Extraction and Exploration of Business Categories Signatures

Leonardo de Assis da Silva and Thiago H Silva

Federal University of Technology - Paraná
Department of Informatics
Curitiba, PR, Brazil
leosil@alunos.utfpr.edu.br, thiagoh@utfpr.edu.br

Abstract. Different business types may have distinct businesses functioning dynamics, i.e., popularity times, that can be dictated not only by the service offered but also due to other aspects. Performing the business popularity time comprehension allows us, for instance, to use this information as a business descriptor that could be explored in new services. Recently, Google launched a service, namely Popular Times, which provides the popularity times of commercial establishments. In this study, we collected and analyzed a large-scale dataset provided by that service for business in different cities in Brazil and in the United States. Our main contributions are: (1) clustering and analysis of the collected business popularity times dataset in each studied city; (2) approach for identifying the signature that represents the behavior of specific categories of venues; (3) training and evaluation of an inference model for categories of establishments; (4) user evaluation of some of our results.

Keywords: Google Popular Times · Time Series · Signature · Large Scale Urban Assessment.

1 Introduction

Urban computing is a field of study that, among others objectives, aims help to understand urban phenomenon envisioning to offer smarter urban services. There are several phenomena worth investigating in the city, for instance, citizens mobility, by any transportation mode, citizens interaction, for example, through phone calls, and businesses functioning dynamics, i.e., their popularity times [21,13].

Different types of business may have distinct popularity times that can be dictated not only by the service offered but also due to diverse economic, social and cultural aspects, such as typical working times and nap periods that might exist in certain cities. Besides, businesses in the same category may also eventually exhibit different peaks of popularity according to their location or particular characteristics of the business. While some types of restaurants might be more popular at night, fast-food venues can present a more evenly distributed popularity throughout the day.

Uncover the functioning dynamics, i.e., the popularity times, of business venues is not a trivial task. However, recently, Google launched a service, namely Popular Times, which provides the popularity times for those type of places. This is possible, among other factors, due to the high number of users who use Google's location-based mobile services, enabling Google to know when users visit a certain venue [4].

In this study, we collected and analyzed a large-scale dataset from Google Popular Times for business related to drinking and food consumption habits in different cities in Brazil and the United States. To the best of our knowledge, the present work is the first to explore this source of information. Also, we study patterns of popularity times considering different cities of the same country. Knowing the different patterns associated with specific business locations, e.g., countries, cities, or neighborhoods, could help to improve the description of the functioning of the business and in the understanding of how people from different geographical areas interpret the usefulness and the use of each business category.

That information could also be used in conjunction with business recommendation algorithms to make them sensitive to the local context, e.g., average less busy times. It could also be explored in a market study, for instance, to open new branches in other countries by studying cultural differences, such as the use of spaces. I.e., while a Brazilian can understand that the primary use of restaurants is for lunch, a user from the United States could assume that dinner is the preferred time to visit these types of place.

The main contributions of this work are: (1) grouping and analysis of popularity time series of business venues. We find, for instance, patterns that are related to local factors coming from cities of the same country; (2) approach to identify the signature that represents the popularity times for categories of business. We show that it was possible to find an association between different signatures of categories that can be used for the proposition of meta-categories of business places, which could be more informative about venues; (3) training and evaluation of a classification model to infer the category of a new establishment given its popularity; (4) user evaluation of some of our results, particularly, the meta-categories of business places proposed in this study.

The remainder of this study is divided as follows. Section 2 presents the related works. Section 3 shows the description of our collected dataset and how it is processed. Section 4 explains the clustering process and the procedure for generating category signatures. Section 5 discusses the results, including a possible application of category signatures in the category inference task given a popularity time series. Section 6 discusses an experiment with users to help to validate the proposed meta-categories. Finally, Section 7 presents the final considerations and future work.

2 Related works

Recently, data extracted from the Web has been explored to help in the better understanding of the urban social behavior and city dynamics. For instance,

check-ins of Foursquare were used to identify the functional use of city areas, e.g., a shopping area [17]. Areas of points of interest, such as sights and popular establishments, were identified from shared photos on Instagram [14]. In addition, check-ins were used to better understand the state of the traffic in urban regions in [16]. Thus, we highlight three groups of studies that are relevant to the present work: distribution of popular elements, clustering of time series and generation of a representative element of the cluster, and semantic extraction of geolocalized data.

