Computers, Environment and Urban Systems 54 (2015) 255–265

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data

Enrico Steiger *, René Westerholt, Bernd Resch, Alexander Zipf

GIScience Research Group, Institute of Geography, Heidelberg University, Germany

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form 8 September 2015
Accepted 17 September 2015
Available online xxxx

Keywords:
Crowdsourcing of human activities
LBSN
Twitter
Spatial autocorrelation
Semantic topic modeling

ABSTRACT

Detailed knowledge regarding the whereabouts of people and their social activities in urban areas with high spatial and temporal resolution is still widely unexplored. Thus, the spatiotemporal analysis of Location Based Social Networks (LBSN) has great potential regarding the ability to sense spatial processes and to gain knowledge about urban dynamics, especially with respect to collective human mobility behavior. The objective of this paper is to explore the semantic association between georeferenced tweets and their respective spatiotemporal whereabouts. We apply a semantic topic model classification and spatial autocorrelation analysis to detect tweets indicating specific human social activities. We correlated observed tweet patterns with official census data for the case study of London in order to underline the significance and reliability of Twitter data. Our empirical results of semantic and spatiotemporal clustered tweets show an overall strong positive correlation in comparison with workplace population census data, being a good indicator and representative proxy for analyzing workplace-based activities.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Cities are multifunctional complex systems serving as major hubs for a number of human social activities. With more than half of the world's population living in urban areas and a continuing urban growth (United Nations Population Fund, 2008), the capability to provide viable service infrastructure (roads, public transport, energy supplies, etc.) for the majority of people is a rising challenge. The characterization of urban structures can facilitate urban and transportation planning processes providing valuable information, which helps to predict the increased pressure on existing urban infrastructures. Regular commuting from workplaces to places of residence, and activities originating from these areas, are major examples of daily routines within urban areas, influencing human mobility and affecting transportation planning. In the UK in 2013, a person on average made 145 trips with 19% of all trip purposes related to business and commuting activities (Department for Transport, 2014).

Determining the frequency and spatial distribution of travel origins and destinations for every trip purpose is a principal quantitative study area currently carried out by mobility surveys (Morris, Humphrey, & Tipping, 2014). However, they are expensive in terms of the required labor input and usually lead to limited sample sizes. Thus, the investigation of typically larger spatiotemporal human activity clusters obtained from crowdsourced information may help to understand commuting patterns and reveal specific urban structures such as workplace concentrations.

In this context, emerging, inexpensive and widespread sensor technologies have created new possibilities to infer mobility data for exploring urban structures and dynamics. This growing availability of mobile devices equipped with GPS sensors having broadband internet access, allows users to actively participate and create content through mobile applications and location-based services (ITU, 2014).

Particularly georeferenced Twitter data is a promising opportunity to understand geographic processes inside online social networks. The enormous potential of interactive social media platforms like Twitter has been increasingly recognized by numerous research domains over the last years. Although there is a growing research body using Twitter data to analyze urban processes, empirical research towards the validation of human social activities revealing urban structures and human mobility patterns using crowdsourced information is still widely unexplored (Resch, Beinat, Zipf, & Boher, 2012).

In a previous study we introduced a semantic and spatial analysis method, through which we were able to extract human mobility flows from uncertain Twitter data (Steiger, Ellersiek, & Zipf, 2014). However, it remains to be investigated whether we can find similar semantic layers that represent collective human behavior in co-occurrence with underlying social activity.

Therefore, research question (RQ1) investigates the possibility of exploring urban structures through characterizing spatiotemporal and semantic patterns of human social activities. Hence, we extract topics covering work-related and home-related activities that reflect typical collective human behavior (e.g., city-scale human mobility). Thus, the

* Corresponding author at: Institute of Geography, Heidelberg University, Berliner Straße 48, D-69120 Heidelberg, Germany.
E-mail address: enrico.steiger@geog.uni-heidelberg.de (E. Steiger).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.09.007
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

first research question aims to find evidence for the reflection of collective behavior in tweets. In a further step, the second research question (RQ2) seeks to validate these findings against reliable census data. In particular, we examine associations and correlations between tweets as a proxy indicator of human social activities and available census populations. Summarizing, the main goal of this paper is to validate the detected human social activity clusters from RQ1 with official UK census data.

We have chosen London to be a reasonable study site, given the vast number of Twitter users in this city, providing us with a large enough data sample for our research. This second research question is particularly important against the background of a broad range of uncertainties that arise with Twitter data analysis (s. sub-section 2.1). To the best of our knowledge, no available study has conducted this kind of validation between semantic information extracted from Twitter data and official census information. We aim to provide a first empirical ground truth on how representative and trustworthy tweets for the inference of social activities indicating human mobility are. We propose a suitable methodological approach for answering these questions.

2. Background

The dataset used in this analysis is collected from Twitter. Within online social networks like Twitter, individuals can create an online profile and communicate with other users by sharing common ideas, activities, events or interests (Boyd & Ellison, 2007). Twitter further enhances existing social networks by adding a spatial dimension becoming a LBSN and allows users to exchange details of their personal location as a key point of interaction (Zheng, 2011). Users can post short status messages, namely tweets with up to 140 characters. With the permission of the user, each tweet contains a corresponding geolocation acquired from the GPS sensor within the mobile device. Therefore, user posts in Twitter represent a spatiotemporal digital footprint (geolocation and timestamp of tweet) with a semantic information layer (content of tweet message).

Georeferenced tweets correspond to particular locations and are influenced by each user's individual perception of urban space (Fig. 1). Thus, Twitter data and specific contextual information might serve as an indicator on how strongly the virtual and physical worlds are connected with each other. However, unlike with Foursquare where users can "check in" at predefined venues (restaurants, hotels, etc.), we do not have any a priori knowledge regarding underlying human social activities in Twitter. One interesting question is therefore concerned with investigating whether single tweets denoting a specific semantic incident tend to co-occur with similar other tweets being close in geographic space and time. Such clustering behavior might provide converging evidence about underlying social activity. Our given example tweet "I'm at work" (see Fig. 1), for instance, indicates a particular human social activity which may characterize an underlying urban structure. In this case a possible indication of a workplace.

2.1. Potential limitations of Twitter data analysis

When analyzing spatiotemporal and semantic information from Twitter, we face several data-specific uncertainties, including a number of components such as the location information, the extracted knowledge and the applied methodology.

2.1.1. Location uncertainty

The location information retrieved by built-in GPS receivers might be inaccurate due to different effects. These include intrinsic effects like adverse mobile device characteristics, but also extrinsic factors such as the built environment or the GPS dilution of precision (Zandbergen & Barbeau, 2011). Furthermore, users can individually choose to add their precise location to a tweet or just a general attached location information (such as a city or neighborhood). This might result in imprecise and coarse location information of geotagged tweets.

2.1.2. Sampling Biases

The spatial distribution of tweets in location-based online social networks is also spatiotemporally heterogeneous as users do not contribute records equally across space and time. Particularly the spatial distribution of tweets strongly varies on different real-world scale levels (country, city, borough, etc.) and might be too sparse in some geographical areas (Sengstock & Gertz, 2012). Moreover, when focusing on the ratio between the number of active Twitter users and the overall population, there is also a mismatch between the population and the sampling frame. This effect might lead to exclusion or under/over-representation of certain population groups (Heckman, 1979). In consequence, unrepresentative subsets and different sample sizes from the whole amount of tweets might be generated depending on the Twitter information and analysis approaches (e.g., only georeferenced tweets).

