings of WebSci'14, ACM, Bloomington, USA, 2014, pp. 220–228.

[15] A. Venerandi, G. Quattrone, L. Capra, D. Quercia, D. Saez-Trumper, Measuring urban deprivation from user generated content, in: Proceedings of CSCW'15, ACM, Vancouver, BC, Canada, 2015, pp. 254–264.

[16] V. Zambaldi, J. P. Pesce, D. Quercia, V. Almeida, Lightweight contextual ranking of city pictures: Urban sociology to the rescue, in: Proceedings of ICWSM'14, Ann Arbor, USA, 2014.

[17] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, A. A. F. Loureiro, Revealing the city that we cannot see, ACM Trans. Internet Technol. 14 (4) (2014) 26:1–26:23.

[18] P. Georgiev, A. Noulas, C. Mascolo, The call of the crowd: Event participation in location-based social services, in: Proceedings of ICWSM'14, Ann Arbor, USA, 2014.

[19] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of WWW'10, ACM, Raleigh, North Carolina, USA, 2010, pp. 851–860.

[20] N. Alsaedi, P. Burnap, O. F. Rana, A combined classification-clustering framework for identifying disruptive events, in: Proceedings of SocialCom'14, Stanford, USA, 2014.

[21] H. Becker, M. Naaman, L. Gravano, Beyond trending topics: Real-world event identification on twitter, in: Proceedings of ICWSM'11, Barcelona, Spain, 2011.

[22] B. Pan, Y. Zheng, D. Wilkie, C. Shahabi, Crowd sensing of traffic anomalies based on human mobility and social media, in: Proceedings of SIGSPATIAL'13, SIGSPATIAL'13, ACM, Orlando, Florida, 2013, pp. 344–353.

[23] J. Gomide, A. Veloso, W. M. Jr., V. Almeida, F. Benevenuto, F. Ferraz, M. Teixeira, Dengue surveillance based on a computational model of spatio-temporal locality of twitter, in: Proceedings of WebSci'11, Evanston, USA, 2011.

[24] C. Robles, J. Benner, A tale of three cities: Looking at the trending feature on foursquare, in: International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing, 2012, pp. 566–571. doi:10.1109/SocialCom-PASSAT.2012.123.

[25] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, S. Ghosh, Extracting situational information from microblogs during disaster events: A classification-summarization approach, in: Proceedings of CIKM'15, ACM, Melbourne, Australia, 2015, pp. 583–592.

[26] A. Chakraborty, J. Messias, F. Benevenuto, S. Ghosh, N. Ganguly, K. P. Gummadi, Who makes trends? understanding demographic biases in crowdsourced recommendations, in: Proceedings of ICWSM'17, Montreal, Canada, 2017.

[27] M. Ciot, M. Sonderegger, D. Ruths, Gender inference of Twitter users in non-English contexts, in: Proceedings of EMNLP'13, Seattle, USA, 2013, pp. 1136–1145.

[28] J. D. Burger, J. Henderson, G. Kim, G. Zarrella, Discriminating gender on twitter, in: Proceedings of EMNLP'11, Edinburgh, United Kingdom, 2011, pp. 1301–1309.

[29] W. Liu, D. Ruths, What's in a name? using first names as features for gender inference in twitter., in: Proceedings of Symp. on Analyzing Microtext, Stanford, USA, 2013.

[30] V. G. David Garcia, Ingmar Weber, Gender asymmetries in reality and fiction: The bechdel test of social media, in: Proceedings of ICWSM'14, Ann Arbor, USA, 2014.

[31] S. Nilizadeh, A. Groggel, P. Lista, S. Das, Y.-Y. Ahn, A. Kapadia, F. Rojas, Twitter's glass ceiling: The effect of perceived gender on online visibility., in: Proceedings of ICWSM'16, Cologne, Germany, 2016, pp. 289–298.

[32] C. Wagner, P. Singer, M. Strohmaier, Spatial and temporal patterns of online food preferences, in: Proceedings of WWW'14, Seoul, Korea, 2014, pp. 553–554.