Check-ins shared on Foursquare were used to investigate the properties of Location-Based Social Networks (LBSN); one of the discoveries is the presence of a power law distribution of the popularity of establishments, that is, a small number of establishments receive a high number of visits, while most establishments have low popularity. Consequently, the analysis of a limited sample of popular establishments is still capable of revealing the behavior of a significant proportion of visitors to a category of establishment. This finding is relevant here as the Google Popular Times service only offers information for establishments above a certain threshold of popularity [4]. Knowing that, Neves et al. [9] evaluated the possibility of reproducing Google Popular Times using data extracted from Foursquare to the cities of Curitiba and Chicago. They found evidence that Google Popular Times is consistent with data voluntarily actively shared through check-ins on Foursquare. That indicates that the reproduction of Popular Times for places that do not have this information might be possible using alternative data sources.

Time series clustering has been applied to find patterns across domains through distinct approaches. The task of identifying clusters of time series requires the proposition of new approaches or the use of conventional algorithms together to a suitable distance measure [8]. The variation patterns of *memes* mention in Twitter messages was observed using a new algorithm specially developed to handle temporal data, called K-Spectral Centroid (K-SC), which is an adaptation of the K-means algorithm [18]. Next, the K-SC algorithm was also explored to understand YouTube videos popularity [3]. The criterion used to define the number of clusters in both studies was the Silhueta method and the Hartigan index, both cluster validity indices that evaluates the similarity between elements of the same cluster compared to the similarity in relation to elements of the other clusters [1].

Regarding the application of conventional algorithms and the need for adequate distance measurements, the Dynamic Time Warping (DTW) measure has obtained satisfactory results in several standard datasets [10]. DTW is a dynamic programming technique similar to the editing distance, or distance of Levenshtein, that seeks to find the optimal global alignment between two time series. The DTW is particularly interesting in the context of discovering the shape of popularity signatures because it helps to find similar time series even when offsets occur, whereas this could eventually not happen when performing the comparison through a point to point distance measure. However, due to its time complexity, the DTW distance usually is only applied on short time se-

ries or small datasets [12]. To minimize this limitation Petitjean *et al.* proposed a heuristic for the calculation of averages called *DTW Barycenter Averaging* (DBA).

The extraction of knowledge from temporal data can be performed based on information from different domains. For example, the inference task of describing a venue, such as its category, can be accomplished by examining the distribution of the number of check-ins in establishments by hour and by day of the week. Such characteristics were used in conjunction with the spatial location and information of each visitor, such as age and gender, to assign a label to locations through a binary support vector machine trained with data from LBSN in [19] and boosted decision trees trained with data collected through surveys in [7].

This work uses the K-means partitioning algorithm, instead of adapting conventional algorithms as in [18], but using the same clustering validation measures of the studies mentioned above to help in the definition of the number of clusters. Different from the other category inference studies, we obtained the data used here from the Google Popular Times, where the sensing process using the user's device occurs through opportunistic sensing, that is, this source is not dependent on the initiative of users. As it is a background service, it has the potential to be less affected by factors such as a desire to omit visits to certain facilities than LBSNs. In addition, as far as we know, we present the first study to discover and explore business categories popularity signature in different cities.

3 Data collection and procedures

In the Google Popular Times service, the distribution of hourly visits, i.e., popularity time, for each day of the week represent the average number of visits to the establishment over several weeks. This information is generated from data sent anonymously by people who have agreed to participate in the Google History Location service by tracking automatic positioning of the device over time via GPS, WI-FI, and mobile network [4].

To collect data from Google Popular Times we explored a web crawler fed with a list of establishments containing the attributes: name, category, city, and country. We created this list with the help of the Yelp Developers API [20]. Explore this source is interesting because it provides information on establishments in a standard format independent of the country. Another strategy would explore open data, which has been increasingly provided by cities. However, the availability of open data is dependent on local policies, so that the scalability of the number of cities and countries could be impaired.

To investigate how people from different geographical areas interpret each category of establishment, we study cities in Brazil (Curitiba, Rio de Janeiro, and São Paulo), and in the United States (Chicago, New York, and San Francisco). One of our goals is to identify the patterns of popularity times exhibited by establishments related to the consumption of food and drink. Specifically, we analyze the following categories from the reference Yelp platform: bakery, bar, coffee, dance club (nightclubs), and restaurant. Look at these type of categories

is interesting because they can represent distinct cultural differences [15]. We show the total number of unique establishments collected in each city in Table 1.