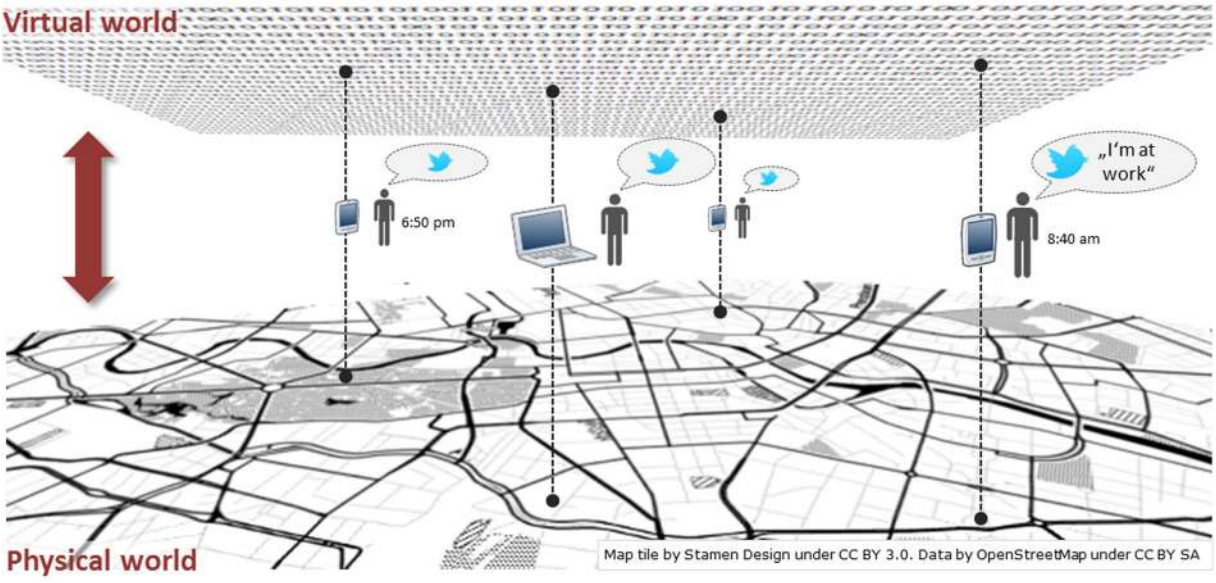

Fig. 1. Information layers according to Resch et al. (2012).

2.1.3. Uncertain semantic knowledge

We are facing a number of uncertainties within the extracted knowledge from Twitter data. Users might not post information in Twitter ad hoc, i.e., they are rather referring to past or future events. Furthermore, the textual component of tweets is a cohesive string of words. These word vectors are vague and semantically uncertain, so they might only be a weak indicator of a positive real-world observation. In other words, people using Twitter have individual motivations to post information and are also perceiving, identifying and interpreting their surroundings differently from each other, deciding *what* is worthwhile tweeting.

2.1.4. Uncertainties caused by applied methods

The exploration of spatial and semantic information from georeferenced Twitter tweets requires methods from spatial statistics as well as semantic analysis, given the dataset and its measurement characteristics (see further up in this sub-section). This results in uncertainties within the application of spatial and semantic methods. Although we do not know about the distribution of specific geographic phenomena and their semantic complexities within tweets, parameter estimations like distance measures, minimum point reachability and number of clusters are required to perform clustering. Moreover, the semantic parameter inference processes involved (especially in non-spatial methods) oftentimes assume stationarity across geographic space. However, this is an unrealistic assumption given the high degree of heterogeneity (e.g., due to topography or socio-economic factors) especially in urban environments (Fotheringham & Wong, 1991).

2.2. Related work

The exploration of spatiotemporal patterns to characterize human mobility and social interactions is one of the main research goals across a variety of disciplines and has become a major study focus due to the increasing, widespread use of wireless communication devices in recent years (Giannotti & Pedreschi, 2008). Investigating processes that influence the interaction between humans and the urban environment is key to understanding specific mobility patterns and their spatiotemporal distribution. Therefore, numerous studies focus on extracting individual and collective human daily activity patterns by analyzing crowdsourced information, such as taxi trip records (Liang, Zheng, Lv, Zhu, & Xu, 2012), GPS traces (Azevedo & Bezerra, 2009; Jiang, Yin, & Zhao, 2009) or large set of mobile phone records (Candia & González, 2008; Gao, 2014). Researchers investigate the spatiotemporal character of human behavior to find daily and weekly reoccurring clusters of frequently visited locations as an indicator of the underlying trip purposes (Bagrow & Koren, 2009; Phithakkitnukoon & Horanont, 2010). More, the estimation and inference of home and workplace locations to derive regular commuting profiles in particular has gained significant research attention due to its impact on transportation planning, reflecting the bimodal nature of human mobility (Kung, Greco, Sobolevsky, & Ratti, 2014).

Furthermore, a significant body of literature exists in using crowdsourced human mobility data from social media to examine the relationship between social activities as indicators for underlying human mobility behavior. Several studies (Cheng, Caverlee, Lee, & Sui, 2011; Cranshaw, Schwartz, Hong, & Sadeh, 2012; Hasan, Zhan, & Ukkusuri, 2013) use social media check-in data from Foursquare to analyze collective human mobility and activity patterns to infer urban (Wakamiya, Lee, & Sumiya, 2011) and user-specific characteristics (Noulas, Scellato, Mascolo, & Pontil, 2011). A validation of these results with cellphone locations revealed similar collective movement patterns of people showing spatial and social proximity (Cho, Myers, & Leskovec, 2011). Kling, Kildare, and Pozdnoukhov (2012) and Ferrari, Rosi, Mamei, and Zambonelli (2011) followed a similar approach with Twitter data and also extracted urban motion patterns with identified spatiotemporal activity hot spots within the city. Regarding human mobility analysis from Twitter, Hawelka et al. (2014) and Li, Goodchild, and Xu (2013) found a correlation between tweet locations and certain socioeconomic characteristics of people. Furthermore, the estimation of work/home locations (Krumm, Caruana, & Counts, 2011) and related mobility flows have revealed similar patterns when compared with community survey data (Gao, 2014). Andrienko and Andrienko (2013) additionally correlated the distributions of places where people tweet with US population densities ($r = 0.52$).

Summarizing, we can state that the majority of studies within the current research focus of human mobility analysis using social media have investigated spatiotemporal distributions to infer social activities by analyzing the textual component of posts without validating these inferred spatiotemporal and semantic human activity clusters. We address this research gap by investigating the *reliability of spatial, temporal and semantic information* indicating specific human social activities and whereabouts of people from georeferenced tweets.

3. Methodology

Several sequential processing steps are necessary for finding overlapping human social activity patterns among tweets. Fig. 2 provides an overview of our analysis framework that comprises three main steps after twitter data retrieval: data pre-processing, semantic similarity assessment, and spatiotemporal autocorrelation analysis. First of all, tweets are collected in real time using the official Twitter streaming API.[1] Since we are interested in geospatial, temporal and semantic analysis, we only queried georeferenced tweets within our study area (Table 1) and did not restrict the data collection by any further type of keyword selection or filter.

3.1. Pre-processing

Initial pre-processing is necessary to reduce the semantic dimension of noisy raw tweets by generating word vectors. The semantic tweet content is therefore processed to remove whitespaces and punctuations. In the next step, all tweet corpora from Twitter undergo a natural language processing step by applying tokenization, stemming and stop word removal. The raw tweets are a cohesive string of words and are therefore split up into single words through tokenization. The advantage and performance of this approach has been described by Metke-Jimenez, Raymond, and MacColl (2011). Afterwards, common stop words are filtered out as many frequently occurring "small" words do not contain any valuable information. This is why they are excluded to reduce the amount of noise in the tweet content. We use a standard stop word list as suggested by Lewis, Yang, Rose, and Li (2004).The remaining tweet corpus act as the input for the following semantic similarity assessment.

3.2. Semantic similarity assessment

Semantic similarity among tweets for a given associated topic is an indicator for coinciding human social activity (Noulas et al., 2011) (RQ1). In order to assess semantic similarity we use Latent Dirichlet Allocation (LDA), a semantic probability-based topic extraction model introduced by Blei, Ng, and Jordan (2003). This unsupervised machine learning model identifies latent topics and corresponding word clusters from our large collection of tweets and has been applied in previous studies (Pan, 2011; Kling et al., 2012) This technique reduces the semantic dimensions and works efficiently, especially on large datasets given a previous training set by clustering co-occurring words into topics ("bag-of-words" model). It is a sophisticated method, particularly in comparison to arbitrary simple keyword filtering techniques which have limited scalability (Becker & Gravano, 2011; Jackoway, Samet, &

[1] https://dev.twitter.com/docs/api/streaming

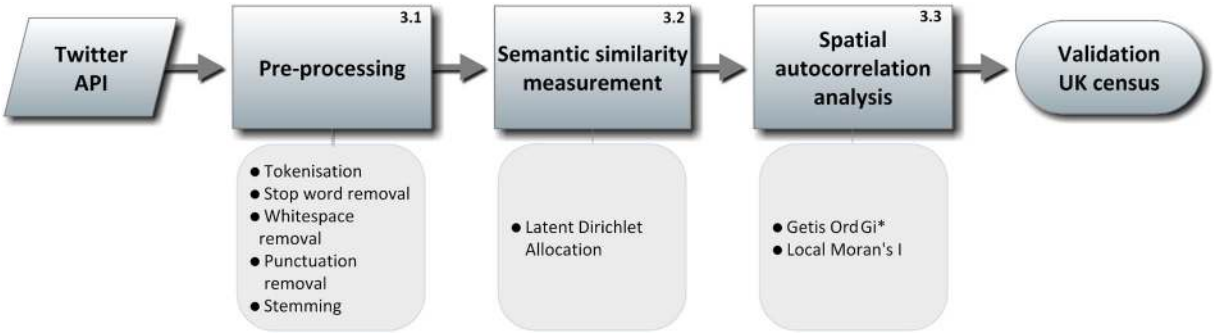

Fig. 2. Overview of the analysis framework.

