This is the "Accepted Version" of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps

Enrico Steiger [a*], Bernd Resch [b,c] João Porto de Albuquerque [d] and Alexander Zipf [a]

[a]*GIScience Research Group, Institute of Geography, Heidelberg University, Germany;*

[b]*Z_GIS, Department of Geoinformatics, University of Salzburg, Austria;*

[c]*Center for Geographic Analysis, Harvard University, USA;*

[d]*Centre for Interdisciplinary Methodologies, University of Warwick, United Kingdom*

[*]Corresponding author: Enrico Steiger, Berliner Strasse 48, 69221 Heidelberg, Germany

Keywords: traffic data; Twitter; self-organizing map; point pattern analysis; human mobility

Abstract

Cities are complex systems, where related Human activities are increasingly difficult to explore within. In order to understand urban processes and to gain deeper knowledge about cities, the potential of location-based social networks like Twitter could be used a promising example to explore latent relationships of underlying mobility patterns. In this paper, we therefore present an approach using a geographic self-organizing map (Geo-SOM) to uncover and compare previously unseen patterns from social media and authoritative data. The results, which we validated with Live Traffic Disruption (TIMS) feeds from Transport for London, show that the observed geospatial and temporal patterns between special events ($r=0.73$), traffic incidents ($r=0.59$) and hazard disruptions ($r=0.41$) from TIMS, are strongly correlated with traffic-related, georeferenced tweets. Hence, we conclude that tweets can be used as a proxy indicator to detect collective mobility events and may help to provide stakeholders and decision makers with complementary information on complex mobility processes.

This is the “Accepted Version” of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

1. Introduction

The complexity of cities with related human activities is becoming an increasingly tough challenge for policy makers and modelers to explore urban dynamics and the study of city-scale mobility patterns. One promising example of available high-granularity information sources is the Transport for London’s (TfL) Traffic Information Management System (TIMS). It provides high resolution, up-to-date disruption information regarding congestions, traffic incidents, events, construction works and other issues affecting traffic. However, these existing traffic measuring systems (e.g., road-side detectors, video surveillance, floating car data, etc.) are resource intensive in terms of ongoing operating and maintenance costs. Furthermore, a complete detection of all traffic and road conditions is simply not feasible.

At the same time, in recent years an increasing amount of information has been generated through mobile devices, becoming a potentially powerful data resource for (geographic) knowledge discovery and human behavior analysis from crowdsourced data (Goodchild, 2007). For a number of disciplines this development opens up enormous potential for various applications, including urban- and traffic planning, disease- and disaster management.

In particular, harnessing human mobility information from social media platforms such as Twitter can potentially lead to new insights into the human mobility process. Due to the high spatiotemporal resolution this may provide complementary information when compared with existing traffic data sources. However, one main challenge when analyzing mobility with officially acquired data is the spatiotemporal complexity of latent processes within traffic events (e.g., effects of incidents such as roadworks on the traffic flow and the correlations with traffic disruptions), hampering the detection of patterns in large road networks (Asif et al., 2014)).

Simultaneously, when using crowdsourced information it is uncertain how representative and trustworthy these new types of geodata are for the inference of human mobility patterns (Steiger et al., 2015c). Thus, research in this area requires new methodological approaches, which consider the high dimensionality and uncertainty of crowdsourced geographic information in the context of a data-driven geography (Miller and Goodchild, 2014). In a previous study, we therefore applied and have demonstrated the efficiency of self-organizing maps (SOMs) to abstract and cluster information from multidimensional Twitter data in a trans-disciplinary approach (Steiger et al., 2015b).

However, it has not been analyzed whether spatiotemporal information from social media data is a suitable proxy for inferring certain traffic-related events. Further the question is whether the results in comparison with official traffic disruption reports lead to new insights regarding the study of human mobility patterns.

In this paper, we use a non-geographic and a geographic self-organizing map (SOM/Geo-SOM) to discover collective human mobility clusters by analyzing similar variances within geospatial, temporal and disruption characteristics from live traffic feeds. The results are correlated with traffic-related georeferenced tweets for a case study in London. We intend to answer the following research questions (RQ):

(RQ1): What is the correlation between inferred spatiotemporal clusters from tweets, as a proxy of collective human mobility patterns and the real time traffic information provided by TIMS?

