IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON SOCIAL COMPUTING APPLICATIONS FOR SMART CITIES

Received April 30, 2018, accepted May 26, 2018, date of publication June 25, 2018, date of current version July 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2850062

Regionalization of Social Interactions and Points-of-Interest Location Prediction With Geosocial Data

ACHILLEAS PSYLLIDIS, JIE YANG, AND ALESSANDRO BOZZON
Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 Delft, The Netherlands

Corresponding author: Achilleas Psyllidis (a.psyllidis@tudelft.nl)

This work was supported in part by the Social Urban Data Lab (SUDL), Amsterdam Institute for Advanced Metropolitan Solutions (AMS), in part by the Dutch national e-infrastructure with the support of SURF Cooperative, and in part by the Onassis Scholars' Association.

ABSTRACT Traditional methods for studying the activity dynamics of people and their social interactions in cities require time-consuming and resource-intensive observations and surveys. Dynamic online trails from geosocial networks (e.g. Twitter, Instagram, Flickr etc.) have been increasingly used as proxies for human activity, focusing on mobility behavior, spatial interaction, and social connectivity, among others. Social media records incorporate geo-tags, timestamps, textual components, user-profile attributes and points-of-interest (POI) features, which respectively address spatial, temporal, topical, demographic, and contextual dimensions of human activity. While the information contained in social media data is complex and high-dimensional, there is a lack of studies exploiting the combined potential of their information layers. This article introduces a framework that considers multiple dimensions (i.e. spatial, temporal, topical, and demographic) of information from social media data, and combines Geo-Self-Organizing Maps (GeoSOMs) in conjunction with contiguity-constrained hierarchical clustering, to identify homogeneous regions of social interaction in cities and, subsequently, estimate appropriate locations for new POIs. Drawing on the discovered regions, we build a Factorization Machine-based model to estimate appropriate locations for new POIs in different urban contexts. Using geo-referenced Twitter records and Foursquare data from Amsterdam, Boston, and Jakarta, we evaluate the potential of machine learning techniques in discovering knowledge about the geography of social dynamics from unstructured and high-dimensional social web data. Moreover, we demonstrate that the discovered homogeneous regions are significant predictors of new POI locations.

INDEX TERMS Geospatial analysis, recommender systems, self-organizing feature maps, social network services.

I. INTRODUCTION

The spatial structure and configuration of different areas (e.g. neighborhoods) comprising cities have traditionally been understood through authoritative urban data, primarily from censuses and household surveys. Given that the majority of these datasets contain sensitive information about individuals, the most common approach to protecting privacy is data aggregation. Data are usually aggregated into predefined administrative units, such as census tracts, enumeration districts, and post code areas. The delineation of these spatial units is defined according to factors of territorial governance, historical or cultural identity, population size and diversity, or according to the efficient performance of certain tasks (e.g. delivery of letters by a postman, in the case of postcode areas) [1].

Despite being extensively used in different types of (statistical) spatial analyses, administrative regions seldom reflect the character and dynamics of human interactions carried out in the underlying geographical space. While administrative regions could facilitate the economic distribution of governance tasks, they prove insufficient when it comes to analyzing the distribution of social interactions in space and time. Aggregating data about human activities and interactions into arbitrarily-defined spatial units could lead to longstanding problems, such as the modifiable areal unit problem (MAUP) [2], [3] and the ecological fallacy [4]. One challenge is,

therefore, to identify spatially contiguous areas in the city that are characterized by similar features regarding social interactions. Another challenge derives from the multidimensional nature social interactions, which encompass several facets of human activity. These include the location where activities take place, the function of activity venues, the type of activity performed, the time period, and demographic features of the individuals carrying out the activities. Fine-grained administrative datasets covering the aforementioned facets are hardly available.

Conversely, the landscape of urban data has been expanded and currently also includes dynamic geo-social records from social media, mobile phones, and sensors. Spatial data generated from these emerging sources have been increasingly used as proxies for human mobility, activity behavior, spatial interaction, and social connectivity. An interesting aspect of social media data, in particular, is the multiplicity of information layers they contain. Posts on social media (e.g. Twitter, Instagram, Flickr etc.) are often geo-referenced and, further, contain timestamps, textual components, and — where applicable — information about points (or places) of interest (POIs). Additional metadata could include user-profile features and information about online friendships. While social media data are prone to several types of biases (e.g. representational, contextual, functional, normative, temporal etc.) [5]–[8], their high-dimensional and fine-grained nature can open up new avenues for studying social dynamics. The extraction of spatial, temporal, topical, demographic, and contextual features from social media data could provide new insights into how social interactions are distributed in space and time. It could further help discover regions in cities that share similar properties of social interaction. Despite this possibility, there is currently a lack of examples of how to simultaneously exploit the different dimensions of social media data.

This article contributes a novel approach to characterizing city areas, as reflected by the online social activity of people. Our approach takes into account multiple dimensions (i.e. spatial, temporal, topical, and demographic) of information from social media data, and employs neural networks, to identify homogeneous regions of social interaction in cities. The discovered regions are, subsequently, used as features, in addition to other parameters, to train a machine learning model for estimating appropriate locations for new POIs in different urban contexts. The proposed approach relies upon geo-referenced Twitter data that are complemented by POI-related features from Foursquare for the city regions of Amsterdam, Boston, and Jakarta. Our goal is to revisit the traditional characterizations of city districts, which frequently draw information from data that pertain predominantly to night-time residence (e.g. census data). Instead, we classify areas according to their character, as reflected by the type of social activities people perform at various places, the function of these places, and the individual characteristics of the people performing the activities. In this way, we first aim to understand the structure of social interactions in cities, as inferred from social media data. Given that the majority of social interactions occur at places where people spend their free time [9], which we define as POIs, we then aim to provide a predictive model that estimates the most appropriate location for such places and examine the contribution of the discovered regions to the predictive power of the model.

We use Geo-Self-Organizing Maps (GeoSOMs) [10], in conjunction with contiguity-constrained hierarchical clustering, to discover multivariate clusters from unstructured social web data. In fact, these clusters reflect the various dimensions of social interaction, and represent city sub regions in close proximity, both geographically and in attribute space. Through a competitive and unsupervised learning process, the GeoSOM neural network has proven to perform better than conventional spatial clustering methods, when it comes to detecting patterns in high-dimensional data [11]. Unlike other popular clustering techniques, such as K-means, GeoSOM responds well to point data with non-Gaussian distributions (e.g. tweets) and also makes no assumptions regarding the distribution of data observations. The outcome of this process is, finally, linked back to geographic space, providing insight into spatiotemporal, semantic, and geo-demographic co-occurrences of social activity. The resulting regions indicate neighborhoods where people of similar demographic characteristics share common types of activity at coinciding times. The clustering validity is evaluated by means of measuring the pairwise difference of between and within-cluster dissimilarities, to ensure that the identified regions show a high degree of attribute homogeneity.

Having gained an understanding of the structure of social interactions, we then focus on places where most interactions occur. We define POIs as point representations of geographical places that host various facilities where people gather, socialize, and spend their free time. Thereby, we focus on what is usually referred to in literature as ''third places'' [12], [13], since it has been shown that the activities performed in these places are essential for social interactions [9]. Drawing on the discovered regions, we build a Machine Learning model to estimate appropriate locations for new POIs in different urban contexts. The model incorporates several POI features relating to category and popularity that are retrieved from Foursquare, in combination with the identified regions, as predictors. We formulate the POI location prediction task as a classification problem, and propose a model based on Factorization Machines. Our model classifies predefined administrative units (i.e. postcode areas) according to their appropriateness for the accommodation of a specific POI. To the best of our knowledge, our predictive model is the first application of Factorization Machines as classifiers for estimating the locations of new POIs in city regions. We demonstrate the influence of each feature on the predicted variable (i.e. the POI location, represented by an administrative unit), and examine the contribution of the identified clusters of social interaction to the classification result. To this end, we compare the predictive performance of

a model that solely uses POI-related features against one that also considers the discovered regions.

II. BACKGROUND AND RELATED WORK

Regionalization is a core concept in this article. The process of regionalization or region building involves the aggregation of spatial units (e.g. administrative areas) or observations (e.g. events with univariate or multivariate information) into a set of geographically connected regions [14], [15]. In deriving new regions, certain constraints — such as spatial contiguity — have to be satisfied, and an objective function — such as attribute similarity of the within-region entities — needs to be optimized. Over the past fifty years, there has been a considerable amount of research on regionalization problems [16]–[34], and the methods used to tackle them have been reviewed in detail by Fischer [35], Murtagh [22], Gordon [36] & [37], and Duque *et al.* [14].

Computational approaches to regionalization, in fact, correspond to methods for cluster analysis with implicit or explicit constraints. Both supervised and unsupervised approaches have been employed. In the first case, certain requirements for region building are pre-specified, such as the number of regions, the size and shape of regions, or the population size, among others [14]. Examples include the automatic zoning procedure (AZP) algorithm proposed by [38], the regionalization algorithm with selective search (RASS) proposed by [39], the SKATER (Spatial Kluster Analysis by Tree Edge Removal) algorithm proposed by [25], and the REDCAP (Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning) family of regionalization methods proposed by [26], all of which have been extensively used in problems of political districting. In the second case, the regionalization process is carried out without requiring predefined information. Examples include algorithms such as the SOM [40]–[42], the GeoSOM [10], the max-*p* [27] & [32], the neural gas [43]–[44].

Various regionalization methods have been applied in domains as diverse as census districting [45], ecoregion building [46], climate zoning [47], health studies [48], social area analysis [49], and spatial optimization [50]. The majority of these studies use structured data from censuses, surveys and various data repositories, while oftentimes simulated datasets are used. Recently, the proliferation of geo-social data, generated from social media platforms (e.g. Twitter, Instagram etc.) has led to the emergence of a limited number of works that use them as input in spatial clustering and regionalization analyses, serving different purposes.

Noulas *et al.* [51] use spectral clustering, drawing on publicly available Foursquare checkins retrieved from Twitter data, to identify clusters of social activity in London and New York. The proposed clustering approach takes into consideration only the place categories, extracted from the Foursquare data, to characterize different areas. Another limitation is that the patterns of point-based observations (i.e. Foursquare check-ins, in this case) are aggregated into predefined equally-sized squares, making them prone to MAUP-related issues. Cranshaw *et al.* [52] follow a similar approach, using spectral clustering to detect neighborhoods of similar activity from Foursquare data. As in the previous work, the features used in the clustering analysis are the venue categories, along with the popularity of places, based on the number of visits by different individuals. Longley and Adnan [53] develop a geo-temporal demographic classification of Greater London, based on modeled individual characteristics of Twitter users. Regions (i.e. clusters) of social connectivity are identified by means of Ward's hierarchical clustering analysis, on the basis of the amount of interactions in terms of tweet activity between two or more areas in the city. Lansley and Longley [54] classify geo-tagged tweets from Inner London to identify clusters of topics over space and time, using topic modeling.

Density-based spatial clustering of applications with noise (DBSCAN) and kernel density estimation have also been used in prior work on clustering analysis of Twitter [55] and Flickr data [56] respectively. While more advanced clustering and regionalization techniques based on artificial neural networks, such as SOM and its variants (e.g. GeoSOM, Hierarchical SOM, Hierarchical Spatiotemporal SOM etc.), have proven to respond well to complex and high-dimensional data [28], [57]–[60], they have rarely been used in the analysis of social media data, to date. One example of SOM application, in combination with principal component analysis (PCA), to Twitter data, is the work of [34] on the detection of regional linguistic variations across the United States.

Prior work focusing specifically on the spatiotemporal variation of social interactions using social media data is, to the best of our knowledge, currently lacking. On the grounds that data about the social dynamics on city-scale were, until recently, rarely available, traditional studies employed time-consuming and resource-intensive methods to discover meaningful patterns [61]–[63].

The work that most closely relates to our paper is from [11], where authors used hierarchical GeoSOMs to detect spatiotemporal and semantic clusters of geo-tagged tweets to characterize areas of Greater London. The method proposed in this article extends the approach of [11] by considering individual geo-demographic characteristics of social media users (i.e. age range, gender, place of residence, country of origin) in addition to spatiotemporal and semantic co-occurrences of social interactions. Moreover, we introduce a novel method that uses the identified regions of social interaction to predict which areas in a city are more appropriate for a new POI of a specific category to be located at.

III. MATERIALS AND METHODS

For the purposes of this work, user-generated geosocial data have been collected from two social media platforms, namely Twitter and Foursquare. Twitter is the primary source of information, whereas Foursquare is used for the extraction of POIs (i.e. venue categories). The latter are used as proxies for urban functions.