Table 1. Number of unique establishments collected by city.

Cities	Curitiba	Rio de Janeiro	São Paulo	Chicago	New York	San Francisco
Yelp	1,755	2,046	2,964	3,652	4,280	3,340
Google	1,089	1,324	2,073	2,672	3,226	2,320

Analyzing Table 1 it is possible to notice that not all the establishments present in the Yelp list returned results of Google Popular Times during the search in Google. That is expected because Google does not offer this service for all venues. The HTML pages collected were processed to extract the popularity values for each day of the week. These values were then modeled as a discrete sequence of values v_k normalized between 0 and 1, where k represents each hour of the day so that a time series of popularity S can be defined as: $S = (v_k | \forall k \in [0..23], 0 \leq v_k \leq 1)$, where for each establishment two time series are generated through the DBA technique, one representing its typical pattern on weekdays and other the pattern of weekends.

As the objective is to determine the most representative pattern exhibited by most of the time series in a cluster, anomalous members were removed by calculating the distance of each series to the centroid of the cluster and cutting the farthest apart. The cutoff threshold applied followed the rule $Quartile_3 + Interquartile * 1.5$, a conventional approach [5], from the centroid distribution, which resulted in the removal around 6%, on average, of the series in a cluster.

4 Clustering and signatures generation

4.1 Time series clustering

Identifying the behaviors, i.e., patterns of popularity, typically exhibited by each category of business may help answer whether the category has a unique homogeneous functioning dynamic or to verify if businesses present different hours of popularity peaks even belonging to the same category. For this purpose, companies of the same category with similar time series were detected and separated into clusters by applying the K-means algorithm with the DTW distance. The K-means algorithm has identified time series clusters more uniformly distributed than clusters generated through hierarchical partitioning. This approach has also been successfully explored in clustering time series of different domains [3] [18].

Choosing the most appropriate number of clusters is a common challenge faced when performing unsupervised learning techniques on unclassified data [2]. The heuristic adopted in this study was the smallest number between the suggestion of the Silhouette method [11] and the Hartigan index [6]. This criterion is necessary due to the possibility that the two indices may not converge to

the same value, and it can be justified by the little variation of signatures shapes resulted by increasing the number of clusters. Such a problem was found in [18] and [3], where the authors also used those two metrics.

We divided the time series clustering experiments into three stages: clustering of the time series of businesses in the same category (explained above), clustering of the signatures of different categories, and clustering of the categories signatures for different cities in a country.

The stage of clustering the signatures of different categories tries to identify behaviors displayed by more than one category, i.e., if it is possible to represent distinct categories with a single signature. For that, it is used the same criterion for choosing the number of clusters in the clustering of the time series of businesses in the same category.

Finally, a process to identify the signatures of countries, by associating the most similar categories of signatures among the cities, was performed. We consider the behavior of a category as representative of the country's behavior if, and only if, such behavior is present in all cities of the sample. For this reason, we determine the number of clusters of country signatures in each category by the number of signatures of the city with the least amount of clusters in that category.

4.2 Signature generation

We can perform the generation of signatures to represent a set of time series in different ways depending on how the similarity between the time series is interpreted in the domain of interest. In the context of business popularity, we assume that time series are similar if they have similar shapes. Thus, a distance such as Euclidean would not be adequate for this scenario as the comparison occurs point by point; and two sets of time series with similar shapes of behavior separated by an offset of one hour might end up being separated when using this distance.

An alternative measure of distance that manages to overcome this problem is the Dynamic Time Warping (DTW). It tries to find the best global alignment between two time series such that, for example, a time series with a narrow peak that starts at 11h00 and ends at 13h00 is aligned with another time series containing a narrow peak that starts at 21h00 and ends at 23h00. To avoid the pitfall of the global alignment in the present scenario where time series with similar shapes in very distinct hours are aligned, a 2-hour window has been applied to limit the alignment between two series.

The signature that represents a cluster of time series is generated using the DBA method which, in our case, starts with the centroid defined as the average observed movement by hour among all the time series of the cluster. Then it proceeds to a phase of iterative calculation of new centroid by aligning each time series of the set and the current centroid, so that at the end of execution the centroid represents an artificial time series that best aligns with the members of the set.

