SPECIAL ISSUE PAPER

Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks

Enrico Steiger [a], Bernd Resch [a,b,c] and Alexander Zipf [a]

[a]GIScience Research Group, Institute of Geography, Heidelberg University, Heidelberg, Germany; [b]Z_GIS, Department of Geoinformatics, University of Salzburg, Salzburg, Austria; [c]Center for Geographic Analysis, Harvard University, Cambridge, MA, USA

ABSTRACT

The investigation of human activity patterns from location-based social networks like Twitter is an established approach of how to infer relationships and latent information that characterize urban structures. Researchers from various disciplines have performed geospatial analysis on social media data despite the data's high dimensionality, complexity and heterogeneity. However, user-generated datasets are of multi-scale nature, which results in limited applicability of commonly known geospatial analysis methods. Therefore in this paper, we propose a geographic, hierarchical self-organizing map (Geo-H-SOM) to analyze geospatial, temporal and semantic characteristics of georeferenced tweets. The results of our method, which we validate in a case study, demonstrate the ability to explore, abstract and cluster high-dimensional geospatial and semantic information from crowdsourced data.

ARTICLE HISTORY

Received 8 April 2015
Accepted 19 September 2015

KEYWORDS

Twitter; location-based social network (LBSN); self-organizing map (SOM); semantic topic model; point pattern analysis

1. Introduction

Analysis of data from social networks has recently become an established research field in a number disciplines including geoinformatics, computational linguistics, computer science, sociology, psychology or urban planning. Location-based social networks (LBSNs) (Roick and Heuser 2013), regarded as a specific sub-domain of social networks, create further research potential by adding a geospatial dimension and providing location-embedded services. This is of central relevance because shared personal locations are becoming more and more a key point of interaction in digital communication (Zheng 2011), whether users are uploading geotagged photos via Flickr or Instagram, checking in at a venue with Foursquare, or commenting on a local event via Twitter.

These emerging, inexpensive and widespread 'human sensor' technologies have facilitated new possibilities to discover (geographic) knowledge and to analyze human behavior from social media data (Miller and Goodchild 2015). Additional value of this kind of data in comparison to traditional geo-data comes from their spatiotemporal resolution, which may complement or validate existing data sources and consequently open up a wide variety of research avenues.

However, the acquisition of crowdsourced data differs from conventional approaches in that underlying measurement processes are not well defined. For instance, the technical properties of remote sensors are known, whereas the motivation, the accuracy or the semantic correctness of a tweet are still unexplored. This results in a number of uncertainties in the data, which have not been quantified yet.

Consequently also the analysis of Twitter data is prone to a variety of geospatial, temporal and semantic uncertainties. First of all, the *geospatial accuracy* of a tweet can be influenced by mobile device characteristics, urban environments or other factors such as the GPS dilution of precision (Zandbergen and Barbeau 2011). Furthermore, users do not contribute records equally in geographic space and time, resulting in a highly *heterogeneous distribution* of tweets inside the LBSN. The geospatial distribution of tweets, for instance, strongly varies on differing scale levels (country, city, neighborhood, etc.) and is prone to be sparse in rural areas (Sengstock and Gertz 2012).

Georeferenced tweets also represent only a *small fraction* (2%–4%) of the entire available tweet set. When focusing on the ratio between Twitter users and the general population numbers, there is a clear mismatch between population and sampling frame (Miller and Goodchild 2015). This effect, known as sampling bias, might lead to exclusion or under/over representation of certain population groups (Heckman 1979). Thus, depending on the Twitter information and analysis the researchers focus on, unrepresentative subsets and different sample sizes might be generated.

Moreover, *semantic* information related to an event in time may refer to past or future topics/activities. Compared to Foursquare posts where users can 'check in' at specific venues (shops, hotels etc.), which already characterize specific categories of activities, within tweets we do not have any a priori knowledge regarding users' activities.

Also, the mentions of activities within tweets are relatively vague and semantically uncertain (use of abbreviations, acronyms, etc., due to character limitations and Internet-specific writing styles), and they might only be a weak indicator of a real-world observation as we have limited knowledge about the underlying social processes.

Even though these shortcomings in Twitter data are known, previous research on spatial analysis methods to extract information from tweets has originally been defined for particular datasets under a number of presumptions: local indicators of spatial association (LISA) (Anselin 1995), geographical weighted regression (GWR) (Fotheringham *et al*. 2003) and others.

In other words, from a GIScience perspective, we are facing a lack of methods for geospatial analysis of crowdsourced data (Steiger *et al*. 2015) and the combination of different methods in the era of data-driven geography (Miller and Goodchild 2015), which handles the high-dimensional uncertainty of spatiotemporal and geographic data, e.g., in tweets.

In order to adapt to this new uncertain type of geo-data and to the shift from a data-scarce to data-rich geographic research environment, Miller and Han (2009) defined the field of geographic data mining, where computational methods for discovering patterns in large, heterogeneous geographic datasets are applied. Yet, methods developed in this field are nearly exclusively of a single-disciplinary nature, i.e., data are analyzed geographically, whereas no inherently trans-disciplinary methods (combining algorithms from several research disciplines) have been defined. In this context, the neural network-based 'self-organizing maps' (SOM) approach has been proven to be useful for analyzing

geographic data (Agarwal and Skupin 2008, Feng *et al*. 2014). However, it has not been investigated yet whether a multi-dimensional SOM-approach combining methods from the disciplines of geography and computational linguistics is usable for detecting spatiotemporal and semantic clusters in Twitter data.

Resulting from the above challenges and the shortcomings identified in previous approaches (Section 2), the goal of this paper is to evaluate the application of SOMs for detecting high-dimensional (geospatial, semantic, temporal) clusters, instead of using traditional spatial autocorrelation methods. This paper shows that SOMs are a promising approach to explore, abstract and cluster geographic data while overcoming some of the previously mentioned limitations (see Steiger *et al*. 2015, p. 20 for a more detailed description). More precisely, the machine-learning process could foster discovery of latent structures from high-dimensional georeferenced Twitter data to gain new insights by utilizing all available semantic geospatial, and temporal information layers.

The research questions this paper answers are as follows:

(RQ1): How can relationships among people be discovered based on mentions of their activities in a trans-disciplinary approach (geospatial, temporal and semantic analysis) by analyzing vast numbers of unstructured georeferenced tweets using SOMs?

(RQ2): How are the explored variations in intensity and similarity of collective human activities related to (dis)similarities in the underlying urban structures?

This paper is structured as follows. We provide a short review of approaches for spatial analysis in Section 2. Section 3 describes our proposed methodology for extracting geospatial, temporal and semantic clusters from tweets. The results of the analysis answering the research questions are presented in Section 4, followed by a discussion of our methods and research results in Section 5. Section 6 has some final remarks and outlines possible future research directions.

2. Related work

This section lays out related work in the areas of spatiotemporal and semantic analysis methods (Section 2.1), Geo-SOM/H-SOM for tweet analysis (Section 2.2) and combined approaches (Section 2.3).

2.1. *Spatiotemporal and semantic analysis of Twitter data*

The exploration of *unstructured textual information* from tweet posts requires text mining methods and has been conducted using numerical statistics intending to create semantic weighting factors such as term frequency (TF) (Hecht *et al*. 2011), term frequency–inverse document frequency (TF–IDF) (Jackoway *et al*. 2011) and term-ranking algorithms (Gupta and Kumaraguru 2012). However, generated term matrices are sparse (Derczynski *et al*. 2013), since the textual information from tweet posts is semantically highly uncertain and therefore requires more sophisticated text-mining algorithms (Gelernter and Balaji 2013). Other approaches include a manual term and keyword filtering (Andrienko *et al*. 2013), dimensionality reduction through latent topic

modeling techniques (Kling *et al*. 2012) and the application of semantic classification algorithms such as named-entity recognition (NER) (Gelernter and Balaji 2013) or naïve Bayes (Zielinski and Bügel 2012).

The exploration of spatial information from georeferenced Twitter tweets requires methods of *spatial statistics* and *spatial analysis*. Point clusters have therefore been assessed by using Kalman filtering (Sakaki *et al*. 2010) and kernel density estimation techniques (Li and Goodchild 2012). Spatial cluster analysis for point data has been applied using density-based spatial clustering (DBSCAN) (Veloso and Ferraz 2011), *K*-means (Pan and Mitra 2011) and spectral clustering (Cranshaw *et al*. 2012).

The *limitations* of data clustering algorithms such as DBSCAN and *K*-means regarding required parameter inference, e.g., distance measures, minimum point reachability and number of clusters, have been illustrated by several studies (Birant and Kut 2007, Jain 2010). Other approaches investigate spatial characteristics of georeferenced social media data by simply aggregating point-based observations into grid cells (Feick and Robertson in press), buffer zones (Lenormand *et al*. 2014), Voronoi polygons (Lee and Sumiya 2010) or administrative bounding polygons (Crooks *et al*. 2015). However, when point-based measures are aggregated into artificial polygons or grid networks, these arbitrary constructs lead to the modifiable areal unit problem (MAUP) with respect to the real-world scale (Fotheringham and Wong 1991).

Therefore, the following introductory paragraph will describe the advantages of SOMs compared to existing clustering approaches by outlining the present research in order to synthesize the SOM methodology closer to the field of GIScience.