[33] D. Fried, M. Surdeanu, S. Kobourov, M. Hingle, D. Bell, Analyzing the language of food on social media, in: Big Data (Big Data), 2014 IEEE International Conference on, IEEE, 2014, pp. 778–783.

[34] S. Abbar, Y. Mejova, I. Weber, You tweet what you eat: Studying food consumption through twitter, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 3197–3206.

[35] Y. Mejova, S. Abbar, H. Haddadi, Fetishizing food in digital age: #foodporn around the world, CoRR abs/1603.00229.

[36] Y. Mejova, H. Haddadi, A. Noulas, I. Weber, #foodporn: Obesity patterns in culinary interactions, in: Proceedings of Conference on Digital Health 2015, DH '15, ACM, New York, NY, USA, 2015, pp. 51–58.

[37] M. De Choudhury, S. Sharma, E. Kiciman, Characterizing dietary choices, nutrition, and language in food deserts via social media, in: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16, ACM, New York, NY, USA, 2016, pp. 1157–1170.

[38] S. S. Sharma, M. De Choudhury, Measuring and characterizing nutritional information of food and ingestion content in instagram, in: Proceedings of WWW '15 Companion, ACM, New York, NY, USA, 2015, pp. 115–116.

[39] R. West, R. W. White, E. Horvitz, From cookies to cooks: Insights on dietary patterns via analysis of web usage logs, in: Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, ACM, New York, NY, USA, 2013, pp. 1399–1410.

[40] N. Hochman, R. Schwartz, Visualizing instagram: Tracing cultural visual rhythms, in: Proceedings of Workshop on Social Media Vis., AAAI, Dublin, Ireland, 2012, pp. 6–9.

[41] B. Poblete, R. Garcia, M. Mendoza, A. Jaimes, Do all birds tweet the same?: characterizing twitter around the world, in: Proceedings of ICIKM'11, ACM, Glasgow, UK, 2011, pp. 1025–1030.

[42] R. García-Gavilanes, Y. Mejova, D. Quercia, Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication, in: Proceedings of CSCW'14, ACM, Baltimore, Maryland, USA, 2014, pp. 1511–1522.

[43] D. Mocanu, A. Baronchelli, N. Perra, B. Gonalves, Q. Zhang, A. Vespignani, The twitter of babel: Mapping world languages through microblogging platforms, PLoS ONE 8 (4).

[44] K. Reinecke, M. K. Nguyen, A. Bernstein, M. Näf, K. Z. Gajos, Doodle around the world: Online scheduling behavior reflects cultural differences in time perception and group decision-making, in: Proceedings of CSCW'13, ACM, San Antonio, Texas, USA, 2013, pp. 45–54.

[45] P. Laufer, C. Wagner, F. Flöck, M. Strohmaier, Mining cross-cultural relations from wikipedia - A study of 31 european food cultures, CoRR abs/1411.4484.

[46] B. State, P. Park, I. Weber, M. Macy, The mesh of civilizations in the global network of digital communication, PLOS ONE 10 (5) (2015) 1–9.

[47] S. P. Huntington, The clash of civilizations?, in: Culture and Politics, Springer, 2000, pp. 99–118.

[48] J. Park, Y. M. Baek, M. Cha, Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis, Journal of Communication 64 (2) (2014) 333–354. doi:10.1111/jcom.12086.
URL http://dx.doi.org/10.1111/jcom.12086

[49] G. Murdock, Social Structure, Macmillan, 1949.

[50] H.-P. Blossfeld, E. Klijzing, M. Mills, K. Kurz, Globalization, uncertainty and youth in society, Routledge, London, 2005.

[51] F. Barth, Ethnic groups and boundaries: the social organization of culture difference, Scandinavian university books, Little, Brown, 1969.

[52] R. Axelrod, The dissemination of culture: A model with local convergence and global polarization, J Conflict Resolut.

[53] L. Festinger, Social pressures in informal groups: a study of human factors in housing, Stanford University Press, 1967.

[54] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks, Annual Review of Sociology 27 (1) (2001) 415–444.

[55] R. Inglehart, C. Welzel, Changing Mass Priorities: The Link between Modernization and Democracy, Perspectives on Politics 8 (02) (2010) 551–567.