The pipeline of our proposed framework is organized around four steps. First, we filter and preprocess the data, using various natural language processing techniques, to retrieve and model the attributes that are used in the regionalization analysis. Second, we assess the spatial, temporal, and topical co-occurrences and demographic similarities of the collected tweets. Third, we use the results of the aforementioned co-occurrence and similarity assessment to train multiple GeoSOM algorithms. We generate several attribute maps for each one of the selected features. To detect regions of homogeneous social interaction, we apply a contiguity-constrained hierarchical clustering technique based on Ward's method, which agglomerates the various feature maps and identifies high-dimensional groupings of similar tweets in attribute space. The detected groups in the feature maps are linked back to geographical maps to extract the latent structure of social interactions. Finally, we build a Factorization Machine-based model to predict appropriate locations for new POIs, and test it in different urban contexts.

A. DATA PREPROCESSING

The first step of the processing pipeline involves the preprocessing of the collected data. Prior to carrying out the regionalization analysis, it is crucial to minimize the uncertainties, inconsistencies, and noise that are inherent to unstructured data, such as tweets.

Specific attention is paid to the textual component of the tweets. To this end, natural language processing (NLP) methods are employed, and in particular, tokenization, feature filtering, and stop-word removal in multiple languages. The tokenization of the tweets' text enables the creation of single-word vectors from the post strings that better facilitate the topical modeling and, therefore, the semantic similarity assessment. A document-feature matrix (DFM) is, then, created from the tokenized tweets, by also removing common stop words. Lists of common stop words are used in this regard, for the entire set of identified languages in the datasets. Given that tweets already comprise shortened texts, words, and abbreviations, further semantic dimension reduction by means of stemming was avoided. The derived DFM is normalized, by assigning each term a weight according to the TF-IDF (term frequency - inverse distance frequency) method. The resulting weighted terms are used as input to the topical modeling process. To prevent skewed topic modeling results, the top ten percent of highly re-occurring — usually city-specific — terms (e.g. 'Amsterdam', '#amsterdam', 'Boston', 'Massachusetts' etc.), based on TF-IDF weights, as well as words containing one or two characters were filtered out.

In total, 9 variables (attributes/features) were retrieved or modeled from the collected data, reflecting characteristics of social interaction. More specifically, the longitude and latitude of tweets, the venue category (retrieved from Foursquare), the topic, hour, and weekday of the post, along with the inferred age, gender, and social category (i.e. residents, commuters/local visitors, foreign visitors) of individual users.

B. MULTIVARIATE ATTRIBUTE SIMILARITY CALCULATION

We define a region of comparable social interactions in geographic space (in this case, the urban fabric of each city) as the co-occurrence of people with similar demographic characteristics in nearby places of similar functions, performing common activities at similar times. This definition implies that social interaction is, in fact, a multidimensional phenomenon, which involves both spatial (i.e. geographic proximity of places) and non-spatial attributes (i.e. function of a place, type of activity performed, time, demographic characteristics of people). Drawing on these features, we can define an attribute space for each city under study. The attribute space consists of a set of variables that are inherent to social interactions, according to the aforementioned definition. The variables pertain to the location of a place where an activity is performed, the function of that place, the type of activity, the time when it is carried out, and to demographic characteristics of individuals.

As proxies for these variables, we use the following attributes: the longitude and latitude of a tweet, extracted from the geo-tag or the POI location, as indicators of the location of a place; the venue category (retrieved from Foursquare) as proxy for the function of a place; the topic of a tweet, modeled through the text corpus, as an indicator of the type of activity; the date and time, retrieved from the timestamps; the age, gender, place of residence, and country of origin, modeled by means of user modeling on user profile metadata, to infer demographic characteristics of the population included in the collected datasets (Table 1).

In our approach, we first identify regions (i.e. clusters) of homogeneous social interactions in attribute space and, then, we link these regions back to the geographic space. The identified regions have to maintain both spatial contiguity and non-spatial attribute similarity, in terms of the data objects they contain. While some of the proxy variables comprising the attribute space are extracted directly from the collected data (e.g. geo-coordinates, timestamps, venue categories), others need first to be modeled (e.g. topics and user demographic characteristics). The following paragraphs describe the attribute modeling process and our approach to assessing the similarity of the selected set of (spatial and non-spatial) attributes.

The geographic longitude and latitude of each tweet are extracted directly from the API. The set of geo-coordinates is, then, used for determining the geographic tolerance parameter of the GeoSOM algorithm, described in the following section.

The temporal co-occurrence of activities is assessed on the basis of timestamps, extracted from each tweet. The timestamp is split into a date and an hour component. This helps aggregate tweets into weekdays and hourly time intervals, respectively. It further allows to assign specific types of activity (inferred from the topic of each tweet) to a certain

TABLE 1. Variables and attributes of social interaction.

Type of Variable	Attribute	Source	Proxy	Method
Spatial	Longitude	Twitter (geotag, POI location)	Location of a place	Extraction from Twitter API
	Latitude	Twitter (geotag, POI location)	Location of a place	Extraction from Twitter API
Temporal	Date	Twitter (timestamp)	Time of activity	Extraction from Twitter API
	Time	Twitter (timestamp)	Time of activity	Extraction from Twitter API
Semantic	Venue category	Foursquare	Function of a place	Extraction from Foursquare API
	Topic	Twitter (text)	Type/topic of activity	LDA
Socio-demographic	Age	Twitter (metadata and picture)	Age of individuals	User modeling (through *SocialGlass*[a])
	Gender	Twitter (metadata and picture)	Gender of individuals	User modeling (through *SocialGlass*[a])
	Place of residence	Twitter (profile metadata)	Home location of individuals	User modeling (through *SocialGlass*[a])
	Country of origin	Twitter (profile metadata)	Tourists or local visitors	User modeling (through *SocialGlass*[a])

[a] www.social-glass.tudelft.nl, and [65–67].

time interval. The function of places where a specific activity is performed are extracted from Foursquare, using the venue root category of a POI. Similar venue categories are, therefore, assumed to correspond to similar POI functions.

Topical features approximating the type of activity performed by people, are modeled by means of unsupervised machine learning techniques. We employ Latent Dirichlet Allocation (LDA) to classify the weighted and tokenized tweets included in the DFM into topics. LDA considers the semantic weighting factors to carry out the classification process. In identifying the optimal number of K topics, we calculated the harmonic mean of a sequence of potential K topics, in combination with Gibbs sampling [64]. The derived topics help subset the initial datasets into categories that represent different types of activities, thereby helping assess the topical similarity of tweets. Combined with the daily and hourly time intervals, they enable correlations between the semantic and temporal frequency of the collected tweets. These correlations indicate how similar types of activity are distributed over time.

Finally, the demographic characteristics of users – namely, the age, gender, place of residence, and country of origin – are derived by means of user modeling techniques. Besides age and gender, we classify individual users into three demographic groups, according to their inferred place of residence: *residents*, *local visitors*, and *foreign visitors*. We employed the user modeling capabilities of the *SocialGlass* platform, which has been developed by the authors. Detailed information on how we implement user modeling in *SocialGlass* can be found in [65]–[67].

C. REGIONALIZATION OF SOCIAL INTERACTIONS

The extracted and modeled features help us define the attribute space for the regionalization analysis. To first identify regions of social interactions in attribute space, we employ the GeoSOM algorithm [10], which is a proximity-aware variant of the standard SOM algorithm [40]–[42] In general, the output of a SOM is an attribute map that organizes data objects with similar attribute values into contiguous regions of neurons. Geographic locations of data objects in the real world usually play no particular role in Kohonen's algorithm [40] and, if considered, are treated equally to other non-spatial attributes. While it is possible to add a weighting factor to geo-coordinates to increase their importance in the training process [68] this would eventually downgrade the role of other non-spatial attributes [59].

Unlike these approaches, the GeoSOM algorithm does not require weights be assigned to input variables to define the relationship between spatial proximity and attribute similarity. Instead, it uses an additional parameter, called geographical tolerance (k), to maintain the geographic vicinity of similar data objects, by mapping them to neurons — usually referred to as best matching units (BMUs) — that are in close proximity in attribute space. The GeoSOM algorithm uses, first, the k parameter to determine a geographical vicinity of neurons, where BMUs of geographically close and similar data objects will be searched. Towards searching for the final location of each data object's BMU in the area defined by k, the algorithm, then, uses the remaining set of non-spatial attributes. This approach accounts for the unique properties of spatial data, such as spatial dependency and heterogeneity [69], given that it assumes no comparability between spatial and non-spatial attributes.

For this study, we mapped the collected data objects onto a 30 × 30 neural network (i.e. a GeoSOM feature map), comprising a regular two-dimensional grid of 900 hexagonal neurons. Given the number of observations (see Sect. IV), the specified GeoSOM feature map is considered medium-sized. This allows multiple observations to be mapped to each neuron, thereby leading to the creation of regions with more general characteristics. The latitude and longitude of the collected tweets, and their distribution over geographic space,

helped define a geographic vicinity on the attribute map for the first step of the BMU search process. In all three cases, we set $k = 2$. We chose to define k in attribute space to avoid creating a fixed radius in geographic space, given the uneven distribution of tweets. For the final search process, the remaining non-spatial attributes described in Sect. III.A are used. To standardize the measures, all variables were first transformed into z-scores. We applied a random map initialization and a Gaussian kernel function to adjust the neurons near each BMU over 100,000 iterations, such that:

$$h_{c,u}(t) = \exp\left(-\frac{\|r_c - r_u\|^2}{2(\sigma(t))^2}\right) \quad (1)$$

where, $h_{c,u}(t)$ is the Gaussian kernel determining the influence of the BMU on the neighboring neurons, r_c represents the center of the BMU, r_u represents the center of neuron u which is neighboring to the BMU, and σ is a time-dependent parameter, responsible for reducing the kernel width during the learning process. In the first training phase, we used an initial kernel width of 10 and a final width of 1. The learning rate α has an initial value of 1.0 and over the course of the training cycles it decreases linearly to 0.0.

To visualize the results of the GeoSOM, we use the unified distance matrix (U-matrix) [70], [71]. The U-matrix illustrates the distance (i.e. the dissimilarity) between the multivariate vectors mapped to adjacent neurons in attribute space. To gain better insight from the U-matrix, we use component planes (CPs) [42] together with a color scheme. Nearby neurons are assigned similar color, therefore allowing clusters to visually emerge. Darker areas in the U-matrix represent larger distances between the multivariate vector — and, therefore, larger dissimilarities — whereas brighter areas represent clusters of similar neurons. The CPs give insight into which attributes contribute more to the overall results, by visualizing neuron weights for each one of the input variables.

The final step involves the actual regionalization of social interactions, by agglomerating the various GeoSOM component planes corresponding to the selected variables. To this end, we apply a spatially-constrained hierarchical clustering algorithm to the various GeoSOM results using Ward's method, to detect multivariate clusters spanning all attributes. Ward's method, first, calculates the dissimilarity among all neighboring neuron clusters identified in the attribute space of the various GeoSOMs and, then, combines together the two neighboring clusters with the lowest dissimilarity into a new cluster [22]. In this case, the dissimilarity between two neighboring neuron clusters in a Geo-SOM is defined as the total within-cluster variance, such that:

$$D_W(C_i, C_j) = \sum_{x \in C_i \cup C_j} \|x - r_{i,j}\|^2 \quad (2)$$

where, D_w is the dissimilarity (i.e. Ward's distance) between the neighboring clusters C_i and C_j, x represents a data objects belonging to both C_i and C_j, and $r_{i,j}$ is the center of cluster resulting from the combination of C_i and C_j. Ward's hierarchical clustering avoids merging clusters that only contain a few neighboring data objects, thereby reducing the influence of the chaining effect [26] & [59].

To evaluate the homogeneity of the resulting clusters (or regions) and to validate their consistency, we assess the within-region attribute similarity. Specifically, we calculate the silhouette index [72], which provides an evaluation of the clustering validity, by measuring the pairwise difference of between and within-cluster dissimilarities, such that:

$$S_i(x) = \frac{b_i(x) - a_i(x)}{\max\{b_i(x), a_i(x)\}} \quad (3)$$

where α_i(x) is the average dissimilarity of a data object i at location $x \in$ x to all other data objects belonging to the same cluster A. b_i(x) denotes the lowest average dissimilarity of a data object i to every other cluster that i is not a member of. The cluster for which b_i(x) is attained, is called the neighbor cluster of i, as it constitutes the second-best choice for i to be assigned to. The overall S_i(x) denotes the average silhouette width of the clusters, and its values are by definition between -1 and 1. The closer the value of S_i(x) is to 1, the better the clustering (i.e. the assignment of each object i to its cluster).