2.2. *Self-organizing maps, Geo-SOM and variants*

SOMs were first introduced by Kohonen (1982, 1990) as a powerful type of artificial neural network (ANN) which abstracts information from multi-dimensional primary signals and represent data properties in a two-dimensional topological connected output space. Openshaw *et al*. (1995) was one of the first geographers demonstrating the advantages of neural networks for geographic analysis in a selected case study of census classifications in Britain. Due to the growing complexity of spatial data and analysis tasks (Miller and Han 2009), SOM as an unsupervised machine-learning algorithm has been proven to be a performant ANN when comparing to classic data mining and cluster analysis approaches (Ultsch and Vetter 1995, Reusch *et al*. 2005, Watts and Worner 2009).

Related to GIScience, SOMs have been applied for spatial pattern detection (Spielman and Thill 2008, Gorricha *et al*. 2013) and spatial clustering (Skupin and Hagelman 2005). Furthermore, SOMs have been used for generalization purposes (Allouche and Moulin 2005, Sester 2005) and for exploratory data visualization (Vesanto 1999, Bruggmann *et al*. 2013, Sagl *et al*. 2014). A broad overview on several applications of SOMs within GIScience has been provided by Agarwal and Skupin (2008).

The Kangas Map (Kangas 1992) represents the first extension of the classic SOM approach by also considering the geographical neighborhood of an input feature. This best matching of geographical closer neurons has been further adapted by Bação *et al*. (2005) introducing a geographical tolerance parameter within their developed Geo-SOM framework.

Lampinen and Oja (1992) train a second-level map from the best-matching unit for each input vector in order to derive a hierarchical SOM (H-SOM). Henriques *et al*. (2012) depict the performance of an H-SOM to detect spatial clusters within a high-dimensional geographical dataset. Hagenauer and Helbich (2013) modified the hierarchical method to discover patterns in spatiotemporal data generating a hierarchical spatiotemporal SOM (HSTSOM). Feng *et al*. (2014) propose a further approach to combine Geo-SOM and H-SOM methods based on a divide and group principle to further explore geographic data.

2.3. *Combined approaches*

In the area of social network analysis, several research articles (Boulet *et al*. 2008, Couronne *et al*. 2013) have studied characteristics of individual users within large complex networks of social media platforms by applying different SOM variations to structure social relationships between users on a two-dimensional SOM output space. Bruggmann *et al*. (2013) detect latent semantic structures and relationships of user-generated content between Wikipedia articles and visualize thematic cluster using an SOM cartogram.

In the context of GIScience, Hagenauer *et al*. (2010) conduct a cluster analysis of point-based crime pattern by applying an SOM and further enhance the approach using an SOM as a text mining tool to classify unstructured citizen provided crime reports (Helbich *et al*. 2013). Sagl *et al*. (2014) demonstrated a combined SOM and spatial autocorrelation approach to explore collective human activities through mobile network traffic data.

However, beside a social network analysis of Twitter metadata (user profile, follower/following analysis), to the best of our knowledge no research articles exist to *use SOMs on highly dimensional, noisy Twitter data* for the exploration of human activities to characterize underlying urban morphological structures.

3. Methodology

The main novelty of the proposed approach is the combination of similarity distance-based concepts (as initially proposed by Resch *et al*. 2015) that exist in the domains of GIScience (i.e., Tobler's First Law) and computational linguistics (Aggarwal and Zhai 2012). Several sequential processing steps are applied to our tweets in order to compute the similarity between the three information layers: linguistic and semantic content, temporal information and geographic location.

The analysis framework described in Figure 1 includes three main steps after Twitter data retrieval: Twitter data pre-processing, similarity assessment for all information layers and the computation of an SOM in an unsupervised learning approach. We collected our own data in the form of georeferenced tweets as only these are usable for applying geographical analysis algorithms (see case study in Section 4). We used the Twitter Streaming Application Programming Interface (API) (https://dev.twitter.com/docs/api/streaming).

3.1. *Pre-processing*

In order to derive meaning from tweets, the text is pre-processed using natural language processing methods including tokenization, stemming and stop word filtering. This reduces the semantic dimension of raw tweets by allowing the creation of word vectors.

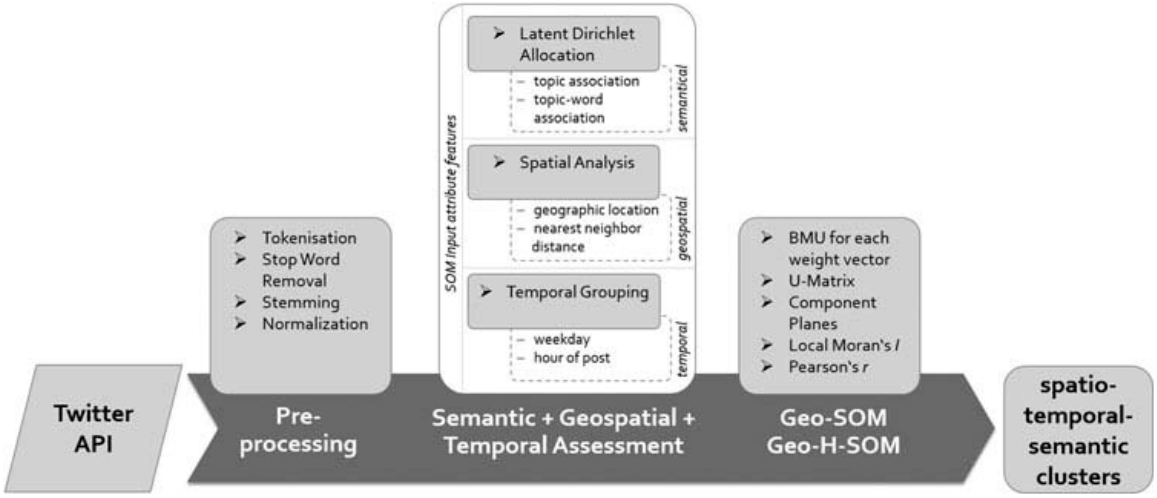

Figure 1. Analysis framework. The preprocessing step appears in Section 3.1. Semantic, geospatial and temporal assessment are described in Section 3.2 with results of the case study presented in Section 4.1. The Geo-SOM/Geo-H-SOM approach appears in Sections 3.3/3.4 and the corresponding results of the case study are described in Sections 4.2 and 4.3.

Within the tokenization process, cohesive strings from tweet posts are split up into single words ('tokens'). The advantages of this method have been pointed out by Metke-Jimenez *et al*. (2011). Afterwards the most common, frequently occurring 'stop words' (short-function words), not containing valuable information, are excluded to reduce the amount of noise among the remaining tokens. We used the standard stop word list from Lewis *et al*. (2004). Subsequently, the stemming process reduces all words to their stem, base or root form to simplify further analysis. The remaining tweet corpora are the input values for the following semantic similarity assessment.

3.2. *Semantic, geospatial and temporal similarity assessment*

We argue that tweets with similar linguistic features will also share high-semantic similarities. Thus, it is likely that they also contain similar semantic information which could be a potential indicator for coinciding social activities (Noulas *et al*. 2011).

In order to assess *semantic similarity*, we apply latent Dirichlet allocation (LDA) – a semantic probability-based topic extraction model (Blei *et al*. 2003). This unsupervised machine-learning model is a sophisticated method compared to previously used arbitrary keyword filtering techniques or word frequency driven approaches (e.g., TF/TF-ID) with a limited scalability (Aggarwal and Zhai 2012), since LDA identifies latent topics by clustering co-occurring words (bag-of-words model) from the given collection of tweets. The LDA model distinguishes between similar phrases with different contexts and assigns them to separate topics. LDA assumes that documents (in our case tweets) contain a random number of topics per document α, where each topic is characterized by a distribution over words β (Figure 2). z is the specific associated topic for an individual word w within each document, while θ denotes the topic distribution for the total number of documents M each of N length.

One of the main challenges when applying LDA is the posterior parameter estimation and computation of variables such as the number of topics k. Therefore, we are using

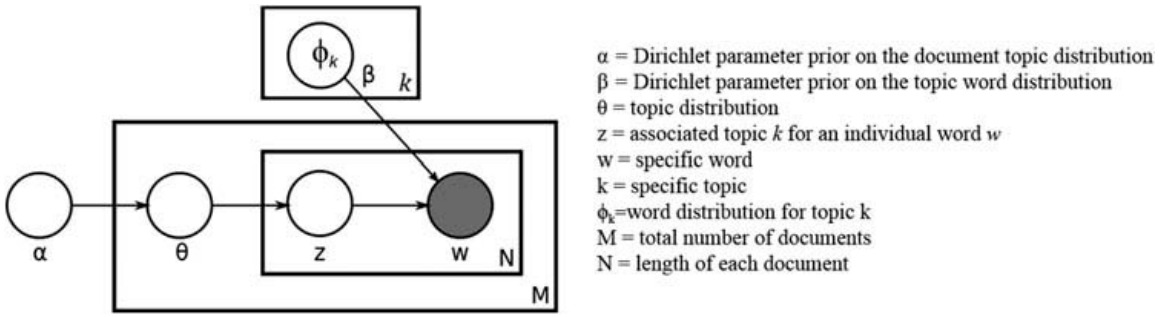

Figure 2. LDA graphical model according to Blei et al. (2003).

