Received November 25, 2021, accepted December 27, 2021, date of publication January 14, 2022, date of current version January 26, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3143503*

# PredicTour: Predicting Mobility Patterns of Tourists Based on Social Media User's Profiles

**HELEN C. MATTOS SENEFONTE**[1,2], **MYRIAM REGATTIERI DELGADO**[1], **RICARDO LÜDERS**[1], **AND THIAGO H. SILVA**[1,3]

[1] Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI), Universidade Tecnológica Federal do Paraná, Curitiba 80230-901, Brazil
[2] Departamento de Computação, Universidade Estadual de Londrina, Londrina 86057-970, Brazil
[3] School of Cities, University of Toronto, Toronto, ON M5S 0C9, Canada

Corresponding author: Helen C. Mattos Senefonte (helen@uel.br)

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; the São Paulo Research Foundation (FAPESP) through the project GoodWeb under Grant 2018/23011-1; and the National Council for Scientific and Technological Development - Brazil (CNPq) under Grant 439226/2018-0, Grant 314699/2020-1, and Grant 310998/2020-4.

**ABSTRACT** This paper proposes PredicTour, an approach to process check-ins made by users of location-based social networks (LBSNs), and predict mobility patterns of tourists visiting new countries with or without previous visiting records. PredicTour is composed of three key parts: mobility modeling, profile extraction, and tourist mobility prediction. In the first part, sequences of check-ins within a time interval are associated with other user information to produce a new structure called "mobility descriptor". In the profile extraction, self-organizing maps and fuzzy C-means work jointly to group users according to their mobility descriptors. PredicTour then identifies tourist profiles and estimates mobility patterns of tourists visiting new countries. When comparing the performance of PredicTour with three well-known machine learning-based models, the results indicate that PredicTour outperforms the baseline approaches. Therefore, it is a good alternative for predicting and understanding international tourists' mobility, which has an economic impact on the tourism industry when services and logistics across international borders should be provided. The proposed approach can be used in different applications, such as in recommender systems for tourists or in decision-making support for urban planners interested in improving tourists' experiences and attractiveness of venues through personalized services.

**INDEX TERMS** Location-based social network, mobility, international tourism, social and urban computing, machine learning.

## I. INTRODUCTION

The tourism industry is essential in several economies. According to the World Tourism Organization (WTO), the flux of tourists around the world generated revenues of more than one billion US dollars in 2019 [1], and it created millions of direct and indirect jobs [1].

In this context, it is relevant to understand patterns of tourists' behaviors to improve the attractiveness of venues with more efficient and personalized services. In particular, the study of tourist mobility is an under-explored aspect of tourism scholarship [2], [3]. Despite previous efforts, very few works have attempted to model the mobility patterns of tourists on large scale [4], [5]. One of the challenges involves finding the appropriate type of data.

Location-based social networks (LBSNs) as Foursquare, Waze, Twitter, and Instagram,[1] provide a new range of possibilities to obtain data on large scale, especially with a considerable increase of social media users. LBSNs have been successfully explored in large scale studies on users' behavioral patterns [6]. They range from identification of specific groups of people with the same interest [7] and study of socio-economic problems in different areas of a city [8], [9], to the understanding of cultural boundaries, and similarities between societies [10]–[12]. In addition, because LBSN data can be obtained from different places around the world, this type of data represents an alternative for studies interested in behavioral patterns of users acting as tourists [7], [13], [14].

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

[1] https://foursquare.com, https://waze.com, https://twitter.com, and https://instagram.com, respectively.

The present work aims to predict international tourists' mobility patterns using LBSNs. We provide a novel approach called PredicTour. First, it models the mobility of users from different perspectives to produce a *mobility descriptor* of each user. Next, it explores this structure to extract profiles of those users classified as tourists with similar characteristics. Finally, taking all the obtained information together, i.e., the model of tourists' previous mobilities associated with their identified profile, the proposed approach predicts the international tourists' mobility pattern in different countries. The experiments show that PredicTour can be extended to cases with no prior information about the tourist's behavior in other countries.

In the present paper, we aim at answering three research questions: (i) what kind of intrinsic relationships can be observed when we group users? (ii) what kind of pattern can be observed in each profile? (iii) how does PredicTour perform when compared with baselines under different difficulty levels? To comparatively evaluate the performance of PredicTour, we consider well-known machine learning-based models – Deep AutoEncoder, Multi-layer Perceptron, and Collaborative Filtering – as comparison approaches. The results indicate that PredicTour outperforms all the baseline approaches, improving the understanding and prediction of international tourists' mobility. The main contributions of the study can therefore be summarized as follows:

- The proposition and exploration of a new structure *mobility descriptor* that considers different data features ranging from straightforward information, such as users' origin and destination countries, to sophisticated information extracted from users' mobility in LBSNs.
- An approach to perform profile extraction which separates groups of tourists with similar mobility patterns and describes each group based on its profile.
- A novel methodology to predict mobility patterns of international tourists when visiting new countries with and without previous information.

Based on these contributions, we believe that PredicTour can be helpful for many applications in tourist planning. For instance, it can build recommender systems of new places for particular groups of international tourists.

The remainder of the paper is organized as follows. Section II discusses related works. Section III details PredicTour. Section IV describes the methodology used in the experiments, with comparison approaches and metrics being discussed in Section V. Results are presented and analyzed in Section VI, with conclusion and future perspectives discussed in Section VII.

## II. RELATED WORKS

In the literature, there are many studies related to our research in different aspects. To summarize the main related topics, we consider the aspects of human mobility using user-generated data, prediction of trends from LBSN data, and characterization of tourist activity patterns.

### A. CHARACTERIZATION OF HUMAN MOBILITY

According to Barbosa *et al.* [15], the study of human mobility is especially important for applications such as estimating migratory flows, traffic forecasting, urban planning, and epidemic modeling. As the authors show, there are several initiatives in this direction. For instance, Pappalardo *et al.* [16] use mobile phone and GPS data to explore patterns of human mobility. They discovered the existence of two distinct classes of individuals: returners and explorers. Lima *et al.* [17] use mobile network data records (from call detail record data) to analyze the behavior of a large number of individuals. According to the authors, human mobility and social structure are important characteristics to understand the diseases spreading. Mourchi *et al.* [18] build a set of features that capture spatial, temporal, and similarity characteristics of user mobility and combine these features for future location prediction.

In the study of Amoretti *et al.* [19], a smart mobility application that recommends points of interest (POI) is proposed based on users' behavior. Aiming to build individual and group behavior profiles, the authors consider user actions, for instance, through check-ins and preferences. Luceri *et al.* [20] investigate the social influence and how it impacts human behavior from event-based social network data. They study how influence propagates among subjects in a social network. In a similar direction, Roy *et al.* [21] quantify the impacts of an extreme event on human mobility, showing that geolocated social media data allow studying socio-economic impacts and help to guide policies toward developing disaster strategies. The study developed by Rajashekar *et al.* [22] shows that the behavioral models proposed by the authors are capable of uniquely identifying each user under a one-class learning constraint. They used smartphone data such as those provided by specific applications, cell towers, and websites to construct a user-specific behavioral model.

### B. PREDICTION OF HUMAN MOBILITY

Regarding the predictability of human mobility, Gonzalez *et al.* [23] present a deep study with 100,000 mobile phone users whose positions have been tracked for six months, presenting evidence supporting such a phenomenon. Connected to that, Song *et al.* [24] raise an important question: To what degree is human behavior predictable? The authors explore the limits of predictability in human dynamics by studying anonymized mobile phone users' mobility patterns. Brockmann *et al.* [25] show that human traveling behavior can be described mathematically on many spatiotemporal scales by a two-parameter continuous-time random walk model with surprising accuracy. Moreover, Zheng *et al.* [4] explore multiple users' GPS trajectories to mine interesting locations and classical travel sequences in a given geospatial region. Ben Zion and Lerner [26] show evidence that by using cellular data, it is possible to identify social lifestyles and extract mobility patterns to better predict the trajectory of users.

Considering related works that also explore LBSN data, Hsieh *et al.* [27] propose a system to recommend time-sensitive trip routes. They use a sequence of locations with associated timestamps, based on the knowledge extracted from large-scale check-in data. In the same direction, Gu *et al.* [28] developed a system that can accomplish fast routing in LBSN, leveraging geographical knowledge predicted from check-in data. Wang *et al.* [29] use LBSN data to construct a prediction model for POI, such as restaurants, stores, popular attractions, and hotels.

Silva *et al.* [30] propose a technique based on transition graphs that summarizes people's movements between location categories of venues; the authors highlight that specific transitions are much more probable than others. Domenico *et al.* [31] focus on the study of the interdependence and predictability of human mobility and social interactions. Senefonte *et al.* [14] propose a novel classification approach $k$-FN to identify the venue categories from unlabeled geolocated check-ins with noisy data using mobility patterns of users. Additionally, D'Silva *et al.* [32] propose a prediction framework able to forecast weekly popularity dynamics of new places by using mobility data from Foursquare and k-nearest neighbor metrics.

### C. TOURISTS' ACTIVITY PATTERNS

There are also studies related to tourists' activity patterns in the literature exploring large-scale mobile data. For example, Vu *et al.* [33] analyze cross-country tourist activities. Grinberger and Shoval [34] explore smartphone data to study tourists' activity patterns and the time-space resource allocation decisions they support. Lozano and Gutiérrez [35] use WTO data to study global tourism network. Ferreira *et al.* [36] consider spatio-temporal aspects of the behavior of tourists and residents to analyze the tourists' behavior. Yochum *et al.* [37] present a current systematic review and map the linked open data, i.e., structured information in a format meant for machines, to a location-based recommendation system in the tourism domain.

More closely related to our present study, Ferreira *et al.* [38] explore LBSN to study the tourist movements through time and space. The authors propose an approach based on a topic model (using Latent Dirichlet Allocation - LDA) to automatically identify mobility pattern themes used to better understand users' profiles.