5 Results

5.1 Clusters and signatures

The number of groups of businesses in a category suggested by the considered heuristic, that is, the number of existing distinct popularity patterns in the same category was usually between 2 and 4 for both weekdays and weekdays. To verify the actual need for more than one group per category a unique signature by category was generated initially.

In general, by generating only one signature representative of all the time series of businesses in a category, that is, without separating them into clusters, similar patterns of behavior still emerge when examining cities within the same country. We show representative signatures of each category on weekdays and weekends in Figures 1 and 2 for Brazil and Figures 3 and 4 for the United States, where the x-axis is the hour of the day, and the y-axis is the popularity. When comparing these results, it is possible to note that one category with great distinction is dance club, which we could expect because those type of places tends to open mainly at night, including on weekends.

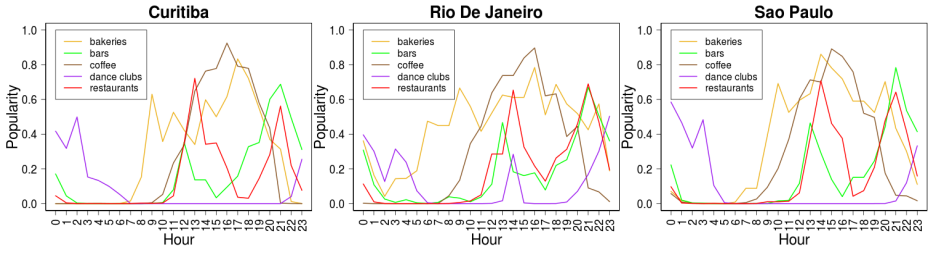

Fig. 1. Signatures for Brazil - weekdays.

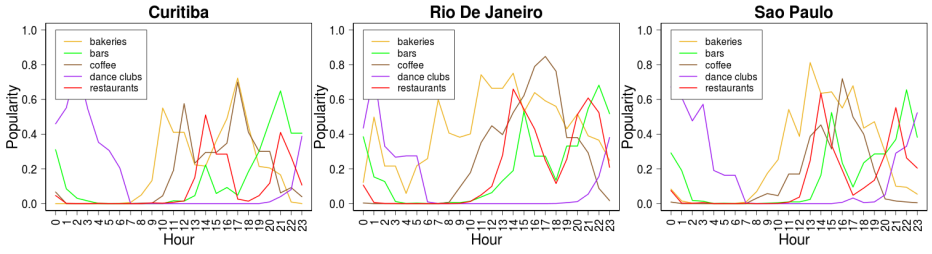

Fig. 2. Signatures for Brazil - weekends.

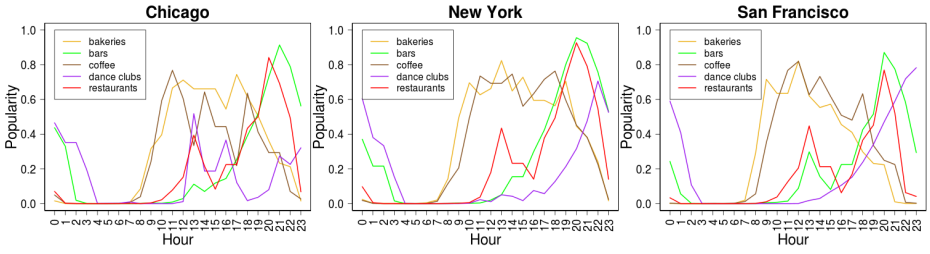

Fig. 3. Signatures for the United States - weekdays.

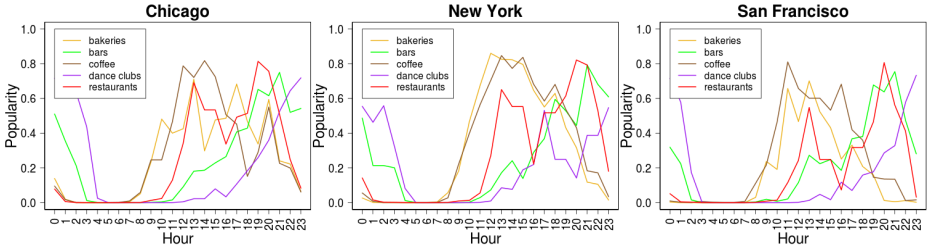

Fig. 4. Signatures for the United States - weekends.