Minimizing the average attribute dissimilarity α_i(x) of the data objects comprising each cluster can ensure internal homogeneity. Similarly, maximizing the average silhouette width S_i(x) would indicate maximum heterogeneity between clusters and, subsequently, a good matching of each data object within its corresponding cluster (or region). The calculation of the average silhouette width is conducted in parallel with the hierarchical clustering, using various numbers of clusters in an iterative fashion, so that the optimal number of regions is defined endogenously. The number that maximizes the intra-region homogeneity (i.e. maximum average silhouette width), constitutes the optimal number of regions.

Given that neighboring GeoSOM neurons contain data objects that are in close proximity in geographic space, the resulting clusters, in fact, represent spatially contiguous regions of homogeneous (i.e. low within-cluster dissimilarity) social interactions. The detected clusters are, finally, mapped onto the geographic space of the cities, using the same color scheme, to better understand the spatial configuration and extent of the derived regions in the real world.

D. POI LOCATION PREDICTION

Drawing on the discovered regions of homogeneous social interaction, and given that the latter mainly occurs around specific places in the city, we develop a model to predict appropriate locations for new POIs and test it in three world cities. As described in the Introduction, we define POIs as point representations of geographical places that host various facilities where people gather, socialize, and spend their free time. We formulate the POI location prediction task as a classification problem, for which we train a machine learning algorithm that takes as input features relating to a POI (e.g. venue category, expected popularity, rating etc.) and,

then, estimates the most appropriate location for a new similar POI, based on the input features.

Classification takes places at the district level. The algorithm classifies predefined administrative units (i.e. post code areas, in this case) according to their appropriateness for the accommodation of a specific POI. Given that the three cities are partitioned into several administrative units, the prediction task in question is effectively a multi-class classification problem. That is, the algorithm assigns a new POI to a single administrative unit. Having already discovered regions of social interaction, where social activity clusters around similar POI types, we aim to explore the extent to which these regions could bring additional predictive power to the model. In other words, our goal is to examine whether the inclusion of regions as additional features in the model could improve the location prediction task. To this end, we compare the performance of a model that only considers POI-related features against a model that further incorporates the discovered regions.

The POI-related features we consider in this study, are retrieved from Foursquare. We particularly focus on category and popularity-related features. The number of categories varies across the different cities (see Sect. IV). Regarding popularity, 10 features are considered in total for each one of the case-study cities. Given that each POI is described by a single category, we encode the category feature as a one-hot representation. That is, we transform each category feature into m possible binary features, so that each one describes whether or not a POI belongs to a specific category. Consequently, for each POI only $1/m$ of the binary category features is active (i.e. represented by 1), whereas all the others are inactive (i.e. represented by 0). As a result, the category features are highly sparse. On the contrary, the popularity-related features are rather dense, given that each POI has values for all the 10 popularity features. Similar to the category, each POI is associated with only one region. Thereby, regions are also transformed into binary features, indicating whether or not a POI belongs to a specific region.

The sparseness of the feature matrix makes the prediction task particularly challenging for conventional machine learning algorithms, such as Logistic Regression (LR) and Support Vector Machines (SVMs). It has been shown that the performance of these algorithms is rather weak, when it comes to highly sparse data [73], [74]. To overcome this limitation, we develop a Factorization Machine-based model to estimate appropriate locations for new POIs in the three cities. Factorization Machines (FMs) [75] have been successfully applied in several cases involving high-dimensional sparse data, especially in the field of collaborative recommendation systems [76]–[78]. Their advantage, compared to other machine learning algorithms such as SVMs, LR, Matrix and Tensor Factorization, lies in their ability to estimate interactions between all the variables of the model, even when the data at hand are highly sparse. While FMs are general predictors, applicable to regression, classification, and ranking tasks, they have rarely been used in cases beyond recommendation systems. Therefore, we aim to investigate the effectiveness of FMs as POI location predictors.

We estimate the appropriate location for a new POI, by fitting the following FM multi-class classification model:

$$\hat{y}_i(x) = w_0 + \sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m} \sum_{j=i+1}^{m} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j \tag{4}$$

where $y_i(x)$ denotes the predicted location x of a new POI y_i. w_0 is the intercept. The importance of each feature x_i is weighted by w_i, which is member of the coefficient vector $w \in \mathbb{R}^m$. $< \mathrm{v}_i, \mathrm{v}_j >$ denotes the dot product of the latent factors of x_i and x_j respectively, and describes the interaction between i-th and the j-th feature. To mitigate the data sparsity problem, the dimensionality $k \in \mathbb{N}_0^m$ of v_i and v_j is set to be small. In this way, a factorization effect is achieved, which leads to improved interaction matrices under sparsity [75]. The multiplication $< \mathrm{v}_i, \mathrm{v}_j >$ by $x_\mathrm{i} x_\mathrm{j}$ determines whether or not to consider the interaction between the i-th and the j-th feature for a specific data object (i.e. POI), in the modeling process. The interaction is considered only when both features are active (i.e. have a value of 1).

The parameters w_0, w, and v of the FM model can be estimated by minimizing the following loss function:

$$\sum_{i=1}^{p} -ln \frac{1}{1 + e^{-\hat{y}_i y_i}} \tag{5}$$

where p is the total number of POIs. The FM model simultaneously estimates the latent factors and the parameters from the training data, with the goal of minimizing the classification error. Therefore, the estimated latent factors are generally more predictive than other low-dimensional representations that could be obtained from mainstream dimensionality reduction algorithms, such as Principal Components Analysis (PCA) [75] & [78].

We use stochastic gradient descent to train the FM model, with a decaying learning rate. All features are standardized in the preprocessing step to accelerate the learning process. The optimal dimensionality of the latent factors is empirically found, using grid search.

To evaluate the prediction performance, we adopt a 5-fold cross-validation, using the micro-average F-score, which is defined as follows:

$$F_{micro} = \frac{\left(1 + \beta^2\right) P \cdot R}{\beta^2 P + R} \tag{6}$$

where β is the parameter to control the weight between precision P and recall R. We set $\beta = 1$, assuming an equivalent importance between precision and recall. We define precision and recall as follows:

$$P = \frac{\sum_{i=1}^{c} (TP)_i}{\left(\sum_{i=1}^{c} (TP)_i\right) + \left(\sum_{i=1}^{c} (FP)_i\right)} \tag{7}$$

$$R = \frac{\sum_{i=1}^{c} (TP)_i}{\left(\sum_{i=1}^{c} (TP)_i\right) + \left(\sum_{i=1}^{c} (FN)_i\right)} \tag{8}$$

where c denotes the number of classes (i.e. administrative units), and $(TP)_i$, $(FP)_i$, and $(FN)_i$ respectively denote the number of true positive, false positive (Type I error), and false negative (Type II error) estimates. We further make use of the F-score(5), which considers a prediction to be true only if the true class falls into the top 5 predicted classes.

IV. RESULTS

To examine and test the adaptability of the proposed framework in different urban contexts, we applied it in three world cities [79], [80], each one belonging to a different continent; namely, Amsterdam (Europe), Boston (N. America), and Jakarta (Asia). We collected geo-referenced Twitter and Foursquare data from the three case-study cities, for the period between May 1 and June 1, 2016. The data collection and crawling process was carried out using the *SocialGlass* platform [66], [67]. *SocialGlass* enables the extraction of each user's post history, by means of backward crawling, which further allows for modeling user-related demographic attributes (e.g. estimated home location, demographic category, age range, gender). Moreover, we used Foursquare to retrieve POI-related features and information (e.g. venue category, popularity, location etc.).

A. REGIONALIZATION: MULTIDIMENSIONAL CLUSTERS OF SOCIAL INTERACTION

After filtering and preprocessing, the remaining number of tweets for the city of Amsterdam is 34,310, for Boston 73,029, and for Jakarta 78,045. We begin by extracting the features of social interaction that will be used as input for training the Geo-SOM models.

First, we extract each tweet's latitude and longitude, as well as the corresponding date and time using the incorporated geo-tag and timestamp respectively. Based on the collected Foursquare records, we then map each tweet to a POI type, which we use as proxy for an urban function (e.g. coffee shop, restaurant, hotel etc.). The classification of urban functions follows the venue category hierarchy of Foursquare.

While urban functions are important indicators of human activities, there could be discrepancies between the function of a place and the type of activity people actually perform. To disambiguate activities from functions, we analyze the textual information of the collected tweets. The semantics embedded in the textual component of tweets can give an indication of the kind of activity carried out at a specific POI type. As described in Sect. III, the identification and extraction of latent topics from the textual information of the collected micro posts is carried out by means of an LDA model. The latter discovers underlying co-occurrences of words and attributes them to a topic, along with a degree of probability. After several experiments, using a series of different K topic values, we identified the optimal number of topics for each city's Twitter dataset as the one that maximizes the log-likelihood of word-topic probability in the collected tweets. By running 1000 iterations of Gibbs sampling, extracting the log-likelihoods from each topic, and calculating the corresponding harmonic means, the optimal K value for the city of Amsterdam is found to be 9 topics, for Boston $K = 8$, and for Jakarta $K = 10$.

We detected a number of common topic categories across the use case cities (Fig. 1–3). For instance, *Topic 8* in Amsterdam and *Topic 5* in Boston both refer to travel and airport-related activities and, as expected, appear in tweets that are either concentrated around airport facilities or distributed along highways and railroads connecting the corresponding city centers with the airport. *Topic 3* in Amsterdam, *Topic 1* in Boston, and *Topic 10* in Jakarta refer to leisure activities, relating to POI categories such as coffee shops, museums, malls, bars, and music theaters, whereas *Topic 1* in Amsterdam and *Topic 9* in Jakarta specifically refer to leisure activities that are performed outdoors (e.g. in parks, canals, and plazas). Conversely, *Topic 5* in Amsterdam, *Topic 4* in Boston, and *Topic 4* in Jakarta relate to work and business activities.

The analysis also detected topics that appear to be prominent in only one — and more rarely, two — of the three cities in question. Activities explicitly relating to food consumption were detected only in Jakarta and were correspondingly classified under *Topic 2*. Contrariwise, *Topic 2* in Amsterdam contains terms that relate to road traffic. *Topic 7* in Amsterdam *Topic 3* in Boston both relate to emergency incidents. While this topical differentiation could demonstrate context-related habits in terms of what type of activities people in different cities/countries are willing to share on social media, it could also be an indication of the demographic bias that characterizes this category of geo-social data.

To further analyze the temporal distribution of the inferred activities, we used the embedded timestamps and aggregated the topics extracted from the LDA model into weekdays and hourly intervals. As shown in Fig. 1–3, leisure-related activities (*Topic 3* in Amsterdam, *Topic 1* in Boston, *Topic 10* in Jakarta) have varying temporal patterns across the three cities. In Amsterdam, there is an evident increase on Fridays and the weekends, as opposed to the other weekdays. Moreover, there is an overall increment of social activity from 3pm throughout the week, with the exception of Sundays, where a peak is observed around 12pm, followed by gradual decrease from 2pm — 5pm, reaching again a high number of tweets from 5pm onwards. Conversely, in Boston and Jakarta the corresponding topic patterns appear quite repetitive over time. When comparing the temporal patterns of the (indoor) leisure topic across the three cities, one may observe relatively low values of social activity during the early hours in Amsterdam, whereas in Jakarta, and most prominently in Boston, leisure-related activities are also carried out after midnight, with relatively stable (Jakarta) or rather high values until 3am (Boston). Activities relating to eating in Jakarta, and represented by *Topic 2*, show repetitive patterns throughout the weekdays, with peaks around 1pm and 3pm. In Amsterdam, road traffic (represented by the corresponding *Topic 2*) is characterized by a high number of related tweets, from early in the morning till late in the afternoon, with peaks around