We explore some of the gaps found in the related works. For example, no other work has explored complex data regarding users' mobility patterns. Moreover, researchers that cluster users by a profile approach usually employ crisp clustering algorithms such as K-means. As human behavior is quite complex, fuzzy clustering methods (as the C-means explored in our approach) tend to be more appropriate because they allow us to capture a more complete and realistic scenario, especially with the idea of one individual belonging to more than one cluster, i.e., the individual might present characteristics of more than one profile. Another important advantage of our method is the use of LBSN data since it improves scalability compared to other sources of data explored in the literature. It is also worth mentioning that, to the best of our knowledge, there is no previous approach proposed to help predict international tourists' mobility in different countries in the way we do in this study. As detailed in the next section, we use mobility descriptors to provide clusters of tourists with similar profiles. Then, the proposal predicts unknown mobility patterns based on those profiles, which can be associated with tourist historical data whenever they are available.

## III. PredicTour

Here we describe the new proposed approach called PredicTour that considers relevant features extracted from LBSN data beyond the trivial tourist origin and destination countries. An overview of the proposed approach is depicted in Figure 1, which includes three main parts: (i) Mobility Modeling, (ii) Profile Extraction, and (iii) Tourists' Mobility Prediction. The first two parts are building blocks for the main task of the third block. Figure 1 also highlights the key outputs of each part of PredicTour. The output of the first block is a mobility descriptor, which aggregates essential features of tourists' behavioral patterns. The second block identifies user profiles based on mobility descriptors. The prediction of user mobility patterns is the output of the last block.

### A. MOBILITY MODELING

One of the contributions of the present work is a structure, namely mobility descriptor, that describes tourists' mobility patterns considering features from several perspectives, providing a rich description of the phenomenon under study. A *mobility descriptor* is a vector $\mathbf{d} = (\mathbf{m}|\mathbf{v}_{class}|\mathbf{v}_{home}|\mathbf{v}_{dest})$ obtained from the concatenation of a *mobility vector* $\mathbf{m} \in \mathbb{N}^{(V \cdot V)}$ with a binary vector $\mathbf{v}_{class} \in \{0, 1\}^2$ composed of two elements, and two binary vectors $\mathbf{v}_{home} \in \{0, 1\}^L$ and $\mathbf{v}_{dest} \in \{0, 1\}^L$, both in one hot codification, with as many components as the number $L$ of locations (countries in our case).

The mobility vector $\mathbf{m}$ contains the number of transitions made by a tourist between two venue categories in a set of $V$ different categories (or types of visited places) according to [7], [13], [14]. The other vectors improve data from different perspectives. Vector $\mathbf{v}_{class}$ expresses user's classification as returners or explorers, a concept and methodology proposed by [16]. The user's country origin $\mathbf{v}_{home}$ and destination $\mathbf{v}_{dest}$ are then aggregated to the previous vectors to complete the information needed. Appendix A details the algorithm of building the mobility descriptor $\mathbf{d}_u^{(dest)}$ of a user $u$ visiting a destination country $dest$.

### B. PROFILE EXTRACTION

The second block of PredicTour extracts profile patterns from mobility descriptors. It encompasses two main tasks:

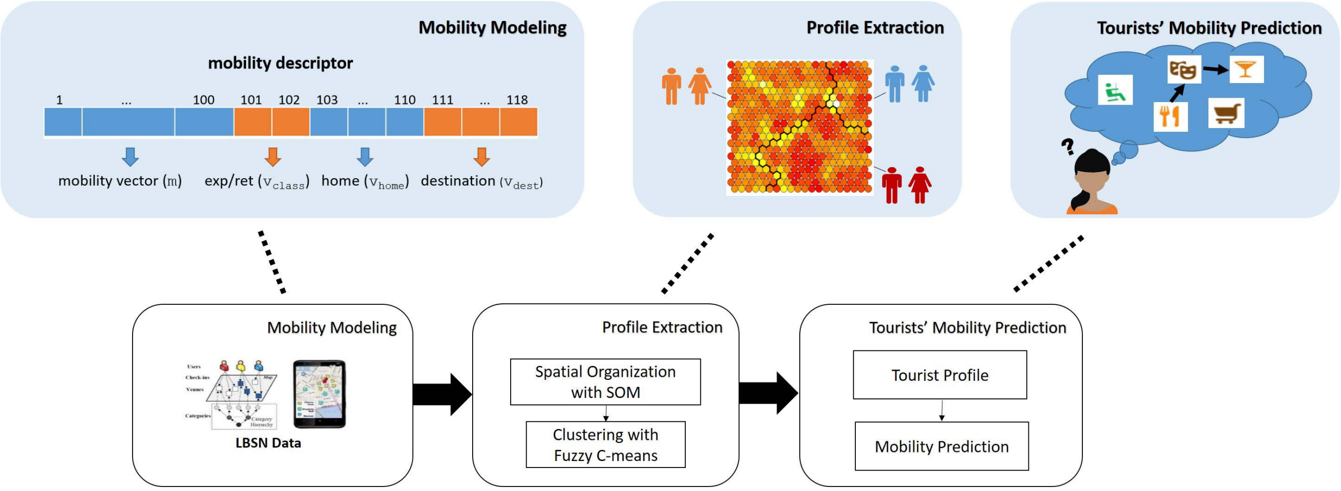

**FIGURE 1. Overview of PredicTour.**

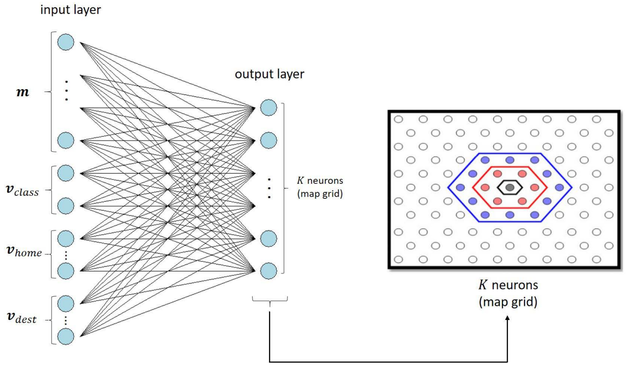

**FIGURE 2. Left side: neural network architecture with the mobility descriptor components ($\mathbf{m}$, $\mathbf{v}_{class}$, $\mathbf{v}_{home}$ and $\mathbf{v}_{dest}$) presented at the input layer and $K$ neurons at the output layer. Right side: example of $K$ neurons structured in a map grid, with the winner neuron and its neighbors $\mathcal{N}$ (different hexagonal neighborhood sizes in red and blue).**

(i) spatial organization of mobility descriptors using a Self-Organizing Map (SOM) [39]; (ii) clustering of outputs provided by SOM into different profiles with the Fuzzy C-means (FCM) algorithm [40]. SOM has been chosen due to its unsupervised learning capability of discovering intrinsic relationships among data and spatially organizing it. FCM generates a membership degree for each input pattern to its corresponding cluster which is further limited by crisp boundaries. The combination of SOM+FCM allows one to use the spatial organization provided by SOM for clustering data in different profiles with membership degrees indicating "how much" a user belongs to a cluster. This information is essential for dealing with cases in a gray zone, i.e., close to cluster boundaries, by providing more robust information.

The mobility descriptor is presented at the SOM's input layer as shown in the architecture map of Figure 2.

The output layer is a map grid of $K$ neurons. A neighborhood $\mathcal{N}$ considered in the map grid defines how weights of each output neuron are updated during the training phase. The neighborhood structures can assume different shapes as the hexagon of Figure 2. The algorithm for building the profile extraction is shown in Appendix B.

### C. TOURISTS' MOBILITY PREDICTION

The last block of PredicTour performs two tasks: (i) identification of the profile most associated with a target tourist; and (ii) prediction of the mobility pattern for unvisited countries based on the tourist's profile. The prediction process considers particular and collective behaviors (or signatures) of users visiting different countries:

- $\bar{\mathbf{m}}_t$ (tourist signature): it is computed either as the average of mobility vectors of tourist $t$ visiting destination countries or the average of mobility vectors of other tourists from the same origin and destination of $t$. Therefore, computing the tourist signature $\bar{\mathbf{m}}_t$ requires the analysis of two different situations: (i) if a tourist has visited other destinations, the calculation assumes previous tourist's visits; (ii) if tourists do not have history, it considers visits of other tourists with the same origin and destination.
- $\mathbf{c}_p$ (profile signature): it is the weighted average of mobility vectors of users in the same profile $p$ of tourist $t$ (weights are membership degrees). The profile signature $\mathbf{c}_p$ represents the collective behavior of users in the same profile.

The prediction of the mobility vector $\hat{\mathbf{m}}_t^{(dest)}$ of a target tourist $t$ in a new destination $dest$ is accomplished by aggregating the tourist and profile signatures according to (1).

$$\hat{\mathbf{m}}_t^{(dest)} = \frac{e^{\log(L_p+1)}\bar{\mathbf{m}}_t + \mathbf{c}_p}{2} \tag{1}$$

In this case, the tourist signature is more relevant when $t$ has non-zero $L_p$ previous visits to other countries. Otherwise, both signatures have the same weight. The algorithm for building the prediction is shown in Appendix C.

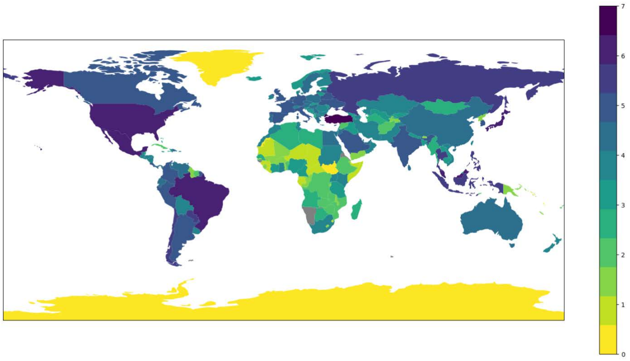

**FIGURE 3.** Number of check-ins per country in a base-10 log scale. Dark blue representing countries with the highest number.

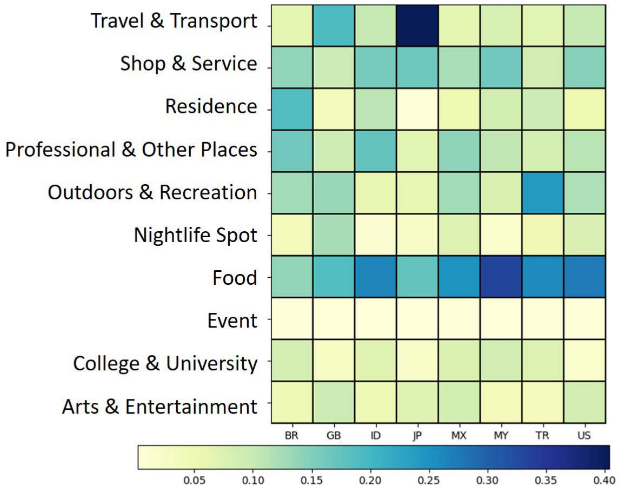

**FIGURE 4.** Number of check-ins normalized per country and category. Dark blue representing the pair *country* x *category* with the highest value (normalized by country).

