Classifying Venue Categories of Unlabeled Check-ins Using Mobility Patterns

Helen C. M. Senefonte, Thiago H. Silva, Ricardo Lüders, Myriam R. B. S. Delgado
CPGEI / PPGCA / DAINF
Federal University of Technology - Paraná (UTFPR)
Curitiba, PR, Brazil
helen@uel.br, {thiagoh,luders,myriamdelg}@utfpr.edu.br

***Abstract*—Some location-based social networks (LBSNs) provide, besides other spatiotemporal data, the category of venues where the data was shared from. This information allows a wide range of semantic analyses which are very useful to understand city dynamics and urban social behavior. Despite being strategic to the study of cities and societies, some LBSNs do not offer the category of venues by default. In this study, we propose an approach to identify the category of venues of unlabeled check-ins according to their geographic locations. This new classification approach is inspired by the classic k-nearest neighbor algorithm improved by mobility pattern information captured through the user's transition information observed in LBSN data. The performance evaluation of the proposed approach is performed with real-world data from different cities: London, New York City, and Tokyo. Experiments show that, for all cities, we can achieve better performances when users' mobility is taken into account. Besides, we have an indication that transfer learning regarding mobility patterns could be feasible between similar cities.**

***Index Terms*—Location-based social networks, Social Sensing, Urban Computing, Category Identification**

I. Introduction

Location-based social networks (LBSNs), a class of social media, allow users to share data containing spatiotemporal information. While some data sources do not scale easily, such as GPS traces, LBSNs present widespread and global scalability as some of their main advantages [1], [2]. As an important data source for *social sensing*, LBSNs have been increasingly used to study several issues of urban societies [3].

Among the information provided by LBSNs, the venue category visited by the user (e.g., school, and coffee shop) may be present in some LBSNs like Foursquare-Swarm[1]. This information allows a wide range of semantic analyses of users' patterns, leveraging the understanding of city dynamics and urban social behavior [1], [4]–[6].

One common way to obtain data from Foursquare-Swarm is by collecting posts shared on Twitter [3]. In spite of its advantage of offering public LBSN data, Twitter does not automatically provide the venue category. To obtain this crucial information for several urban computing studies, an additional collection may be required, reducing the scalability of the process, or even turning it prohibitive due to the huge number of data typically obtained.

In this paper, we propose an approach to identify, based on the geographic location, the venue category of unlabeled check-ins[2], like ones coming from Twitter or any other LBSN information with missing venue category data. This new classification approach is inspired by the classic k-nearest neighbor algorithm (k-NN), and improved by information regarding mobility patterns. The main idea is to decide the new label considering among the nearest venues from the unlabeled geolocated check-in, which one is preferable. Instead of looking only for the destine, as it is usually performed by popularity-based approaches, in this work the classification is made based on the origin-destine information available in the city's mobility pattern database.

In a similar work, the authors in [7] used classic machine learning algorithms to classify venue category of unlabeled check-ins, obtaining the best results considering only some categories of venues individually. Previous related efforts that explored mobility patterns use traditional mobility features to conclude their experiments, such as volume and number of check-ins at venues, which are explored in [8] and [9]. Inspired by [10], in this paper we use the idea of transition graphs, yet, here, we extend the original idea by considering subcategories of places to favor a more comprehensive mobility view. Mobility features such as the number of check-ins at venues are also used in [9] but the information about users transition has not been previously explored. Therefore, innovatively, our study explores users' transition information in the problem under study. In addition, differently from all previous effort, we propose a new classifier that is based on the k-NN, and this approach explores the mobility patterns of users observed in a specific city to help the classification decision. Besides, our approach is not dependent on historical activities of users in venues, i.e., our model does not use specific patterns of individual venues in the classification process. Our approach is simple and suitable to deal with large data-sets; also it is easy to incorporate new correctly labeled information.

Our main contributions are summarized as follows:

- A new classification approach to identify the venue categories from unlabeled geolocated check-ins. This new approach is inspired by k-NN, which is modified to deal with the particular characteristics of the problem.

[1] https://www.swarmapp.com.

[2] Check-ins with GPS location without venue category information.

The decision of using a k-NN as an inspiration has been supported by previously performed comparisons, in which a k-NN based approach was competitive with random forest and outperformed neural network and support vector machine-based models. Simplicity and straightforward update processes are some of the main advantages of the proposed approach when compared with traditional training-based classification approaches.

- Identification of population mobility patterns of the addressed cities using a large scale dataset of Foursquare-Swarm and the idea of transition graphs.
- Results from experiments considering the proposed approach applied to real-world data of different cities in different countries. Satisfactory results have been obtained for the investigated problem, particularly those in which we observe a performance increase when users' mobility is taken into account. The achieved results are reliable because they reflect the mobility patterns of real users.

Therefore, our main contributions rely on: (i) the classification method based on an unexplored feature (probability of transition) which has shown in some preliminary experiments, to perform better than the venue number of check-ins, i.e., the popularity of places; (ii) the simplicity of the proposed approach, what, differently from training-based approaches, turns it interesting for our problem since straightforward update processes could take place whenever we need to incorporate new correctly labeled information (ie. retraining is not necessary); (iii) experiments considering a large scale real-world dataset.