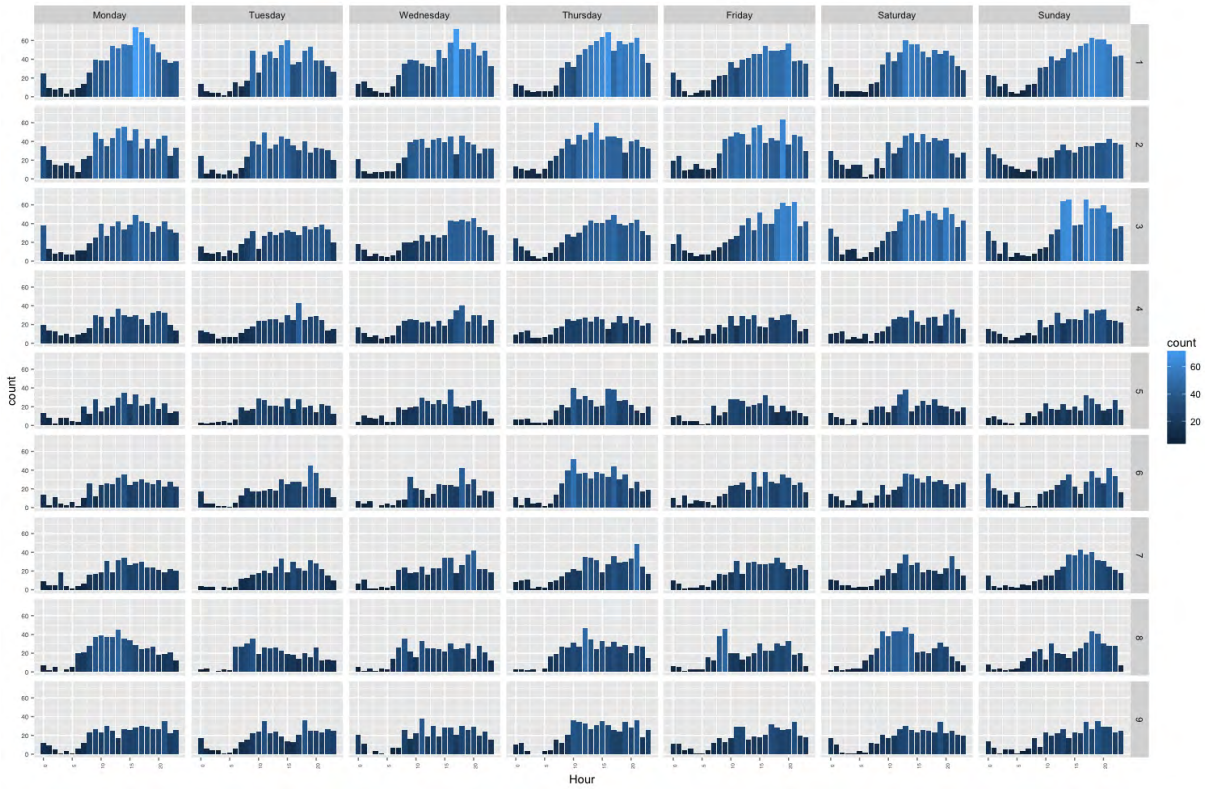

FIGURE 1. Hourly temporal-semantic frequencies of LDA-classified topics from geosocial media in Amsterdam, the Netherlands.

9am — 11am and 6pm — 8pm. The observed variability in social activity patterns could be an indicator of varying human activity habits across the three cities.

To better understand the contextual aspects of social interactions, we further annotate the collected tweets with demographic characteristics (i.e. age, gender) of the users and, additionally, classify individuals into residents, visitors, and foreign visitors, based on their inferred location of residence. In obtaining these features, we made use of the built-in user modeling components of the *SocialGlass* system, which are described in detail in [65]–[67]. Even though it is difficult to assess how representative the identified social media users are of the actual urban populations, the modeled individual characteristics allow us to complement the social activity data with demographic information at a fine-grained level. It further enables us to gain insight into the spatial dynamics of individual behaviors, relating to demographic groups that are not included in conventional census data (e.g. visitors and foreign visitors), yet constitute an important component of the social interactions in a city.

Having extracted, inferred, and modeled the features relating to social interactions, we train several Geo-SOM models to identify areas in the three cities that are characterized by both locational and feature similarity. As described in Sect. III, we use unified distance matrices — or, short, U-matrices — to represent the feature space of the GeoSOM output layer. While U-matrices help illustrate cluster structures of similar (light-colored neurons) or dissimilar (dark-colored neurons) observations on input variables in feature space, they provide little insight on the spatial distribution of the detected clusters. To this end, and to facilitate the interpretation of the observed structures, we map the GeoSOM results onto the geographic space of each city under study.

The various GeoSOMs and the corresponding U-matrices of their output layer detect several sub-clusters across the three cities that manifest how each of the features in focus is structured in geographic space (Fig. 4). Due to limited space, we will showcase indicative examples of the individual GeoSOM results. Fig. 5–6 illustrate the existence of a large cluster in both Amsterdam and Boston, corresponding to individuals of the age group between 25–35, in the center-west and south-west areas, respectively. The aforementioned clusters are indicated in dark red, in both the U-matrices (Fig. 4) and the geographic maps (Fig. 5–6). Adjacent to the above cluster in both cities, is the one indicated in light orange, which corresponds to the age group between 35–45 and contains fewer observations (i.e. tweets) than the former. The vicinity of two clusters further manifests that the respective age groups perform similar activities in both space and time.

On the contrary, in Jakarta (Fig. 7), the population group between 25-35 years appears to cluster around several areas of the city, as illustrated by the patchy dark red pattern in the

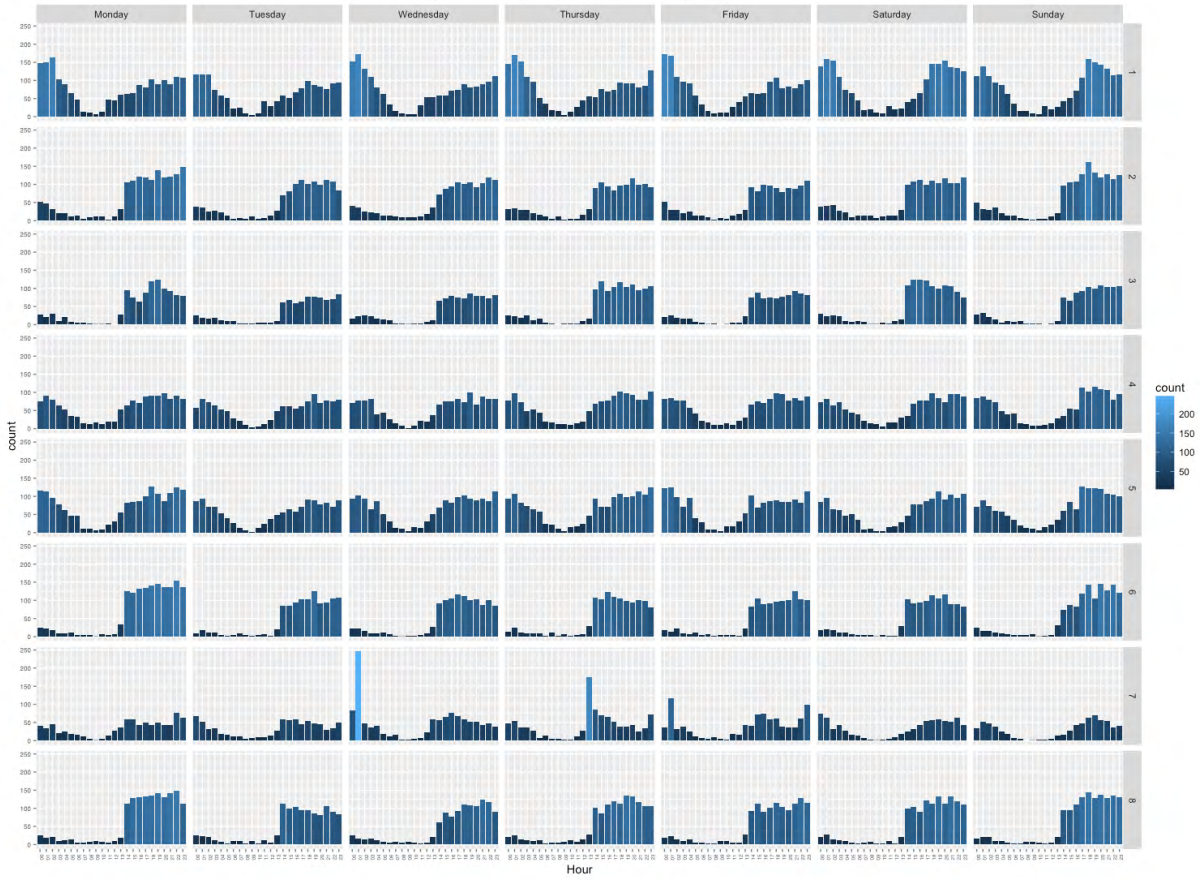

FIGURE 2. **Hourly temporal-semantic frequencies of LDA-classified topics from geosocial media in Boston, MA, USA.**

U-matrix and the geographic map. The dark grey neurons represent areas with high dissimilarity (i.e. high weight vectors) in terms of input variables, as compared to their neighboring neurons on the U-matrix. Therefore, they act as separators between areas that contain observations with similar feature values but are, however, spatially dissimilar.

Looking further into the demographic features of social activity, the component plane corresponding to the social category feature, indicates in the city of Amsterdam the existence of a cluster, which corresponds to foreign tourists' activity. Part of this cluster appears to overlap with the area dominated by the population age group between 25–35 years, thereby, signifying areas in the city where young foreign tourists choose to perform similar types of activity at similar times. On the other hand, a large cluster covering an extensive area in the city of Boston, corresponds to the activities performed by residents. The tweets assigned to the neurons forming the aforementioned cluster are generated from residential areas of the city.

While the various GeoSOM models offer insights into the way in which input features with similar values group together across space, the identification of multivariate regions of social interaction needs to be more explicitly defined. The application of a spatially-constrained hierarchical clustering method to the GeoSOM results — as described in Sect III.B — helps address this issue. The outcome of this process is a higher-level neural network, the attribute space of which summarizes the hierarchical structure of all GeoSOMs. The clusters (indicated with the letter C and the corresponding cluster number in Fig. 8–10) represent regions that are characterized by similar feature values across the 7 variables of social interaction (the latitude and longitude variables are used to define the proximity of the observations and ensure the contiguity of the resulting clusters/regions). A color scheme is used to indicate the assignment of a neuron to a multivariate cluster. Given that the hierarchical clustering method used is spatially constrained, the discovered regions are subsequently geographically contiguous. Each of the resulting regions is characterized by internal homogeneity with regard to the 7 non-spatial feature values. As described in Sect. III.B, the homogeneity and consistency of the resulting clusters is assessed by means of the silhouette index. After performing a silhouette analysis with fifty different cluster values for each city, the optimal number of clusters that maximizes the intra-region homogeneity (i.e results in the maximum average silhouette width) is 16 ($s.i. = 0.725$) for the city of Amsterdam, 15 ($s.i. = 0.692$) for Boston, and 42 ($s.i. = 0.789$) for Jakarta.

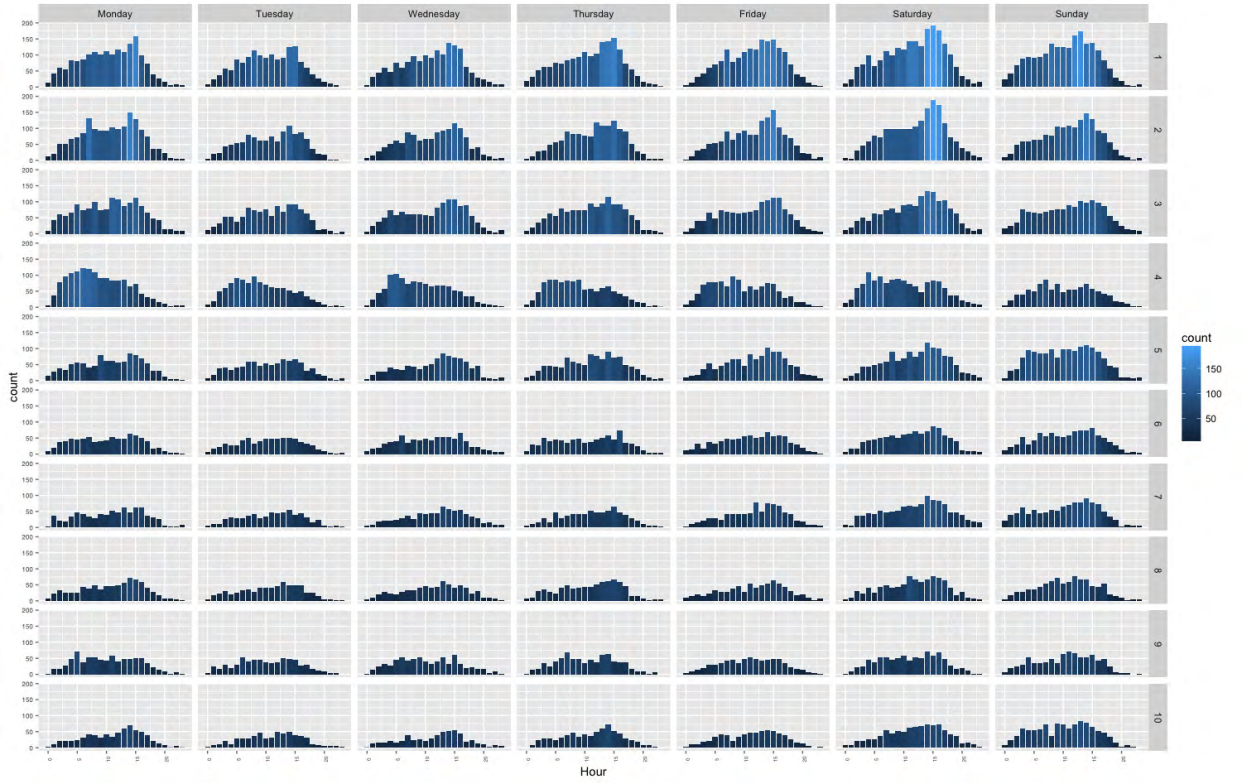

FIGURE 3. Hourly temporal-semantic frequencies of LDA-classified topics from geosocial media in Jakarta, Indonesia.