Gibbs sampling considering a Markov chain Monte Carlo for the LDA parameter inference, as proposed by Griffiths and Steyvers (2004). This sampling from probability distributions to obtain target distributions solves a key inference problem and optimizes parameter values (Figure 2). Each tweet's topic association (θ_M) and topic-word association (ϕ_k) are the semantic input components to the Geo-SOM.

For assessing *spatial similarity* of processed tweets, we analyze their geographic location and their geospatial distribution as follows. The geographic position (latitude and longitude) of every tweet is used as a geospatial input component to the Geo-SOM. Additionally, the nearest neighbor (NN) distance for each tweet is computed to assess whether tweets indicating similar semantic activities cluster geographically. We therefore consider the general topology or geospatial relationships between the tweets shown by point patterns and their mutual interaction distances as a representation of the underlying geospatial context. The Euclidean distance d_i between every feature i and its NN for each given topic is a measure of statistical spatial dispersion and constitutes a second Geo-SOM input component representing the geographic information layer,

$$NN = \sum_{i=1}^{n} d_i, \tag{1}$$

where n is the number of geographic features. Finally, the Geo-SOM considers the temporal component in order to assess temporal similarity of activities mentioned in tweets. Other than in exploratory time series analysis with time-related geospatial input vectors, i.e., trajectories, we do not have any prior knowledge regarding the observable underlying phenomena and their spatiotemporal patterns when using point observations like tweets (Hagenauer and Helbich 2013). Therefore, the main approach to discover temporal structures among our observations lies in hierarchically aggregating Geo-SOMs to observe textual dependencies within time periods. For this purpose, we create time bins covering (1) every hour of the day and (2) each day, as categorical variables to weight tweets higher when sharing similar activities in a geographical proximity and close in time. We derived these bin widths from our data sample size (in order to have a large enough number of points in each bin for the local clustering and subsequent Geo-SOM processes). However, our method is generically applicable to any aggregation interval (e.g., daily, weekly, monthly etc.) since all lower level Geo-SOMs are combined into one upper level Geo-H-SOM (see analysis framework Figure 1) to detect matching temporal patterns (Guimaraes 2000).

3.3. *Geographical SOM (Geo-SOM)*

The applied SOM learning approach (Kohonen 1990, Agarwal and Skupin 2008) uses the set parameter of Kohonen's standard algorithm and the Geo-SOM parameter extension (Bação *et al*. 2005). For the initial training process with random weights using input components for every information layer from the high-dimensional, geospatial, temporal and semantic feature attribute space, a 15 × 15 neuron network was setup, having fewer dimensions and within the limit of observations from the input space, as suggested by Kohonen (2001). The chosen network size has been validated in an iterative performance test following Feng *et al*. (2014). Based on an extracted random training cluster sample set for each Geo-SOM, the spatial cluster proximity (average NN distance) and attribute proximity have been computed together with the total amount of tweet points which fall within (commission) and outside (omission) these clusters. For the whole case study, the 15 × 15 neuron network size setup has shown the least amount of quantified omission errors.

For each input vector, the map node with the smallest Euclidean distance within a certain spatial neighborhood is defined, also known as best-matching unit (BMU). Additionally, the nodes in the neighborhood of the BMU are also updated and moved towards the direction of the input units (Figure 3) by repeating 10,000 training iterations and a fine tuning phase until convergence is reached, where the learning rate α decreases linearly from 0.5 to 0 and the final radius r is 0. During the Geo-SOM training phase, variations of the tolerance parameter k = [0,4] have been tested to expand the search radius for potential BMUs in order to increase/decrease the importance of geographical coordinates (see Bação *et al*. 2005 for tests on artificial datasets). A final tolerance radius of k = 2 appears to be the most suitable distance when weighting between attribute and geographic features having a tolerable distance between input and mapped output attribute space (quantization error) with the least amount of commission and omission errors within geographical space.

After training, the two-dimensional Geo-SOM output space with final BMU vectors represents fractions of the input space, while topologically preserving geographic,

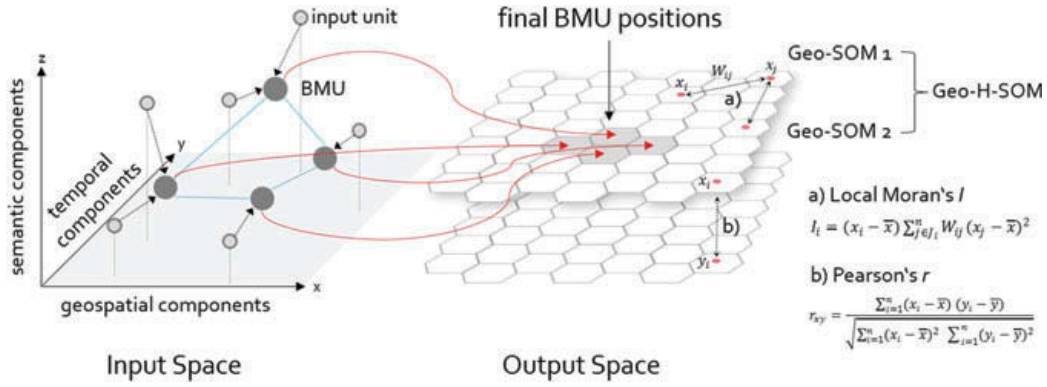

Figure 3. Input space (three-dimensional) with units (light gray dots) and map node's weight vectors having the smallest Euclidean distance (BMU as dark gray dots). The final BMU positions on the 2 × 2 sample unit Geo-SOM U-matrix output space (two-dimensional) are highlighted. Two Geo-SOMs weight vectors are bivariately (Pearson's r) correlated and also univariate spatially (Local Moran's I) autocorrelated.

semantic and temporal properties of the original Twitter dataset for each individual LDA topic. Geo-SOM parameters have been chosen by considering the feature characteristics of the dataset having the main goal to observe local cluster pattern on a medium size Geo-SOM (Kohonen 2001, Helbich *et al*. 2013).

The point features (tweets) are not aggregated into artificially delineated boundaries or polygons, which would lead to the MAUP (Fotheringham and Wong 1991). Thus, the Geo-SOM algorithm constitutes an analytical exploratory data mining process by abstracting the main input data's component characteristics themselves. The Geo-SOM also facilitates the consideration of numerous additional input variables (e.g., more topics). One of the main advantages of the Geo-SOM approach is the high scalability of the framework which enables us to assess input component characteristics and to incorporate different geographic scales during the analysis. As a result, the approach is able to overcome the limitations laid out in the introduction (e.g., high-dimensional uncertainty of data or multi-dimensional cluster characteristics) since the combined methodology also shows robust performance in handling noisy and uncertain Twitter data.

3.4. *Hierarchical Geo-SOM (Geo-H-SOM)*

In order to facilitate pattern detection over the whole dataset considering different levels of granularity (and due to a better computational efficiency tested by Feng *et al*. 2014), we use a thematic agglomerative H-SOM in the last step (see Figure 1), aggregating and merging all previous Geo-SOM results. By design, an SOM compresses high-dimensional input data into a two-dimensional map-like representation, where a Geo-SOM considers the first law of geography by geographically restricting the search radius and an H-SOM reveals information at different levels of detail. The combination of both approaches by ordering and dividing the input vectors enables the identification of meaningful cluster patterns. Therefore, we have chosen a combined Geo-SOM and H-SOM approach since the main goal of our research is to detect latent structures of locally occurring activities on a large geographic scale with the ability to assess these cluster characteristics. A classic SOM would infer structures regardless of their geospatial proximity, and one Geo-SOM over all tweets is not able to detect distinct local clusters since the input signal from tweets is high-dimensional. The H-SOM combines the separate Geo-SOM training results, where each Geo-SOM training result represents a semantic topic. Since the search of the BMU is limited to a defined maximal geographical neighborhood aiming to detect local clusters, all Geo-SOM BMU weight vectors are compared and grouped within the final merging H-SOM approach depicted by Henriques *et al*. (2012) to detect intra-urban human activity patterns on a city scale level.

The evaluation of our proposed Geo-H-SOM algorithm underlines the great potential of neural networks to perform spatiotemporal and semantic analysis on high-dimensional and large volume data types such as tweets, in order to reveal complex latent structures.

For a simpler interpretation of the Geo-SOMs' structures, hexagonal U-matrices are used as a visual representation of the distances between adjacent neighboring neurons using the trained vector derived from the input data dimension space. All trained Geo-SOM output neurons are converted and plotted as U-matrices into geospatial vector data structures (shapefiles). In this way, all calculated neuron distances can be easily interpreted and visually presented with commonly available desktop GIS.

Furthermore, the resulting SOM output space with each BMU's weight is correlated using R packages (cor, spdep) by computing Pearson's *r* as a spatial bivariate association measure, and Local Moran's *I* (Anselin 1995) as a univariate spatial association measure. That way, locations of spatial clusters and spatial outliers as well as the topological relationship among spatial entities are identified. The Local Moran's neighborhood size is defined as the distance between each neuron and its neighboring neurons.

4. Case study and results

This section presents the results of the analysis framework (Section 3), which has been applied to the case study described below (see Figure 1 for a visualization of the whole analysis framework). We first illustrate the results of the semantic, geospatial and temporal assessment (Section 4.1), then we lay out the results of the Geo-SOM analysis (Section 4.2) and the Geo-H-SOM analysis (Section 4.3).