The rest of the work is organized as follows. Section II presents the related works. Section III describes the explored dataset. Section IV describes the considered problem. The proposed approach is detailed in Section V. Section VI shows results and discussions. Finnaly, conclusion and future work are presented in Section VII.

II. Related Work

Several urban computing studies explore LBSN data in different aspects. For instance, [5] and [10] provide better understanding of city dynamics, while [11] studies features of groups with the same interest. Also [12] and [13] present an approach for detecting features associated with socioeconomic issues in different areas of a city, and [3], [14] and [15] aim to understand cultural boundaries and similarities between societies. More related to challenges derived from urban concentration, [16], [17], [18], [19] present an approach for detecting indicators of pollution, noise, and traffic for regions in a city, whereas [20] proposes mechanisms for updating control systems on traffic accidents and disasters.

Kounev [21] proposes two simple predictors to model future geo-contextual user behavior. The model is constructed to predict the future location of users by suggesting the most likely category of place and the expected time frame a visit may occur in. Silva et al. [10] propose a technique, based on transitions graphs, that summarizes people's movements between location categories of venues. The results confirm the capability of the proposed approach in finding similarities and differences in human dynamics across the different addressed cities. Gu et al. [18] present a home location identification method based on a proposed model, and evaluate it on a large real-world LBSNs dataset. Mart et al. [22] propose a methodology to identify successful public spaces through LBSNs data from Foursquare and to analyze user's urban position using morphological and historical cartographies.

In the context of unsupervised learning and more related to physical-social dependencies of information embedded in location-based social networks (LBSN), Huang et al. [12], [23] exploit the physical and social dependence between users under a maximum likelihood estimation framework. In [12], the authors' primary goal was to discover which places attract the interest of users most. In [23], the authors developed a scheme to infer if a user is a resident in a city and a venue's attractiveness by information provided in LBSNs.

Regarding supervised learning, several approaches have been presented in the literature considering different machine learning approaches such as SVM, Naive Bayes, Random Forest, and others [7]–[9], [24], [25]. Torabi et al. [24], for example, apply Decision Trees, SVM, Naive Bayes, and Random Forest to predict, based on nodes' position, the future connections in LBSNs. Every learning algorithm has its parameters configuration fine-tuned to improve its performance. The final approach is obtained from an ensemble of classifiers divided into distinct groups (partitions). Mourchi et al. [8] build a set of features that capture spatial, temporal and similarity characteristics of user mobility and combine these features for future location prediction. Wang et al. [25] use LBSN data to construct a prediction model for points of interest (POIs), such as popular attractions, and hotels.

More closely related to our work, Falcone et al. [7] presented a learning model to infer the category of a venue. Their model explores tweets spatiotemporal features and determines dynamics observed by the activities of the users, such as the duration of the stay, the time of day of the typical visit, among others. The authors use some classical classifiers to discriminate the categories of places based on those features. Ye et al. [9] proposed an approach to identify the category of places from LBSN data. They explore patterns observed at particular venues and also implicit relatedness among similar venues. For that, they use users' check-in activities to extract descriptive features for venues. They develop a semantic annotation technique for LBSNs to annotate all places with a category automatically. Their work considers feature selection and supervised learning phases to train and build the prediction model. The extracted features are used as inputs for the semantic annotation phase to learn a binary SVM for each tag. All places are used for each binary SVM training, i.e., an instance labeled with the specific semantic tag under examination is considered as a positive example, while places without this label serve as negative examples. Some gaps in these previous works are explored in this study, as pointed out in Section I.

III. Dataset Description

In this study, we explore a dataset from Foursquare concerning the year of 2014. Foursquare is an online social media[3] that uses the geographic location (a venue - building or business), in a way that users can log in to the app on their mobile phones to check-in to various locations they visit; these checkpoints can be shared with users' friends between multiple social media sites, including Facebook and Twitter.

Each venue in Foursquare has fields for category and subcategory. For instance, a given venue in Foursquare may have *Food* as category and *Burger Place* as subcategory, or *Food* and *Sushi Place* for category and subcategory, respectively. A complete list of venue categories and subcategories in Foursquare is available in the corresponding website[4]. For the sake of simplicity, we refer to the category or subcategory of Foursquare as simply category in this work.

Our Foursquare data collection was compiled from Twitter because check-ins from Foursquare are not public by default. This dataset represents approximately 4.7 million tweets containing check-ins. Data from Twitter (a tweet) contains: a location (l) where data has been shared from, composed by a tuple l=(*latitude,longitude*) of geographic coordinates; a timestamp (t) representing the time when the data was shared.

Each tweet provides a URL to the Foursquare website with information about venue category. This dataset has been previously explored in [6]. Each check-in is composed of 6 fields: user ID (*iduser*); date and time of the check-in (*date*); latitude (*latitude*); longitude (*longitude*); venue ID (*idvenue*); venue category (*categorievenue*).