[56] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, A. A. F. Loureiro, Challenges and opportunities on the large scale study of city dynamics using participatory sensing, in: Proceedings of ISCC'13, Split, Croatia, 2013, pp. 528–534.

[57] T. H. Silva, C. Celes, J. N. andVincius Mota, F. Cunha, A. Ferreira, A. Ribeiro, P. V. de Melo, J. Almeida, A. Loureiro, Users in the urban sensing process: Challenges and research opportunities, in: Pervasive Computing: Next Generation Platforms for Intelligent Data Collection, Academic Press, 2016, pp. 45–95.

[58] A. Pentland, To signal is human, American scientist 98 (3) (2010) 204–210.

[59] M. E. Newman, Assortative mixing in networks, Phy. rev. let. 89 (20) (2002) 208701.

[60] J. Watson, Golden arches east: McDonald's in East Asia, Stanford University Press, 2006.

[61] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, A. A. F. Loureiro, A picture of Instagram is worth more than a thousand words: Workload characterization and application, in: Proceedings of IEEE DCOSS'13, Cambridge, USA, 2013, pp. 123–132.

[62] C. E. Shannon, A mathematical theory of communication, Bell system tech. jour. 27.

[63] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (4) (2010) 433–459.

[64] I. T. Jolliffe, Principal component analysis and factor analysis, in: Principal component analysis, Springer, 1986, pp. 115–128.

[65] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on twitter, in: Proc. of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), Redmond, USA, 2010.

[66] S. Yardi, D. Romero, G. Schoenebeck, et al., Detecting spam in a twitter network, First Monday 15 (1).

Appendix A. Further Information on Entropy Results

This section presents the complementary results for the entropy evaluation performed on Section 5.3. Tables A.9, A.10, and A.11 show the entropy for all the subcategories of the Drink, Fast Food, and Slow Food classes, respectively.

Table A.9: Entropy results for subcategories of the Drink class. The minimum entropy is 0 and the maximum is 4.76 for cities and 4 for countries.

	Country (max = 4)	**Cities (max = 4.76)**
Bar	3.74	4.54
Beer Garden	3.73	4.24
Brewery	3.59	3.97
Cocktail Bar	3.60	3.92
Coffee Shop	3.87	4.45
Distillery	3.03	2.75
Dive Bar	3.20	3.67
Gay Bar	3.64	3.75
Hookah Bar	3.31	3.79
Hotel Bar	3.79	4.04
Juice Bar	3.46	4.18
Karaoke Bar	3.16	3.68
Piano Bar	3.13	3.45
Pub	3.50	4.19
Sake Bar	1.14	1.63
Speakeasy	3.56	2.81
Sports Bar	3.31	3.85
Tea Room	3.45	3.46
Whisky Bar	3.77	4.14
Wine Bar	3.68	4.31
Winery	3.30	4.01

Table A.10: Entropy results for subcategories of the Fast Food class. The minimum entropy is 0 and the maximum is 4.76 for cities and 4 for countries.

	Countries (max = 4)	Cities (max = 4.76)
Bagel Shop	2.96	3.92
Bakery	3.56	4.29
Breakfast	3.61	4.04
Burger	3.66	4.35
Burrito	2.28	2.91
Cafe	3.96	4.62
Cupcake	3.67	4.25
Delis	0.00	0.00
Dessert	3.72	4.54
Donut	3.20	3.86
Fast Food	3.77	4.59
Fish & Chips	2.77	2.95
Food Court	2.92	3.96
FoodTru	3.07	4.07
FriedChi	3.47	4.04
HotDog	2.29	3.63
IceCream	3.47	4.30
Mac& Cheese	3.13	3.08
Pizza	3.48	4.37
Ramen	2.29	3.28
Salad	3.18	3.65
Sandwich	3.20	4.17
Snack	3.67	4.02
Soup	2.81	3.14
Taco	1.18	3.07
Tapas	1.84	3.30
Wings	1.82	2.91

Table A.11: Entropy results for subcategories of the Slow Food class. The minimum entropy is 0 and the maximum is 4.76 for cities and 4 for countries.