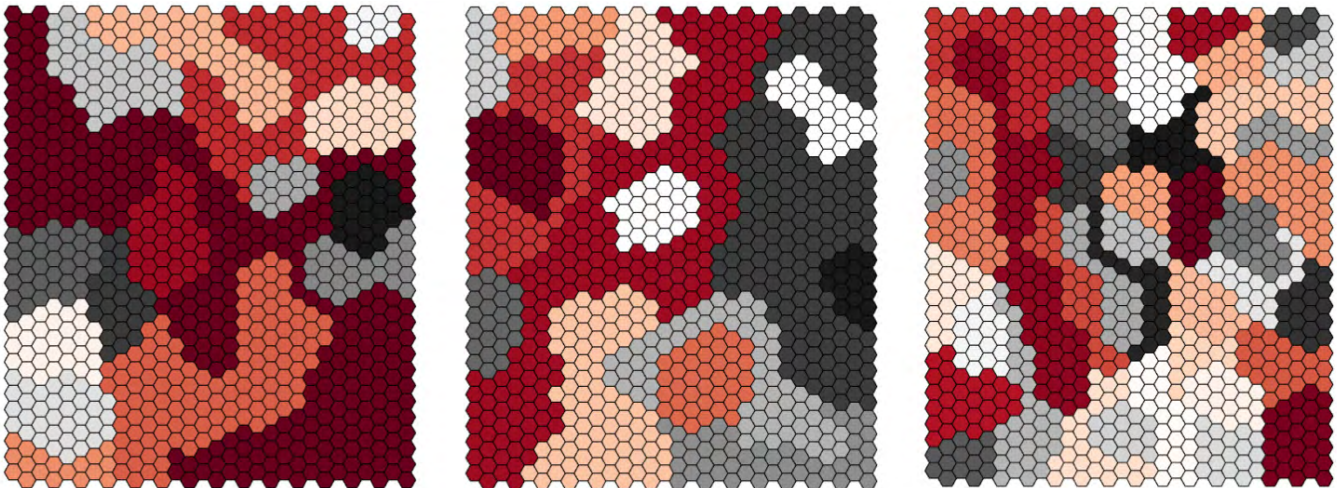

FIGURE 4. U-matrices of the GeoSOM component plane corresponding to the age feature for the cities of (a) Amsterdam, (b) Boston, and (c) Jakarta.

Fig. 8 illustrates the discovered regions in the city of Amsterdam. Clusters 3 and 6 comprise the largest amount of tweets relating to nightlife activities, performed by population groups between 25 and 45 years, who are predominantly foreign tourists. The two aforementioned clusters cover the districts south of the central train station. Cluster 4 accumulates daytime outdoors activities, mainly performed by residents between the age of 25 and 55. Two of the city's main parks (i.e. the Vondel park and the Rembrandt park) are included in its boundaries. On the other hand, its neighboring Cluster 15 contains tweets relating to cultural activities, mainly performed by foreign tourists, and is delineated around the city's museum district. Activities relating to work appear to agglomerate into Cluster 8, which geographically

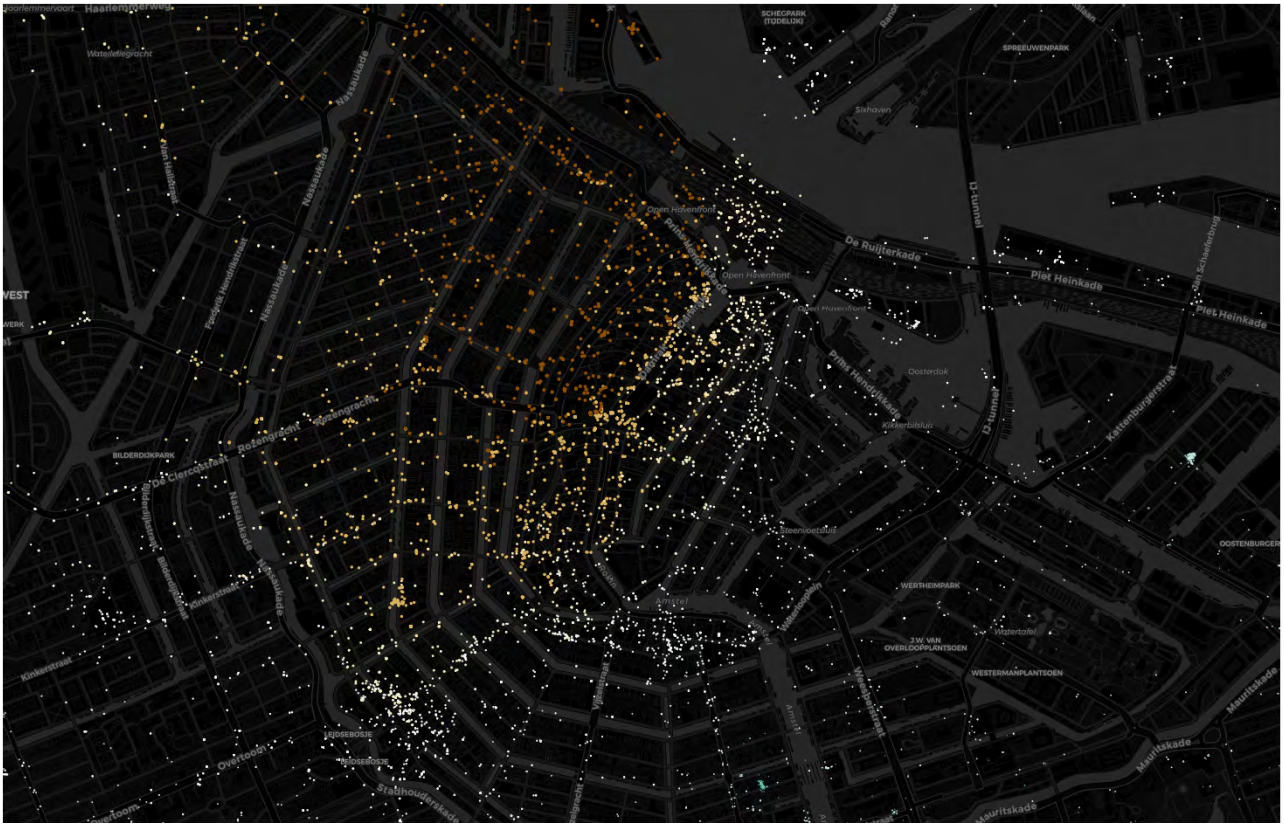

FIGURE 5. GeoSOM clusters in Amsterdam (base map: CARTO).

FIGURE 6. GeoSOM clusters in Boston (based map: CARTO).

covers the southern districts (Zuidas) of the city. Another predominant region discovered by the model is Cluster 1, which is mainly characterized by activities performed by residents, of various age groups ranging from 25 to 55 years of age.

FIGURE 7. GeoSOM clusters in Jakarta (based map: CARTO).

The majority of activities in this cluster is linked to POIs relating to cafés and restaurants. Most of nightlife activities of residents and local visitors accumulate around Cluster 2, whereas the region defined by Cluster 11 accommodates social activities of residents, local visitors, and foreign tourists relating to sports. Finally, a distinctive region (Cluster 10) is formed around the Airport area and its surrounding road network, accommodating the majority of travel-related tweets.

Fig. 9 displays the discovered multidimensional clusters (regions) in the city of Boston. A large number of tweets relating to travel and outdoor activities during daytime, across all three demographic groups used in this study, are part of the region delineated by Cluster 2. On the other hand, nighttime activities appear to distribute along the neighboring north and northeastern districts of the city, and are represented by Cluster 3. Similarly to the city of Amsterdam, the Airport area is identified as a separate region (Cluster 10) by the model, comprising tweets relating to Topic 5 (i.e. Airport/Transport). Cluster 11 encompasses the MIT campus, whereas the neighboring region defined by Cluster 12 is characterized by micro-posts generated primarily by residents, within the age range between 25 and 45, linked to POIs such as coffee shops, restaurants, and shopping centers, and mainly classified under Topic 1 (Leisure/Outdoor activities) and Topic 8 (Retail). The majority of social activity generated by residents spatially clusters in the south and southeast districts of Boston forming a region defined by the boundaries of Clusters 14 and 15, as well as in the northern districts of Cambridge, within the region delineated by Cluster 13. On the other hand, Cluster 6 in the central district of Cambridge gathers the largest number of tweets relating to retail activities and generated primarily by residents and local visitors.

A significantly larger number of regions was discovered by the hierarchical GeoSOM model in the city of Jakarta, as illustrated in Fig. 10. The regions defined by the neighboring Clusters 6 and 22 are characterized by a large number of tweets linked to POIs such as retails stores, malls, and restaurants. Cluster 6 mainly contains tweets classified under Topic 10 (i.e. indoor leisure activities, relating to cafés, malls, and bars), whereas Cluster 22 accumulates several tweets under Topic 2 (i.e. eating and food-related activities). In close proximity to the previous, the region defined by Cluster 4 primarily gathers social activity relating to outdoor activities (abundance of tweets under Topic 9; activities relating to parks and plazas), performed by various demographic groups

FIGURE 8. Regions resulting from the projection of the hierarchical GeoSOM clusters onto the geographic space of Amsterdam (an indicative number of clusters is highlighted) (base map: Mapbox).

of a quite broad age range. Outdoor social activities relating to sports are gathered in Clusters 31 and 7 in the central and north-west parts of the city. On the other hand, educational activities accumulate around the regions defined by Clusters 10 and 12, in the central districts of the city. The majority of social activity generated by residents appears cluster in the southern (Clusters 21, 27, and 18) and western (Clusters 19 and 23) districts of the city. A large number of tweets generated by tourists geographically accumulates around the neighboring Clusters 24, 36, 39, and 42.

B. PREDICTION OF NEW POI LOCATIONS

After identifying latent neighborhood structures within the three cities, we then investigate the extent to which these regions could help discover appropriate areas for siting new POIs. As mentioned in Sect. III, we formulate the prediction task as a multi-class classification problem, following a factorization machine approach. We use the identified regions as a new input feature in the classification process, and compare it with features associated with popularity and POI category, to evaluate its predictive power. We, further, test the effectiveness of our approach to the classification problem by comparing it with three baseline classifiers. First, the random classifier, which randomly assigns an administrative unit to a POI according to a uniform distribution — that is, there is an equal probability of siting a POI at any administrative unit. Second, the stratified random classifier, which assigns an administrative unit to a POI according to a stratified distribution — that is, administrative units with a larger number of POIs in the training data have a higher probability. Third, the logistic regression (LR) classifier, which classifies POIs according to a linear decision function, the parameters of which are learned from the training data. It is worth noting that both the random and stratified classifiers do not make use of any features in the prediction process, whereas the LR and FM classifiers are both feature-based.

Results of the different experimental configurations are presented in Table II. Feature-based classifiers (i.e. LR and FM), trained with either POI popularity or category features, appear to outperform both the random and the stratified random classifiers. This is a clear indication that the use of POI features plays a significant role in the prediction process. The superior performance of the stratified random classifier against the random classifier implies the imbalanced distribution of POI classes, especially in the case of Boston.