Sankaranarayanan, 2011). The LDA model is able to distinguish between and assign similar phrases with different context into separate topics. Other than quantified weighted word vectors representing term-frequency and inverse document frequency (TF-IDF) scores, LDA assumes that documents (in our case tweets) contain a random number of latent topics per document α, whereby each topic is characterized by a distribution over words β (Fig. 3). Parameter z is the specific associated topic for an individual word w within each document. In contrast, θ denotes the topic distribution for the overall number of documents M, each being of length N.

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 use Gibbs sampling (a Markov chain Monte Carlo method) for LDA parameter inference. This sampling method solves a key inferential problem and optimizes parameter values with respect to the number of topics (Fig. 4). Results show the highest log likelihood over all tweets with 11 topics ($logMLE_{k=2} = -1{,}189{,}078$) following the topic model selection by Griffiths and Steyvers (2004).

3.3. Spatial autocorrelation analysis

Since our main objective is to extract statistically significant spatio-temporal and semantic human activity clusters from Twitter (RQ1), we need to analyze the degree of dependency among similar/dissimilar semantic observations in geographic space. To assess whether observed nearby tweets cover the same topics and show similar associated topic indicator values or not, we apply local measurements of spatial association (Fischer & Wang, 2011). Spatial processes are indicated by the presence of significant spatial autocorrelation (Getis & Ord, 1992), which quantifies Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 234). All test statistics for measuring spatial autocorrelation assess the extent to which empirically inferred spatial associations deviate from the null hypothesis of complete spatial randomness (CSR). In our study we use local indicators of spatial association (LISA) as introduced by Anselin (1995). These enable the identification of spatial clusters and spatial outliers denoting our targeted human social activities.

Table 1
Meta information for our selected Twitter dataset.

Dataset	Greater London (UK)
Bounding box (WGS 84)	−0.5543, 51.2386, 0.3038, 51.731
Timespan	31/07/2013–31/07/2014
Covered area	3265.387 km^2
Number of geotagged tweets after pre-processing	20.4 million
Number of individual users	476,071

3.3.1. Analysis scale

Any spatial analysis requires appropriate scale adjustment. This adjustment is usually reflected by a spatial weight matrix (termed W). This matrix models spatial relationships between all tuples of observations. In our case we define a neighborhood size of 250 m, since we compute the average distance from every polygon line segment to the centroid of this areal polygon for all statistical polygonal units in London. The output units are electoral wards derived from London boroughs, representing a certain neighborhood area and constitute the observation scale level for our research. Since the two observed topics each are inherently diverse and theoretical scale limits are greatly unknown, we argue that this region-oriented point of view is the most suitable approach with respect to the examined modeled processes.

3.3.2. Local Moran's I

The *Local Moran's* I_i statistic (Anselin, 1995) investigates correlations between spatial units and their surrounding spatial lags. Spatial lags are formed by observations having a nonzero spatial weight W_{ij}, i.e., having some spatial relationship with the focal observations i. Variable x is the estimated sample mean of X, which in our case represents the probabilistic LDA topic association indicator.

$$I_i = \frac{(x_i - \overline{x})}{1/n \sum_{j=1}^{n} (x_j - \overline{x})^2} \sum_{j=1}^{n} w_{ij}(x_j - \overline{x}) \quad (1)$$

$$z_{(I_i)} = \frac{I_i - E(I_i)}{\sqrt{V(I_i)}} \quad (2)$$

Given a sufficiently large number of samples n, the resulting test statistic I_i is asymptotically Gaussian. We therefore conduct significance testing by means of normal approximation. Hence, we have to transform I_i to standard derivatives (z-scores, see Eq. 2). A positive/negative value for I_i indicates positive/negative spatial autocorrelation, i.e., spatially close features tend to have similar/dissimilar attribute values. By arranging the obtained z-scores on a Moran scatterplot

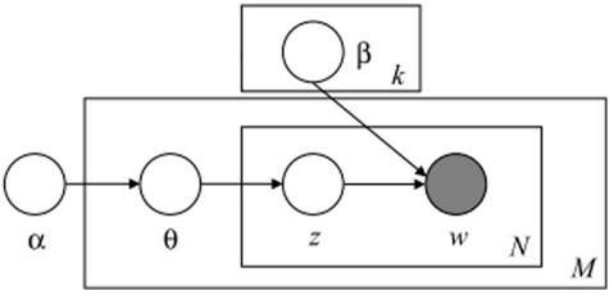

Fig. 3. LDA graphical model according Blei et al. (2003).

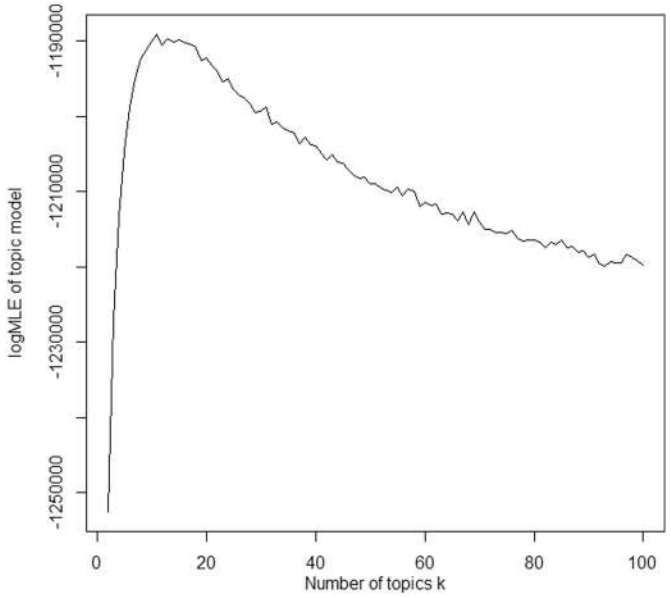

Fig. 4. Topic Model results, showing the log-likelihood (*logMLE*) of all tweets for different numbers of topics (k).

(Anselin, 1993), we can distinguish between two different kinds of local spatial autocorrelation clusters:

- *HH and LL*. Observations showing a high (HH) or low (LL) z-score and being surrounded by neighbors with similar characteristics. These observations are part of a larger spatially homogenous region.
- *HL and LH*. Observations with z-scores being dissimilar to those of their neighbors. These observations are part of a larger spatially heterogeneous region and can thus be considered spatial outliers.

Since the *Local Moran's* I_i statistic identifies spatial clusters of features with attribute values similar in magnitude (HH and LL), we are able to detect distinct human activity patterns over geographic space (RQ1).

3.3.3. Getis Ord G_i^*

Local Moran's I can detect neighborhoods comprising similar values. However, it is not able to distinguish local pockets of either high or low attribute values. Thus, we additionally perform a hot spot analysis by using the G_i^* method (Ord & Getis, 1995) This method is able to:

$$G_i^*(d) = \frac{\sum_{j\neq 0}^{n} w_{ij}x_j - W_i^*\bar{x}}{SD(x)\left\{\left[(nS_{1i}^*) - W_i^{*2}\right]/(n-1)\right\}^{1/2}} \quad (3)$$

where $S_{1i}^* = \sum_{j=1}^{n} w_{ij}^2$ and $W_i^* = \sum_{j}^{n} w_{ij}$.

In 3, W_{ij} represents the spatial weight shared by points i and j, and x represents the variable value (the probabilistic LDA topic association indicator) for location j. Significance testing with G_i^* can also be done by using normal approximation. The statistic as presented in 3 is already in the form of a z-score and must therefore not be further converted.