(RQ2): Which official traffic events along with their individual traffic disruption characteristics (category, severity, duration) are reflected in traffic-related tweets and have a dissimilar/similar spatiotemporal distribution?

2. Background

This section summarizes the characteristics of both datasets used in this analysis (sub-sections 2.1 and 2.2). Then, related work in the area of spatiotemporal mobility analysis is depicted in sub-section 2.3, followed by a description of the current state of the art regarding the application of SOMs for mobility analysis in sub-section 2.4.

2.1 Comparative reference dataset: TIMS disruption messages

The Transport for London (TfL) authorities provide real time open traffic disruption data for the area of Greater London as part of their Traffic Information Management System (TIMS). TIMS contains a broad range of a priori known information concerning road disruptions, such as the location of occurrence, details regarding road closures and more in-depth categorization of the cause of a disruption. Our research solely focuses on analyzing active persisting traffic event messages within the five categories of traffic incidents, traffic volume, hazards, and special and planned events (Transport for London, 2007). As persisting traffic disruptions are repeated until cleared, we can compute the length of every incident and have, additional categorical information regarding the severity of a traffic event by combining it with the provided level of interest and priority status. All observed traffic disruption messages and the

identified general spatiotemporal patterns of occurrences in London provide a reliable ground truth of mobility events. Thus, they serve as a reference dataset for our comparison with Twitter data.

2.2 Social media dataset: Twitter

Within the online social network and microblogging system Twitter, currently more than 288 million users post 500 million short status messages (tweets) with up to 140 characters per day[1]. As a further option, users can geotag their tweets with a geo-location acquired by their mobile devices. Therefore tweets are provided in high spatiotemporal resolution (geolocation and timestamp of the tweet) and include a semantic information layer (message content of the tweet). Since georeferenced tweets are to a certain extent a proxy of real-world observations (Hawelka et al., 2014), they represent a valuable opportunity for studying human mobility dynamics. Frequently repeating patterns of contextually similar tweets over geographic space and time might serve as an indicator to characterize human activity and to detect traffic-related events (exemplary georeferenced tweet in Figure 1).

Figure 1. Exemplary traffic-related, georeferenced tweet.

2.3 Spatiotemporal and semantic human mobility analysis from social media

Recent research efforts on harnessing human mobility related information have focused on extracting individual and collective human daily activity patterns, such as taxi trip records (Jiang et al., 2009, Liu et al., 2012, Peng et al., 2012), GPS trajectories (Gonzalez et al., 2008, Wang et al., 2014), or large sets of mobile phone records (Song et al., 2010, Gao, 2014). All

[1]https://about.twitter.com/de/company

studies generally conclude the possibility to infer and predict collective human mobility patterns by analyzing travel distances, staying times and collective displacements.

Regarding human mobility analysis from social media, a number of studies, (Cranshaw et al. 2012, Hasan et al., 2013, Liu et al., 2014) have used check-in data from Foursquare to analyze collective human activity patterns to infer urban- (Wakamiya et al., 2011) and user-specific travel characteristics (Noulas et al., 2012). Foursquare is a location and recommendation service where users can check in with their mobile phones at specific venues (shops, hotels etc.). Liu et al. (2014) extracted trips and spatial interactions from Foursquare check-in data to detect inter-urban movements. A further validation of check-in data compared to mobile phone locations revealed similar collective movement patterns of people showing spatial and social proximity (Cho et al., 2011).

As for human mobility analysis from Twitter data, several studies estimated individual travel behavior with urban motion patterns (Krumm et al., 2011, Ferrari et al., 2011) and also found a correlation between tweet locations and certain socioeconomic characteristics of people (Li et al., 2013, Hawelka et al., 2014). Tweets are a proxy for tracking and predicting human movement and also have similar features compared with mobile phone records (Jurdak et al., 2015). With respect to our research questions, Gao (2014) estimated spatiotemporal mobility flows from Twitter to infer origin-destination trips. Results have shown similar patterns when compared with community survey data. Lenormand et al. (2014) investigated the use of Twitter in transport networks for Europe by comparing the number of georeferenced tweets with average annual daily traffic reports. The authors were able to detect frequent user transport modalities along with overall traffic congestions. Yet, in both studies, tweets have simply been aggregated and matched onto the transportation networks on different scale levels (leading to modifiable areal unit problem ,Fotheringham & Wong, 1991) without analyzing the textual components of tweets. Moreover, no study currently focuses on analyzing, which events and information are reflected within tweets. This is an important aspect in order to assess the reliability as a proxy source of human mobility activities, adding complementary information to existing knowledge.