There is no specific information about the city and country in each post of the original data. Thus, we use filters on latitude and longitude to identify city and country information. In this study, we explore popular cities in different continents according to check-ins from Tokyo ($27,541$ check-ins), New York City ($20,998$ check-ins), and London ($2,968$ check-ins).

IV. Problem Definition

Due to its relevance, the problem of inferring venue categories from unlabeled geolocated check-ins is getting attention over the past years [7], [9]. Predicting venue categories visited by the users enables, for example, a better understanding about user preferences, and better explanations concerning user mobility, which ultimately lead to new services and applications that might improve life in urban settings.

In this study, we present the problem assuming a specific LBSN (Twitter), but the idea can be generalized to other LBSNs as well. Given a check-in containing only GPS coordinates representing a particular venue in a city, our goal is to identify the venue category based on an existing semantic mapping *SM* and user mobility patterns *MP* observed in the city.

The city's *SM* can be obtained in several ways. One example is using open data about establishments of the city. LBSNs do not offer data for venue categories by default. In this work, we use the venue categories provided by a Foursquare dataset to construct SM and MP data, both necessary for our classifier to infer venue categories from unlabeled data.

SM can be understood as follows. In a particular city, each unique venue is represented by a category, whereas, each category has a central coordinate. In our case, the SM for a particular city is composed of all unique venues in our Foursquare-Swarm dataset, and for every unique venue location, we have a category representing the venue. In the illustrative scenario of Figure 1, the *SM* for this partial view of the city is composed of two unique venues: *Supermarket* and *Bakery*. Each of them has a central coordinate representing the venue (red dot).

A labeled check-in at a specific supermarket unit, for example, always has the same coordinate associated with the category *Supermarket*, independently of which part of the establishment the user is located at. Figure 1 illustrates this scenario. Let's suppose that there are three different users in a specific supermarket unit. One of them is located at position number 10 in *Apparel* department, the second user is situated at position 18 in *Electronics* and the third user is located at position 13 in *Home*. Although all of them are in very distant departments inside the supermarket, the coordinate of their check-ins will be the same (the central coordinate of that supermarket unit).

On the other hand, an unlabeled check-in requires additional features to be classified into a particular venue category instead of using only distance-based classification techniques. For instance, an unlabeled check-in at position 17 in *Lawn and Garden* can be closer to *Bakery* than to *Lawn and Garden* itself by using distances to the unit central coordinates. As previously pointed out, the classic k-NN approach will require some adaptation to deal with this situation as detailed in Section V.

The problem of labeling venues reaches a variety of rich applications such as recommendation, advertisement, and space planning systems. Labeled data improve uses of social media data. For example, by knowing that a place has a certain category might help people to identify leisure places for tourists.

V. Classification Approach

In this section, we present the basis of the proposed approach (Section V-A) as well as the characterization of the necessary input sets (Section V-B).

A. The Proposed Classification Approach (k-FN)

The k-NN principles inspire the proposed classification approach. Algorithm 1 presents the main steps of the proposed approach. For each city c, it receives the set U_c of unlabeled check-in coordinates, the semantic map SM_c, the mobility pattern MP_c, and an integer value representing k neighbors to be considered as inputs. At the end the algorithm outputs the set L_c with the labeled check-ins.

[3] https://socialmedia.colostate.edu/using-foursquare/

[4] https://developer.foursquare.com/docs/resources/categories.

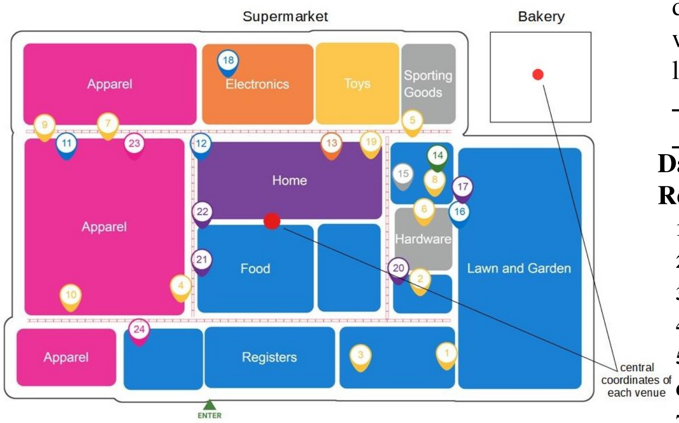

Fig. 1. Illustration of a scenario regarding unlabeled check-ins.

It starts by measuring the distance between each unlabeled check-in $u \in U_c$ to each element $sm \in SM_c$ to find the k-nearest neighbors using the Haversine distance (on a sphere). After collecting the k nearest neighbors, the classic k-NN algorithm would consider the venue label of the majority of neighbors as the label of the query instance. However, in this work, we typically do not have "a majority" of neighbors with the same label as any k nearest neighbor commonly has a different venue label. It would be possible in certain situations, e.g., a business region containing most venues with category *Office*. However, even in this case, it is not adequate to consider that an unlabeled check-in is *Office* instead of other any other category, unless we are considering a specific region in the city that only has those type of places, which is rare on practice.