The incorporation of the discovered regions as additional features leads both LR and FM to achieve the best performance, when compared to other feature sets. More specifically, it leads to a 5.14% increase in F-measure and a 3.19% in F-score(5), when LR is applied, as compared to the

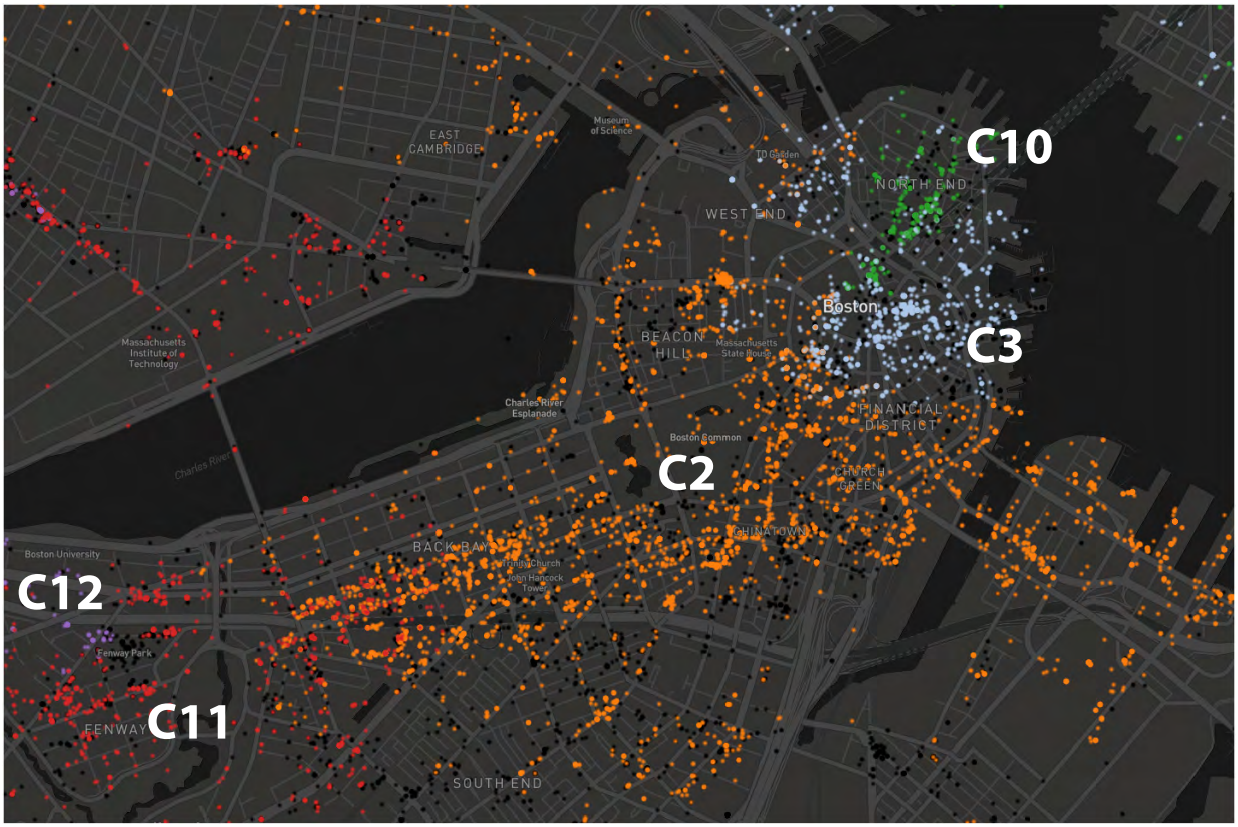

FIGURE 9. Regions resulting from the projection of the hierarchical GeoSOM clusters onto the geographic space of Boston (an indicative number of clusters is highlighted) (base map: Mapbox).

TABLE 2. Performance and comparison of features and classifiers for the prediction of appropriate POI locations in the administrative units of the Amsterdam, Boston, and Jakarta, measured by the micro-averaged F-measure and F-score(5).

		Random	Stratified	Logistic Regression				Factorization Machine			
	Feature	NA	NA	Cat.	Popularity	Category & Popularity	Cat. & Pop. & Region	Category	Popularity	Cat. & Pop.	Cat. & Pop. & Reg.
F-measure	Amsterdam	0.0105	0.0140	0.0524	0.0619	0.0651	0.0680	0.0608	0.0709	0.0752	0.0794
	Boston	0.0356	0.0660	0.1905	0.1542	0.1955	0.2009	0.1987	0.1670	0.2037	0.2066
	Jakarta	0.0239	0.0288	0.0534	0.0528	0.0549	0.0594	0.0584	0.0550	0.0604	0.0645
F-score(5)	Amsterdam	0.0137	0.0189	0.0699	0.0826	0.0887	0.0913	0.0829	0.0974	0.1045	0.1113
	Boston	0.0498	0.0863	0.2541	0.2144	0.2698	0.2739	0.2716	0.2229	0.2827	0.2908
	Jakarta	0.0325	0.0381	0.0706	0.0702	0.0738	0.0792	0.0816	0.0749	0.0844	0.0908

performance achieved when only taking into account popularity and category features. In the case of FM, the incorporation of the regions leads to a 4.59% increase in F-measure and a 5.65% (significant, paired t-test with p-value <0.01) in F-score(5). The FM approach consistently outperforms LR across all feature sets, reaching an increase of 9.40% in F-measure and 14.71% in F-score(5) (significant, paired t-test with p-value <0.01).

By adding up the coefficients learned in the prediction process, we analyze the contribution of each feature. Fig. 11 illustrates the coefficients of select top-ranked features in the prediction of appropriate areas for siting

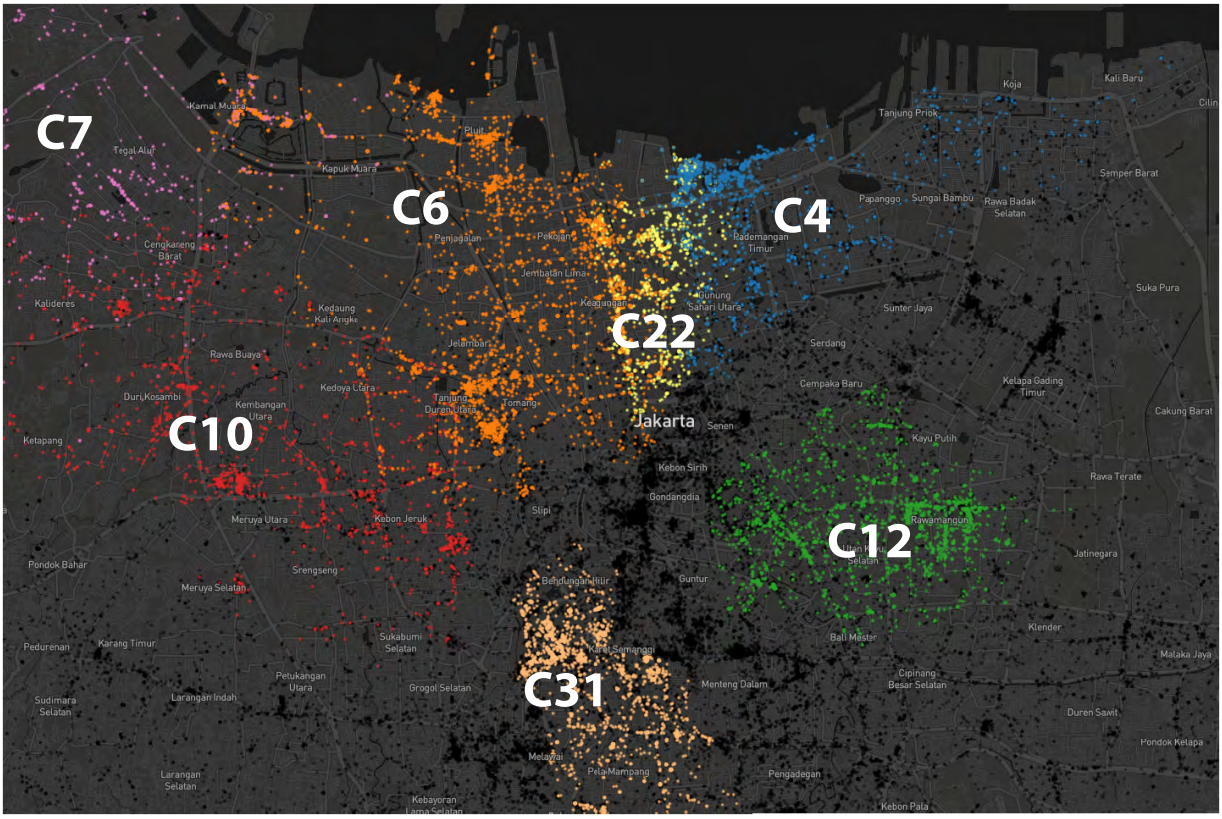

FIGURE 10. Regions resulting from the projection of the hierarchical GeoSOM clusters onto the geographic space of Jakarta (an indicative number of clusters is highlighted) (base map: Mapbox).

specific POIs. As an example, POIs associated with the category of restaurant appear, in all three cities, to be more suitable to be placed in areas that are attractive (i.e. have a high number of visits by different individuals), yet at a distance from the city center. Conversely, administrative units within the city center appear more suitable to host hotel functions.

V. DISCUSSION

This section presents a further discussion of the results associated with the regionalization of social interactions and the prediction of POI locations described previously.

Emphasis is on limitations and uncertainties relating to the nature of user-generated data, as well as to the methods presented in this paper.

A. DEFINING REGIONS OF SOCIAL INTERACTION FROM USER-GENERATED CONTENT

Social media data, being one of the prominent examples of user-generated content, offer a new lens through which the notion of place, as well as the way people experience and interact with it and other people can be revisited. Similarly, the availability of this new type of data allows us to re-examine and potentially enrich the traditional approaches to delineating urban districts. However, conducting a regionalization analysis using social media data comes with a set of limitations, which mostly derive from the particular nature of this new data source. Although the content of geosocial data is more closely aligned with the way we colloquially describe and communicate about places, uncertainties associated with natural language and different types of bias could impact the analysis results.

Specifically, in the preprocessing step, the semantic uncertainty, ambiguousness, and noise common to Twitter data could affect the LDA-based topical classification and, subsequently, the clustering of similar topics in the GeoSOM model. This can be further affected by the variety of languages, which the collected microposts are written in. Although the LDA model is capable to classify terms into topics in several languages, the absence of a semantic matching framework among the various terms could lead to some of them being mis-classified. In addition, terms with different meanings in different contexts, such as the term *house*, which could, for instance, be associated with a person's home location or refer to the music genre (i.e. *house music*), could respectively relate to different activity types. Despite these limitations, the prevalence of certain topics in particular cities (e.g. food-related activities in Jakarta), as well as the association of the semantic attributes with their distribution over time, give us insight into local habits of interaction with different places and the corresponding communication of these activities online.