By correlating the results of both methods explained above we are able to find spatially autocorrelated values of high magnitude (correlation between HH and significant positive G_i^* values), medium magnitude (HH, but no significant G_i^* value) and low magnitude (correlation between HH and significant negative G_i^* values). The detected tweet clusters showing high I_i and G_i^* z-scores indicate spatially homogenous patterns (*Local Moran's* I_i) comprising a strong dominance of semantic topics (G_i^*). These observed clusters are a proxy indicator of "home" and "work" related social activities and are correlated in the following section with data obtained from available UK census populations (RQ2, s. Section 4).

4. Results

In the following section we summarize the results of the previously described analysis framework Section (3), which we applied in our case study, as described below.

4.1. Case study Greater London

Our case study is based on a one-year sample of georeferenced tweets from the area of Greater London. We have chosen this particular case study due to the availability of reliable open access census data, allowing us to answer RQ2. Table 1 shows further details regarding the retrieved Twitter data.

4.2. Results of semantic topic modeling with LDA

Figs. 5 and 6 show the temporal decomposition of the LDA-based probabilistic topic extraction for the highest assigned (>0.03) topic-associated words "work" and "home" over all tweets. For simplification in the descriptive parts of the paper, we only refer to each topic by labeling it with the most dominantly associated word. Furthermore, for visualization purposes, we only show the five words with the highest probabilities of being assigned to a topic in Figs. 5 and 6. As the LDA model embodies textual documents as a mixture of topics, these words are the most likely to have generated the original text corpora. Other words also appear and show lower probabilities of being assigned to a topic (<0.03). By visualizing all home-associated tweets we can detect a cluster C_1 of average low home topic counts between 12 pm (noon) and 7 pm ($n < 200$). In contrast, a lot of home-related tweets ($n > 400$) have been posted from 9 pm onwards until 1 am during the whole week (Cluster C_2) and between 9 am and 3 pm on the weekend. Cluster C_3 corresponds to medium topic intensities ($x = 0.2$) with a concentration of work-related topics ($n \sim 300$) between 7 am and 8 pm

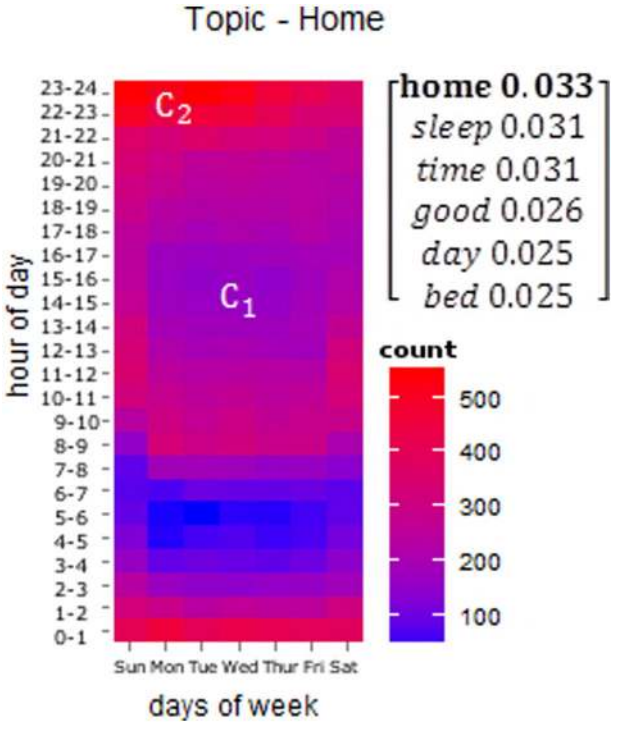

Fig. 5. Temporal distribution of home-related topics.

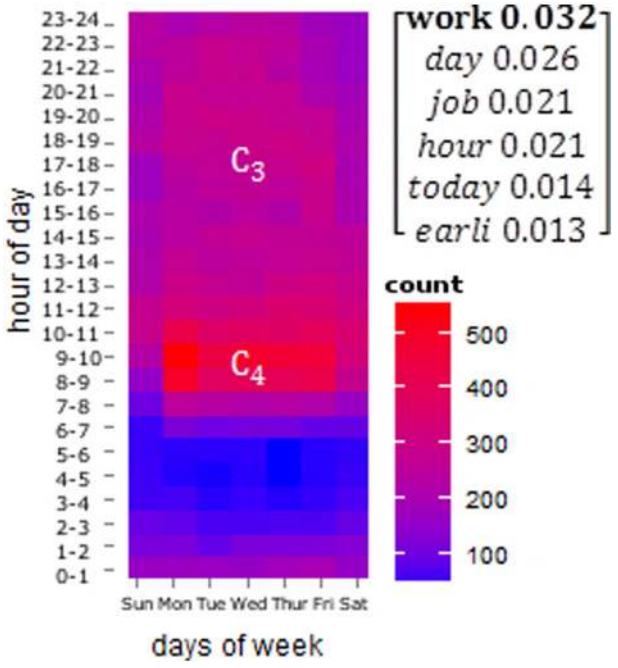

Fig. 6. Temporal distribution of work-related topics.

during weekdays. While Tuesdays and Wednesdays show significant work topic associations until 11 pm ($n > 200$) these counts decrease on Thursdays and Fridays ($n < 200$). Cluster C_4 contains a high amount work associated topics with increased topic indicator values during weekday periods between 8 am and 11 am.

The graph in Fig. 7 visualizes the frequency distribution of work and home topic-associated tweets aggregated over all seven weekdays binned into hourly slots. Looking for association rules and terms which show a significant correlation with the extracted home and work topics, a term adjacency matrix represented as a graph highlighting the most frequently occurring terms is additionally shown in the figure. The more terms mutually correlate, the higher the edge weight is and the more closely they appear in our graph. For instance, the terms "home"/"weekend" and "work"/"today" correlate by more than 0.5 and are therefore associated. When focusing on the most frequently occurring terms within work- and home-classified tweets over time, we can detect semantically stronger associations for specific words revealing human activities.

4.3. Results of spatial autocorrelation analysis

After previously extracting temporal-semantic information, which can be seen as an indicator for social activities, we now focus on the spatial aspects of the data and whereabouts of people for different time periods. With the applied *Local Moran's* I_i measure we can assess the degree of spatial association for all geographic neighborhoods of topic-classified tweets. This way, local spatial clusters (high I_i z-scores) of human social activities can be inferred. Subsequently, we assess the dominance of topics (hot spots) across geographic space using the G_i^* statistic.

Having a closer look at the spatiotemporal distribution of work-associated tweets during weekdays (Fig. 8), we can detect time-dependent work cluster patterns with dominant topic clusters of high