For our case study, we use a dataset containing 41.2 million georeferenced tweets from the area of Greater London for one year. Table 1 provides further details regarding the used Twitter data. Since we are interested in geospatial, temporal and semantic analysis, we only query georeferenced tweets and did not restrict the data collection by any further type of keyword selection or (language) filter.

4.1. *Results semantic, geospatial and temporal similarity assessment*

In the first step of this assessment (see method Section 3.2), we classified all collected georeferenced tweets using LDA to analyze the *temporal-semantic characteristics*. The corresponding histograms, aggregating tweets in hourly intervals, show a periodical daily repeating signal of tweets classified according to the LDA extracted topics ($k = 8$) (Figure 4). The LDA model assigns probabilities of mutual word occurrences for each tweet to detect the most likely word overlaps. Figure 4 shows the words with the highest probabilities for each topic. For purposes of visualization efficiency, we only show the three words with highest probabilities.

From Figure 4, it is evident that topic 3 (T3) shows the highest amount of associated tweets after 6 pm, especially on weekends. The words 'game,' 'cup' and 'match' are selected by the LDA model as the most dominant words for describing this specific topic. These varying temporal-semantic frequencies indicate a linkage to real-world sports events since the temporal distribution is characteristically dispersed with single occurring peaks (league matches at the weekend, European cup matches during the week on Tuesdays and Wednesdays).

Table 1. Meta information for our selected Twitter dataset.

Dataset	Greater London (UK)
Bounding Box (WGS 84)	−0.303, 51.238, 0.554, 51.731
Time span	1 January 2014–31 December 2014
Covered area	3.265,387 km^2
Number of geotagged tweets after pre-processing	41.2 million
Number of individual users	476,071

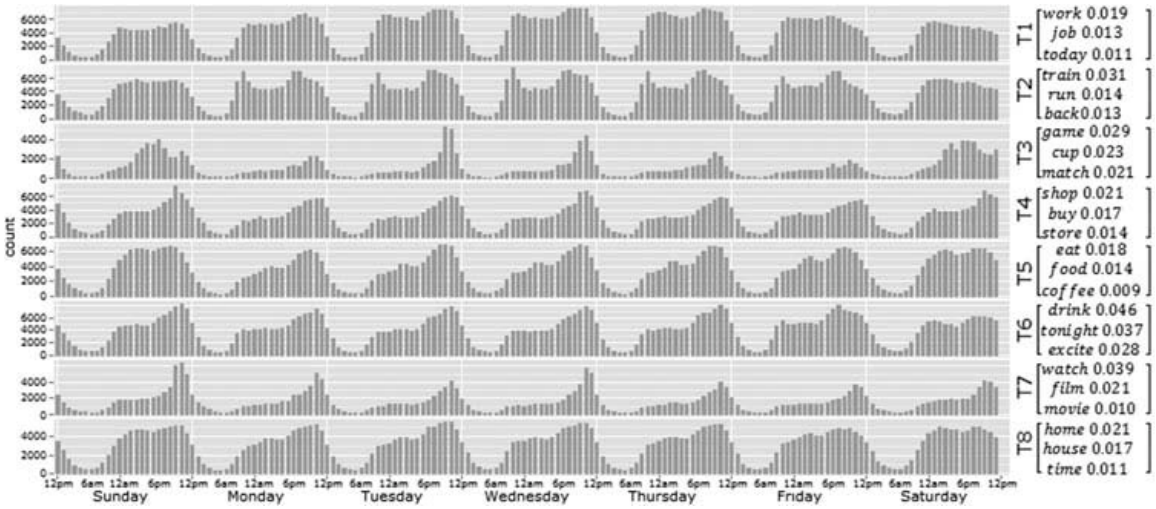

Figure 4. Temporal-semantic frequencies of all tweets posted within one year aggregated into weekdays consisting of one hour time intervals for eight LDA classified exemplarily topics.

In contrast, T2 (dominant words 'train,' 'run' and 'back') is characterized by a periodically repeating tweet signal. The majority of classified tweets are posted during weekdays between 8 am–10 am and 5 pm–8 pm, indicating morning and evening tweeting patterns. Note that for simplification in the descriptive parts of the paper we are only referring to each topic by labeling it with the most dominantly associated word.

In a second step, we analyze the *geospatial-semantic characteristics* of the dataset by exploring a distinctive geographical distribution for every topic. Figure 5 exemplarily illustrates the geospatial-semantic point densities for topic T2-train and T3-game. T2-train shows a more dispersed distribution and geospatially concentrates along the railway segments of London, whereas T3-game tweets mainly cluster within central London and at few distinct locations. The point densities reveal high peaks of T2-train in the vicinity of main public transport hubs.

4.2. *Results of the Geo-SOM*

As mentioned in Section 3, the geospatial, temporal and semantic assessment presented in Section 4.1 constitutes the basis for further analysis to uncover relationships and to

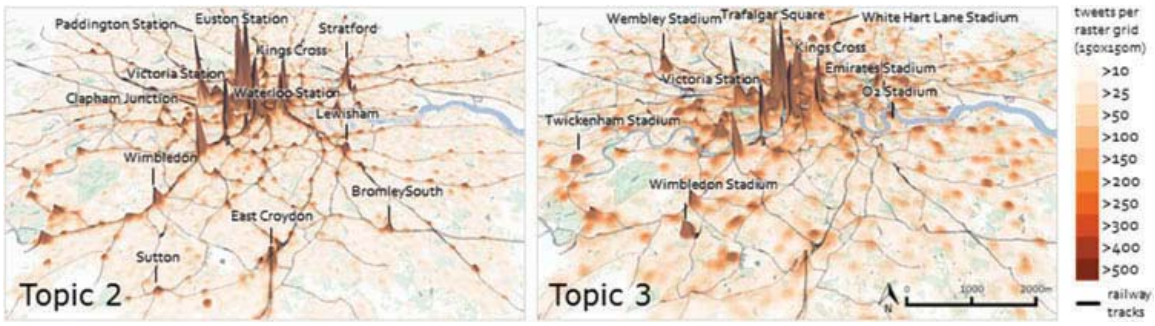

Figure 5. Geospatial-semantic point densities of LDA-classified tweets posted within one year vertically extruded as spikes and aggregated for T2-train and T3-game (base map: Stamen Design CC BY 3.0, data by OpenStreetMap CC BY SA).

identify cluster patterns of human activities using the Geo-SOM algorithm. As mentioned above, we used the following input features as distinct input components for the Geo-SOM analysis (see method Section 3.3): each tweet's day of post, hour of post, topic association, topic-word association, geographic location (latitude and longitude), and Euclidean NN distance.

For a simpler interpretation of the Geo-SOMs' structures, hexagonal U-matrices are used to plot the two-dimensional distances between neighboring neuron vectors. The distance between the adjacent neurons is indicated by intensities of gray: light gray (0.45–0.75 range) means that codebook vectors are close to each other in the input space; dark gray (>0.75) between the neurons indicates a large distance of codebook values. Thus, light gray (smaller values) indicate the presence of clusters since similar tweets are closer on the Geo-SOM compared to dissimilar ones. Dark U-matrix areas in between light areas can be interpreted as cluster separators and allow the visual distinction of characteristic tweet patterns. Component planes (CPs) (Kohonen 2001) show the relative values of each input variable's codebook vectors in order to identify correlations between input attributes. One can thus assess which components mainly characterize and contribute to the inferred Geo-SOM clusters.

Figure 6 shows the results of the Geo-SOM analysis: Figure 6(a) illustrates the U-matrix of the Geo-SOM (clusters are shown in different colors. Due to space constraints, we have only shown detailed visualizations of Geo-SOM G-T2 and G-T3, Figure 7(c) shows an overview of all Geo-SOM U-matrices) and the number of assigned tweets for the selected topics T2-train/ T3-game; Figure 6(b) presents a map of three exemplary clusters for G-T2 (Geo-SOM, Topic 2) and G-T3 with the highest amount of associated tweets visualized in geographical space (base map: Stamen Design CC BY 3.0, Data by OpenStreetMap CC BY SA); Figure 6(c) depicts the CPs for the Geo-SOMs; and Figure 6(d) shows the correlation matrix (each BMU's Pearson's *r* and Local Moran's *I* value) between all CP input variables, where red indicates strong positive correlation and blue stands for negative correlation. Crosses denote CP correlation values which are below the significance level of $p = 0.05$. The correlation dendrogram represents the distance or dissimilarity on the horizontal axis between each cluster grouped on the vertical axis.

The U-matrix of G-T2 reveals several separate clusters within the Geo-SOM output space which correspond to the geographical locations of major transportation hubs for Greater London (G-T2-$C_{1\text{-}29}$). For instance, the input attributes of clusters with the highest number of features (G-T2-$C_{10/16/18}$) are highly similar in time, geographical space and semantics. Hence, they have very close BMU weight vectors, whereas the surrounding neurons' weights are higher, which is indicated by darker hexagonal grids. The results of G-T3 indicate the presence of fewer clusters in comparison to G-T2. Clusters G-T3-$C_{1/2/5}$ are large and geospatially concentrated in the vicinity of major sports venues. A large cluster (G-T3-C_5) also converges within the areas of major public squares and public transport hubs. When directly comparing G-T2 and G-T3, overlapping cluster areas within the SOM, U-matrix output space can be visually identified. This indicates a link between activities related to sports events and major public transportation hubs (G-T2-$C_{10/11/12/16}$ and G-T3-C_5). When visualizing the geographic extent of three exemplary Geo-SOM clusters with the highest number of assigned tweets (Figure 6 (b)), a link between the real-world geospatial objects can be drawn.