The drawback of using only one signature per category is the loss of the possibility of identifying distinct behaviors exhibited by each category. We illustrate that for the bar category in São Paulo. Figures 5 and 6 display several time series of businesses in color and the generated signature in dotted black line for the two distinct groups discovered through the *K-means* algorithm for the category bar in the city of São Paulo on weekdays. It is noticeable that bars indeed have two distinctive types of behavior, the first presents more expressive peaks at 12h00 and 21h00 (9 p.m.) and the second with only one peak at 21h00 that decreases through the rest of the night. When studying those signatures for bar in that city, the signature containing only one expressive peak of visitation was not particularly visible in Figure 1 (without clustering).

Analyzing the different signatures presented by a category, we can note that some are similar even though they belong to different categories. That can indicate the existence of businesses which, although officially declared as a specific category, have characteristics that are closer to another category. To investigate this phenomenon we clustered the category signatures to find out if signatures of the same category would form the clusters or resulted from the merging of different categories. As shown in Figures 7 and 8, for example, the categories signatures of bar and dance club were grouped in the same set for the cities of Curitiba and Chicago.

This information could be useful in several tasks, for example, in the detection of establishments mistakenly registered in an inappropriate category. Also in the conception of meta-categories, or mixed categories, that more accurately describe the real behavior of a type of business, as the meta-category bar-dance club for

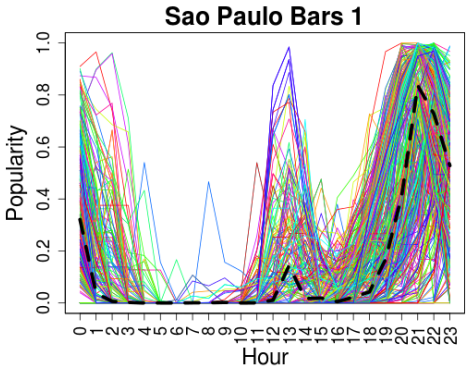

Fig. 5. Bars 1 - weekdays.

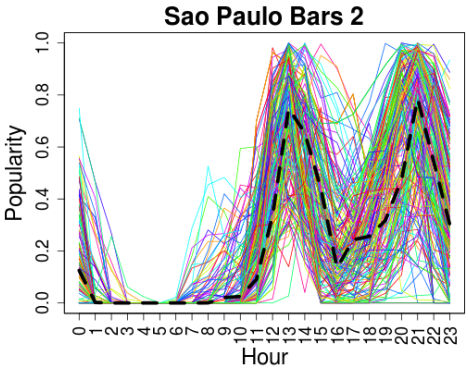

Fig. 6. Bars 2 - weekdays.

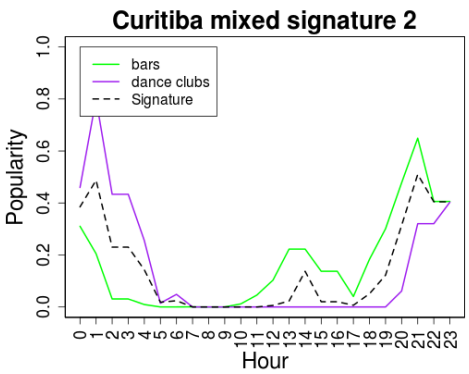

Fig. 7. Signature bar-dance clubs in Curitiba.

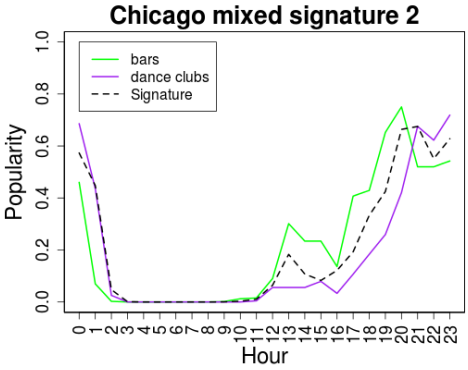

Fig. 8. Signature bar-dance clubs in Chicago.

establishments that although are initially bars are recognized as dance clubs by the visitors.

To confirm that the behaviors found in each city are consistent within a country, that is, if there are patterns of popularity that typically occurs in a country's urban scenarios, the category signatures of each city were associated by minimizing the DTW distance among them. Studying these signatures generated, we can see that there are behavioral patterns that almost not differ among different urban scenarios of a country, independently of the city.