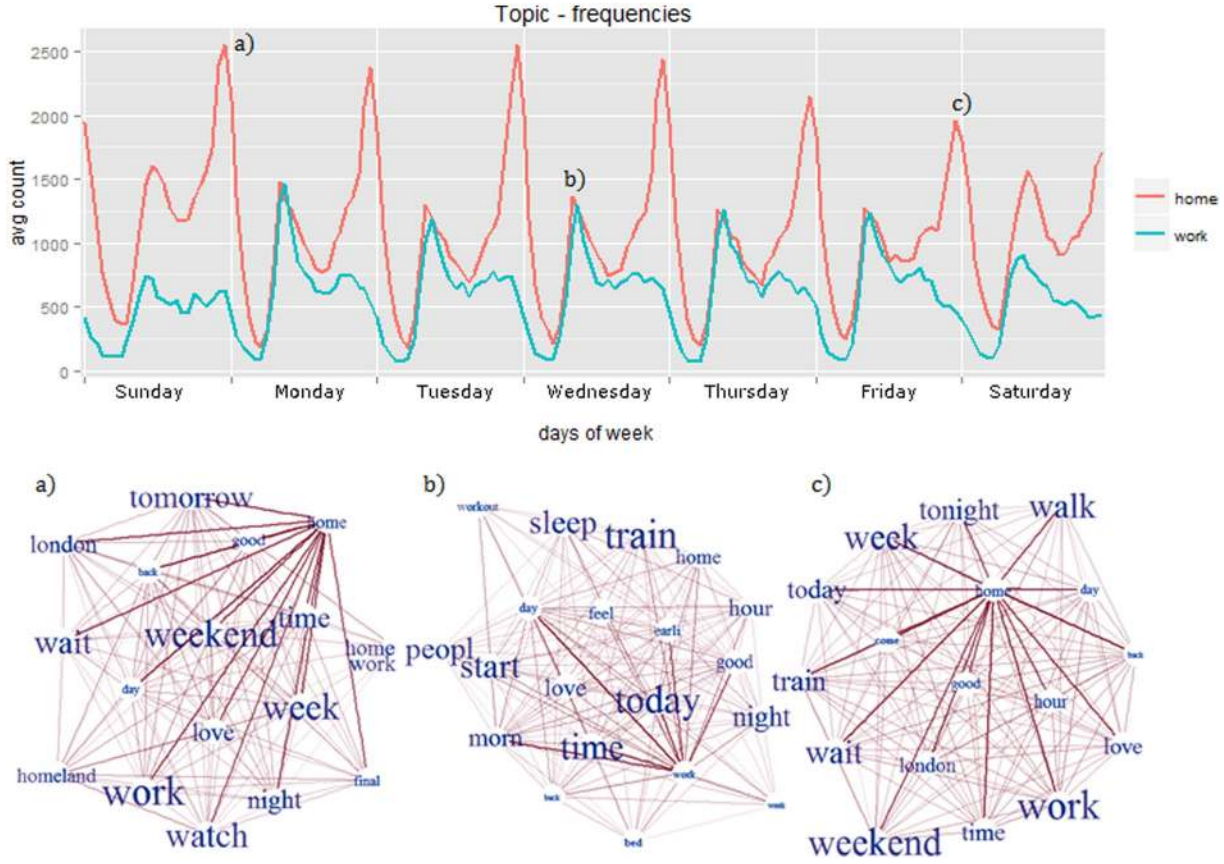

Fig. 7. Spatiotemporal topic frequencies of work- and home-classified topics during weekdays and most frequent associated occurring words for the exemplary peaks a), b) and c) shown as a weighted graph network.

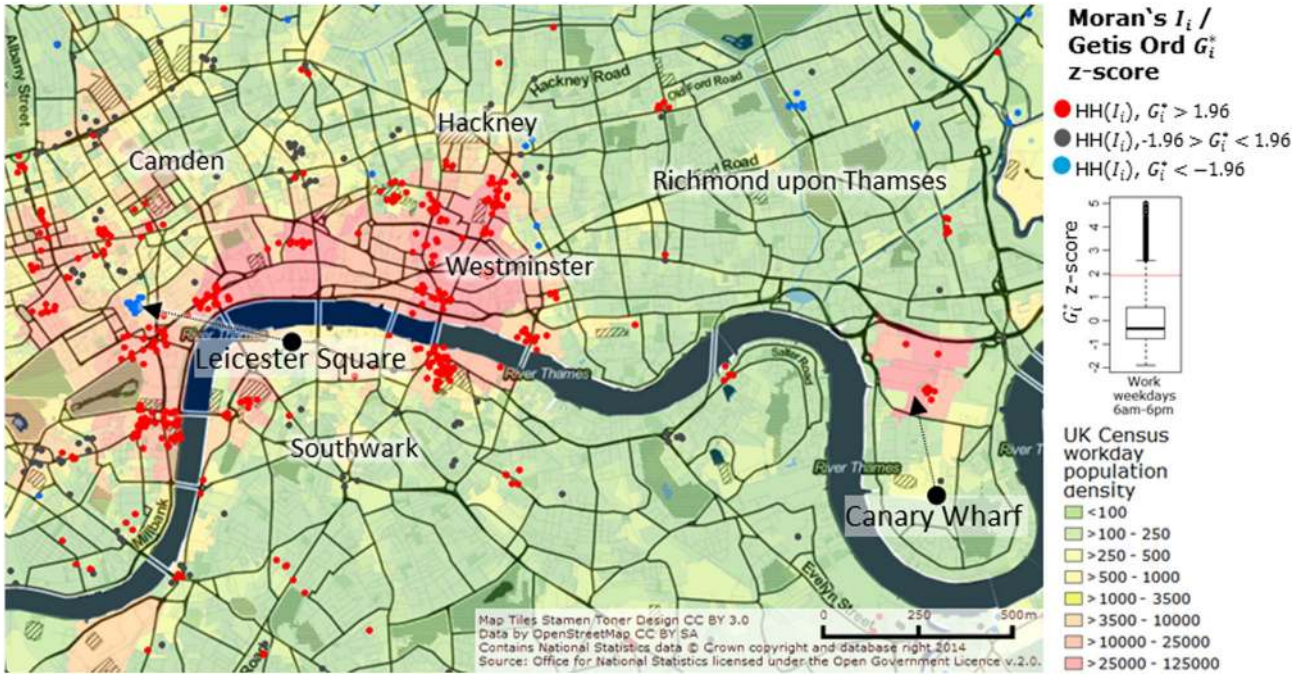

Fig. 8. Local spatial clusters (HH I_i) of work-dominant topics (G_i^*) with significance test.

topic intensities inside the City of London. Spatial clusters that are characterized by a high *Local Moran's* I_i z-score (HH) and a significant positive G_i^* z-score with a given p-value below 0.05, are colored in red on the map. In total, 85% of highly positive G_i^* z-scores, indicating work-dominant topics, fall within spatial clusters (HH I_i). The distribution of observed G_i^* z-scores and their quartiles inside the boxplot shows a proximate normal value distribution with a mean close to zero. Very high G_i^* z-scores within the given adjusted confidence level (95%) above 1.96 standard deviates or 1.5 interquartile range (appearing in the upper box-and-whisker plot in Fig. 8) are statistically significant hot spots of work-dominant topics. Therefore, we can reject the null hypothesis (CSR) for our observations. Since testing multiple hypotheses on a single sample of data increases the risk of obtaining false-positive results (type I errors), the significance levels in our research are adjusted by applying the False Discovery Rate (FDR) (Benjamini & Hochberg, 1995).

As a result, the highest positive G_i^* z-scores, indicating high work topic-associated tweets, occur during the period between 6 am and 6 pm, spatially clustering (HH I_i) inside the City of London. Furthermore, we observe that the high work-related values are clustering as spatial patterns nearby the city center (Westminster) and the financial business district (Canary Wharf). These clusters with a high positive autocorrelation visually match with the workday population density in the 2011 UK census data. In contrast, work-related tweets spatially cluster

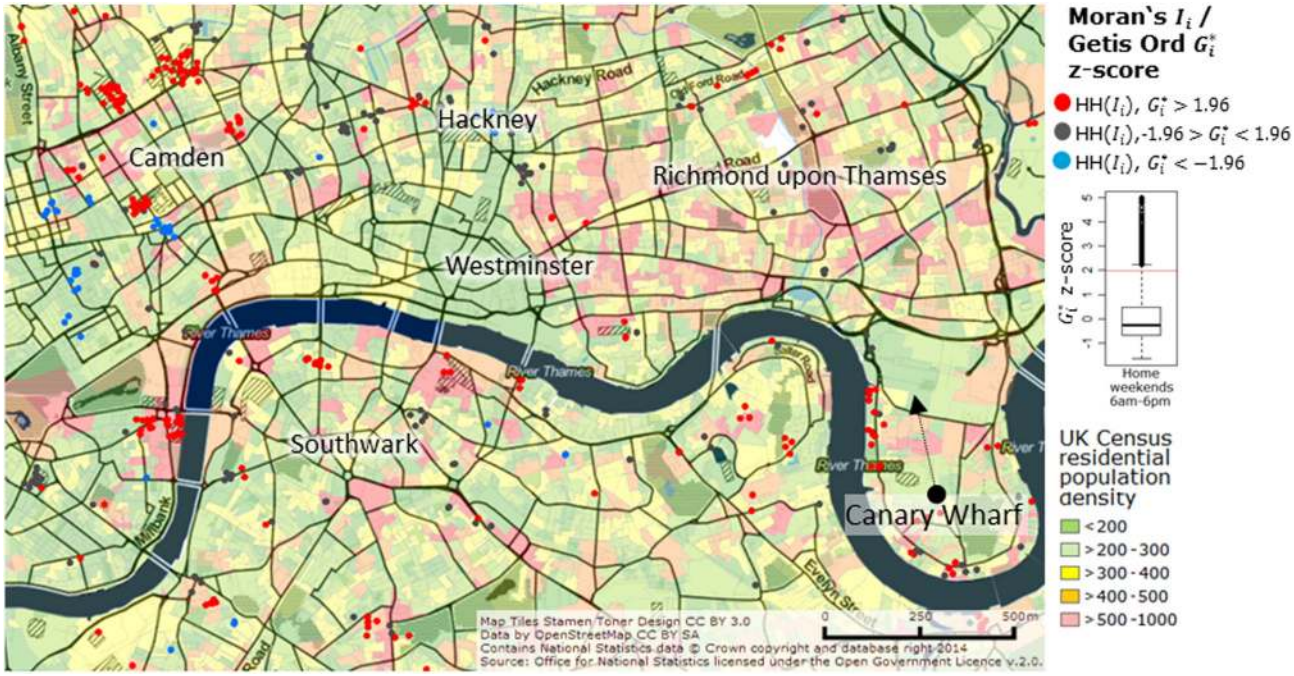

Fig. 9. Local spatial clusters (HH I_i) of home-dominant topics (G_i^*) with significance test.

in the vicinity of Leicester Square, but show significant negative G_i^* z-scores (cold spots). This means that tweets posted at the square are less dominantly covering work-related topics and are rather a mixture of different topics.