	Countries (max = 4)	Cities (max = 4.76)
African	3.39	3.51
American	3.45	3.94
Arepa	3.06	3.09
Argentinian	1.39	2.95
Asian	3.26	3.92
Australian	1.07	2.93
BBQ Joint	3.65	4.37
Brazilian	1.05	3.04
Cajun or Creole	2.16	2.80
Caribbean	2.53	3.02
Chinese	3.36	4.21
Cuban	3.07	1.77
Dim Sum	2.62	3.55
Diner	3.55	4.31
Dumpling	2.21	3.61
Eastern European	2.03	2.56
Ethiopian	2.62	2.73
Falafel	3.15	3.44
Filipino	2.48	0.37
Food	3.45	4.06
French	2.69	3.57
Gastropub	2.66	3.77
German	3.04	4.03
Gluten-free	3.21	2.58
Greek	2.44	3.34
Indian	2.66	3.29
Indonesian	1.10	1.80
Italian	3.57	4.41
Japanese	3.32	4.23
Korean	1.96	3.80
Latin American	1.76	3.10
Malaysian	1.68	1.66
Mediterranean	2.92	3.69
Mexican	2.77	3.54
Middle Eastern	2.47	3.21
Molecular	2.96	3.27
Mongolian	1.92	1.86
Moroccan	2.65	2.61
New American	3.10	3.06
Paella	1.89	2.67
Peruvian	1.42	2.05
Portuguese	1.09	1.83
Scandinavian	2.86	3.03
Seafood	3.42	4.08
South American	1.50	2.86
Southern	2.71	2.73
Spanish	1.70	3.59
Steakhouse	3.71	4.56
Sushi	3.30	4.35
Swiss	2.39	3.27
Thai	2.31	3.23
Vegetarian or Vegan	3.51	4.19
Vietnamese	2.23	3.31

Appendix B. Similarity Score

Algorithm 1 shows the steps to calculate the similarity score c. This algorithm analyzes all pairs of clusters that we want to compare. For each pair, it calculates the number of similar elements (*hit*) between clusters, and the number of different elements (*miss*). The algorithm uses this values to calculate a discount factor. The discount factor is used to penalize bad clustering, i.e, clusters with a low value of "hit" and high value of "miss". The result of c is a value up to 1. The closer to 1 the more similar are the compared clusters. The following examples consider two hypothetical set of clusters, Clusters1 and Clusters2, to help us to understand the algorithm.

Algorithm 1: Steps to calculate the cluster similarity score c.

```
1  listMaxs = []
2  foreach c_1 in clusterSet1 do
3      max = 0
4      discount = 0
5      foreach c_2 in clusterSet2 do
6          hit = c_1 ∩ c_2
7          if hit == 0 then
8              continue
9          end
10         miss = lenght(c_2) − hit
11         if miss ≠ 0 then
12             discount = miss / hit
13         end
14         calc = hit − discount
15         if calc > max then
16             max = calc
17         end
18     end
19     listMaxs.append(max)
20 end
21 c = sum(listMaxs)/numTotalElementsClusters
```

1. Clusters1: $(x, y, z), (a, b, c, d), (e, f)$. Clusters2: $(x, y, d), (a, b, c, z), (e, f)$. Result: $c_{1,2} = 0.68$. Explanation: $(2 - 1/2) + (3 - 1/3) + (2 - 0)$ (sum of maximum intersection with its respective discount factor) divided by 9 (number of total elements);
2. Clusters1: $(x, y, z), (a, b, c, d), (e, f)$. Clusters2: $(x, y, d, a, b, c, z), (e), (f)$. Result: $c_{1,2} = 0.47$. Explanation: $[(4 - 3/4) + (1 - 1/1) + 1(1/1)]/9$;
3. Clusters1: $(x, y, z), (a, b, c, d), (e, f)$. Clusters2: $(x, y, z), (a, b, c, d), (e, f)$. Result: $c_{1,2} = 1$. This case shows a perfect match. Explanation: $[(3 - 0) + (4 - 0) + (3 - 0)]/9$.