## IV. DATASET AND PredicTour SETUP

This section describes the dataset and setup considered in the experiments conducted to address the questions raised in Section I.

### A. MODELING MOBILITY WITH FOURSQUARE CHECK-INS

Here we present the characteristics of the addressed dataset and how data is used to provide the mobility descriptor of any tourist in the dataset.

#### 1) DATASET

In the experiments, we explore the same Foursquare dataset of [12] and [41]. Foursquare is a location-based social network where users can use a mobile app to perform a check-in, which is the act of disclosing their current location to friends.

Each venue in Foursquare has a category with subcategories according to a hierarchical taxonomy provided by Foursquare. For instance, a given venue may have *Food* as category with *Burger Place* being its subcategory in the hierarchy. A complete list of venue categories and subcategories in Foursquare is available on the corresponding website.[2]

As check-ins from Foursquare are not public by default, useful data are compiled from Twitter. The dataset addressed in this paper represents around 15 million tweets containing publicly available URLs of check-ins from Foursquare. Each check-in in the raw dataset is composed of seven fields: user ID (*iduser*); date and time of the check-in (*date*); latitude (*latitude*); longitude (*longitude*); venue ID (*idvenue*); venue category (*categorievenue*); and venue subcategory (*subcategorievenue*).

There is no specific information on the city and country of each post in the raw dataset. This information can be retrieved using standard reverse geolocation procedures. In the present paper, we focus on popular countries of different continents. Figure 3 shows the number of check-ins worldwide on a base-10 log scale.

The eight selected countries of this study are very popular in Foursquare: Brazil (BR), Great Britain (GB), Indonesia (ID), Japan (JP), Mexico (MX), Malaysia (MY), Turkey (TR), and the United States (US).

Figure 4 shows the distribution of check-ins for the main categories of the eight most popular countries. We can observe similarities and differences between the behaviors of users in different countries. For example, *Residence* category is vastly shared by Brazilians, but the Japanese almost avoid it, preferring *Travel & Transport* category to perform a check-in. We can also observe that *Food* is a very popular category to share check-ins for all addressed countries.

#### 2) MOBILITY DESCRIPTORS

We explore transitions between Foursquare's venue categories of the set $\mathbb{V}$ = {<Travel & Transport>, <Shop & Service>, <Residence>, <Professional & Other Places>, <Outdoors & Recreation>, <Nightlife Spot>, <Food>, <Event>, <College & University>, <Arts & Entertainment>} with $V = 10$ categories. If a user makes consecutive *check-ins* at venues $v_i$ and $v_j$, respectively, within no more than two periods of the day, a transition links $v_i$ to $v_j$ (see Appendix A for details). We use the same division proposed by Veiga *et al.* [42] to define a period of the day: (i) morning between 6:00 am and 9:59 am; (ii) noon between 10:00 am and 2:59 pm; (iii) afternoon between 3:00 pm and 6:59 pm; (iv) night between 7:00 pm and 11:59 pm; and (v) dawn between 00:00 am and 05:59 am. In the performed experiments, we consider the set of locations $\mathbb{L}$ = {BR, GB, ID, JP, MX, MY, TR, US} encompassing $L = 8$ countries with the highest numbers of check-ins.

In the experiments, the dataset is filtered by considering only users with a minimum of two check-ins and one visiting country other than their home country (see Appendix A for this classification). A minimum of two check-ins guarantees

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

at least one transition between a pair of categories, and one visiting country guides the analysis to the international flow of tourists, which is the main interest of this work. The filtered dataset is then composed of 974 mobility descriptors. Each one has dimension $\dim(\mathbf{d}) = \dim(\mathbf{m}) + \dim(\mathbf{v}_{class}) + \dim(\mathbf{v}_{home}) + \dim(\mathbf{v}_{dest}) = 10^2+2+8+8 = 118$ components. A mobility descriptor of 118 components is computed for each LBSN user visiting a specific country. The number of check-ins among venue categories is normalized by user, avoiding scale problems when comparing users with different numbers of check-ins.

### B. DISCOVERING TOURIST PROFILES

Aiming to discover tourists' profiles, we adopt a self-organizing map to obtain intrinsic relationships among the whole set $\{\mathbf{d}\}$ of $|\{\mathbf{d}\}| = 974$ mobility descriptors, and to spatially organize such relationships, as mentioned in Section III. The SOM's architecture is defined by 118 input neurons, each one associated with an component of $\mathbf{d}$ presented at the input layer. The SOM's input encompasses features like: normalized mobility transitions $t_{ij}$ between each pair of venue categories $(v_i, v_j)$, $i, j = 1, \ldots, 10$, returner/explorer classification (two binary inputs with one-hot encoding) and the origin and destination countries (eight binary inputs with one hot encoding).

A 2-dimensional array of $K = \left\lfloor \sqrt{5\sqrt{|\{\mathbf{d}\}|}} \right\rfloor^2 = 144$ units (neurons), organized in a hexagonal topology, forms the SOM's output (see Figure 2 for more details on SOM's architecture and hexagonal topology). This setup provides a suitable distribution of samples among the output neurons, more than those proposed by other works [43].

Then, FCM groups the output neurons into $C = 3$ clusters (or profiles) defined experimentally. The membership degree $\mu_{up}$ establishes how much the mobility descriptor of user $u$ belongs to profile $p$. It can be characterized by the average membership degree $\bar{\mu}_p = \frac{1}{N_p} \sum_u \mu_{up}$ of $N_p$ mobility descriptors of cluster $p$.

### C. PREDICTING MOBILITY PATTERNS FOR DIFFERENT DIFFICULT LEVELS

To assess the PredicTour performance when predicting the mobility pattern of users with and without previous information, we separate the set of users into two groups: (i) set $\mathbf{u_h}$ of users with history i.e., users that have visited more than one destination country (2 to 5 in our case); (ii) set $\mathbf{u}_\emptyset$ of users with empty history, i.e., users that have visited only one country (which is the target destination when they form the test set). In the addressed dataset, $|\mathbf{u_h}| = 36$ users generate a set $\{\mathbf{d}\}_h$ of $|\{\mathbf{d}\}_h| = 83$ mobility descriptors, and $\mathbf{u}_\emptyset$ generates a set $\{\mathbf{d}\}_\emptyset$ of $|\{\mathbf{d}\}_\emptyset|$ mobility descriptors. The dataset used for training and testing sets has size $|\{\mathbf{d}\}_h| + |\{\mathbf{d}\}_\emptyset|$.

According to Table 1, we can evaluate PredicTour considering several difficult levels from 1 to 20, each one with a different dataset size. For instance, the difficulty level 2 is set with $|\{\mathbf{d}\}_\emptyset| = \lfloor 10\% \cdot |\{\mathbf{d}\}_h| \rfloor = 8$ mobility descriptors. Therefore, the dataset size is $83 + 8 = 91$.

**TABLE 1.** Difficulty level of experiments and dataset partitions.

| Difficulty level | $\lvert\{\mathbf{d}\}_\emptyset\rvert$ | Size of Tr+Ts Dataset | Total repetitions ($R$) |
|---|---|---|---|
| 1 | 0 | 83 | 1 |
| 2 | 8 (10%) | 91 | 111 |
| 3 | 16 (20%) | 99 | 56 |
| 4 | 25 (30%) | 108 | 36 |
| 5 | 33 (40%) | 116 | 27 |
| 6 | 42 (50%) | 125 | 21 |
| 7 | 50 (60%) | 133 | 18 |
| 8 | 58 (70%) | 141 | 15 |
| 9 | 66 (80%) | 149 | 14 |
| 10 | 75 (90%) | 158 | 12 |
| 11 | 83 (100%) | 166 | 11 |
| 12 | 166 (200%) | 249 | 5 |
| 13 | 249 (300%) | 332 | 4 |
| 14 | 332 (400%) | 415 | 3 |
| 15 | 415 (500%) | 498 | 2 |
| 16 | 498 (600%) | 581 | 2 |
| 17 | 581 (700%) | 664 | 2 |
| 18 | 664 (800%) | 747 | 1 |
| 19 | 747 (900%) | 830 | 1 |
| 20 | 830 (1000%) | 913 | 1 |

Depending on the difficulty level and for $|\{\mathbf{d}\}_\emptyset| \neq 0$, there are $R$ repetitions (alternatives) in Table 1 for choosing elements of $\{\mathbf{d}\}_\emptyset$ among the whole set $\{\mathbf{d}\}$ according to (2).

$$R = \left\lfloor \frac{(|\{\mathbf{d}\}| - |\{\mathbf{d}\}_h|)}{|\{\mathbf{d}\}_\emptyset|} = \frac{974 - 83}{|\{\mathbf{d}\}_\emptyset|} = \frac{891}{|\{\mathbf{d}\}_\emptyset|} \right\rfloor \tag{2}$$

For each repetition $rep$ out from $R$, we perform the well-known 10-fold cross-validation method to evaluate the PredicTour (or any comparison approach) performance for different training and test sets. The dataset $\{\mathbf{d}\}_h \bigcup \{\mathbf{d}\}_\emptyset$ is partitioned into one fold for test $\{\mathbf{d}\}^{ts}$ and nine folds using the remaining mobility descriptors for training $\{\mathbf{d}\}^{tr}$. This strategy results in a pair $(\{\mathbf{d}\}^{tr}, \{\mathbf{d}\}^{ts})_{it,rep}$ which is repeated for iterations (folds) $it = 1, \ldots, 10$, and repetitions $rep = 1, \ldots, R$.

## V. COMPARISON APPROACHES

The comparison of approaches is an additional challenge as we did not find a proposal like ours in the literature. We compare PredicTour with five different standard baseline approaches. The baselines range from simple statistics taken from destination information to more complex machine learning-based models (Deep AutoEncoder, Multi-Layer Perceptron, and Collaborative Filtering). Trivial statistics of users with same destination are considered for Baselines 1 (random choice) and 2 (average behavior). The recommender system (Baseline 3) is a natural choice for predicting preferences of users visiting new countries. Additionally, we consider two machine learning techniques (Baselines 4 and 5) successfully applied to regression/prediction problems.