While the signatures representing bakery and coffee shop, Figures 9 and 10, respectively, denote similar behaviors between Brazil and the United States, the signatures for bar and dance clubs, Figures 11 and 12, respectively, display more differences between the two countries. Brazil exhibits a bar signature with the highest peak during the afternoon, and the United States presents a dance club signature with a peak earlier in the evening.

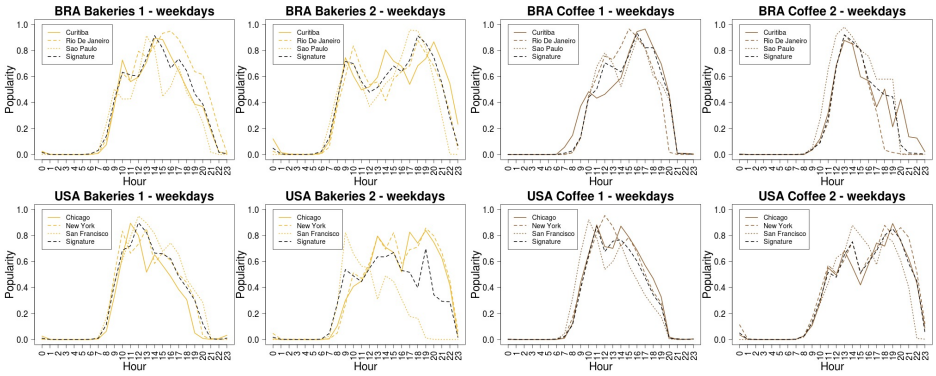

Fig. 9. Bakery signatures - weekdays.

Fig. 10. Coffee signatures - weekdays.

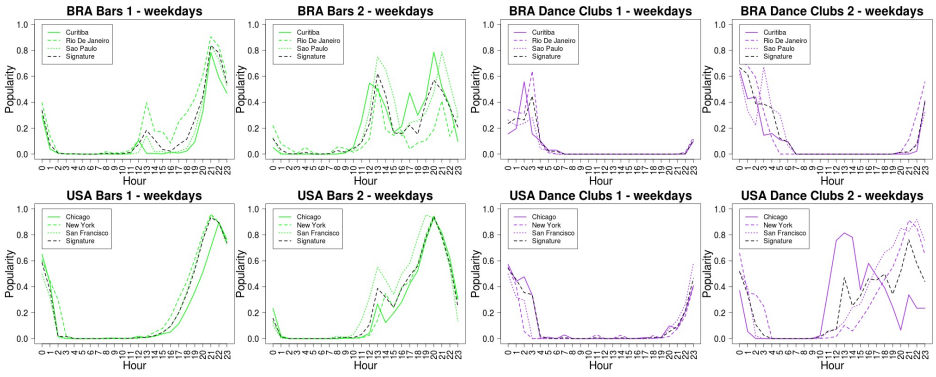

Fig. 11. Bar signatures - weekdays.

Fig. 12. Dance club signatures - weekdays.

The restaurant category revealed two very distinct behaviors between the two countries. According to Figures 13 and 14, while Brazilians tend to frequent restaurants mostly during lunch time, Americans prefer to visit this type of business at dinner. The same information was also observed using Foursquare check-ins [15].

These similarities and differences have important implications. The way in which people from different countries use certain businesses categories can be affected by many economic and social factors. Therefore the distance between country signatures could be explored when studying the culture of each country by using it as an index of similarity between countries.

5.2 Category inference

In this section, we present the task of category inference of businesses using its time series and location as an example of application. To evaluate how much

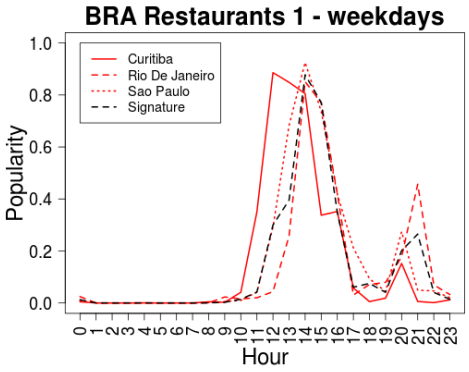

Fig. 13. Signature of restaurant in Brazil - weekdays.