Therefore, in accordance with our research goals stated in section one, the following sub-section summarizes the recent state of the art concerning the advantages of applying SOMs. The section further outlines the potential of SOMs for detecting and comparing human mobility clusters from official real-time traffic information and tweets in a combined approach.

2.4 Application of self-organizing maps, Geo-SOM and variants

A self-organizing map is an unsupervised artificial neural network (ANN) learning algorithm, introduced by Kohonen (1982, 1990), that produces a two-dimensional topological connected output map from multi-dimensional input data properties. The advantages of neural networks for geographic analysis were first demonstrated by Openshaw et al. (1995). Facing a growing complexity of spatial data and analysis tasks (Miller and Han, 2009), SOMs have been further evaluated as a high-performance data mining method to cluster high-dimensional data (Ultsch and Vetter, 1995, Watts and Worner, 2009). Thus, SOMs are of great interest for the field of GIScience for spatial data mining and pattern detection (Spielman and Thill, 2008, Gorricha et al., 2013), and spatial clustering (Jiang and Harrie, 2004, Skupin and Hagelman, 2005). Agarwal and Skupin (2008) provide a broad summary of SOM applications within GIScience. Geospatial attributes of temporal observations have been first considered by Kangas (1992) (“Kangas Map”). Bação et al. (2005) further modified SOMs by only matching geographically close neurons in their Geo-SOM framework.

Various studies have proven the ability to infer and predict real-world traffic clusters inside road networks from official traffic data by using SOMs (Asamer et al., 2007, Asif et al., 2014, Feng et al., 2014). Sagl et al. (2014) further evaluated spatially autocorrelating SOMs to foster the exploration of collective human activities from mobile phone records. A combination between Geo-SOM and a hierarchical SOM to further explore a spatial motorcycle flow dataset has been described by Feng et al. (2014).

Regarding the application of SOM for social media analysis, several studies (Boulet et al., 2008, Couronne et al., 2013) investigated relationships inside social networks with neural networks to study user characteristics and collective similarities. Related to knowledge discovery from geographic data, Hagenauer et al. (2010) applied SOMs to cluster point-based crime patterns and later use a text mining approach to classify unstructured citizen crime reports (Helbich et al., 2013).

To the best of our knowledge no past research exists on validating and comparing the characteristics between inferred spatiotemporal clusters from official traffic disruption information and tweets as a proxy of collective human mobility patterns.

This is the “Accepted Version” of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