Looking at home-related spatial autocorrelation results (Fig. 9), spatial clusters (HH) of high G_i^* z-scores are more dispersed over the inner City of London. In total, 87% of highly positive G_i^* z-scores, indicating home-dominant topics, fall within spatial clusters (HH I_i). Previously detected work-related local spatial clusters within the Westminster borough and along the Canary Wharf district show no positive autocorrelation with home-related topics. In contrast, the area southwards of Canary Wharf is characterized by a high amount of spatial tweet clusters along the Thames river, and visually matches the residential population density according to the UK census. G_i^* z-scores of home-related topics have a lower amount of values appearing in the upper box-and-whisker plot (see Fig. 9), exceeding the critical z-scores (with a 95% confidence), indicating less significant hot spots. These hot spots of home-related topics between weekend periods have a more widely dispersed distribution over whole area of Greater London including the outskirts of the city.

4.4. Correlation with UK census Data

For our case study we use census data provided by the UK Office for National Statistics (ONS) under Open Government License v. 2.0 (Office for National Statistics, 2012). The first visual comparison of the spatial autocorrelation results in Section 4.3 with census data revealed specific similarities and overlapping patterns.

In the next step, we quantify the statistical relationship between outcomes of the spatial autocorrelation analysis with workday and residential census data to investigate how reliable derived home and workplace clusters from Twitter data are. The UK census collects population statistics for statistical output areas (OA) – the last collection took place in 2011 – in which output areas have a defined minimum and maximum population criterion, mainly for privacy reasons with an average cell size of roughly 1 km^2. These statistical areas constantly change and are aggregated or disaggregated with adjacent vicinities according to demographic factors such as the population distribution.

Within the published datasets we extracted output areas for workday populations and residential populations. Workplace population statistics are an estimate of the overall population working in an area during the working day. Residential populations include all residents aged 16 to 74 whose living places are in the area. In order to correlate census populations with single tweet observations, we normalize all population counts over the size of the regions to minimize differences in values based on the size and the number of features in each output area. Afterwards, we spatially aggregate all tweets belonging to spatial clusters of high *Local Moran's* I_i z-score (HH) and having a significant positive G_i^* z-score (indicating clusters of high topic values) into overlapping output areas. In the next step, we normalize all selected tweets over each user to avoid an over- or underrepresentation of particular Twitter users in the census output areas as a result from varying tweet frequencies. Finally, we Studentize both variables (census population counts and aggregated home/work tweet counts) by subtracting the arithmetic mean and dividing by the respective standard deviation to be able to compare these two different distributed variables and their varying intervals with each other.

Figs. 10 and 11 provide normalized scatter plots representing the specific correlation of counts of work- and home-classified tweets against the residential and workplace population for each census output area. Fig. 10 illustrates the Pearson correlation ($r_{work} = 0.75$ slope, $m_{work} = 0.87$) between the amounts of semantic highly associated tweets to work-related LDA-classified topics having a positive autocorrelation and real-world workplace densities in the UK census data. The statistical significance of this value is given since a t-test reveals that the empirical t-value (20.46, $\alpha = 0.01$, $v = 2$) exceeds the corresponding critical value (9.925). Therefore, we can reject the null hypothesis (no significant relationship between work topic clusters and census workplace densities). Furthermore, the F-test results confirm the hypothesis that our given set of Twitter and census populations are equally dispersed since $0.779 > 0.773_{(f = 2,\ P = 99\%)}$. Output areas with less workplace population, visible in the left lower corner of the correlation plot, denote mixed urban areas of residential and workplace population with a subsequently weaker correlation for work-related tweets. However, only an extremely weak statistical correlation ($r_{home} = 0.08 / r_{S\ home} = 0.11$) is present for home topic classified and spatially clustered tweets, as shown in Fig. 11.

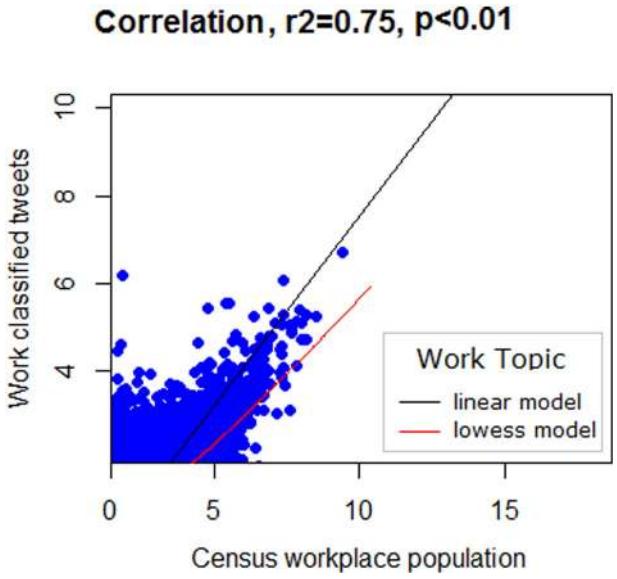

Fig. 10. Scatterplot results correlating work-clustered tweets with workplace census.

A first hypothesis for this discrepancy of expected correlation results is supported by Fig. 12, when analyzing human mobility patterns and human behavior. Out of 37,765 clustered tweets (HH I_i) having a significant positive G_i^* z-score, 19,679 posts (51%) spatially match public transport polygons of station buildings derived from OpenStreetMap. All major transportation hubs within London show a high amount of home-classified tweets and constitute significant spatial clusters with positive autocorrelation. Referring to the initially mentioned

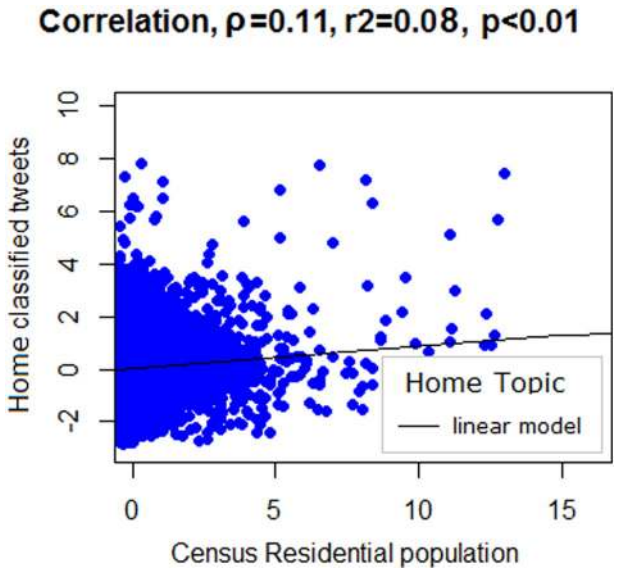

Fig. 11. Scatterplot results correlating home-clustered tweets with residential census.