This justifies the proposition of a novel approach k-FN (k-Favorable Neighbor) by modifying the classic k-NN algorithm. In the proposed approach, we also consider k to select nearest neighbors candidates. The main difference relies on the fact that k-FN explores the mobility pattern MP information to predict each label instance, instead of using the majority vote among the k competing venue labels. As previously commented, this is due the fact that there is no majority among the votes since they are all different from each other.

The approach selects $label(query) = label(j^*)$, the label of the neighbor j^* with the highest score $favorable$ among $j = 1, \ldots, k$ neighbors as in (1).

$$j^* = \arg\max_j \{favorable_j = 1/d_j + tp_j\}, \qquad (1)$$

This score takes into account the normalized distance d_j between GPS coordinates of the query and the neighbor j as well as the transition probability tp_j from MP_c of the neighbor venue j.

Therefore, the decision of our algorithm takes into account both distances among k neighbors and information about user movements in the city to select the most *Favorable Neighbor*, which inspires the name of the proposed approach. Note that if there is no information about the previous check-in in the dataset, the transition probability is zero, and only distances will be considered. At the end, all unlabeled check-ins will be labeled.

Algorithm 1: Main steps of the proposed k-FN algorithm.

```
Data: $U_c$, $SM_c$, $MP_c$, $k$
Result: $L_c$
/* Copy the unlabeled set of check-ins*/
$L_c \leftarrow U_c$
for $l \in L_c$ do
    $d \leftarrow 0$
    /* Dictionary to store $u$, distance and an extra value*/
    $N \leftarrow$ empty dictionary of size $k$
    for $sm \in SM_c$ do
        /* Distance using Haversine*/
        $d \leftarrow haversine(l, sm)$
        for $n \in N$ do
            if $d < n$ then
                /* delete $n$ neighbor*/
                $N \leftarrow delete(N, n)$
                /* Update $N$: new neighbor and $d$*/
                $N \leftarrow update1(N, n, d)$
                $break$
            end
        end
    end
    for $n \in N$ do
        /* Find the previous check-in of the user in the day*/
        $source \leftarrow source(n, U_c)$
        /* Find the transition probability for the n neighbor*/
        $tp \leftarrow transition(MP, source, n)$
        $favorable \leftarrow 1/d + tp$
        /* Update $N$: score calculated for $n$*/
        $N \leftarrow update2(N, n, favorable)$
    end
    /* Finds neighbor with the biggest score calculated*/
    $n \leftarrow favorableNeighbor(N)$
    /* Update the category for $l$ according to $n$*/
    $L_c \leftarrow update3(n, sm, l)$
end
```

B. Building Sets SM_c, U_c, and MP_c

The semantic map SM_c is the ground-truth view of a particular city c. It is obtained from labeled check-ins of Foursquare using latitude, longitude, and venue category.

In the experiments we considered that the set U_c of unlabeled check-ins is a synthetic dataset generated by perturbing the original venue coordinates of SM_c. It simulates new data coming from LBSN sources containing check-ins with only GPS coordinates (without any venue category information) as in Twitter. The synthetic dataset is created by considering uniformly random generated noise added to the original coordinates of SM_c. This approach simulates perturbations that might occur on latitude and longitude of the original venue coordinates [26].

Besides, each city is characterized by an aggregated mobility pattern MP describing users' transition patterns manifested in Foursquare. The mobility pattern is a directed weighted graph $G = (V, E)$ with a set V of vertices corresponding to venue categories, and a set $E = V \times V$ of arcs such that $(i, j) \in E$ represents two consecutive check-ins (a transition) made by the same user from venues i to j. The weight of arc (i, j) is the total number of transitions computed from i to j in the whole dataset within a given time interval. This MP is represented by an n-square matrix M whose entry (i,j) is the probability of having category j as the consecutive check-in from category i. Each entry (i,j) of M is computed as a relative frequency between the number of transitions from i to j and the total number of transitions having j as a destination venue.

Despite testing different time intervals during the computation of the number of transitions between venues i and j, we decided to use a 'social day' which ranges from 5:00 am to 4:59 am of the next day. By doing that, we increase the chance of considering transitions that started before midnight and ended late at night. Therefore, only check-ins (origin-destine) performed in the same social day are considered when computing every transition.

VI. Experiments and Results

In this section, we present the main characteristics of the data from studied cities (Section VI-A), detail the method used to evaluate the accuracy of classification from synthetic datasets (Section VI-B), and show the results of our approach to classify unlabeled check-ins into venue categories (Section VI-C).

A. City data characteristics and users' mobility

The exploratory analysis performed to obtain the main sets considered in the proposed approach (U_c, SM_c, MP_c) turned possible a better understanding of important information about the studied cities and their population. For instance, there are more than 130 unique categories in each city, Tables I, II and III, show the most popular venue categories chosen by the users to perform their check-ins in Tokyo, New York City, London, respectively.