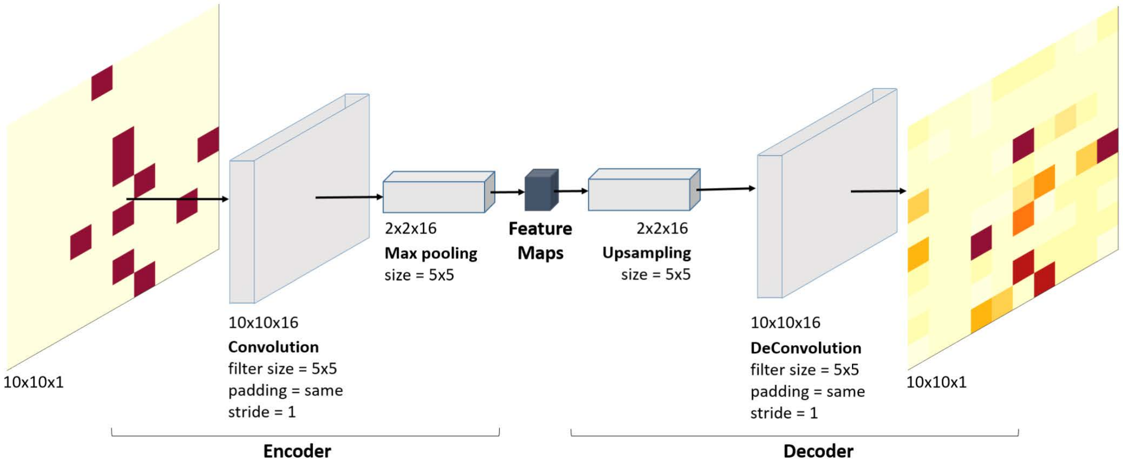

**FIGURE 5.** **Architecture of the baseline 5 (Deep AutoEncoder).**

*Baseline 1 (Random):* This first baseline simulates a simple case where the prediction $\hat{\mathbf{m}}_t^{(dest)}$ is performed by a random choice among users' mobility vectors. As the prediction of the tourist's behavior should occur in a new destination, it considers a randomly chosen user with the same destination country, simulating a more likely situation. Then, $\hat{\mathbf{m}}_t^{(dest)} = random_u\{\mathbf{m}_1^{(dest)}, ., \mathbf{m}_{U_{dest}}^{(dest)}\}$, $U_{dest}$ is the total of users in the training dataset with same destination of tourist $t$.

*Baseline 2 (Mean):* This approach considers the average mobility vectors of users visiting the same destination country. In other words, instead of taking a random mobility vector of a user with the same destination of $t$, it calculates the average of mobility vectors of those users. Therefore, $\hat{\mathbf{m}}_t^{(dest)} = \sum_{u=1}^{U_{dest}} \mathbf{m}_u^{(dest)}/U_{dest}$, with $U_{dest}$ defined as in Baseline 1.

*Baseline 3 (Collaborative Filtering):* This baseline adapts the k-Nearest Neighbors, a standard collaborative filtering method. The original algorithm finds the most similar (nearest neighbors) users to the target tourist $t$ and uses their behavior to predict $t$'s behavior. However, as the target tourist $t$ might have no historical information, this approach calculates a mobility vector based on origin and destination information. Therefore, $\hat{\mathbf{m}}_t^{(dest)} = \sum_{u=1}^{U_{nn}} \mathbf{m}_u/U_{nn}$, where $U_{nn}$ is the total of estimated neighbors in the training data, i.e., mobility descriptors with the most similar (based on the cosine similarity) origin and destination of the target tourist $t$.

*Baseline 4 (MLP):* This baseline is introduced to allow a comparison of PredicTour with one of the most well-known machine learning techniques. It encompasses a multi-layer perceptron (MLP) composed of 100 neurons in the input layer, one hidden layer with 50 neurons, and 100 neurons in the output layer. By using MLP as an auto-regression model, the pairs $(\mathbf{m}_u^{(dest)}, \mathbf{m}_u^{(dest)})$, for $u = 1, \ldots |\{\mathbf{d}\}^{tr}|$ are used as a pair (input, target_output) in the training phase. For the testing phase, the input of MLP is the same used by PredicTour, i.e., the tourist signature $\bar{\mathbf{m}}_t$ considered in (1). The testing output is then $\hat{\mathbf{m}}_t^{(dest)} = \text{MLP}\,(\bar{\mathbf{m}}_t)$. It is important to point out that we also trained MLP considering the pairs $(\bar{\mathbf{m}}_u, \mathbf{m}_u^{(dest)})$ i.e., using the same method adopted in the test phase to calculate the inputs. However, as the results became worse, we decided to disregard this method and maintain $\mathbf{m}_u^{(dest)}$ as the input in the training phase.

*Baseline 5 (DeepAC):* In this last approach, a deep auto-encoder addresses the problem considering the original representation of mobility information: the transition matrix (see Appendix A for details). This baseline provides an estimate computed with a more complex and recent technique of machine learning. As depicted in Figure 5, instead of using mobility vectors like the other methods, it considers transition matrices: as input, it receives tourist signature in a matrix form, and as output, it produces the matrix representing the predicted transition, which can be further submitted to a linearization process to provide the predicted mobility vector $\hat{\mathbf{m}}_t^{(dest)}$. Notice in Figure 5 that the model's architecture is composed of convolutional, max pooling, upsampling, and deconvolution layers.

### A. METRICS FOR PERFORMANCE EVALUATION

The mobility vector $\hat{\mathbf{m}}_t^{(dest)}$ estimated by each approach for every user in the test set can be compared with the actual $\mathbf{m}_t^{(dest)}$ from $\mathbf{d}_t^{(dest)} \in \{\mathbf{d}\}^{ts}$ to compute the performance $\mathcal{P}$ in terms of the Root Mean Square Error (RMSE) as in (3).

Assuming that the test set has cardinality $S = |\{\mathbf{d}\}^{ts}|$, the performance is measured for each approach $appr$ at iteration $it$ and repetition $rep$:

$$\mathcal{P}(appr, it, rep) = \sqrt{\frac{1}{S}\sum_{t=1}^{S}\left\|\mathbf{m}_t^{(dest)} - \hat{\mathbf{m}}_t^{(dest)}\right\|^2} \tag{3}$$

A total of 10 iterations (folds) and $R$ repetitions are made. The overall performance is then calculated considering the average over all iterations and repetitions:

$$\mathcal{P}(appr) = \frac{1}{R}\sum_{rep=1}^{R}\left(\frac{1}{10}\sum_{it=1}^{10}P(appr, it, rep)\right) \tag{4}$$

Aiming to avoid biased results, particularly for Baseline 1, the prediction error $\left\|\mathbf{m}_t^{(dest)} - \hat{\mathbf{m}}_t^{(dest)}\right\|^2$ in (3) is calculated according to (5).

$$\left\|\mathbf{m}_t^{(dest)} - \hat{\mathbf{m}}_t^{(dest)}\right\|^2 = \frac{1}{100}\sum_{rand=1}^{100}\left\|\mathbf{m}_t^{(dest)} - \mathbf{m}_{rand}\right\|^2 \tag{5}$$

We also perform an additional evaluation of the prediction results using normalized discounted cumulative gain (nDCG). It is a technique proposed by Wang *et al.* [44] to measure ranking quality. The goal here is to evaluate the results incorporating the idea of ranking relevance for each mobility vector of every target tourist $t$ in the test set. Assuming that we have the actual mobility vector $\mathbf{m}$ and the predicted mobility vector $\hat{\mathbf{m}}$, we sort $top \in \{1, \ldots, |\mathbf{m}|\}$ features of $\mathbf{m}$ in a decreasing way. Then, we attribute levels of relevance $rel_i(\mathbf{m})$ to each feature $i = 1, \ldots, top$, reinforcing that high relevance features represent more active transitions. For example, assuming $top = 10$, we would have for a linear scale: $rel_1(\mathbf{m}) = 10$, and $rel_{10}(\mathbf{m}) = 1$. Next, we find the respective $rel_i(\hat{\mathbf{m}})$ for the same $top$ features found in $\mathbf{m}$.

Based on this ranking applied to a tourist $t$ at destination $d$, we calculate $nDCG_{top}(t)$ by Equations (6) to (8).

$$nDCG(t)_{top} = \frac{DCG(t)_{top}}{IDCG(t)_{top}} \tag{6}$$

$$DCG(t)_{top} = \sum_{i=1}^{top}\frac{rel_i(\hat{\mathbf{m}}_t^{(dest)})}{\log_2(i+1)} \tag{7}$$

$$IDCG(t)_{top} = \sum_{i=1}^{top}\frac{rel_i(\mathbf{m}_t^{(dest)})}{\log_2(i+1)} \tag{8}$$

The overall performance of an approach is then averaged over $S$ tourists of the test set:

$$\overline{nDCG}(appr) = \frac{1}{S}\sum_{t=1}^{S}nDCG(t)_{top}.$$

## VI. RESULTS

Here, results are presented and discussed according to the three research questions stated in the beginning.

### A. WHAT KIND OF INTRINSIC RELATIONSHIPS CAN BE OBSERVED WHEN WE GROUP USERS?

As discussed in Sections III-B and IV-B, an important result of SOM is the neighborhood map generated from the set of weights of neurons that compose the output layer. Figure 6 presents the neighborhood map generated by the proposed model applied to the whole set $\{\mathbf{d}\}$. Red color represents close neurons, i.e., neurons whose neighbor neurons' weights are similar, whereas light yellow colored neurons represent high distances from their neighbors.