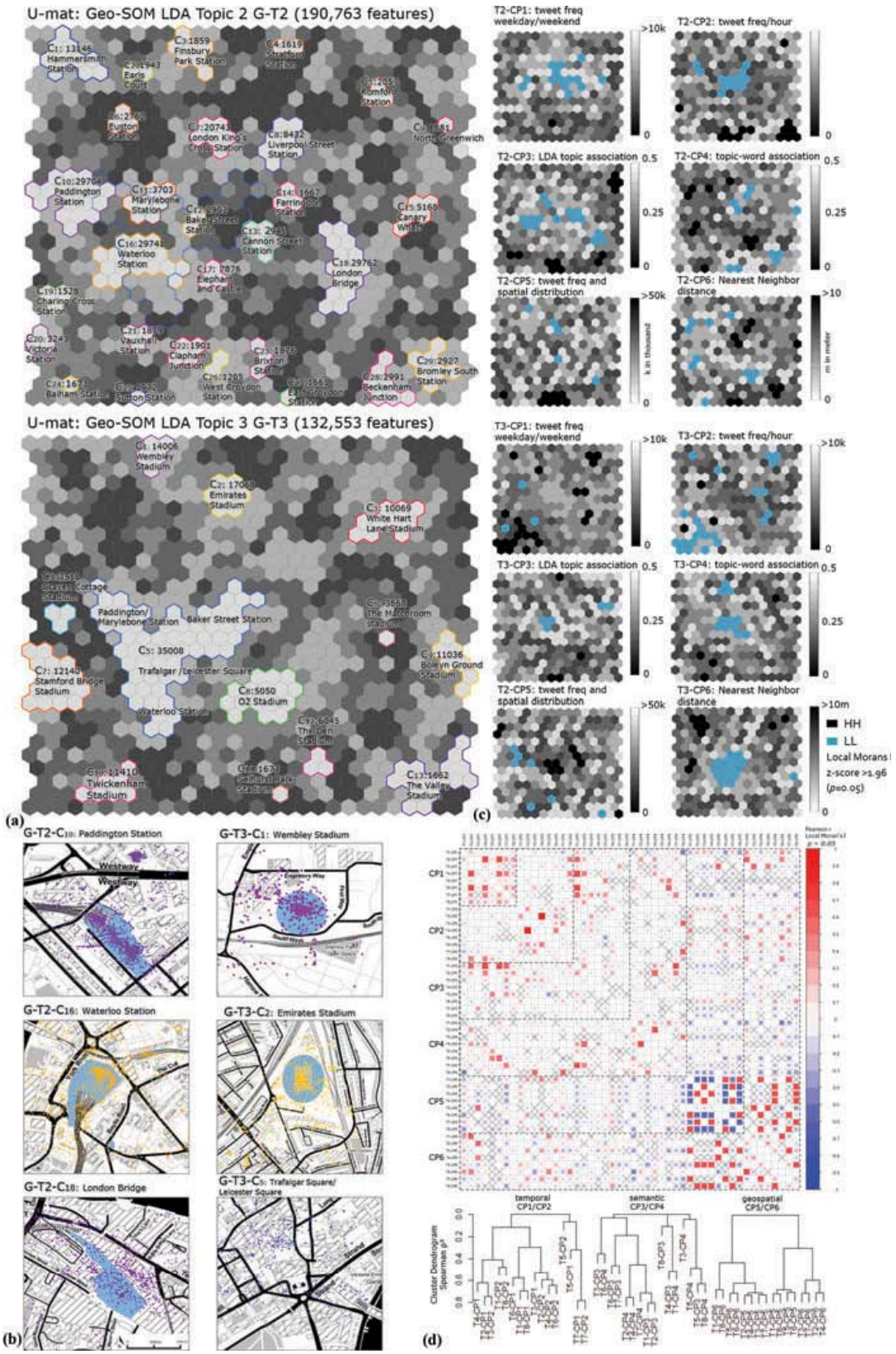

Figure 6. Results of the Geo-SOM analysis: (a) U-matrix of two exemplary Geo-SOMs G-T2 and G-T3; (b) map of three exemplary clusters for G-T2 and G-T3; (c) component planes for the Geo-SOMs; and (d) correlation matrix (each BMU's Pearson's r and Local Moran's I value). SOM G-T2 results represent 190,763 features (tweets), G-T3 132,553 features.

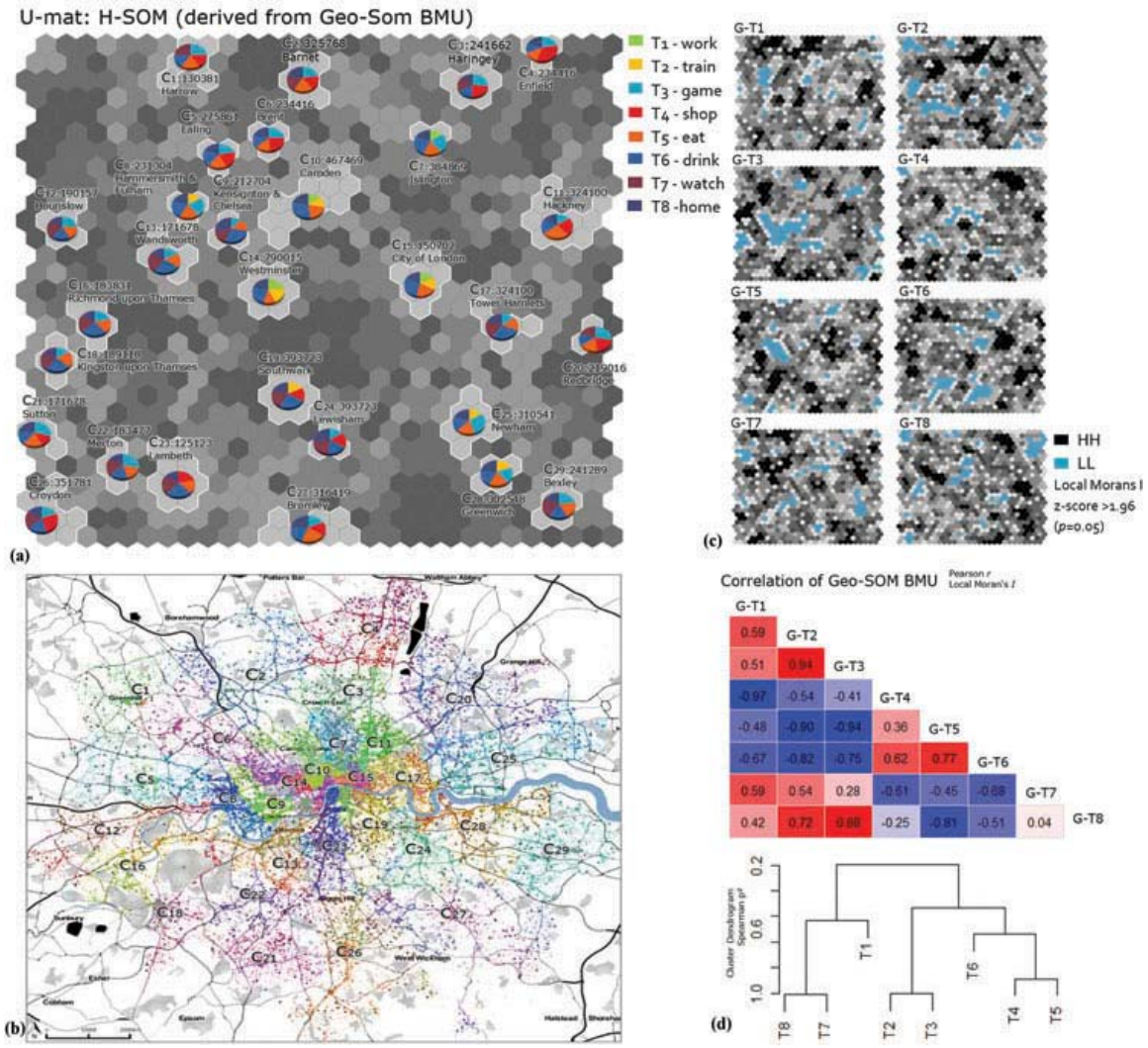

Figure 7. (a) Geo-H-SOM U-matrix results with derived clusters. (b) Links back the derived clusters of the Geo-H-SOM to the geographic space (base map: Stamen Design CC BY 3.0, data by OpenStreetMap CC BY SA). (c) All previously computed Geo-SOM U-matrices and hot- and cold-spots. (d) Correlation matrix of derived codebook vector distances from neighboring neurons (BMU) for each Geo-SOM LDA topic.