As we can see, the top category of venues to perform check-ins varies considerably among cities. It is also possible to note that the popularity reflects, somehow, the cultural habits associated with the users. "Ramen or Noodle House" in Tokyo is very present in the Japanese culture. Pub is also a typical place in the United Kingdom; similarly, Bar venues tend to be more common among users from the United States. It is also interesting to note that performing check-ins at "Home (private)" venues are not as popular in Tokyo as in London and New York. This might suggest that users in Japan are less concerned with privacy then users from the UK or the USA. These observations suggest significant differences regarding the behavior of users in certain places of the world, inspiring further investigation on these findings.

TABLE I
Top categories in Tokyo.

Category	Number of Check-ins
Train Station	10023
Subway	1361
Ramen or Noodle House	1165
Japanese Restaurant	960
Park	844
Mall	732
Convenience Store	642
Electronics Store	618
Supermarket	611
Coffee Shop	517

TABLE II
Top categories in New York City.

Category	Number of Check-in
Bar	1153
Park	911
Home (private)	775
Coffee Shop	746
American Restaurant	705
Office	467
Airport	444
Train Station	438
Italian Restaurant	435
Gym	424

TABLE III
Top categories in London.

Category	Number of Check-in
Pub	357
Train Station	266
Bar	161
Home (private)	155
Hotel	124
Coffee Shop	121
Park	105
Airport	92
Stadium	84
Supermarket	79

Next, we investigated the transition (movements) between venue categories performed by each user. This is important because the extraction of urban user movements can help to understand the dynamics of each city [10]. To accomplish this analysis, we followed the steps to generate the mobility pattern MP described in Section V.

The total number of transitions extracted in Tokyo is $17,206$. Table IV presents the most popular transitions between pairs of categories, which are *Train Station* to *Train Station* ($3,038$ transitions), *Subway* to *Subway* (317 transitions), and *Train Station* to *Subway* (225 transitions). Transitions involving the same categories were considered because they represent different venues (with distinct *idvenue*). There are more than 600 different train stations in Tokyo, each one identified with a unique *idvenue*, and all of them are in the same category "Train Station". Since these transitions represent valid ones, they are important to be accounted.

TABLE IV
MOST POPULAR TRANSITIONS BETWEEN CATEGORIES IN TOKYO.

From	To	# of Transitions
Train Station	Train Station	3,038
Subway	Subway	317
Train Station	Subway	225
Subway	Train Station	204
Train Station	Ramen or Noodle House	120
Train Station	Electronics Store	107
Train Station	Japanese Restaurant	105
Train Station	Convenience Store	100
Train Station	Mall	98
Ramen or Noodle House	Train Station	90

As expected, the categories with the highest number of transitions are correlated with the most popular categories in Tokyo, shown in Table I.

The most popular transitions between categories for New York City and London are available in Table V. Although the rank of transitions for New York and London is different, there are more similar transitions between New York and London than between these two cities and Tokyo.

TABLE V
MOST POPULAR TRANSITIONS BETWEEN CATEGORIES IN NEW YORK CITY AND LONDON.

New York City		London	
From	To	From	To
Bar	Bar	Train Station	Train Station
Train Station	Trains Station	Pub	Pub
Home (private)	Home (private)	Pub	Bar
Park	Park	Subway	Subway
Subway	Subway	Pub	Train Station
Art Gallery	Art Gallery	Train Station	Pub
Highway-Road	Highway-Road	Pub	Stadium
Supermarket	Home (private)	Subway	Train Station
Beer Garden	Bar	Home (private)	Home (private)
Playground	Playground	Park	Park

B. Accuracy Evaluation

The accuracy[5] evaluation of the proposed classifier is made on two different synthetic datasets of unlabeled check-ins U_c. *Synthetic-Dataset1* is built from the original venues of SM_c by adding into their coordinates a uniformly random generated noise of a maximum of 15 meters. Similarly, *Synthetic-Dataset2* is built by adding a uniformly random generated noise of a maximum of 50 meters. These values try to simulate typical errors expected in urban scenarios and semi-urban scenarios, respectively [26]

In addition, we use a 5-fold cross-validation approach [27]. It means that five sets are created from each dataset *Synthetic-Dataset1* and *Synthetic-Dataset2*. Each set corresponds to $1/5$ of the respective datasets. The accuracy of the classifier is then given by the average over five accuracy values obtained for each set. The average is also characterized by a confidence interval using t-Student distribution with a significance level of $\alpha = 5\%$.

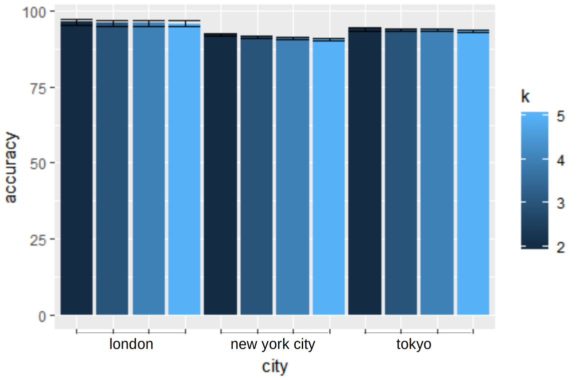

Fig. 2. Accuracy results (average, and confidence intervals) for the classification of unlabeled check-ins of *Synthetic-Dataset1* using our proposed approach for all cities and different values of k. Each value in the legend refers to results using a particular value of k.