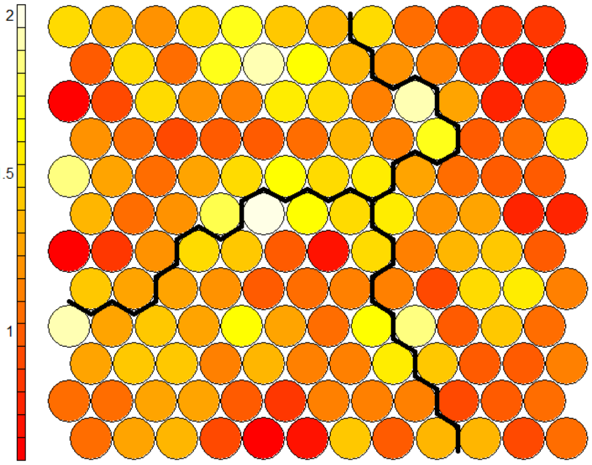

**FIGURE 6.** Resulting SOM from the whole set $\{\mathbf{d}\}$ as input.

After the fuzzy clustering performed by FCM, we obtain a fuzzy partition of the grid map. Crisp partitions can also be obtained when considering the maximum membership degree for each element in all clusters. The black line of Figure 6 illustrates a crisp partition of neurons into three clusters. From the mapping $f_{SOM+FCM} : \mathbf{d}_u^{(dest)} \mapsto \mu_{up}$, data mapped into regions located near the boundaries have a lower membership degree $\mu_{up}$. High membership degrees occur on data mapped into the cluster centers. The average membership degrees $\bar{\mu}_p$ (with respective confidence interval as subscript) are $\{0.673_{0.012}, 0.675_{0.007}, 0.655_{0.007}\}$ of $N_p = \{280, 378, 316\}$ mobility descriptors associated to each profile $p = \{1, 2, 3\}$, respectively. Notice that all profiles present similar average membership degrees, but profiles 2 and 3 encompass most users. High values of $\bar{\mu}_p$ indicates that data are properly clustered. Other algorithms such as K-means and hierarchical clustering were tested without good clustering indicators. The same was observed for a number of clusters other than three.

### B. WHAT KIND OF PATTERN CAN BE OBSERVED IN EACH PROFILE?

Clustering mobility descriptors of Foursquare users in different profiles (see Appendix B for more details) allows us to characterize each group according to specific patterns. Here we highlight some of the most distinct characteristics to better understand each profile. Considering only users with membership degrees higher than 0.75, we investigate separately the following information: $\mathbf{m}$ (user's mobility vector), $\mathbf{v}_{home}$ (user's origin information), $\mathbf{v}_{dest}$ (user's destination information), and $\mathbf{v}_{class}$ (user's returner/explorer information). Moreover, we focus the analysis on transitions with high values because they represent more relevant users' preferences. The number of transitions is normalized in [0,1] in each profile to provide a fair comparison between profiles.

Figure 7 shows important transition preferences of each profile regarding the 100 elements of $\mathbf{m}$, i.e., only transitions with values higher than the profile's average are depicted.

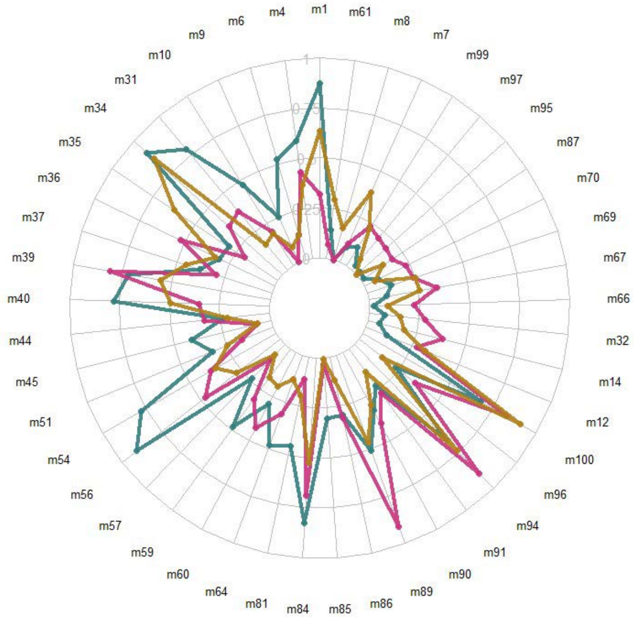

**FIGURE 7.** Key characteristics (signature) of profile 1 (yellow), 2 (blue), and 3 (pink) taken from the mobility vector (m).

**TABLE 2.** Top 3 transitions by profile, with category$^2$ meaning transitions between the same category, e.g. TT$^2$ means Trav/Trans → Trav/Trans.

| | Popularity | | |
|---|---|---|---|
| Profile | First | Second | Third |
| 1 | TT$^2$ | Food$^2$ | TT → Food |
| 2 | Food$^2$ | Out/Recr$^2$ | Arts/Entert$^2$ |
| 3 | Shop/Serv$^2$ | TT → Food | TT$^2$ |
| Global | Food$^2$ | TT$^2$ | Shop/Serv$^2$ |

Different line colors represent each profile: yellow for profile 1, blue for profile 2, and pink for profile 3. It is possible to note that, in terms of mobility, profiles have key characteristics or signatures. For instance, whereas transition m56 is high for users of Profile 2, it is low for users of Profile 1. On the other hand, transition m89 is preferred in Profile 3, and it is not as popular in Profiles 1 and 2. Profile 1 shares some characteristics with Profile 2 and others with Profile 3. Therefore, the second block of PredicTour could extract some particular mobility patterns from each profile.

Table 2 presents the top three category transitions (those with the highest number of transitions between the subcategories) for each profile, with TT representing Trav/Trans category. The last line of this table describes the preferences of transitions among the whole set of users (global), i.e., without considering users' profiles. By observing the top three characteristics, we notice that most transitions, particularly all the Top 1, are between the same category - represented by category$^2$. Moreover, according to Top 1, each profile could be taken as representative of one preferential transition. We can also see that most preferences in each profile are different compared to global choices, especially in Profile 3, where distinct patterns are observed for Top 1 and Top 2.

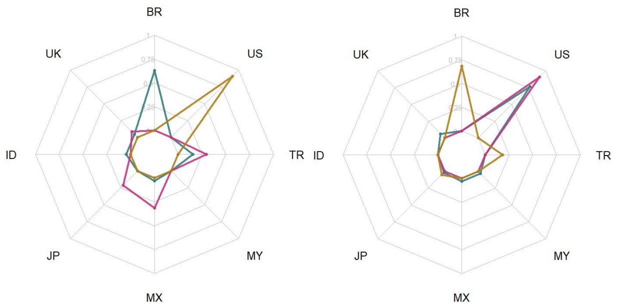

**FIGURE 8.** (a) Origin countries by profile. (b) Destination countries by profile.

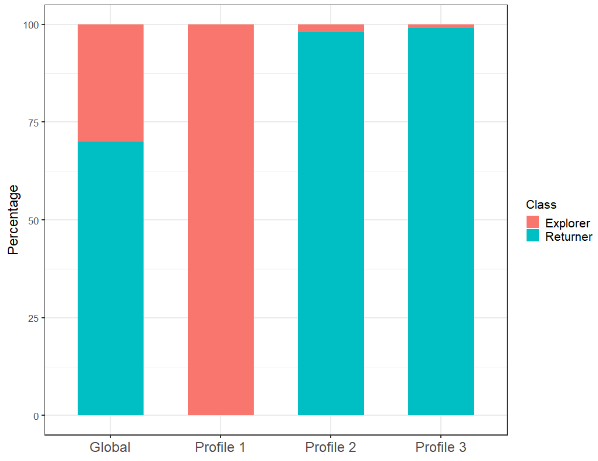

**FIGURE 9.** Classification of users (returner or explorer).

Considering $\mathbf{v}_{home}$ and $\mathbf{v}_{dest}$ to characterize users of each profile, Figure 8 highlights the user's origin and destination countries, respectively. According to Figure 8(a), regarding the origin country of users, we can see that most users of profile 1 (yellow) are from the United States. Most Brazilian users are in profile 2 (blue). Profile 3 (pink) is mostly composed of people from Mexico, Turkey, and Japan. We can see from Figure 8(b) that the United States is the favorite destination country among all tourists except for users classified as profile 1 (yellow) because it is composed by many American users and we only consider international tourism, i.e., data from users visiting countries different from their origin.

By considering the *returner* and *explorer* characteristics ($\mathbf{v}_{class}$), the results depicted in Figure 9 show that the proposed approach distinguishes profile 1 from the other two.

We can see that profiles 2 and 3 have *returner* users mostly, with more than 98% and 99%, respectively. *Explorer* users are mainly included in profile 1 (100%). Figure 9 also shows that considering all data (without profile classification) there are 70% of *returners* users and 30% of *explorers*. Therefore, we reinforce that PredicTour can appropriately split users into particular mobility patterns.

The results presented in this section show that discovering tourists' profiles (task performed by block 2 of PredicTour) highlights particular mobility patterns of users by transition

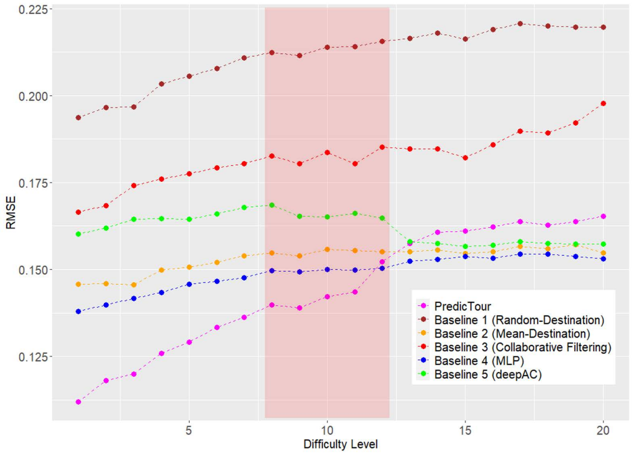

**FIGURE 10.** RMSE of PredicTour and baselines for different difficulty levels.

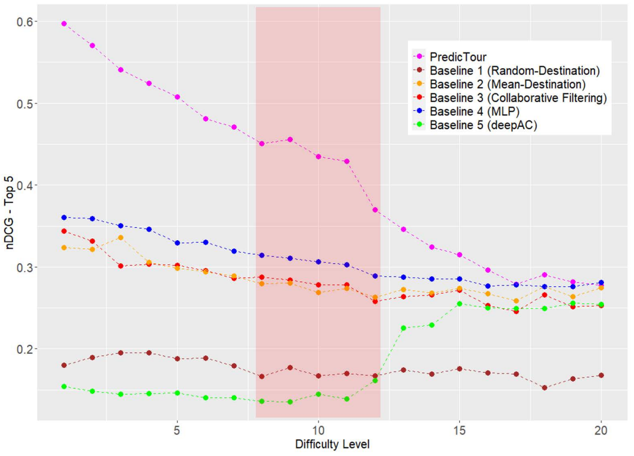

**FIGURE 11.** Average nDCG-top 5 of PredicTour and baselines for different difficulty levels.

preferences, country, and explorer or returner behaviors grouped into profiles 1, 2, and 3.