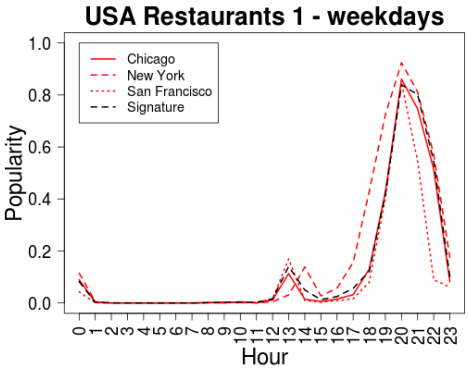

Fig. 14. Signature of restaurant in the United States - weekdays.

the signatures can generalize the behaviors of the categories, we compared two classification models.

The model $M1$ is based on a decision tree classifier that infers the category of an establishment given as attributes its time series, whether it represents the weekday or weekend, and the city from which it was obtained. We used the 10-fold cross-validation method to check the performance of the classifier in the set of all $12,704$ time series. A second inference model, $M2$, was developed using the 87 signatures of categories and meta-categories, where an establishment is categorized according to the signature closest to its time series using the DTW distance, but with a small tolerance that favors the classification as a meta-category.

Table 2 presents the results for the performance comparison between $M1$ and $M2$ models. According to those values, it is noticeable that model $M1$ presented better performance in the F1-score, while model $M2$ obtained better precision on weekdays and accuracy on weekends. We might explain this result by the small loss of generality produced by the process of assigning a signature to a large set of time series. It is also possible to note that the performance in both models is slightly worse for weekends, which may indicate the existence of a lower consistency in the behavior performed by visitors compared to working days.

Table 2. Performance comparison between M1 and M2 models.

	Weekdays		Weekends	
	M1	M2	M1	M2
Accuracy	0.70	0.67	0.59	0.65
F1-score	0.69	0.63	0.66	0.60
Precision	0.64	0.66	0.67	0.62
Recall	0.76	0.60	0.66	0.58

As the performance presented by both models are similar, the use of signatures allows us to represent several different popularity times of commercial establishments using a reduced number of time series with little loss of information.

6 Meta-category validation

One question that naturally emerges is whether meta-categories are useful to represent the perception of users regarding certain businesses. For that, an experiment was conducted to study the meta-category signatures with the help of ten volunteers aged between 18 and 30 years. They were instructed to respond to a multiple choice online questionnaire in which they should select categories from a group of five that could best describe the business. We also provided a free text option so that users were able to point out any supplementary information they thought necessary. Volunteers were free to interpret what defines each category. For each country, we randomly selected ten businesses that had been classified by a meta-category during the inference task, and, for each one, links for Facebook, Foursquare and Google were pointed to the volunteers so they could perform brief research about the venue to help them classify the venues.

Considering only the answers with the highest number of votes for the 20 analyzed businesses, 17 corresponded to the category considered as our ground truth, that is, the businesses category on the Yelp platform, while for the three remaining businesses the right category received the second highest number of votes. Therefore, according to the sample, the main categorization corresponds to what we would expect.

However, when analyzing all the response options that more than three people agreed on, we noticed that 14 businesses were classified under more than one category, corresponding to our meta-category classification. Therefore, the main category is not able to completely describe the business as the users perceive them. We can observe that in Table 3 which shows the evaluation of the volunteers about a bakery business that ended up receiving six votes for the coffee category.

Table 3. Classification of a bakery by volunteers

Category	Bakery	Bar	Coffee	Dance Club	Restaurant
Votes	9	1	6	0	1

The six cases in which categories attributed by volunteers did not entirely match the meta-category appear to be due to confusion in the interpretation of the café category, in which the free-text option included comments such as “tea house”, “ice cream shop”, as well as questions about the definitions of the café category. This phenomenon sometimes also occurs in the classification used in websites, for example, we found cases where Google classifies certain businesses

as a particular category, and the same place is described as a different category in Foursquare.

During the validation step, we discovered that the semantic of each category is difficult to define. Also, as we can note by the signatures of each country shown in Section 4, the interpretation of the meaning of each category seems to be affected by the cultural context of each country, so that only translating category names might not be sufficient to transmit the desired meaning completely. Therefore, the use of approaches such as the one presented in this study can be used to enrich the description of businesses by including knowledge about the behavior of their users.