Figure 6(c) shows positive LISA among each CP neuron's weights with similar neighboring features. Considering the given Local Moran's $Z(I_i)$ scores and significance tests, one can distinguish between local spatial autocorrelation clusters of high values (HH) indicating distant neuron weight vectors, and low values (LL) indicating close neuron weight vectors. The latter ones are the intended clusters since similar input features are mapped close together whereas dissimilar ones are further apart within the CP space. A further comparison of the CPs' low-value spaces reveals strong resemblance between the geographic location (CP5 and CP6) and the semantic attribute features (CP3 and CP4). Several component structures appear: The majority of tweets during weekdays are geospatially widespread within the city center, and the amount of tweets decreases the further the inner city is away (T2-CP1/T2-CP2). In comparison, T3-CP1 shows fewer tweets during weekday and weekend periods with a dispersed geospatial structure and only a few distinct clusters. The topic and word association indicator for locations in the outskirts of London is very high, especially for the event-related topic T3-game

with distinct clusters surrounded by areas with a lower amount of tweets. The NN distance (CP6) reveals distinct clusters of landmarks (stadiums, train stations, squares, etc.) with a high amount of tweets being very close in geographic and semantic space (CP3/CP4) surrounded by more dispersed tweet point patterns. In general, we can conclude that tweets are more clustered within central London, which is highlighted by a strong similarity between the CPs T2-CP1/2/3 and CP6.

The correlation matrix in Figure 6(d) statistically proves the strength of the relationship between input CPs. Thus, we can conclude that all geographically correlating features either correlate within the semantic or the temporal domain. Features having either similar (positive *r*) or dissimilar (negative *r*) temporal attributes (CP2) also correlate correspondingly within the semantic dimension (CP3 and CP4). Observed topics with similar temporal characteristics (CP1 and CP2) also correlate within their geospatial distribution (CP5 and CP6).

Strong positive correlations can be discerned for all time-related topic attributes in CP1, except between T1-work and T3-game, T2-train and T3-game as well as T6-drink and T7-watch. Temporal attributes of CP1 for T1-work, T2-train, T3-game and T4-shop also show a link between their semantic attributes in CP3 and CP4. Furthermore, we observe a high temporal correlation between the hourly classified topic T2-train in CP2 and topic T4-shop. The semantic component plane CP3 shows a strong mutually positive correlation between T1-work and T2-train. CP4, which incorporates the LDA word association indicator, reveals similarly strongly semantic associations between T2 and T4-shop as well as T3-game and T8-home. CP5 and CP6, describing the geospatial attributes of the dataset, show a strong positive correlation between all topics. T1-work and T2-train, as well as T3-game and T7-watch highly correlate mutually in geographic space. Topics T3-game, T7-watch and T8-home along with T5-eat and T8-home have a strong tendency to form dense clusters in similar geographic places.

The hierarchical cluster analysis using Euclidean distances based on the correlation variables between all Geo-SOM CPs confirms that specific topics' attributes appear as a cluster branch within the temporal and geospatial clustered dendrogram (e.g., T2-CP1/CP2 with T3-CP1/2, and T2-CP5/6 with T3-CP5/6 plus T1-CP1/2 with T8-CP1/2 and T1-C5/6 with T8-CP5/6). Comparing all vertical dendrogram distances of these correlations, one can detect that hierarchical clusters of the geospatial CP are more similar to each other, compared to temporal and semantic CPs. Cluster branches within the temporal and semantic CPs are thus grouped highly similarly or dissimilarly (e.g., outlier T5-CP1/2 and T8-CP3/T3-CP4).

4.3. *Results of the Geo-H-SOM*

In the following H-SOM, the attributes of the input patterns (tweets) represented as an output subspace within each computed Geo-SOM are used to train and group a secondary unsupervised neural network (see method Section 3.4). This results in a Geo-H-SOM that incorporates a fraction of each Geo-SOM and the previously observed characteristic Twitter data pattern. Different clustering perspectives are extracted in the lower level Geo-SOM, which can then be merged into a global H-SOM, allowing the user to better understand and explore the overall emerging patterns (Henriques *et al*. 2012).

The U-matrix of all Geo-SOMs resulting in the final Geo-H-SOM are depicted in Figure 7: Figure 7(a) shows the H-SOM U-matrix results with derived clusters including the number of assigned tweets and the corresponding pie chart of the five most frequently occurring topics (shown in different colors) within the observed U-matrix clusters covering the administrative boroughs accordingly. Figure 7(b) links back the obtained clusters of the H-SOM to the geographic space (base map: Stamen Design CC BY 3.0, Data by OpenStreetMap CC BY SA). Figure 7(c) illustrates all previously computed Geo-SOM U-matrices and positive autocorrelating hot- and cold-spots. Figure 7(d) visualizes the correlation matrix of derived codebook vector distances from neighboring neurons (BMU) for each Geo-SOM LDA topic. The colors represent the resulting correlation coefficient $r_{combined}$ (each BMUs Pearson's r and Local Moran's I value), where red indicates a strong positive correlation and blue stands for a negative correlation.

The Geo-H-SOM in Figure 7(a) is an upper level SOM combining the attribute information from the previous Geo-SOMs shown in Figure(c) and enables the inference of larger geographical clusters and their underlying characteristics on a borough level. The following U-matrix structures appear.

A high amount of tweets classified as T1-work and T2-train geographically cluster in central London (H-C9/10/14/15) and are the most dominant set of topics. T4-shop/ T7-watch have been the most frequently occurring classified topics outside the city center, distributed across suburban boroughs in the peripheral area of Greater London. H-C10/14/15 contain the highest frequency of T5-eat and T6-drink classified tweets.

Specific H-C8/14/19/25/28 cover the highest amount of human mobility-related activities accompanied by 'leisure time'-related activities. Topics T3-game, T4-shop and T7-watch mostly occur in geospatial proximity of T8-home, whereas T1-work appears within clusters that are characterized by a high share of topics T2-train and T5-eat. The resulting H-SOM clusters are geographically visualized in b).

The comparison between the overall Geo-SOM U-matrices in Figure 7(c) shows that, for instance, G-T1 and G-T2 have a notable wider geospatial distribution than G-T3 where less BMU clusters exist. G-T7 and G-T8 have similar and close BMU's appearing in the outer U-matrix space. G-T1 (work), G-T2 (home) and G-T3 (train) visually match each other, emphasizing the indication of a latent urban activity structure among different SOM clusters within the study area.

In order to quantify the statistical relationship between each Geo-SOM and to investigate how the visually observed cluster structures of Twitter data correspond, all computed BMU distances for every hexagonal grid cell have been studentized and correlated. Figure 7(d) illustrates a positive Local Moran's and Pearson's correlation ($r_{combined} > 0.5$) between: G-T1 and G-T2; G-T2 and G-T3; G-T5 and G-T6; G-T3 and G-T8. These correlation values are shown in the upper and lower left corners of the correlation plot. The following pairs show a strong negative correlation ($r_{combined} < -0.5$): G-T1 and G-T4; G-T2 and G-T5; G-T2 and G-T6; G-T3 and G-T5; G-T3 and G-T6. This indicates a dissimilar BMU distance distribution, which is visible in the middle part of the correlation plot. Furthermore, we observe a strong bivariate correlation ($r > 0.8$) between the following pairs: G-T1 and G-T2; G-T3 and G-T8. These also show a similar geospatial distribution pattern as $Z(I_i) = 1.96$. The final hierarchical cluster analysis performed on the combined Pearson's and Moran's correlation matrix reveals three cluster branches in the dendrogram at about the same vertical distance.

5. Discussion

Summarizing the CP results (Section 4) for each retrieved topic-specific Geo-SOM, one can discern that the observed clusters have individually varying temporal, semantic and geospatial characteristics, which can be investigated using our proposed Geo-H-SOM approach (RQ1). The majority of input tweets share strong similarities across their temporal and semantic as well as geographical and semantic characteristics. The geographic space over all tweets is greatly homogenous compared to the other CP input layer. This supports the need to perform a cluster inference by incorporating all available geospatial, temporal and semantic attribute features. As one example, tweets classified as T4-shop and T5-eat related topics show a strong mutual correlation within their temporal-semantic attributes and form a distinct cluster outlier compared to all other topics. T1-work and T8-home related topics instead are highly correlating within their spatiotemporal attributes (but less in their semantic attributes), suggesting that people tweet about these activities at similar times in similar locations. Thus, these features would not be detectable as a cluster by solely focusing on the geographic space. Furthermore, some cluster characteristics might be omitted by leaving out one dimension. This demonstrates the usefulness of our combined approach for discovering relationships of human activities within all available input dimensions (RQ1).

The Geo-SOM results and their mutual correlation, e.g., between T1-work and T2-train, T3-game, T7-watch and T8-home or between T4-shop, T5-eat and T6-drink, revealed matching spatiotemporal and semantic latent urban activity structures across London (RQ2). In the final H-SOM, features with similar characteristics appear as merged clusters within the U-matrix space allowing for the investigation of predominant interests and activities of people around various public locations.

Due to space constraints, we have only shown exemplary visualizations of our Geo-SOM approach since further additional visualizations do not lead to more information content (other results show similar correlations). For reasons of rigorousness and completeness, the observed clusters have been statistically evaluated by comparing correlation measures of each individual Geo-SOM neuron grid weight. Nonetheless, several limitations during the conducted analysis need to be addressed.

One clear limitation of the current methodology for extracting urban activity clusters from social media is the assumption that tweets are written in situ, i.e., the posts' semantic content concerns the location and time at which the posts are published. The spatiotemporal and semantic input signal from tweets might also be too sparse since clusters can only be detected when there is a notable amount of observable tweets. Therefore only the most significant clusters (T2-train/T3-game) have been inferred in our research.