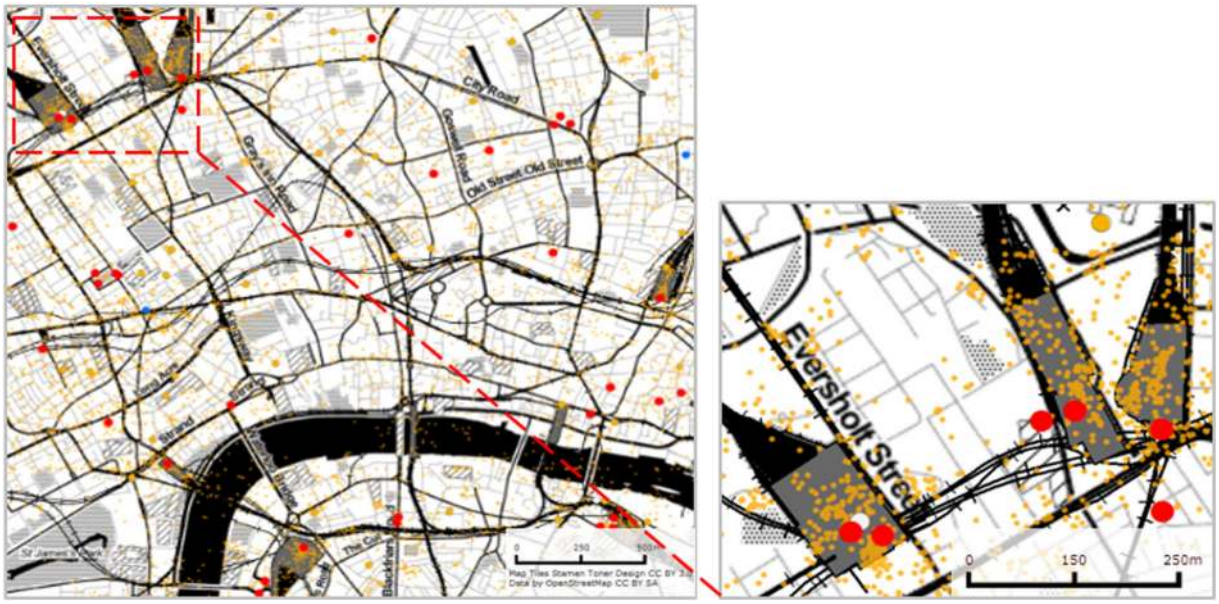

Fig. 12. Local spatial clusters (HH I_i) of home-dominant topics (G_i^*) and overlay of public transport network with exemplary stations Euston, St. Pancras and King's Cross.

characteristic of Twitter data (2.1), there is a strong indication that Twitter users post home-related information in the proximity of public transportation facilities, referring to past or future events.

5. Discussion

In this paper, we analyzed if georeferenced tweets cover collective real-world characteristics within urban space. We also cross-validated these patterns against authoritative census data. We analyzed detected semantic topics indicating human social activities in terms of spatial autocorrelation to detect specific clusters. Mainly, we investigated the correlation between tweet activity clusters and census population densities. Our results show that the conducted study has several limitations that need to be addressed in future research efforts.

Within the initial pre-processing step (s. sub-section 3.1) the semantic dimension of raw tweets is reduced to word vectors. Substantial semantic information remains part of the tweet corpora, while additional information that is irrelevant to our research questions is filtered out. Since the dataset is highly uncertain involving colloquial writing styles and unpredictable Internet-oriented terms, the efficiency of natural language processing is likely to vary.

Regarding the applied methodologies when measuring semantic similarities with Latent Dirichlet Allocation (Section 3.2), there are some pitfalls within the bag-of-words assumption in particular. Words that should be associated with the same topic might be assigned to several different topics. Additionally, considering syntactical structures such as n-grams would enable word sequences to be allocated to several topics, thus better accounting for real-world semantic complexities. During the posterior parameter estimation and computation of input variables, the LDA model assumes that, for instance, the number of topics k have a probabilistic Dirichlet prior topic distribution. However, this approach only infers LDA input parameters by considering the semantic dimension. The computation of probability-based parameters needs to be further adapted by taking the spatial dependence of semantic topics over different geographic areas (chosen study area) and geographic scales (extent of the study area and interaction distances of the observations) into account. Further, modified LDA versions like Twitter-LDA (each tweet as a single document) and the author-topic model (treating all tweets of the same user as a single document) (Zhao, Jiang, Weng, He, & Lim, 2011) have not been included within this study due to missing statistical validation and benchmarking process of these methods' results.

The results of the spatial autocorrelation analysis (Section 3.3) show a strong indication for hot spots of semantic similarity occurring over time and space. Spatial autocorrelation statistics assess the degree of spatial associations for an observed attribute by testing against a random distribution. In our analysis we detected non-random behavior, which is an indicator for systematic spatial patterns. Therefore, it can be assumed that collective human behavior in urban spaces is reflected in the semantics of Twitter, at least to a certain degree. However, we are still lacking more detailed knowledge about such phenomena with respect to their precise geographic distribution and stochastic properties such as measures of dispersion (e.g., variance) (Goodchild, 2009). More, we do not have knowledge regarding extrinsic factors driving the geography of human behavior. These might impose effects such as trends or intertwined global and local effects (Ord & Getis, 2001).

Furthermore, that autocorrelation results depend on the chosen parameters like the neighborhood definition. With respect to geographic scale effects (Getis, 1999), we might have missed out on effects at different scales. In our study we have chosen our analysis scale based on empirical considerations, i.e., observations in neighborhoods of a certain scale are compared against observations within the overall dataset on a different scale.

Considering the advantages of spatial autocorrelation analysis, the *Local Moran's* I_i statistic helps to identify the characteristics of local spatial patterns. Since the degree of spatial association is evaluated by using normal theory (Anselin, 1995), results can be easily compared and extended with other test statistics, such as the applied G_i^*. Therefore, combined *Local Moran's* I_i and G_i^* represent an efficient exploratory data mining method for the detection of similar spatially homogenous patterns with significant value accumulations.

When looking at our selected case study in London (Section 4.1), we have chosen a popular well-researched study area regarding Twitter analysis. However, the previous research is not exhaustive in terms of the geographical distribution of Twitter usage, which highlights the need for future studies to span a wider geographic coverage. Nevertheless, this research serves as a first benchmark answering how trustworthy inferred social activities from Twitter based on spatial characteristics are.

The observed spatial patterns of inferred human social activities are a proxy indicator for the identification of urban residential areas and

workplace areas. The results presented in Section 4.2 have revealed typical daily commuting patterns and temporal differences of home- and work-related activities between weekdays and weekends, reflecting the bimodal nature of collective human mobility. These observed urban dynamics, of how people live, work and move in the city, are also in line with similar studies analyzing mobile phone traffic (Sagl, Loidl, & Beinat, 2012; Yuan, Raubal, & Liu, 2012).

This research can potentially be conducted in other regions with limited access to official data and knowledge about socioeconomic processes within urban structures. When tweets are compared with official continuous census data, these two different ascertained datasets might lead to a possible sampling bias. Although we know the average estimated census residential and workplace population having a minimum and maximum set age limitation, we do not have any Twitter demographics and know who exactly tweets. Also, the time span of collected tweets differs from the census data creating a sampling bias (Section 2.1) and might have an influence on the overall correlation results. However, since the census data are recorded only every ten years, this factor cannot be dedicatedly excluded. Our results have empirically shown the reliability of georeferenced tweets as a proxy of human social activities, since tweet locations autocorrelate with real world census observations and revealed similar, overlapping patterns.

Regarding the characterization of urban structure, the issue arises on how to divide a city into areal units. Existing approaches include division into arbitrary grid cells, Voronoi polygons or administrative boundary polygons. When point-based measures of spatial phenomena (tweets) are aggregated into population density polygons, which are themselves aggregated from census data records, cohesive regions might be artificially split up. With respect to real-world scales these arbitrary constructs inevitably lead to the Modifiable Areal Unit Problem (MAUP) Fotheringham & Wong, 1991).

6. Conclusion and future research

In this paper we presented a spatiotemporal and semantic analysis framework for georeferenced tweets. Concluding the results answering our first research question (RQ1 — assessing work-related and home-related activities that reflect typical collective human behavior), we have been able to detect significant clusters indicating home- and work-related human social activities. Analyzing the temporal distribution of LDA-classified topics (Section 4.2) and most frequently associated words, we were able to infer further information regarding collective time-dependent human behavior. Using semantic topic modeling and autocorrelation analysis (Section 4.3), we extracted spatial, temporal and semantic clusters of human activities from tweets for our selected case study in London.

The second research question (RQ2) of this paper attempted to investigate to what degree observed tweet clusters can be regarded as a proxy of human social activities, i.e., to what extent they correlate with available residential and workplace populations from official census data. Semantically and spatiotemporally clustered work-related tweets have shown strong positive correlation in comparison with workplace population census data (Section 4.4), indicating that topic-classified tweets are a potential proxy for real-world workplace-related activities. In contrast, classified home-related topics from tweets are only weakly correlating with residential populations from census data. This might have been caused by either the involved uncertainties of Twitter data (Section 2.1) or by the semantic complexity and diversity of the home-related topics that might therefore be considered being a less representative proxy for residential activities.