C. Results for Unlabeled Check-ins Classification

As described in Section V, our classification approach explores the mobility pattern expressed by the transitions between different venues. In this section, we present the results for the mechanisms discussed in Sections V-A and V-B. Thus, the classification tests were performed in different scenarios.

Figure 2 shows the accuracy for the classification of unlabeled check-ins of *Synthetic-Dataset1* (15 meters) using our proposed approach for all cities and different values of k. According to Fig. 2, there is no statistically significant differences between different values of k for the same city. This means that k-FN may be less sensitive to different values of $k \geq 2$, although further experiments should be done. In this case, we have chosen $k = 3$ with the highest average accuracy for all cities.

Figure 3 shows the accuracy for the classification of unlabeled check-ins of *Synthetic-Dataset2* (50 meters) using our proposed approach for all cities and different values of k. Similarly to Fig. 2, Fig. 3 shows no statistically significant differences between different values of k for the same city. However, $k = 3$ provides the highest average accuracy for all cities. In this case, the k-FN algorithm obtained a slightly better performance with more neighbors when considering higher noise. Nevertheless, k-FN may be still less sensitive to different values of $k \geq 2$.

It is important to emphasize that the proposed approach uses only the distance to make its decision for $k = 1$, i.e., the k-FN algorithm ignores the mobility pattern when $k = 1$, as there is no tie in the voting process and the query label is the one associated with the closest neighbor. This result is used as a reference value (baseline) to compare the performance of our experiments. When $k \geq 2$, the algorithm uses both distance and mobility pattern (expressed in terms of transition probability) of k-neighbors according to subsection V-A. The results for $k = 1$ are thus omitted in Figures 2 and 3 and showed only in Table VI for comparison.

In order to better understand the influence of transitions on

[5] Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

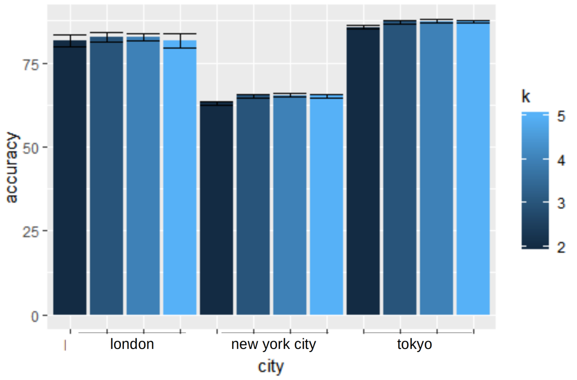

Fig. 3. Accuracy results (average and confidence interval) for the classification of unlabeled check-ins of *Synthetic-Dataset2* using our proposed approach for all cities and different values of k. Each value in the legend refers to results using a particular value of k.

the performance of our k-FN algorithm, we summarize the main results of classification in Table VI (average accuracy and the respective confidence interval in subscript as in x_y which means $x \pm y$). The results of Table VI are presented for both *Synthetic-Dataset1* and *Synthetic-Dataset2* (column SD) using the baseline and taking into account distances and transitions (dist+trans in column Approach). It is possible to observe that there are improvements for all cities using both datasets. Precision, recall and F1-score[6] reflect the message given by accuracy in all cases. Because of space limitation they were omitted in the text for this analysis.

TABLE VI
ACCURACY (%) OF k-FN. SD STANDS FOR SYNTHETIC DATASET.

Approach	SD	Tokyo	New York	London
baseline ($k = 1$)	1	$88.63_{0.55}$	$87.32_{0.33}$	$94.82_{1.36}$
dist+trans ($k = 3$)	1	$93.86_{0.37}$	$91.28_{0.43}$	$96.18_{1.19}$
baseline ($k = 1$)	2	$77.56_{0.72}$	$55.74_{0.55}$	$76.48_{1.27}$
dist+trans ($k = 3$)	2	$87.21_{0.57}$	$65.23_{0.53}$	$82.49_{1.19}$

Although the main contribution of our results came from the addition of transition information, it is important to point out the use of synthetic datasets. The *Synthetic-Dataset1* was tested to simulate problems with GPS coordinates in an urban setting. For this case the accuracy is considerably high, indicating that this strategy could work well in practice. Since the performance results were good for this scenario, we decided to increase the noise, trying to simulate a semi-urban scenario to evaluate how the accuracy would be impacted. As expected using *Synthetic-Dataset2*, this scenario is more affected, decreasing the accuracy from around 90% to 70%.