Regarding the pre-processing step (Section 3.1) and the following semantic similarity assessment (Section 3.2), the issue arises how efficiently natural language processing performs on highly semantically uncertain Twitter data, considering Internet-specific writing styles including abbreviations, acronyms, slang, etc. The LDA model also requires an initial parameter inference phase to define the number of topics assuming they have a probabilistic Dirichlet topic distribution. Tweets classified to T3-game in particular have shown the highest topic association indicators since the textual information covering sports event-related topics are more easily distinguishable from other tweets. This is a

critical aspect as the efficiency of the semantic classification process has an effect on the subsequent cluster assessment when analyzing other temporal and geospatial attributes. The generative LDA model considers tweets written in different languages since these posts would be associated into thematically coherent topics across multiple languages. However, within the Geo-SOM results these topics would form distinctive clusters, semantically covering the same topic and therefore require an ontology in order to match and link identical semantic topics written in different languages.

We referred to different kinds of uncertainty during Twitter analysis (Section 1) in order to state where outcomes might have an undesired effect or bias. The proposed framework performs robust with uncertain data, since, e.g., LDA considers the high-dimensionality of textual information, where a single tweet usually expresses a complex semantic topic and the observed semantic clusters are always modelled as a mixture of topics and associated words.

The result of the Geo-H-SOM algorithm (Sections 3.3 and 3.4) always depends on the given input attribute space as it considers the chosen network size with set training parameters and always needs to pre-determine which factors are relevant. The initial set radius of the neighborhood function k and the size of the SOM affect the output of the neural network. The Geo-H-SOM representation of geographical locations and the corresponding attribute space depends on the k parameter and varies from $k = 0$, (a proportional representation of solely geographical locations) via $k = 2$, (a mediation between observed local geographic clusters and their attribute space) up to $k \geq 4$, approximating the standard SOM by only preserving topological properties of the input attributes and thus neglecting the geographic space.

Since the Geo-SOM algorithm assesses similarity among nearby points and their attribute space, the robustness and performance of the technique to develop meaningful clusters depends highly on the spatiotemporal distribution of georeferenced tweets and their grade of sparseness.

Features that share similar attribute values will appear as one cluster and are therefore suitable to explore latent structures of highly multivariate input dimensions. However, the results cannot be considered to represent absolute real-world clusters of human activity patterns.

Furthermore, we compared the resulting Geo-SOM output space and each BMU's weight by computing Pearson's r and Moran's I. The detected links, e.g., between activities related to sports events and major public transportation hubs, have been inferred by statistically comparing spatial associations between Geo-SOMs and the correlation of their CPs representing temporal, semantic and geospatial characteristics of the input dataset. Therefore, the issue arises how comparable multiple SOM results are, since they only preserve topological relationships and a fraction of the original input space. During the analysis we used constant parameters with a random initialization for all Geo-SOM's. Moreover, since we applied the Geo-SOM algorithm, the BMU search is limited to a tolerance $k = 2$, where geographically distant features are less likely to be part of the same cluster (Bação *et al*. 2005). Thus, only geographically close neighboring neurons and their topological relationships are compared, indicating that the geographical context of the dataset is predominantly considered. Moreover, within those geospatially centered BMUs, we consider whether tweets are part of a geographically clustered or more dispersed set of observations by using the NN analysis results as an

additional geospatial Geo-SOM input component. Consequently, we further incorporate the spatial nature of geographic data and our empirical results for the given Twitter dataset have shown that the topological structures of different resulting Geo-SOMs correspond to each other since the observed clusters from the attribute space occur at similar geographic locations (Figure 6(c)). The distance between every input unit and the mapped training pattern after each iteration (quantization error, topographical error and geographical error) for every Geo-SOM were constant. Nevertheless, with our approach we can just detect coincidences of geospatial and non-geospatial input attributes having similar characteristics, not absolute causal relationships. Thus, further research in this area needs to be done in order to assess how the derived topologically close latent intra-urban human activity patterns cluster at similar geographic locations and to what degree they still reflect the original geographic properties of the input data.

6. Conclusion and future work

In this paper, we applied an explorative data mining approach to extract hidden relationships and latent structures of information regarding human activities to characterize urban activity structures from unstructured georeferenced Twitter data. As a result, our combined Geo-H-SOM model considers tweets to be 'similar' if the distance to each other is small in semantic space, in geographic space and in the time domain, following Resch *et al*. (2015). Thus, the paper demonstrates that 'human sensors' have the potential to become a powerful source of information towards the inference of human activities and the study of urban morphology (Crooks *et al*. 2015).

Answering RQ1 we have shown for our case study that similarities among spatiotemporal and semantic information reveal latent human activity patterns and are a proxy indicator for the characterization of underlying urban structures. The derived SOM results are dominated by the relationship between the identified semantic components and their temporal characteristics within geographic space.

As for RQ2, we can state that the extracted spatiotemporal and semantic activity clusters allow for inferring latent patterns with (dis)similar spatiotemporal characteristics. The applied unsupervised machine-learning approach enables clustering of high-dimensional uncertain georeferenced Twitter data without requiring any a priori information. The Geo-SOM algorithm is robust, reflecting the heterogeneous spatiotemporal distribution of tweets and is able to reduce the high-dimensional uncertain input attributes down to an easily interpretable data visualization and representation. Even in cases where the Geo-SOM initialization has been randomized, the output SOM weight vectors and their topological relationships were constant. We assessed the statistical dependence of obtained U-matrices and CPs by combining the bivariate (Pearson's r) and spatial association (Local Moran's I) of all neuron weights.

The NN analysis is an effective instrument to measure the degree of local spatial dispersion and allows also a distinction of close spatial clusters within the SOM neuron weights. The proposed analysis framework is therefore generic, scalable and provides a better understanding of human spatial interaction and latent urban activity structures by simultaneously exploring geographic space, time and the semantic attribute space from tweets. Furthermore we analyzed point observations intensities and varieties in

geographic- and attribute feature space and investigated their relationship in a combined approach.

It could potentially be used in other regions where limited knowledge about socio-economic processes within urban structures exists. Furthermore, since the proposed framework can handle high-dimensional datasets and extract semantic information, it might also be beneficial for other datasets containing unstructured textual information with similar characteristics like tweets.

The fundamental concept behind SOMs to learn and recognize patterns on any given dataset, together with the ability to handle multiple input variables from diverse information sources (like Twitter) without having explicit knowledge about urban structures, has a great potential for modeling and predicting certain behavior and relationships for various application domains, including the investigation of user activities and collective activity structures, the study and forecast of human mobility flows or the event detection and prediction within the application of disease-, health- and disaster management.

As one future research direction, exploring urban social dynamics using semantic information from social media in a more context-sensitive manner with metadata extracted from Twitter, such as user profiles, followers/following information, should also been considered. This additional knowledge would enable researchers to detect possible clusters of social interaction (i.e., people who share common interests at similar times and places) gaining further insights into urban morphology by exploring human behavior.

Acknowledgements

This research has been funded through the graduate scholarship program 'Crowdanalyser-spatiotemporal analysis of user-generated content,' supported by the state of Baden Wurttemberg. This research has been supported by the Klaus Tschira Stiftung gGmbH. We thank the anonymous reviewers for their constructive and helpful suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Enrico Steiger http://orcid.org/0000-0002-8648-9817

References

Agarwal, P. and Skupin, A., 2008. *Self-organising maps: applications in geographic information science*. Chichester: John Wiley & Sons.

Aggarwal, C. and Zhai, C., 2012. *Mining text data*. New York: Springer Science & Business Media.

Allouche, M. and Moulin, B., 2005. Amalgamation in cartographic generalization using Kohonen's feature nets. *International Journal of Geographical Information Science*, 19 (8–9), 899–914. doi:10.1080/13658810500161211

Andrienko, G., *et al*., 2013. Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, 15 (3), 72–82. doi:10.1109/MCSE.2013.70

Anselin, L., 1995. Local indicators of spatial association-LISA. *Geographical Analysis*, 27 (2), 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Bação, F., Lobo, V., and Painho, M., 2005. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computers & Geosciences*, 31 (2), 155–163. doi:10.1016/j.cageo.2004.06.013

Birant, D. and Kut, A., 2007. ST-DBSCAN: an algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60 (1), 208–221. doi:10.1016/j.datak.2006.01.013

Blei, D., Ng, A., and Jordan, M., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Boulet, R., *et al*., 2008. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71 (7–9), 1257–1273. doi:10.1016/j.neucom.2007.12.026

Bruggmann, A., Salvini, M.M., and Fabrikant, S.S.I., 2013. Cartograms of self-organizing maps to explore user-generated content. *In*: *26th international cartographic conference*, 25–30 August, Dresden. doi:10.5167/uzh-80972

Couronne, T., Beuscart, J., and Chamayou, C., 2013. Self-organizing map and social networks: unfolding online social popularity. *In*: *IEEE 24th of the international symposium on computer and information sciences*, 14–16 September, Nicosia. http://arxiv.org/ftp/arxiv/papers/1301/1301.6574.pdf

Cranshaw, J., *et al*., 2012. The livehoods project: utilizing social media to understand the dynamics of a city. *In*: *ICWSM*. AAAI.

Crooks, A., *et al*., 2015. Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, 29 (5), 720–741. doi:10.1080/13658816.2014.977905

Derczynski, L., *et al*., 2013. Twitter part-of-speech tagging for all : overcoming sparse and noisy data. *In*: *Proceedings of recent advances in natural language processing*, 7–13 September, Hissar, 198–206.