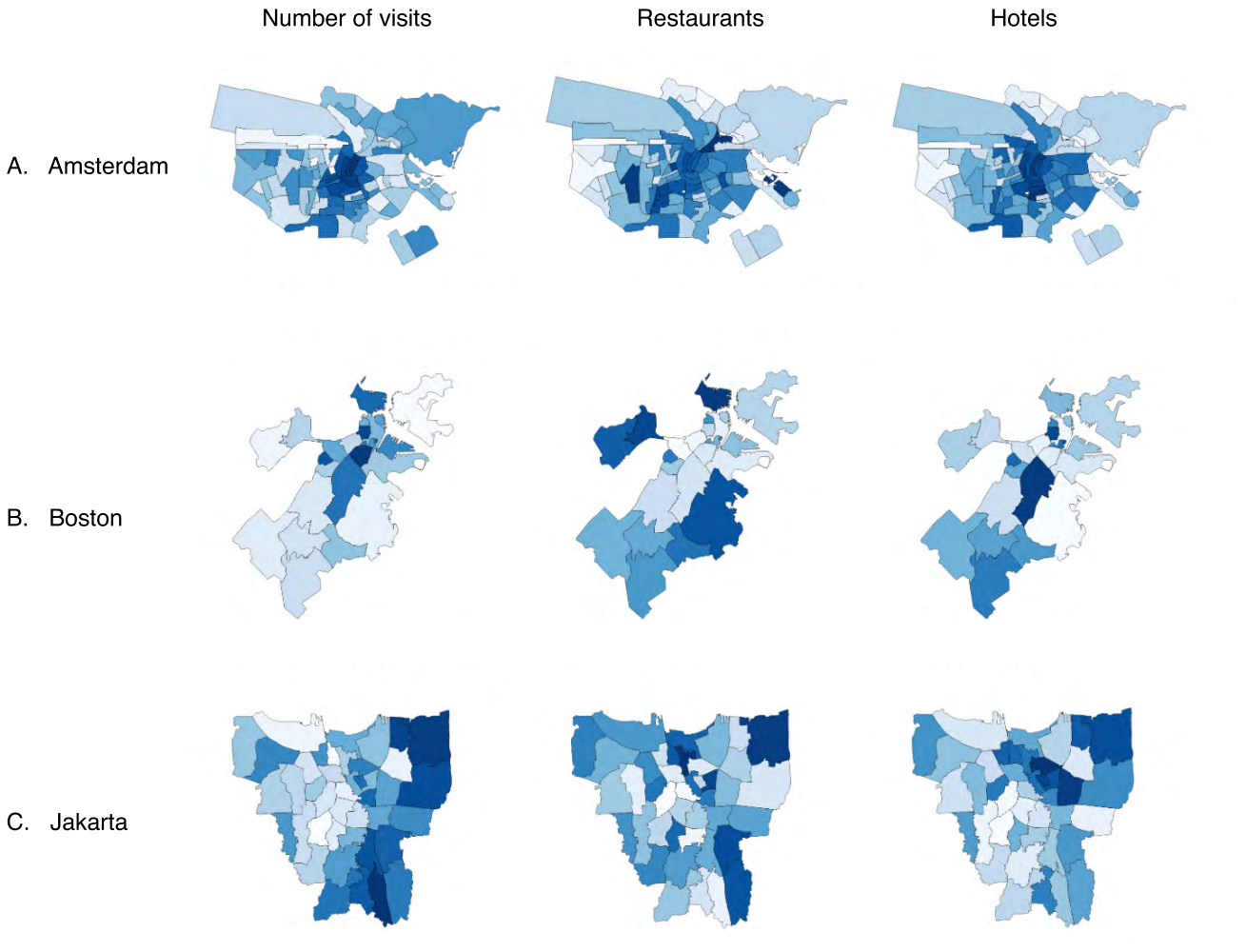

FIGURE 11. Visualization of coefficients of important features that contribute to the prediction of appropriate areas for siting different POIs. For each feature, the darker the shade of each administrative unit is, the more appropriate is this unit for hosting a POI characterized by that specific feature.

The spatial and temporal attributes associated with a tweet present another kind of uncertainty. Geo-location and time intervals play a key role in the study of human interactions with places over time. However, a discrepancy between the actual time of an activity and the timestamp of a micropost is common in social media data. Moreover, it is possible that the content of a tweet refers to an activity that is not necessarily related to the place at which it is posted, or even describes a past or future activity.

One of the attributes used in the training of the GeoSOM model is associated with the function of a POI, which is based on the place type taxonomy of Foursquare. Given that the upper-level classification of place types in Foursquare contains ambiguous categories, such as *Professional and Other Places*, we have decided to use the extended taxonomy, which comprises 421 POI types. Even with such an extended classification, uncertainties might occur in the delineation of functional regions, especially when it comes to places of different types that, however, share similar activities (e.g. bars, pubs, nightclubs) [81].

In addition to the above, a well-known issue with social media data is the demographic bias. That is, certain demographic groups – predominantly young adults – contribute most of the content on social media platforms and, subsequently, the activities they perform are overrepresented in the collected datasets. Besides stressing the activities of a particular group of people, the demographic bias has an additional aspect. It also relates to the posting habits these people have. Young adults not only perform different activities from more elderly populations, they also tend to post about certain types of them online (e.g. visiting a bar, watching a movie, eating out etc.). This, in turn, influences the resulting clusters of social interaction. The latter should, therefore, be interpreted as reflections of the activity dynamics pertaining to a limited set of demographic categories and their corresponding posting habits.

By following a multi-dimensional approach in our proposed framework, we aimed to mitigate the effects of the aforementioned uncertainties associated with user-generated content. The simultaneous consideration of spatial, temporal, semantic, and demographic attributes helps minimize the limitations pertaining to each individual attribute type. Of course, the outcomes of our proposed regionalization analysis of social interactions are also dependent on the initial training parameters of the GeoSOM model. These include the size of the network, the value of the geographic tolerance (k), the initialization function ($h(t)$), the kernel width, and learning rate (α), which were constant throughout the analysis. By setting the geographic tolerance k at 2, we ensured that tweets with similar non-spatial attributes (e.g. topic, POI type, demographic features etc.) would belong to the same cluster (region) only if they are also in close proximity in geographic space. The obtained high values (i.e. close to $+1$) of the average silhouette width in the three cities under examination, indicate a good performance of the model, in terms of cluster consistency and internal homogeneity.

B. POI LOCATION PREDICTION

In predicting appropriate locations for the placement of new POIs, we employed a factorization machine approach. Our results have shown that this prediction approach consistently outperforms baseline methods, such as random and stratified random classifiers, as well as logistic regression, across the three cities. Factorization machines have been recently proven to be effective in tasks such as recommendation and click-through rate prediction for online advertising [82], where the data are often highly sparse and multi-dimensional. Although the application presented in our work is of different nature, the data used have similar characteristics to the ones employed in the aforementioned application domains, especially in terms of sparsity. The improved performance of our model against popular classifiers indicates that factorization machines hold promise as an appropriate model for predicting new POI locations.

To comprehensively evaluate the performance of our model, we used the F-measure as a metric that describes both precision and recall. Similarly to other multi-class problems, such as object recognition in images [83], we further considered the top-5 predictions by using the variant F-score(5) metric. The absolute predictive performance of our model as measured by both F-measure and F-score(5) is relatively low; that is, within the range between 0.06 and 0.29. This is primarily caused by the complexity of the problem, given that the number of potential units (i.e. administrative areas) within which a POI can be placed is relatively high (e.g. 96 administrative units in the city of Amsterdam). This is further confirmed by the low performance achieved by random classifiers. However, our results do not contradict the levels of absolute performance achieved in studies using factorization machines for other multi-class problems. For instance, the achieved F-scores in recommendation problems are typically at the order of 0.01 to 0.1 [75].

Despite the complexity of the problem, we have shown that the identified regions have a strong positive impact on the prediction of new POI locations, when considered as additional features in the modeling process. This is demonstrated by the improved performance of our model, which is 7.6 and 8.1 times higher than the performance of the respective random classifiers.

VI. CONCLUSIONS

Identifying spatially contiguous regions with homogeneous characteristics within a city is a challenging topic. It becomes even more challenging when the aim is to understand latent characteristics, such as social interactions, which are however integral to the study of human activity dynamics. This work introduced a novel framework that combines several machine learning techniques and simultaneously considers spatial, temporal, thematic, and demographic information from unstructured user-generated content to identify homogeneous regions of social interaction in cities. It, further, showed how the regions identified by the proposed model can be used to predict appropriate locations for new POIs. Our analysis illustrated how the various spatiotemporal, semantic, and demographic features characterizing human activities could help indicate latent patterns associated with social interactions in different urban contexts. The ability of our proposed framework to cater to the several uncertainties of high-dimensional and unstructured user-generated data, along with its application in different cities, contribute to its generalizability and scalability. That is, both in terms of using data from different sources (e.g. Instagram) with similar characteristics and in terms of its application to other urban contexts.

In future research, we will examine the incorporation of morphological features of the urban environment – extracted from street-level imagery – and its potential to enhance our understanding of places and latent structures in the urban fabric, as well as the corresponding interactions of people with them.

ACKNOWLEDGMENT

The authors would like to thank Erik Boertjes for his help in visualizing the multi-dimensional clusters (regions) on the maps of the three case-study cities.

REFERENCES

[1] Eurostat, "Regions in the European Union: Nomenclature of territorial units for statistics NUTS 2013/EU-28," Luxembourg, Publications Office European Union, 2015.
[2] S. Openshaw and P. J. Taylor, "The modifiable areal unit problem," in *Quantitative Geography*, N. Wringley and R. J. Bennett, Eds. London, U.K.: Routledge, 1981, pp. 60–70.
[3] S. Openshaw, *The Modifiable Areal Unit Problem*. Norwich, U.K: GeoBooks, 1984.
[4] W. S. Robinson, "Ecological correlations and the behavior of individuals," *Amer. Sociol. Rev.*, vol. 15, no. 3, Jun. 1950.
[5] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," in *Proc. ICWSM*, Ann Arbor, MI, USA, 2014, pp. 505–514.
[6] L. Hong, G. Convertino, and E. H. Chi, "Language matters in Twitter: A large scale study," in *Proc. ICWSM*, Barcelona, Spain, 2011, pp. 518–521.