### C. HOW DOES PredicTour PERFORM WHEN COMPARED WITH BASELINES ON DIFFERENT DIFFICULT LEVELS?

To compare the results of block 3 of PredicTour, which uses profile information to predict mobility patterns, we consider five baseline approaches (see Section IV-C). The overall RMSE performance is shown in Figure 10 for different difficulty levels. It compares the performance of PredicTour with all baselines, which consider as input an approximated mobility pattern of a tourist $t$ and as output: Baseline 1) the mobility vector of a randomly chosen tourist with the same destination; Baseline 2) an average of mobility vectors among tourists with the same destination; Baseline 3) the output vector provided by collaborative filtering; Baseline 4) the output vector provided by a multi-layer perceptron; Baseline 5) the linearization of the matrix provided by a deep autoencoder applied to the input transition matrix.

The results of RMSE show that PredicTour produces smaller errors for predicting tourists' mobility when they visit different countries, especially when the number of tourists with historical information is representative. This characteristic can be better analyzed if we consider three different scenarios when observing the red region of Figure 10:

- **optimistic**: the left side of the red region, where the majority of users have historical information and the performance of PredicTour is far better than the others;
- **pessimistic**: the right side of the red region, where the majority of users do not have historical information, PredicTour performance is similar to the other approaches;
- **realistic**: the red region, where the proportion between non-historical and historical information is balanced, PredicTour still outperforms the remaining approaches.

Aiming to understand if the prediction errors are occurring for more relevant components of the mobility vectors, we consider the nDCG metric as in (6). The average nDCG over the whole test set is presented in Figures 11 and 12 for top 5 and top 10 more relevant components, respectively.

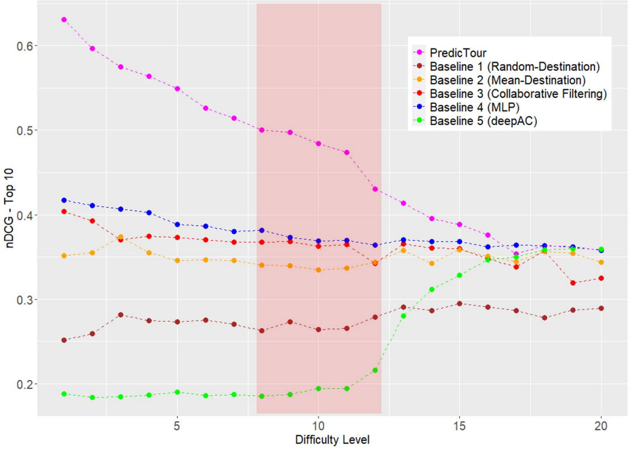

**FIGURE 12.** Average nDCG-top 10 of PredicTour and baselines for different difficulty levels.

It is possible to see that the results are significantly better for PredicTour in almost all cases. Therefore, the general prediction of tourist' behavior is satisfactory if we consider more important components (top 5 and top 10) that should be correctly predicted in practice.

Another important observation present in Figures 10, 11 and 12 regards the evolution of PredicTour in different scenarios. Even in more challenging scenarios, PredicTour can effectively predict users' preferences. We can thus affirm that prediction with profile information accomplished by PredicTour obtains better results than those obtained by approaches that do not apply user profiling (Baselines 1 to 5). We notice that the performance of deep-AC increases as the training process accesses more information, even when the percentage of users without history increases. This is a characteristic that can be further evaluated in another work, being out of the scope of this present study.

## VII. CONCLUSION

This study aimed to extract and explore patterns available on LBSN data to improve the understanding of users' behavior and use this information to predict international tourists' mobility patterns in different countries. In the present paper, the proposed approach, PredicTour, explored LBSN data to construct a *mobility descriptor* necessary to express non-trivial information regarding international tourists' mobility. PredicTour is then capable of predicting the mobility of tourists at unvisited countries based on the tourists' profiles extracted from Self Organizing Maps (SOMs) and Fuzzy C-means clustering.

We showed evidence that PredicTour can extract important characteristics of each user profile, which can be further explored in a tourism context. In this paper, we evaluated the performance of PredicTour in different scenarios and against relevant baselines. The results showed that our approach could achieve satisfactory performance, providing smaller RMSE than the addressed baselines, especially for non-pessimistic scenarios. Furthermore, if we focus on the performance for the top 5 and top 10 features, the most important ones, PredicTour outperformed the baselines in virtually all test cases, except on the extreme ones expected to occur less in practice.

Our approach can be helpful in different kinds of applications for tourists. For instance, in the construction of specialized place recommendation systems, the suggestion of attractive services and products, and improvement of transport and attractions strategies. The map of intrinsic relationships provided by SOM could also be useful to show (in a visual tool for tourism planning, for example) tourists with similar behaviors, which could be grouped into the same activities. Although this paper has focused on international tourism, we believe that PredicTour could also provide results for intern tourism with slight adaptations in home and destination locations.

This study can be expanded in numerous ways. In future works, we intend to address larger datasets to evaluate the impact on the prediction performance of all the considered approaches. We can also pay more attention to outliers, i.e., tourists whose behavior is far from the behavior represented by the profile centroid. Another possible expansion is the idea of grouping users by geographic proximity, considering spatial distances of places. In the same way, the availability of places that are part of tourists' routine should also be studied in the future. These factors can influence their choices. For example, a tourist from Indonesia may have difficulty finding Indonesian cultural preferences in a western country due to religious or gastronomic differences. However, it is the opposite when visiting a country with a similar culture to the tourist's home.

## APPENDIX A
## MOBILITY MODELING

Algorithm 1 describes the main steps performed in the first block of PredicTour for building mobility descriptors.

**Algorithm 1** PredicTour Block1 Mobility Modeling

```
Inputs: time window for computing transitions among venue categories; set $\{\mathbf{x}\}$ of check-ins $\mathbf{x} = (ck_{id}, u_{id}, l, v_{cat}, date)$;
    // $ck_{id}$=check-in id, $u_{id}$=user id, $l$=country
    // $v_{cat}$=venue category, and $date$
for each user $u$ and country $dest$ visited by $u$ do
    Compute $\{\mathbf{x}\}_{u,dest} = \{\mathbf{x} \mid u_{id} = u, l = dest\}$;
    Calculate the mobility vector $\mathbf{m}$ according to (9);
    Calculate the binary vector $\mathbf{v}_{class}$ according to (10);
    Define the binary vector $\mathbf{v}_{home}$ with the one hot encoding of the origin country
    Define the binary vector $\mathbf{v}_{dest}$ with the one hot encoding of the visited country
    Concatenate $\mathbf{m}$, $\mathbf{v}_{class}$, $\mathbf{v}_{home}$,$\mathbf{v}_{dest}$ to get $\mathbf{d}_u^{(dest)}$;
end for
```

The algorithm receives a time window and check-ins data as inputs. The time window defines the period elapsed between two consecutive check-ins made by the same user for characterizing a transition between two venue categories. Check-ins data of a user contain information about the country and venue category where a check-in was made. A user is classified as a resident or tourist in a country depending on how much time is spent in each country. This procedure is key for setting $\mathbf{v}_{home}$ and $\mathbf{v}_{dest}$.

The time of stay is computed from check-in sequences performed in each country. For example, if a user did a check-in on May 5th and another on June 10th of the same year in Brazil, we assume the user stayed 35 days in Brazil (BR). Therefore, BR is considered the user's home country if the user stays more than 30 days (for instance) posting from BR and stays no longer in any other country. All remaining countries with check-ins other than BR are considered destinations $d$ of user $u$. Note that users are only considered tourists if we can spot a home country according to our approach. Previous studies in the literature have successfully applied similar strategies [13], [42], [45], and [46].

The mobility vector $\mathbf{m}$ is computed from the adjacency matrix $T_{V \times V}$ of a transition graph $G$ whose elements are $w_{ij}$. $G$ is a directed graph $G(\mathbb{V}, \mathbb{E})$ with a set $\mathbb{V}$ of $V$ venue categories, and a set $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ of edges or transitions. An edge $\mathbf{e}_{ij} = (v_i, v_j) \in \mathbb{E}$ with weight $w_{ij}$ represents the number of transitions made within a given time interval from venue $v_i$ to $v_j$. The row vectors $\mathbf{t}_i$ of the transition matrix $T$ are concatenated to generate $\mathbf{m}$ of dimension $V^2$ according to (9).

$$\mathbf{m} = (\mathbf{t}_1 | \mathbf{t}_2 | \dots | \mathbf{t}_V) \tag{9}$$

The classification $\mathbf{v}_{class}$ of users into returners or explorers is defined by how far a tourist goes around most visited places [42], [47]. It is based on the gyration radius information according to (10).

$$\mathbf{v}_{class} = \begin{cases} Returner & \textbf{if } r_g(M) > 0.5\ r_g(n_a) \\ Explorer & \textbf{otherwise} \end{cases} \tag{10}$$

The gyration radius $r_g(M)$ is computed by (11) over $M$ most visited venues of a user making $n_i$ visits to venue $i$ with coordinate $\mathbf{l}_i$, $i = 1, \ldots, M$.

$$r_g(M) = \sqrt{\frac{1}{\sum_i n_i} \sum_i n_i \cdot \mathrm{dist}(\mathbf{l}_i - \mathbf{l}_M)^2} \tag{11}$$

The gyration radius $r_g(n_a)$ is taken over all $n_a$ venues of same user. The term dist($\mathbf{l}_i - \mathbf{l}_M$) can be calculated by the *Haversine* distance metric, which measures the distance between venue $i$ and the center $\mathbf{l}_M$ given by (12).

$$\mathbf{l}_M = \frac{\sum_i n_i \mathbf{l}_i}{\sum_i n_i} \tag{12}$$

Finally, the vector $\mathbf{d}_u^{(dest)} = (\mathbf{m}|\mathbf{v}_{class}|\mathbf{v}_{home}|\mathbf{v}_{dest})$ describing a user $u$ visiting a particular country *dest* is a concatenation of the previous vectors.