Feick, R. and Robertson, C., in press. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*.

Feng, -C.-C., Wang, Y.-C., and Chen, C.-Y., 2014. Combining Geo-SOM and hierarchical clustering to explore geospatial data. *Transactions in GIS*, 18 (1), 125–146. doi:10.1111/tgis.2014.18.issue-1

Fotheringham, A., Brunsdon, C., and Charlton, M., 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: John Wiley & Sons.

Fotheringham, A. and Wong, D., 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23 (7), 1025–1044. doi:10.1068/a231025

Gelernter, J. and Balaji, S., 2013. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17 (4), 635–667. doi:10.1007/s10707-012-0173-8

Gorricha, J., Lobo, V., and Costa, A., 2013. A framework for exploratory analysis of extreme weather events using geostatistical procedures and 3D self-organizing maps. *International Journal on Advances in Intelligent Systems*, 6 (1), 16–26.

Griffiths, T. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the national academy of sciences of the United States of America*, 101 (1), 5228–5235. doi:10.1073/pnas.0307752101

Guimaraes, G., 2000. Temporal knowledge discovery for multivariate time series with enhanced self-organizing maps. *In*: *Proceedings of the IEEE-INNS-ENNS international joint conference on neural networks*, 24–27 July, Como. IEEE, 165–170.

Gupta, A. and Kumaraguru, P., 2012. Credibility ranking of tweets during high impact events. *In*: *Proceedings of the 1st workshop on privacy and security in online social media*, 17 April, Lyon. ACM. doi:10.1145/2185354.2185356

Hagenauer, J. and Helbich, M., 2013. Hierarchical self-organizing maps for clustering spatiotemporal data. *International Journal of Geographical Information Science*, 27 (10), 2026–2042. doi:10.1080/13658816.2013.788249

Hagenauer, J., Helbich, M., and Leitner, M., 2010. Visualization of crime trajectories with self-organizing maps : a case study on evaluating the impact of hurricanes on spatio- temporal crime hotspots. *In*: A. Ruas, ed. *Proceedings of the 25th conference of the international cartographic association*, 3–8 July 2011, Paris.

Hecht, B., *et al*., 2011. Tweets from Justin Bieber's heart: the dynamics of the 'location' field in user profiles. *In*: *Proceedings of the SIGCHI conference on human factors in computing systems*, 7–12 May, Vancouver, BC. ACM, 237–246.

Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica*, 47 (1), 153–161. doi:10.2307/1912352

Helbich, M., *et al*., 2013. Exploration of unstructured narrative crime reports : an unsupervised neural network and point pattern analysis approach. *Cartography and Geographic Information Science*, 40 (4), 326–336. doi:10.1080/15230406.2013.779780

Henriques, R., Lobo, V., and Bação, F., 2012. Spatial clustering using hierarchical SOM. *In*: M. Johnsson, ed. *Applications of self-organizing maps*, Rijeka: INTECH Open Access, 231–250.

Jackoway, A., Samet, H., and Sankaranarayanan, J., 2011. Identification of live news events using Twitter. *In*: *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks - LBSN '11*, 1 November, Chicago, IL. New York: ACM, 248–260.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31 (8), 651–666. doi:10.1016/j.patrec.2009.09.011

Kangas, J., 1992. Temporal knowledge in locations of activations in a self-organizing map. *In*: I. Aleksander and J. Taylor, eds. *Artificial neural networks*. Vol. 2. Amsterdam: Elsevier, 117–120.

Kling, F., Kildare, C., and Pozdnoukhov, A., 2012. When a city tells a story: urban topic analysis. *In*: *Proceedings of the 20th international conference on advances in geographic information systems*, 7–9 November, Redondo Beach, CA. New York: ACM, 482–485.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43 (1), 59–69. doi:10.1007/BF00337288

Kohonen, T., 1990. The self-organizing map. *Proceedings of the IEEE*, 78 (9), 1464–1480. doi:10.1109/5.58325

Kohonen, T., 2001. Self-organizing maps. *In*: *Springer Series in Information Sciences*, Vol. 30. Berlin: Springer-Verlag. doi:10.1007/978-3-642-56927-2

Lampinen, J. and Oja, E., 1992. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2 (2–3), 261–272. doi:10.1007/BF00118594

Lee, R. and Sumiya, K., 2010. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. *In*: *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks – LBSN '10*, 3–5 November, San Jose, CA. New York: ACM, 1–10. doi:10.1145/1867699.1867701

Lenormand, M., *et al*., 2014. Tweets on the road. *PloS One*, 9 (8), e105407.

Lewis, D.D., *et al*., 2004. RCV1: a new benchmark collection for text categorization re-search. *The Journal of Machine Learning Research*, 5, 361–397.

Li, L. and Goodchild, M.F., 2012. Constructing places from spatial footprints. *In*: M. Goodchild, D. Pfoser, and D. Sui, eds. *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, 7–9 November, Redondo Beach, CA. 15–21.

Metke-Jimenez, A., Raymond, K., and MacColl, I., 2011. Information extraction from web services: a comparison of Tokenisation algorithms. *In*: *2nd international workshop on software knowledge*, 26 October, Paris. doi:10.5220/0003698000120023

Miller, H. and Han, J., 2009. Geographic data mining and knowledge discovery. *In*: *Handbook of geographic information science*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 352–366.

Miller, H.J. and Goodchild, M.F., 2015. Data-driven geography. *GeoJournal*, 80 (4), 449–461. doi:10.1007/s10708-014-9602-6

Noulas, A., *et al*., 2011. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *In*: *The social mobile web 11*, 21 July, Barcelona. AAAI.

Openshaw, S., Blake, M., and Wymer, C., 1995. Using neurocomputing methods to classify Britain's residential areas. *Innovations in GIS*, 2, 97–111.

Pan, -C.-C. and Mitra, P., 2011. Event detection with spatial latent Dirichlet allocation. *In*: *Proceeding of the 11th annual international ACM/IEEE joint conference on digital libraries – JCDL '11*, 13–17 June, Ottawa, 349.

Resch, B., *et al*., 2015. Urban emotions – geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced Data. *In*: G. Gartner and H. Huang, eds. *Progress in location-based services 2014*. Cham: Springer International Publishing, 199–212.

Reusch, D., Alley, R., and Hewitson, B., 2005. Relative performance of self-organizing maps and principal component analysis in pattern extraction from synthetic climatological data. *Polar Geography*, 29 (3), 188–212. doi:10.1080/789610199

Roick, O. and Heuser, S., 2013. Location based social networks - definition, current state of the art and research agenda. *Transactions in GIS*, 17 (5), 763–784.

Sagl, G., Delmelle, E., and Delmelle, E., 2014. Mapping collective human activity in an urban environment based on mobile phone data. *Cartography and Geographic Information Science*, 41 (3), 272–285. doi:10.1080/15230406.2014.888958

Sakaki, T., Okazaki, M., and Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *In*: *Proceedings of the 19th international conference on World wide web*, 26–30 April, Raleigh, NC. New York: ACM, 851–860.

Sengstock, C. and Gertz, M., 2012. Latent geographic feature extraction from social media. *In*: *Proceedings of the 20th international conference on advances in geographic information systems - SIGSPATIAL '12*, 7–9 November, Redondo Beach, CA. New York: ACM Press, 149.

Sester, M., 2005. Optimization approaches for generalization and data abstraction. *International Journal of Geographical Information Science*, 19 (8–9), 871–897. doi:10.1080/13658810500161179

Skupin, A. and Hagelman, R., 2005. Visualizing demographic trajectories with self-organizing maps. *GeoInformatica*, 9 (2), 159–179. doi:10.1007/s10707-005-6670-2

Spielman, S. and Thill, J., 2008. Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, 32 (2), 110–122. doi:10.1016/j.compenvurbsys.2007.11.004

Steiger, E., Albuquerque, J.P.D., and Zipf, A., 2015. An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS*. doi:10.1111/tgis.12132

Ultsch, A. and Vetter, C., 1995. *Self-organizing-feature-maps versus statistical clustering methods: a benchmark*. Research Report. Marburg: University of Marburg.

Veloso, A. and Ferraz, F., 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *In*: *Proceedings of the 3rd international web science conference*, 15–17 June, Koblenz. ACM, Article No. 3.

Vesanto, J., 1999. SOM-based data visualization methods. *Intelligent Data Analysis*, 3 (2), 111–126. doi:10.1016/S1088-467X(99)00013-X

Watts, M. and Worner, S., 2009. Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. *Ecological Modelling*, 220 (6), 821–829. doi:10.1016/j.ecolmodel.2008.12.016

Zandbergen, P.A. and Barbeau, S.J., 2011. Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. *Journal of Navigation*, 64 (03), 381–399. doi:10.1017/S0373463311000051

Zheng, Y., 2011. *Location-based social networks: users. Computing with spatial trajectories*. New York, NY: Springer.

Zielinski, A. and Bügel, U., 2012. Multilingual analysis of Twitter news in support of mass emergency events. *In*: L. Rothkrantz, J. Ristvej, and Z. Franco, eds. *ISCRAM'12: Proceedings of the 9th international ISCRAM conference*, 22–25 April, Vancouver, 1–5.