The outcomes of this study may be considered in future research work regarding the inference and trustworthiness of human mobility patterns from crowdsourced data. Location inference of residential and workplace areas are a key factor of the given transportation demand. LBSN can help to better understand these processes and explore the impact of urban spatial structures on travel demand and human mobility as a future research direction. The presented approach can be generalized to study human dynamics not only at an urban level, but also on a regional or national level.

Further potentially interesting social activities related to home and work might also be inferred. This may constitute a useful source of information to determine, which social activities correspond to which underlying urban structures, such as points of interest, landmarks, etc. Since our results concerning home-related topics have shown spatial clustering in the vicinity of highly frequented squares and major transportation hubs, other social activities (e.g., mass events like concerts or games) might also show a similar pattern and could thus lead to new insights in characterizing urban mobility. The detection of human mobility patterns without a priori categorical information regarding certain human social activities, however strongly depends on the reliability of given semantic information.

Finally, after we were able to answer our initial research questions (Section 1) and to provide new insights on human social activities in LBSN, we hope to further encourage and foster new research.

Acknowledgments

This research has been funded through graduate scholarship program "Crowdanalyser — spatiotemporal analysis of user-generated content" supported by the State of Baden Wurttemberg. We also thank the British Office for National Statistics for publishing UK Census data licensed under the Open Government License v. 2.0. This research has been supported by the Klaus Tschira Stiftung gGmbH.

References

Andrienko, G., & Andrienko, N. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, *15*(3), 72–82.

Anselin, L. (1993). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *GISDATA Specialist Meeting on GIS and Spatial Analysis, Amsterdam*.

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

Azevedo, T., & Bezerra, R. (2009). An analysis of human mobility using real traces. *Wireless Communications and Networking Conference, WCNC* (pp. 1–6).

Bagrow, J., & Koren, T. (2009). Investigating bimodal clustering in human mobility. *International Conference on Computational Science and Engineering* (pp. 944–947).

Becker, H., & Gravano, L. (2011). Beyond trending topics: real-world event identification on Twitter. *ICWSM*, *11*, 438–441.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 289–300.

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

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x.

Candia, J., & González, M. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, *41*(22).

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. *ICWSM, Vol. 2010.* (pp. 81–88).

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining — KDD '11* (pp. 1082–1090). New York, NY: ACM. http://dx.doi.org/10.1145/2020408.2020579.

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The Livehoods Project: Utilizing social media to understand the dynamics of a city. *ICWSM*. AAAI.

Department for Transport (2014). National Travel Survey: England 2013 trip rates have been falling steadily since (July). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/337473/nts2013-01.pdf

Ferrari, L., Rosi, A., Mamei, M., & Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks — LBSN'11* (pp. 9–16). New York, NY: ACM. http://dx.doi.org/10.1145/2063212.2063226.

Fischer, M. M., & Wang, J. (2011). Spatial data analysis: Models, methods and techniques. *Springer briefs in regional science* (pp. 82).

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

Gao, S. (2014). Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation*, *15*(2), 86–114. http://dx.doi.org/10.1080/13875868.2014.984300.

Getis, A. (1999). Spatial statistics. *Geographical Information Systems*, *239–251*.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association. *Geographical Analysis*, *24*(3), 189–206.

Giannotti, F., & Pedreschi, D. (2008). *Mobility, data mining and privacy*. ACM.

Goodchild, M. F. (2009). What problem? Spatial autocorrelation and geographic information science. *Geographical Analysis*, *41*(4), 411–417. http://dx.doi.org/10.1111/j.1538-4632.2009.00769.x.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America. Vol. 101*. (pp. 5228–5235).

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing — UrbComp '13* (pp. 1). New York, NY: ACM Press. http://dx.doi.org/10.1145/2505821.2505823.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. http://dx.doi.org/10.1080/15230406.2014.890072.

Heckman, J. J. (1979). Sample selection bias as a specification error. *The Econometric Society Stable*, *47*(1), 153–161.

International Telecommunication Union (ITU) (2014). *Measuring the information society. report.* (Retrieved from http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2014/MIS2014_without_Annex_4.pdf).

Jackoway, A., Samet, H., & Sankaranarayanan, J. (2011). Identification of live news events using Twitter. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks — LBSN '11* (pp. 248–260). New York, USA: ACM. http://dx.doi.org/10.1145/2063212.2063224.

Jiang, B., Yin, J., & Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Physical Review E*, *80*(2).

Kling, F., Kildare, C., & Pozdnoukhov, A. (2012). when a city tells a story: Urban topic analysis. *Proceedings of the 20th international Conference on Advances in Geographic Information Systems* (pp. 482–485). New York, USA: ACM. http://dx.doi.org/10.1145/2424321.2424395.

Krumm, J., Caruana, R., & Counts, S. (2011). Learning likely locations. *User modeling, adaptation, and personalization* (pp. 64–76). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-38844-6_6.

Kung, K., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS One*, *9*(6)http://dx.doi.org/10.1371/journal.pone.0096180.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, *5*, 361–397.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. http://dx.doi.org/10.1080/15230406.2013.777139.

Liang, X., Zheng, X., Lv, W., Zhu, T., & Xu, K. (2012). The scaling of human mobility by taxis is exponential. *Physica A: Statistical Mechanics and Its Applications*, *391*(5), 2135–2144.

Metke-Jimenez, A., Raymond, K., & MacColl, I. (2011). Information extraction from web services: a comparison of Tokenisation algorithms. *2nd International Workshop on Software Knowledge*.

Morris, S., Humphrey, K. P., & Tipping, S. (2014). *National Travel Survey technical report.* (Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/337263/nts2013-technical.pdf).

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *The Social Mobile Web 11*. AAAI.

Office for National Statistics (2012). *Statistical Bulletin 2011 Census — Population and household estimates for England (July)*. , 1–36 (Retrieved from http://www.ons.gov.uk/ons/guide-method/census/2011/index.html).

Ord, J., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, *27*(4), 286–306. http://dx.doi.org/10.1111/ j.1538-4632.1995.tb00912.x.

Ord, J. K., & Getis, A. (2001). Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, *41*(3), 411–432. http://dx.doi.org/10.1111/0022-4146.00224.

Pan, C. (2011). Event detection with spatial latent Dirichlet allocation. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 349–358).

Phithakkitnukoon, S., & Horanont, T. (2010). Activity-aware map: Identifying human daily activity pattern using mobile phone data. *Human Behavior Understanding* (pp. 14–25). Springer Berlin Heidelberg.

Resch, B., Beinat, E., Zipf, A., & Boher, M. (2012). Towards the live city — Real-time urbanism. *International Journal on Advances in Intelligent Systems*, *5*(3), 470–482.

Sagl, G., Loidl, M., & Beinat, E. (2012). A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. *ISPRS International Journal of Geo-Information*.

Sengstock, C., & Gertz, M. (2012). Latent geographic feature extraction from social media. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems — SIGSPATIAL '12* (pp. 149). New York, New York, USA: ACM Press. http://dx.doi.org/10.1145/2424321.2424342.

Steiger, E., Ellersiek, T., & Zipf, A. (2014). Explorative public transport flow analysis from uncertain social media data. *Third ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD)*http://dx.doi.org/10.1145/2676440.2676444.

Tobler, W. (1970). On the first law of geography : A reply. *94*, 304–310 (2003, July).

United Nations Population Fund (2008). *The state of the world population 2007: Unleashing the potential of urban growth*, *Vol. 45*, (Retrieved from http://www.unfpa.org/webdav/site/global/shared/documents/publications/2007/695_filename_sowp2007_eng.pdf).

Wakamiya, S., Lee, R., & Sumiya, K. (2011). Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from Twitter. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (pp. 77–84). New York, NY: ACM. http://dx.doi.org/10.1145/2063212.2063225.

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior — A case study of Harbin, China. *Computers, Environment and Urban Systems*, *36*(2), 118–130. http://dx.doi.org/10.1016/j.compenvurbsys.2011.07.003.

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

Zhao, W. X., Jiang, J., Weng, J., He, J., & Lim, E. (2011). Comparing Twitter and traditional media using topic models. *Advances in information retrieval*. Springer Berlin Heidelberg.

Zheng, Y. (2011). Location-based social networks: Users. *Computing with spatial trajectories*. Springer New York.