7 Conclusion

In this study, information of the distribution of visits to commercial establishments in different cities in the United States and Brazil was collected from the Google Popular Times and used to discover patterns of popularity (signatures) that represent a group of establishments. When we analyzed the categories signatures, we could observe an association between different categories that we could use for the creation of meta-categories, perhaps more informative about the venue. Through an experiment with volunteers, we obtained evidence that those meta-categories make sense based on the perception of users. As one of the possible applications, we show that it is possible to infer the category of an establishment from its time series and city of location. For this purpose, we compared two inference models: one using the all the time series of commercial establishments and the second only the signatures of categories and meta-categories. Although using only the signatures did not have significant performance gains, using them reduces the required number of time series to perform this task. Besides, having signatures representing popularity times in different geographical regions, such as countries or cities, could enable cross-cultural studies regarding the visits time people tend to frequent business.

As future work, we propose the application of the signature discovery approach in a more substantial number of countries and cities, in order to identify possible similarities between them and explore the insights, for instance, in the better business description in different places. Another opportunity is to develop new services, for example, one that focuses on identifying and predicting low-movement hours, which can be exploited to facilitate the access to several services in fields such as banking, healthcare, and shopping.

Acknowledgements

This work was partially supported by the project CNPq-URBCOMP (process 403260/2016-7), CAPES, and Fundação Araucária. The authors would also like to thank all the volunteers for the valuable help in this study.

References

1. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition **46**(1), 243–256 (2013)
2. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (1979)
3. Figueiredo, F., Almeida, J.M., Gonçalves, M.A., Benevenuto, F.: On the dynamics of social media popularity: A youtube case study. ACM Transactions on Internet Technology (TOIT) **14**(4), 24 (2014)
4. Google: Google popular times. `https://support.google.com/business/answer/6263531` (2017), acessed em: 2017-09-10
5. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
6. Hartigan, J.A.: Clustering algorithms, vol. 209. Wiley New York (1975)
7. Krumm, J., Rouhana, D.: Placer: semantic place labels from diary data. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, Zurich, Switzerland. pp. 163–172. ACM (2013)
8. Liao, T.W.: Clustering of time series data—a survey. Pattern recognition **38**(11), 1857–1874 (2005)
9. Neves, Y.C., Sindeaux, M.P., Souza, W., Kozievitch, N.P., Loureiro, A.A., Silva, T.H.: Study of google popularity times series for commercial establishments of curitiba and chicago. In: Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, Teresina, Piauí, Brazil. pp. 303–310. ACM (2016)
10. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition **44**(3), 678–693 (2011)
11. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
12. Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis **11**(5), 561–580 (2007)
13. Silva, T.H., Loureiro, A.A.: Users in the urban sensing process: Challenges and research opportunities. In: Pervasive Computing: Next Generation Platforms for Intelligent Data Collection, pp. 45–95. Academic Press (2016)
14. Silva, T.H., Vaz de Melo, P.O.S., Almeida, J.M., Salles, J., Loureiro, A.A.F.: A picture of Instagram is worth more than a thousand words: Workload characterization and application pp. 123–132 (May 2013)
15. Silva, T.H., de Melo, P.O.V., Almeida, J.M., Musolesi, M., Loureiro, A.A.: A large-scale study of cultural differences using urban data about eating and drinking preferences. Information Systems **72**(Supplement C), 95 – 116 (2017). https://doi.org/https://doi.org/10.1016/j.is.2017.10.002, `http://www.sciencedirect.com/science/article/pii/S0306437917300261`
16. Tostes, A.I.J., Silva, T.H., Assuncao, R., Duarte-Figueiredo, F.L.P., Loureiro, A.A.F.: Strip: A short-term traffic jam prediction based on logistic regression. In: 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall). Montreal, Canada (2016)
17. Vaca, C.K., Quercia, D., Bonchi, F., Fraternali, P.: Taxonomy-based discovery and annotation of functional areas in the city. In: Proc. of ICWSM'15. Oxford, UK (2015)

18. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining, Kowloon, Hong Kong. pp. 177–186. ACM. Kowloon, Hong Kong. (2011)
19. Ye, M., Shou, D., Lee, W.C., Yin, P., Janowicz, K.: On the semantic annotation of places in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 520–528. ACM. San Diego, California. (2011)
20. Yelp: Yelp developers. `https://www.yelp.com/developers/documentation/v3` (2017), acessed: 2017-09-10
21. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **5**(3), 38 (2014)