Additional experiments were done trying to understand the influence of mobility patterns in different cities. To evaluate these experiments we tested London MP in New York City (and vice-versa), London MP in Tokyo (and vice-versa) and Tokyo MP in New York City (and vice-versa). Table VII compares the results using the original MP and the MP taken from another city. As expected, all metrics are better for SD1 than for SD2, i.e., high noise level in GPS coordinates imposes additional challenge into the classification process. Precision and accuracy metrics have better results when using the original MP for all cities but London. Results for London are not affected by any of the MP considered (rows in light blue). The recall metric is affected by using the original MP only for NYC in SD2. It means that positive predictive values (precision and accuracy) are benefited by using the original MP, but false negatives (recall) are not in general.

VII. CONCLUSION AND FUTURE WORK

This study tackled the problem of classifying venue categories from unlabeled check-ins with geographic location. We introduced a novel classification algorithm, k-FN to deal with the particular characteristics of our problem. In our new approach, we explored mobility patterns that were observed through a mobile crowdsensing fashion with LBSN data. It is important to point out that the proposed classification method is based on an unexplored feature (probability of transition) and its simplicity allows straightforward update processes whenever new correctly labeled information tuns out available (ie. retraining is not necessary). To evaluate k-FN performance we used a large scale real-world dataset representing three different cities of different continents. The results using k-FN were satisfactory for all considered cities, reaching, in some cases, around 96% accuracy (in more realistic simulated scenarios). It suggests that users' mobility is relevant to be taken into consideration for the studied problem, and also that our proposed approach could be an interesting alternative in practice. Another important conclusion is that the proposed approach presents a low sensibility to its main parameter (k), what is an important achievement for any machine learning-based approach. Future work could address the investigation of other types of users' mobility data in the decision process and also the inclusion of weights based on the time stamp for the transition values.

ACKNOWLEDGMENT

This work was partially supported by the project CNPq-URBCOMP (process 403260/2016-7), and individual grants from CAPES, CNPq (process 309935/2017-2), and Fundação Araucária. H. C. M. S. thanks Londrina State University.

REFERENCES

[1] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," in *Proc. of ICWSM'12*, (Dublin, Ireland), 2012.

[2] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, and A. A. F. Loureiro, "Large-scale study of city dynamics and urban social behavior using participatory sensing," *IEEE Wireless Comm.*, vol. 21, pp. 42–51, 2014.

[3] T. Silva, A. Viana, F. Benevenuto, L. Villas, J. Salles, A. Loureiro, and D. Queercia, "Urban computing leveraging location-based social network data: a survey," *ACM Computing Surveys*, p. 37, 2019.

[4] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An Empirical Study of Geographic User Activity Patterns in Foursquare," in *Proc. of ICWSM'11*, (Barcelona, Spain), 2011.

[6] A model that produces no false positives has a precision of 1.0, while the absence of false negatives means recall 1.0. F1 Score is the weighted average of precision and recall.

TABLE VII

ACCURACY OF k-FN WITH DIFFERENT MPS. SD STANDS FOR SYNTHETIC DATASET, NYC = NEW YORK CITY, LND = LONDON, TKO = TOKYO.

Approach	SD	Precision	Recall	F1-score	Accuracy
NYC - original MP	1	$93.49_{0.87}$	$90.48_{1.12}$	$91.69_{0.65}$	$91.28_{0.43}$
NYC - LND MP	1	$89.86_{0.79}$	$88.12_{0.66}$	$88.64_{0.68}$	$88.53_{0.35}$
NYC - TKO MP	1	$90.90_{0.96}$	$88.76_{0.69}$	$89.54_{0.68}$	$89.21_{0.52}$
LND - original MP	1	$96.45_{1.61}$	$95.35_{1.55}$	$96.34_{1.41}$	$96.18_{1.19}$
LND - NYC MP	1	$96.34_{1.24}$	$94.54_{1.72}$	$96.15_{1.09}$	$96.06_{0.84}$
LND - TKO MP	1	$96.01_{0.87}$	$94.36_{1.51}$	$95.83_{1.15}$	$95.90_{0.97}$
TKO - original MP	1	$95.15_{0.49}$	$91.84_{1.36}$	$96.34_{1.41}$	$96.18_{1.19}$
TKO - NYC MP	1	$93.04_{0.82}$	$91.12_{1.57}$	$91.66_{0.82}$	$92.44_{0.51}$
TKO - LND MP	1	$91.86_{0.73}$	$90.62_{0.73}$	$90.82_{0.89}$	$91.26_{0.54}$
NYC - original MP	2	$70.37_{1.62}$	$62.88_{1.55}$	$68.28_{1.16}$	$65.23_{0.53}$
NYC - LND MP	2	$61.69_{1.24}$	$57.28_{1.40}$	$61.51_{1.00}$	$58.89_{0.40}$
NYC - TKO MP	2	$63.04_{1.41}$	$58.40_{1.53}$	$62.87_{1.10}$	$60.22_{0.60}$
LND - original MP	2	$83.80_{1.44}$	$79.53_{0.76}$	$84.46_{2.07}$	$82.49_{1.19}$
LND - NYC MP	2	$83.50_{2.23}$	$77.40_{3.08}$	$84.51_{1.85}$	$81.71_{0.94}$
LND - TKO MP	2	$82.50_{2.49}$	$77.01_{2.09}$	$83.87_{1.29}$	$81.45_{1.75}$
TKO - original MP	2	$88.89_{0.83}$	$83.85_{1.38}$	$86.69_{0.78}$	$87.21_{0.57}$
TKO - NYC MP	2	$84.86_{0.67}$	$82.71_{1.19}$	$83.76_{0.83}$	$83.91_{0.84}$
TKO - LND MP	2	$82.66_{1.20}$	$81.79_{1.35}$	$82.20_{1.15}$	$82.68_{0.83}$