## APPENDIX B
## PROFILE EXTRACTION

The main steps performed in the second block of PredicTour are shown in Algorithm 2.

It is divided into two different parts: the task performed by SOM and the task performed by FCM. Algorithm 2 starts with random initialization of all weights $\mathbf{w}_k$, $k = 1, \ldots K$. Then it enters the main loop by setting the training data and subsequently starts the inner loop. It randomly picks an input ($\mathbf{d}_u^{(dest)}$) from the current training dataset. Considering the strategy "winner takes all," it identifies the output neuron with the most similar weights to the picked input. As an attempt to provide region maps in the output grid, it updates not only the winner's weights but also those of its neighborhood $\mathcal{N}$. At each iteration, the process of updating weights is repeated for all inputs of the training dataset, which are chosen in random order without repetition. Then, the algorithm updates the neighborhood size (usually using a non-increasing update function) and starts the next iteration. This process repeats until it achieves the maximum number of 100 iterations.

After finishing the SOM task, FCM receives the resulting set of weights as input for the second task of Algorithm 2. It starts by randomly initializing membership degrees of all weights to $C$ clusters, with $C$ set as a parameter of FCM. The algorithm enters the main loop by computing the centroid $\mathbf{c}_j$ of cluster $j$ according to to (13).

$$\mathbf{c}_j = \frac{\sum_{k=1}^{K} \mu_{kj}^{z} \mathbf{w}_k}{\sum_{k=1}^{K} \mu_{kj}^{z}} \tag{13}$$

The parameter $z = 2$ controls how fuzzy the cluster will be. The process is repeated for all clusters. Subsequently, the membership degree $\mu_{kj}$ of neuron $k$ in cluster $j$ is updated according to (14).

**Algorithm 2** PredicTour Block2 Profile Extraction

```
// Task 1: SOM spatially organizes the input data
Inputs: training set ${\mathbf{d}}^{tr}$ of mobility descriptors;
Randomly initialize the weights $\mathbf{w}_k$ of SOM neurons;
while (SOM stop condition is FALSE) do
    $S = \{\mathbf{d}\}^{tr}$; // start with the whole training dataset
    while $S \neq \emptyset$ do
        Randomly pick (without reposition) one input pattern $\mathbf{d}_u^{(dest)} \in S$;
        Track the output neuron that produces the smallest distance $dist(\mathbf{d}_u^{(dest)}, \mathbf{w}_k)$ (the winner node);
        Update weight vectors of the winner's neighborhood $\mathcal{N}$ to approximate them to the input;
    end while
    Update size($\mathcal{N}$);
end while
// Task 2: FCM clusters the output map of SOM
Inputs: weights of SOM output neurons;
Randomly initialize the membership degrees;
while (FCM stop condition is FALSE) do
    for ($j = 1 : C$ number of clusters) do
        Compute the centroid of cluster $j$ as in (13);
    end for
    for ($k = 1 : K$ number of neurons) do
        for ($j = 1 : C$) do
            Update the membership degree of neuron $k$ to cluster $j$ according to (14);
        end for
    end for
end while
```

$$\mu_{kj} = \left( \sum_{c=1}^{C} \left( \frac{\|\mathbf{w}_k - \mathbf{c}_j\|}{\|\mathbf{w}_k - \mathbf{c}_c\|} \right)^{\frac{2}{z-1}} \right)^{-1} \tag{14}$$

The update of membership degree is performed for all $K$ neurons of the output map. The main loop repeats until the stop condition is achieved (the maximum number of iterations is 100 or when the difference between updated values of $J$ obtained by (15) in two consecutive iterations is arbitrarily small.

$$J = \sum_{k=1}^{K} \sum_{j=1}^{C} \mu_{kj}^{z} ||\mathbf{w}_k - \mathbf{c}_j||^2 \tag{15}$$

In summary, we assume a dataset $\{\mathbf{d}\}^{tr}$ of $N$ mobility descriptors $\mathbf{d}_u^{(dest)}$ $(u = 1, \ldots, N)$. Then, $f_{SOM+FCM}$ provides the mapping $\mathbf{d}_u^{(dest)} \mapsto \mathbf{w}_k \mapsto \mathbf{c}_j \mapsto \mu_{kj}$ which characterizes, by setting the membership degree $\mu_{kj}$, how fairly $\mathbf{d}_u^{(dest)}$ belongs to a profile with centroid $\mathbf{c}_j$. The complete description of the technical parameters is available in the GitHub repository.[3]

[3] https://github.com/helen-senefonte/PredicTour

## APPENDIX C
## TOURISTS' MOBILITY PREDICTION

The main steps accomplished in the last block of PredicTour are described in Algorithm 3.

**Algorithm 3** PredicTour Block3 Mobility Prediction

```
// Profile Identification
1: Obtain vectors $\mathbf{v}_{home}$, $\mathbf{v}_{dest}$;
2: Calculate the average $\bar{\mathbf{m}}_t$ of the queried/target tourist $t$ using (16);
3: Calculate $\mathbf{v}_{class}$ of the queried/target tourist $t$;
4: Concatenate all vectors in $\tilde{\mathbf{d}}_t = (\bar{\mathbf{m}}_t|\mathbf{v}_{class}|\mathbf{v}_{home}|\mathbf{v}_{dest})$;
5: Input $\tilde{\mathbf{d}}_t$ to SOM and obtain the winner output neuron;
6: Identify the profile $p$ of the winner neuron;
// Mobility Prediction
7: Calculate the profile signature $\mathbf{c}_p$ using (17);
8: Calculate $\hat{\mathbf{m}}_t^{(dest)}$ according to (1);
```

It starts by defining the origin and destination of the target tourist $t$. Then, it computes the remaining information of $t$, considering that:

- when $t$ has previous information, the algorithm calculates $\bar{\mathbf{m}}_t$ using (16) as the average of mobility patterns of $t$ in other previously visited locations. It also calculates the classification vector $\mathbf{v}_{class}$ (returner or explorer) as in Algorithm 1;
- when $t$ has no previous information, the algorithm computes $\bar{\mathbf{m}}_t$ using (16) based on the average mobility vectors of other tourists with same origin and destination of $t$, and the classification vector $\mathbf{v}_{class}$ is the average of classifications of those tourists.

Formally, the tourist signature $\bar{\mathbf{m}}_t$ is calculated by (16).

$$\bar{\mathbf{m}}_t = \begin{cases} \sum_{l=1}^{L_p} \mathbf{m}_t^{(l)}/L_p & \text{if } L_p \neq 0 \ (t \text{ has historical data}) \\ \sum_{u=1}^{U_t} \mathbf{m}_u/U_t & \text{otherwise} \end{cases} \tag{16}$$

where $\mathbf{m}_t^{(l)}$ is the mobility vector of tourist $t$ visiting country $l$ among $L_p$ countries previously visited by $t$, and $\mathbf{m}_u$ is the mobility vector of every user $u$ in $U_t$, the set of tourists with same origin and destination of $t$.

Further, the algorithm aggregates all information in the approximated vector $\tilde{\mathbf{d}}_t = (\bar{\mathbf{m}}_t|\mathbf{v}_{class}|\mathbf{v}_{home}|\mathbf{v}_{dest})$ of the target tourist $t$. Then it presents $\tilde{\mathbf{d}}_t$ at the input layer of a trained SOM which maps it into the output grid region. The output neuron with the most similar weights to the approximation $\tilde{\mathbf{d}}_t$ is activated as the winner neuron. The profile $p$ of the winner neuron is chosen as the profile of $t$, and the profile signature $\mathbf{c}_p$ is calculated by (17),

$$\mathbf{c}_p = \frac{\sum_{i=1}^{N_p} \mu_{ip}\mathbf{m}_i}{\sum_{i=1}^{N_p} \mu_{ip}} \tag{17}$$

with $N_p$ mobility vectors $\mathbf{d}_i$ clustered in the same profile $p$, $\mathbf{m}_i$ mobility vectors extracted from $\mathbf{d}_i$, and $\mu_{ip}$ defined as the membership degree of the neuron fired when $\mathbf{d}_i$ is presented at SOM input. Joining $\mathbf{c}_p$ calculated by (17), with $\bar{\mathbf{m}}_t$ calculated by (16) we obtain $\hat{\mathbf{m}}_t^{(dest)}$ according to (1).

## ACKNOWLEDGMENT

The authors thank David Veiga. Myriam Regattieri Delgado and Ricardo Lüders thank the Project Smart City Concepts in Curitiba. Thiago Helen Senefonte thanks the State University of Londrina (UEL).

## REFERENCES

[1] *International Tourism Highlights*, World Tourism Organization, Madrid, Spain, 2021.

[2] A. Lew and B. McKercher, "Modeling tourist movements: A local destination analysis," *Ann. Tour. Res.*, vol. 33, no. 2, pp. 403–423, 2006.

[3] D. A. Fennell, "A tourist space-time budget in the shetland islands," *Ann. Tourism Res.*, vol. 23, no. 4, pp. 811–829, 1996.

[4] Y. Zheng, L. Zhang, X. Xie, and W. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. WWW*, Madrid, Spain, 2009, pp. 791–800.

[5] H. Yoon, Y. Zheng, X. Xie, and W. Woo, "Smart itinerary recommendation based on user-generated GPS trajectories," in *Ubiquitous Intelligence and Computing. UIC* (Lecture Notes in Computer Science), vol. 6406, Z. Yu, R. Liscano, G. Chen, D. Zhang, and X. Zhou, Eds. Berlin, Germany: Springer, 2010, doi: 10.1007/978-3-642-16355-5_5.

[6] T. H. Silva, A. C. Viana, F. Benevenuto, L. Villas, J. Salles, A. Loureiro, and D. Quercia, "Urban computing leveraging location-based social network data: A survey," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 17:1–17:39, Feb. 2019.

[7] A. P. G. Ferreira, T. H. Silva, and A. A. F. Loureiro, "Uncovering spatiotemporal and semantic aspects of tourists mobility using social sensing," *Comput. Commun.*, vol. 160, pp. 240–252, Oct. 2020.