[7] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquis, "Understanding the demographics of Twitter users," in *Proc. ICWSM*, Barcelona, Spain, 2011, pp. 554–557.
[8] E. Hargittai, "Whose space? Differences among users and non-users of social network sites," *J. Comput.-Mediated Commun.*, vol. 13, no. 1, pp. 276–297, 2008, doi: 10.1111/j.1083-6101.2007.00396.x.
[9] M. S. Rosenbaum, "Exploring the social supportive role of third places in consumers, lives," *J. Service Res.*, vol. 9, no. 1, pp. 59–72, 2006, doi: 10.1177/1094670506289530.
[10] F. Bação, V. S. Lobo, and M. Painho, "Geo-self-organizing map (Geo-SOM) for building and exploring homogeneous regions," in *Proc. GIScience*, Adelphi, MD, USA, 2004, pp. 22–37.
[11] E. Steiger, B. Resch, and A. Zipf, "Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks," *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 9, pp. 1694–1716, 2015, doi: 10.1080/13658816.2015.1099658.
[12] R. Ahas *et al.*, "Modelling home and work locations of populations using passive mobile positioning data," in *Location-Based Services and Telecartography II*, G. Gartner and K. Rehrl, Eds. Berlin, Germany: Springer, 2009, pp. 301–315.
[13] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *J. Roy. Soc. Interface*, vol. 10, no. 84, p. 20130246, 2013, doi: 10.1098/rsif.2013.0246.
[14] J. C. Duque, R. Ramos, and J. Suriñach, "Supervised regionalization methods: A survey," *Int. Regional Sci. Rev.*, vol. 30, no. 3, pp. 195–220, 2007, doi: 10.1177/0160017607301605.
[15] D. Guo and H. Wang, "Automatic region building for spatial analysis," *Trans. GIS*, vol. 15, no. 1, pp. 29–45, 2011, doi: 10.1111/j.1467-9671.2011.01269.x.
[16] B. J. L. Berry, "A method for deriving multifactor uniform regions," *Przeglad Geograficzny*, vol. 33, no. 2, pp. 263–282, 1961.
[17] N. A. Spence, "A multifactor uniform regionalization of British counties on the basis of employment data for 1961," *Regional Stud.*, vol. 2, no. 1, pp. 87–104, 1968.
[18] P. M. Lankford, "Regionalization: Theory and alternative algorithms," *Geograph. Anal.*, vol. 1, no. 2, pp. 196–212, Apr. 1969.
[19] M. Monmonier, "A comparison of quantitative regionalization methods," *Geograph. Rev.*, vol. 62, no. 3, pp. 426–428, Jul. 1972.
[20] P. Haggett, A. D. Cliff, and A. Frey, *Locational Analysis in Human Geography*. London, U.K.: Edward Arnold, 1977.
[21] M. F. Goodchild, "The aggregation problem in location allocation," *Geograph. Anal.*, vol. 11, no. 3, pp. 240–255, 1979.
[22] F. Murtagh, "A survey of algorithms for contiguity-constrained clustering and related problems," *Comput. J.*, vol. 28, no. 1, pp. 82–88, 1985.
[23] F. Murtagh, "Contiguity-constrained clustering for image analysis," *Pattern Recognit. Lett.*, vol. 13, no. 9, pp. 677–683, 1992.
[24] S. Openshaw, "Developing GIS-relevant zone-based spatial analysis methods," in *Spatial Analysis: Modelling in a GIS Environment*, P. A. Longley M. Batty, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1996, pp. 55–73.
[25] R. M. Assunção, M. C. Neves, G. Câmara, and C. Da Costa Freitas, "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees," *Int. J. Geograph. Inf. Sci.*, vol. 20, no. 7, pp. 797–811, Aug. 2006, doi: 0.1080/13658810600665111.
[26] D. Guo, "Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)," *Int. J. Geograph. Inf. Sci.*, vol. 22, no. 7, pp. 801–823, Jul. 2008, doi: 10.1080/13658810701674970.
[27] J. C. Duque, L. Anselin, and S. J. Rey, "The max-*p*-regions problem," *J. Regional Sci.*, vol. 52, no. 3, pp. 397–419, Jan. 2012, doi: 10.1111/j.1467-9787.2011.00743.x.
[28] J. Hagenauer and M. Helbich, "Hierarchical self-organizing maps for clustering spatiotemporal data," *Int. J. Geograph. Inf. Sci.*, vol. 27, no. 10, pp. 2026–2042, 2013, doi: 10.1080/13658816.2013.788249.
[29] M. Helbich, W. Brunauer, J. Hagenauer, and M. Leitner, "Data-driven regionalization of housing markets," *Ann. Assoc. Amer. Geographers*, vol. 103, no. 4, pp. 871–889, 2013, doi: 10.1080/00045608.2012.707587.
[30] S. Sobolevsky, M. Szell, R. Campari, T. Couronné, Z. Smoreda, and C. Ratti, "Delineating geographical regions with networks of human interactions in an extensive set of countries," *PLoS ONE*, vol. 8, no. 12, p. E81707, 2013, doi: 10.1371/journal.pone.0081707.
[31] S. E. Spielman and J. R. Logan, "Using high-resolution population data to identify neighborhoods and establish their boundaries," *Ann. Assoc. Amer. Geographers*, vol. 103, no. 1, pp. 67–84, Jan. 2013, doi: 10.1080/00045608.2012.685049.
[32] D. C. Folch and S. E. Spielman, "Identifying regions based on flexible user-defined constraints," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 1, pp. 164–184, 2014, doi: 10.1080/13658816.2013.848986.
[33] D. Arribas-Bel, "The spoken postcodes," *Regional Stud., Regional Sci.*, vol. 2, no. 1, pp. 458–461, Aug. 2015, doi: 10.1080/21681376.2015.1067151.
[34] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, "Understanding U.S. regional linguistic variation with Twitter data analysis," *Comput., Environ. Urban Syst.*, vol. 59, pp. 244–255, Jan. 2016, doi: 10.1016/j.compenvurbsys.2015.12.003.
[35] M. M. Fischer, "Regional taxonomy: A comparison of some hierarchic and non-hierarchic strategies," *Regional Sci. Urban Econ.*, vol. 10, no. 4, pp. 503–537, 1980.
[36] A. D. Gordon, "A survey of constrained classification," *Comput. Statist. Data Anal.*, vol. 21, no. 1, pp. 17–29, 1996.
[37] A. D. Gordon, *Classification*, 2nd ed. Boca Raton, FL, USA: Chapman & Hall, 1999.
[38] S. Openshaw, "A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling," *Trans. Inst. Brit. Geographers*, vol. 2, no. 4, pp. 459–472, 1977.
[39] J. C. Duque, "Design of homogeneous territorial units: A methodological proposal and application," Ph.D. dissertation, Univ. Barcelona, Barcelona, Spain, 2004.
[40] T. Kohonen, "Clustering, taxonomy, and topological maps of patterns," in *Proc. 6th Int. Conf. Pattern Recognit.*, Munich, Germany, 1982, pp. 114–128.
[41] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.
[42] T. Kohonen, *Self-Organizing Maps*. 3rd ed. Secaucus, NJ, USA: Springer, 2001.
[43] T. M. Martinetz and K. J. Schulten, "A 'neural-gas' network learns topologies," in *Proc. Artif. Neural Netw.*, Amsterdam, The Netherlands, 1991, pp. 397–402.
[44] J. Hagenauer and M. Helbich, "Contextual neural gas for spatial clustering and analysis," *Int. J. Geograph. Inf. Sci.*, vol. 27, no. 2, pp. 251–266, 2013, doi: 10.1080/13658816.2012.667106.
[45] S. Openshaw and L. Rao, "Algorithms for reengineering 1991 census geography," *Environ. Planning A, Econ. Space*, vol. 27, no. 3, pp. 425–446, 1995.
[46] R. N. Handcock and F. Csillag, "Spatio-temporal analysis using a multiscale hierarchical ecoregionalization," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 1, pp. 101–110, 2004.
[47] R. G. Fovell and M.-Y. C. Fovell, "Climate zones of the conterminous United States defined using cluster analysis," *J. Climate*, vol. 6, pp. 2103–2135, Nov. 1993.
[48] F. Wang, D. Guo, and S. McLafferty, "Constructing geographic areas for cancer data analysis: A case study on late-stage breast cancer risk in Illinois," *Appl. Geogr.*, vol. 35, nos. 1–2, pp. 1–11, 2012.
[49] S. E. Spielman and J.-C. Thill, "Social area analysis, data mining, and GIS," *Comput., Environ. Urban Syst.*, vol. 32, no. 2, pp. 110–122, Jan. 2008, doi: 10.1016/j.compenvurbsys.2007.11.004.
[50] S. E. Spielman and D. C. Folch, "Reducing uncertainty in the American community survey through data-driven regionalization," *PLoS ONE*, vol. 10, no. 2, p. E0115626, Feb. 2015, doi: 10.1371/journal.pone.0115626.
[51] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in *Proc. Workshop Social Mobile Web ICWSM*, Barcelona, Spain, 2011, pp. 570–573.
[52] J. Cranshaw et al, "The Livehoods project: Utilizing social media to understand the dynamics of a city," in *Proc. 6th Int. AAAI Conf. Weblogs Social Media*, Dublin, Ireland, 2012, pp. 58–65.
[53] P. A. Longley and M. Adnan, "Geo-temporal Twitter demographics," *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 2, pp. 369–389, 2015, doi: 10.1080/13658816.2015.1089441.
[54] G. Lansley and P. A. Longley, "The geography of Twitter topics in London," *Comput., Environ. Urban Syst.*, vol. 58, pp. 85–96, Jun. 2016, doi: 10.1016/j.compenvurbsys.2016.04.002.
[55] A. Veloso and F. Ferraz, "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter," in *Proc. 3rd Int. Web Sci. Conf.*, Koblenz, Germany, 2011, pp. 15–17.
[56] L. Li and M. F. Goodchild, "Constructing places from spatial footprints," in *Proc. 1st ACM SIGSPATIAL Int. Workshop Crowdsourced Volunteered Geogr. Inf.*, Redondo Beach, CA, USA, 2012, pp. 15–21.

[57] P. Agarwal and A. Skupin, *Self-Organizing Maps: Application in Geographic Information Science*. Chichester, U.K.: Wiley, 2008.
[58] R. Henriques, V. Lobo, and F. Bação, "Spatial clustering using hierarchical SOM," in *Applications of Self-Organizing Maps*, M. Johnsson, Ed. Rijeka, Croatia: InTech, 2012, pp. 231–250.
[59] C.-C. Feng, Y.-C. Wang, and C.-Y. Chen, "Combining geo-SOM and hierarchical clustering to explore geospatial data," *Trans. GIS*, vol. 18, no. 1, pp. 125–146, 2014, doi: 10.1111/tgis.12025.
[60] J. Hagenauer and M. Helbich, "SPAWNN: A toolkit for spatial analysis with self-organizing neural networks," *Trans. GIS*, vol. 20, no. 5, pp. 755–774, 2016, doi: 10.1111/tgis.12180.
[61] S. Milgram, *The Individual in a Social World: Essays and Experiments*. London, U.K.: Longman Education, 1977.
[62] R. Oldenburg, *The Great Good Place*. New York, NY, USA: Paragon House, 1989.
[63] W. H. Whyte, *The Social Life of Small Urban Spaces*. New York, NY, USA: Project Public Spaces, 2001.
[64] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.
[65] A. Psyllidis, "Revisiting urban dynamics through social urban data: Methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities," Ph.D. dissertation, Faculty Archit. Built Environ., Delft Univ. Technol., Delft, The Netherlands, 2016.
[66] S. Bocconi, A. Bozzon, A. Psyllidis, C. T. Bolivar, and G.-J. Houben, "Social glass: A platform for urban analytics and decision-making through heterogeneous social data," in *Proc. WWW*, Florence, Italy, 2015, pp. 175–178.
[67] A. Psyllidis, A. Bozzon, S. Bocconi, and C. T. Bolivar, "A platform for urban analytics and semantic data integration in city planning," in *Proc. 16th Int. Conf. Comput.-Aided Archit. Design Futures-New Technol. Future Built Environ. (CAAD)*, G. Celani, D. M. Sperling, and F. J. M. Santos, Eds. Berlin, Germany: Springer, 2015, pp. 21–36.
[68] V. Lobo, F. Bação, and M. Painho, "Regionalization and homogeneous region building using the spatial kangas map," in *Proc. 7th AGILE Conf. Geograph. Inf. Sci.*, Heraklion, Greece, 2004, pp. 301–313.
[69] L. Anselin, "What is special about spatial data? Alternative perspectives on spatial data analysis," presented at the Symp. Spatial Stat., Past, Present Future, New York, NY, USA, 1989.
[70] A. Ultsch and H. P. Siemon, "Kohonen's self-organizing feature maps for exploratory data analysis," in *Proc. Int. Neural Netw. Conf.*, Paris, France, 1990, pp. 305–308.
[71] A. Ultsch, "Self-organizing neural networks for visualisation and classification," in *Information and Classification*, O. Opitz, B. Lausen, R. Klar, Eds. Berlin, Germany: Springer-Verlag, 1993, pp. 307–313.
[72] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
[73] J. Zhang, Z. Ghahramani, and Y. Yang, "Flexible latent variable models for multi-task learning," *Mach. Learn.*, vol. 73, no. 3, pp. 221–242, Dec. 2008.
[74] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. NIPS*, Vancouver, BC, Canada, 2007, pp. 1257–1264.
[75] S. Rendle, "Factorization machines," in *Proc. IEEE ICDM*, Sydney, NSW, Australia, Dec. 2010, pp. 995–1000.
[76] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Comput. Surv.*, vol. 47, no. 1, Apr. 2014, Art. no. 3.
[77] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering," in *Proc. RecSys*, Barcelona, Spain, 2010, pp. 79–86.
[78] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 42–49, Aug. 2009.
[79] P. Hall, *The World Cities*. New York, NY, USA: McGraw-Hill, 1966.
[80] J. V. Beaverstock, R. G. Smith, and P. J. Taylor, "A roster of world cities," *Cities*, vol. 16, no. 6, pp. 445–458, 1999, doi: 10.1016/S0264-2751(99)00042-6.
[81] G. McKenzie, K. Janowicz, S. Gao, J.-A. Yang, and Y. Hu, "POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data," *Cartographica*, vol. 50, no. 2, pp. 71–85, 2015.
[82] Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin, "Field-aware factorization machines for CTR prediction," in *Proc. RecSys*, Barcelona, Spain, 2010, pp. 43–50.
[83] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, Miami, FL, USA, Jun. 2009, pp. 248–255.

ACHILLEAS PSYLLIDIS is currently a Post-Doctoral Researcher in spatial data science and urban analytics with the Web Information Systems Group, Delft University of Technology. He is also a Research Fellow with the Amsterdam Institute for Advanced Metropolitan Solutions. His research focuses on data-driven approaches to characterizing places, and the activities, experiences, and interactions of people in them and the processes that drive their behavior. His work involves the acquisition, processing, analysis, and the enrichment of large, unstructured, and user-contributed geosocial data from a wide range of sources, with the aim to extract rich descriptions of and gain insight into the dynamics of human activities in urban places, and the dimensions that contribute to the character of a place. To this end, he combines spatial analysis and geographic information retrieval techniques with (deep) machine learning and neural networks to build scalable methods and software tools that facilitate people extract knowledge from and maximize the value of unstructured data.

JIE YANG received the Ph.D. degree from the Web Information Systems Group, Delft University of Technology, The Netherlands, in 2017. He is currently a Senior Researcher with the eXascale Infolab, University of Fribourg. He conducted the research of this paper while being the Ph.D. student within the Web Information Systems Group, Delft University of Technology. His research focuses on building effective human-machine loop systems that combine human intelligence with machine scalability to solve complex tasks at scale. The topic lies at the intersection of human computation, machine learning, recommendation, and user modeling. His work finds its natural application in human computation, recommendation, question answering, and urban computing systems.

ALESSANDRO BOZZON is currently an Assistant Professor with the Web Information Systems Group, Delft University of Technology. He is also a Research Fellow with the Amsterdam Institute for Advanced Metropolitan Solutions, and a Faculty Fellow with the IBM Benelux Center of Advanced Studies. His research lies at the intersection of crowdsourcing, user modeling, and web information retrieval. He studies and builds novel social data science methods and tools that combine the cognitive and reasoning abilities of individuals and crowds, with the computational powers of machines, and the value of big amounts of heterogeneous data.