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

[6] W. Mueller, T. H. Silva, J. M. Almeida, and A. A. Loureiro, "Gender matters! analyzing global cultural gender preferences for venues using social sensing," *EPJ Data Science*, vol. 6, no. 1, p. 5, 2017.

[7] D. Falcone, C. Mascolo, C. Comito, D. Talia, and J. A. Crowcroft, "What is this place? inferring place categories through user patterns identification in geo-tagged tweets," in *Proc. of MobiCASE'14.*, (Austin, Texas, United States.), pp. 10–19, 2014.

[8] F. Mourchid, A. Habbani, and M. Elkoutbi, "Mining user patterns for location prediction in mobile social networks," in *Proc. of IEEE CIST'14*, pp. 213–218, 2014.

[9] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz, "On the semantic annotation of places in location-based social networks," in *Proc. of KDD '11*, (San Diego, California, USA), pp. 520–528, ACM, 2011.

[10] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, J. F. S. Salles, and A. A. F. Loureiro, "Revealing the city that we cannot see," *ACM Trans. Internet Techn.*, vol. 14, pp. 26:1–26:23, 2014.

[11] A. P. G. Ferreira, T. H. Silva, and A. A. F. Loureiro, "Beyond sights: Large scale study of tourists' behavior using foursquare data," in *Proc. of ICDM'15 work.*, (Atlantic City, United States.), pp. 1117–1124, 2015.

[12] C. Huang and D. Wang, "Unsupervised interesting places discovery in location-based social sensing," in *Proc. of DCOSS'16*, (Washington, DC, United States), 2016.

[13] D. Hristova, M. J. Williams, M. Musolesi, P. Panzarasa, and C. Mascolo, "Measuring urban social diversity using interconnected geo-social networks," in *Proc. of WWW16*, (Montreal, Canada, 2130), 2016.

[14] R. O. G. Gavilanes, Y. Mejova, and D. Quercia, "Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication," in *Proc. of CSCW14*, (Baltimore, USA), 2014.

[15] G. L. Falher, A. Gionis, and M. Mathioudakis, "Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities," in *Proc. of ICWSM'15*, (Oxford, UK), 2015.

[16] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, pp. 38:1–38:55, 2014.

[17] A. I. J. T. Ribeiro, T. H. Silva, F. de L. P. Duarte-Figueiredo, and A. A. F. Loureiro, "Studying traffic conditions by analyzing foursquare and instagram data," in *Proc. of PE-WASUN*, (Montreal, Canada), 2014.

[18] Y. Gu, W. Liu, Y. Yao, and J. Song, "Fast routing in location-based social networks leveraging check-in data," in *Proc. of iThings, GreenCom and CPSCom*, (Taipei, Taiwan), pp. 428–435, 2014.

[19] H.-P. Hsieh, C.-T. Li, and S.-D. Lin, "Exploiting large-scale check-in data to recommend time-sensitive routes," in *Proc. of UrbComp'12*, (Beijing, China), p. 5562, 2012.

[20] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proc. of SIGSPATIAL13*, (Orlando, Florida, United States), p. 344353, 2013.

[21] V. Kounev, "Where will i go next?: Predicting future categorical check-ins in location based social networks," in *Proc. of CollaborateCom'12*, (Pittsburgh, United States), pp. 605–610, 2012.

[22] P. Mart, L. Serrano-Estrada, and A. Nolasco-Cirugeda, "Using locative social media and urban cartographies to identify and locate successful urban plazas," *Cities*, vol. 64, pp. 66 – 78, 2017.

[23] C. Huang, D. Wang, J. Tao, and B. Mann, "On physical-social-aware localness inference by exploring lbsn," *IEEE Transactions on Big Data. Volume: PP Issue: 99. 1-14.*, 2017.

[24] N. Torabi, H. Shakibian, and N. M. Charkari, "An ensemble classifier for link prediction in location based social network," in *Proc. of ICEE*, (Shiraz, Iran), pp. 529–532, 2016.

[25] Z.-S. Wang, J.-F. Juang, and W.-G. Teng, "Predicting poi visits with a heterogeneous information network," in *Proc. of TAAI*, (Tainan, Taiwan), pp. 388–395, 2015.

[26] M. Lehtinen, A. Happonen, and J. Ikonen, "Accuracy and time to first fix using consumer-grade gps receivers," in *Proc. of SoftCOM'08*, (Dubrovnik, Croatia), pp. 334–340, Sep. 2008.

[27] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.