[8] C. Huang and D. Wang, "Unsupervised interesting places discovery in location-based social sensing," in *Proc. DCOSS*, Washington, DC, USA, 2016, pp. 67–74.

[9] D. Hristova, M. J. Williams, M. Musolesi, P. Panzarasa, and C. Mascolo, "Measuring urban social diversity using interconnected geo-social networks," in *Proc. WWW*, Montreal, QC, Canada, 2016, pp. 21–30.

[10] R. O. G. Gavilanes, Y. Mejova, and D. Quercia, "Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication," in *Proc. CSCW*, Baltimore, MD, USA, 2014, pp. 1511–1522.

[11] G. Le Falher, A. Gionis, and M. Mathioudakis, "Where is the soho of Rome measures and algorithms for finding similar neighborhoods in cities," in *Proc. ICWSM*, Oxford, U.K., 2015, pp. 1–5.

[12] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, M. Musolesi, and A. A. F. Loureiro, "A large-scale study of cultural differences using urban data about eating and drinking preferences," *Inf. Syst.*, vol. 72, pp. 95–116, Dec. 2017.

[13] H. C. M. Senefonte, G. Frizzo, M. R. B. S. Delgado, R. Lüders, D. Silver, and T. Silva, "Regional influences on tourists mobility through the lens of social sensing," in *Proc. SOCINFO*, Pisa, Italy, 2020, pp. 312–319.

[14] H. C. M. Senefonte, T. H. Silva, R. Lüders, and M. R. B. S. Delgado, "Classifying venue categories of unlabeled check-ins using mobility patterns," in *Proc. DCOSS*, Santorini, Greece, 2019, pp. 562–569.

[15] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Phys. Rep.*, vol. 734, pp. 1–74, Mar. 2018.

[16] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A. Barabási, "Returners and explorers dichotomy in human mobility," *Nature Commun.*, vol. 6, no. 1, pp. 1–8, 2015.

[17] A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi, "Disease containment strategies based on mobility and information dissemination," *Sci. Rep.*, vol. 5, no. 1, 2015, Art. no. 10650.

[18] F. Mourchid, A. Habbani, and M. El Koutbi, "Mining user patterns for location prediction in mobile social networks," in *Proc. CIST*, Tetouan, Morocco, 2014, pp. 213–218.

[19] M. Amoretti, L. Belli, and F. Zanichelli, "UTravel: Smart mobility with a novel user profiling and recommendation approach," *Pervasive Mobile Comput.*, vol. 38, pp. 474–489, Jul. 2017.

[20] L. Luceri, T. Braun, and S. Giordano, "Social influence (deep) learning for human behavior prediction," in *Proc. CompleNet*, Boston, MA, USA, 2018, pp. 261–269.

[21] K. C. Roy, M. Cebrian, and S. Hasan, "Quantifying human mobility resilience to extreme events using geo-located social media data," *EPJ Data Sci.*, vol. 8, p. 45, Oct. 2019.

[22] D. Rajashekar, A. N. Zincir-Heywood, and M. I. Heywood, "Smart phone user behaviour characterization based on autoencoders and self organizing maps," in *Proc. ICDMW*, Barcelona, Spain, 2016, pp. 319–326.

[23] M. C. C. A. R. González Hidalgo and A. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[24] C. Song, Z. Qu, N. Blumm, and A. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[25] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–465, Oct. 2006.

[26] E. B. Zion and B. Lerner, "Identifying and predicting social lifestyles in people's trajectories by neural networks," *EPJ Data Sci.*, vol. 7, p. 45, Dec. 2018.

[27] H. Hsieh, C. Li, and S. Lin, "Exploiting large-scale check-in data to recommend time-sensitive routes," in *Proc. UrbComp*, Beijing, China, 2012, pp. 55–62.

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

[29] Z. Wang, J. Juang, and W. Teng, "Predicting poi visits with a heterogeneous information network," in *Proc. TAAI*, Tainan, Taiwan, 2015, pp. 388–395.

[30] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "Revealing the city that we cannot see," *ACM Trans. Internet Technol.*, vol. 14, no. 4, pp. 1–23, Dec. 2014.

[31] M. De Domenico, A. Lima, and M. Musolesi, "Interdependence and predictability of human mobility and social interactions," *Pervasive Mobile Comput.*, vol. 9, no. 6, pp. 798–807, 2013.

[32] K. D-Silva, A. Noulas, M. Musolesi, and C. Mascolo, "Predicting the temporal activity patterns of new venues," *EPJ Data Sci.*, vol. 7, p. 45, Apr. 2018.

[33] H. Q. Vu, G. Li, and R. Law, "Cross-country analysis of tourist activities based on venue-referenced social media data," *J. Travel Res.*, vol. 59, no. 1, pp. 90–106, Feb. 2020.

[34] A. Y. Grinberger and N. Shoval, "Spatiotemporal contingencies in Tourists' intradiurnal mobility patterns," *J. Travel Res.*, vol. 58, no. 3, pp. 512–530, Mar. 2019.

[35] S. Lozano and E. Gutiérrez, "A complex network analysis of global tourism flows," *Int. J. Tourism Res.*, vol. 20, no. 5, pp. 588–604, Sep. 2018.

[36] A. P. G. Ferreira, T. H. Silva, and A. A. Ferreira Loureiro, "Beyond sights: Large scale study of tourists' behavior using foursquare data," in *Proc. ICDMW*, Atlantic City, NJ, USA, 2015, pp. 1117–1124.

[37] P. Yochum, L. Chang, T. Gu, and M. Zhu, "Linked open data in location-based recommendation system on tourism domain: A survey," *IEEE Access*, vol. 8, pp. 16409–16439, 2020.

[38] A. P. G. Ferreira, T. H. Silva, and A. A. F. Loureiro, "Uncovering spatiotemporal and semantic aspects of tourists mobility using social sensing," *Comput. Commun.*, vol. 160, p. 240– 252, 2020.

[39] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer, 2012.

[40] J. C. Bezdek, W. Full, and R. Ehrlich, "FCM: The fuzzy C-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.

[41] 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 Sci.*, vol. 6, no. 1, p. 5, Dec. 2017.

[42] D. A. M. Veiga, G. B. Frizzo, and T. H. Silva, "Cross-cultural study of tourists mobility using social media," in *Proc. WebMedia*, Rio de Janeiro, Brasil, 2019, pp. 313–316.

[43] J. Tian, M. Azarian, and M. Pecht, "Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm," in *Proc. PHME*, Nantes, France, 2014, pp. 1–5.

[44] Y. Wang, L. Wang, Y. Li, D. He, and T. Liu, "A theoretical analysis of NDCG type ranking measures," in *Proc. COLT*, Princeton, NJ, USA, 2013, pp. 25–54.

[45] S. Paldino, I. Bojic, S. Sobolevsky, C. Ratti, and M. C. González, "Urban magnetism through the lens of geo-tagged photography," *EPJ Data Sci.*, vol. 4, no. 1, pp. 1–17, 2015.

[46] A. P. G. Ferreira, T. H. Silva, and A. A. F. Loureiro, "Profiling the mobility of tourists exploring social sensing," in *Proc. DCOSS*, Santorini, Greece, 2019, pp. 522–529.

[47] D. Veiga, G. Frizzo, H. C. M. Senefonte, R. Lüders, M. R. B. S. Delgado, and T. Silva, "Influências regionais na mobilidade de turistas e residentes usando dados de mídia social," in *Proc. SBRC*, Rio de Janeiro, 2020, pp. 589–602.

**HELEN C. MATTOS SENEFONTE** received the M.S. degree in electronic and computer engineering from the Aeronautics Institute of Technology (ITA), Brazil, in 2009. She has been an Assistant Professor with the Computer Science Department, State University of Londrina (UEL), Brazil, since 2006. She has been on a doctoral program in electrical and computer engineering at the Federal University of Technology—Paraná (UTFPR), Brazil, since 2018. Her doctoral research englobes machine learning and social computing fields, and it concerns an evaluation, characterization, and prediction of users' behavior from LBSNs in the context of tourists' mobility patterns. In her master's research, she proposed the acceleration of reinforcement learning in multiple objective systems, and she received a fellowship from the São Paulo Research Foundation (FAPESP). Her current research interests include machine learning, natural language processing, and social computing.

**MYRIAM REGATTIERI DELGADO** received the graduate degree in electrical engineering from the Federal University of Goiás, in 1990, and the M.S. degree in electric engineering and the Ph.D. degree in computer engineering from the State University of Campinas, in 1993 and 2002, respectively. From 2015 to 2016, she was a Visiting Professor (postdoctoral stage) at the University of Kent, Canterbury, U.K. She also visited Heriot-Watt University, Edinburgh, U.K., in 2016 and the University of Lille, France, from 2015 to 2017. She is the Head of the group of research in parallel and natural computing (ComPaNat), developing algorithms inspired by Nature. She is currently a Full Professor of electrical and computer engineering with the Department of Informatics and Graduate Program, Federal University of Technology—Paraná, Brazil. Her current research interests include machine learning–deep learning, particularly on intelligent optimization.

**RICARDO LÜDERS** received the M.S. and Ph.D. degrees in electrical engineering from the State University of Campinas (Unicamp), in 1992 and 2001, respectively. His sabbatical leave was with the University of Michigan, USA, in 2009. He is a Full Professor at the Universidade Tecnológica Federal do Paraná (UTFPR), Brazil. His current research interests include operations research topics as metaheuristics for multiobjective optimization, and mathematical modeling of transportation systems and urban mobility using complex networks with applications to smart cities.

**THIAGO H. SILVA** received the M.Sc. and Ph.D. degrees in computer science from the Federal University of Minas Gerais, in 2009 and 2014, respectively. He was also a Postdoctoral Researcher in urban computing at the Federal University of Minas Gerais, from 2014 to 2015. From 2019 to 2021, he was a Visiting Professor at the University of Toronto. He was a Research Intern at Telecom Italia, Italy, in 2011, researching energy efficiency in smart homes. He was also a visiting student at the University of Birmingham, U.K., in 2012, and at INRIA, France, in 2013, studying large-scale urban-related problems with social media data. He is a Professor of computer science at the Universidade Tecnológica Federal do Paraná, Curitiba, Brazil, where his current research explores data mining to the study of urban societies. He has substantial experience in the industry and academia in urban computing, data mining/machine learning, and user behavior modeling, working on those topics, since 2005.