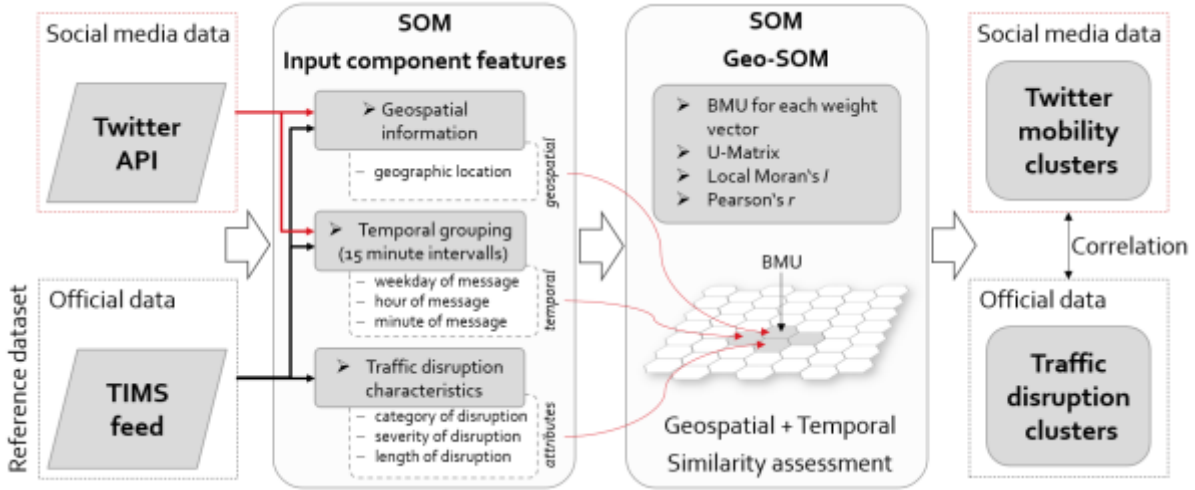
3. Methodology

Our methodological approach intends to leverage existing knowledge about the spatiotemporal characteristics of traffic disruptions from official data in order to compare, which patterns of inferred disruption clusters are similarly reflected within georeferenced tweets. Therefore, we compute the similarity between the three available information layers for traffic disruptions: the categorical attributes for each disruption (category of event, severity and duration), the geographic location and the temporal information (for a detailed methodological evaluation see (Steiger et al., 2015b)).

Figure 2 visualizes the overall computational framework. First, we apply a standard Kohonen SOM in order to observe and to analyze the general topological relationships of our reference dataset (official traffic disruption messages).

Second, a geographic self-organizing map (Geo-SOM) is computed for the identification of similar overlapping traffic disruption patterns. Finally, the resulting traffic disruption clusters and their corresponding Geo-SOM weight vectors are correlated with the computed Geo-SOM weight vectors from all retrieved georeferenced tweets, which semantically cover traffic-related information.

Figure 2. Analysis framework. As proposed in (Steiger et al., 2015b).



For assessing the geospatial similarity we are considering the obtained geographic location for all traffic disruptions. The geospatial position of every observation is used within our Geo-SOM to only cluster geographical close input vectors, whereas within the applied standard

Kohonen SOM we omit the geographic space and solely focus on exploring the topological relationship between the non-geographic input components (see sub-section 3.3 below).

The temporal information within each SOM/Geo-SOM is considered by analyzing the attribute characteristics and dependencies within the given time periods. Since tweets posted at a similar time does not imply that they also correspond to each other, we cannot infer any absolute temporal associations between tweets. Unlike GPS trajectories which are a reliable sequence of locations and timestamps, within tweets we do not have any a priori, sequential information. Thus, each tweet is treated as an independent observation. Since official traffic disruptions are retrieved every 15 minutes, we created categorical time bins for every 15 minutes, every hour and every day to cluster traffic disruptions/tweets when sharing similar information close in geographic space and close in time. Note that all derived 15 minute time bins are grouped into either weekday or weekend periods to reflect the well-known bimodal nature of human mobility (Kung et al., 2014). However, the bin width of time intervals can be adjusted in a generic manner (consideration of any given interval, such as day, weeks, months etc.).

3.1 Self-organizing map (SOM) and geographical self-organizing map (Geo-SOM)

In order to spatially organize the topological representation of our data input features, we consecutively apply the standard Kohonen SOM algorithm (Kohonen, 1990, Agarwal & Skupin, 2008) and a Geo-SOM (Bação et al., 2005). To cluster the dataset, it is necessary to determine a reasonable map size. For this reason, we trained random weight maps with different sizes and parameter settings in the initial learning phase. To measure the quality of the standard SOM and the Geo-SOM, we compute the average distance between every input unit and the mapped training pattern after each iteration, known as quantization error (QE) (Table 1). Further, topographical errors (TE) are evaluated to measure the topological preservation and continuity of mapping by assessing the distances between all nonadjacent best matching units and the second best matching unit for all input features (Uriarte and Martín, 2005). Last, we determine the geographical error (GE) by computing the average distance between each geographic input component and the final mapped output neuron (Agarwal and Skupin, 2008). The clustering results show that a 10x10 neuron network entails the least errors (see Table 1).

This is the "Accepted Version" of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

Table 1. Comparison of computed QE, TE, GE between Kohonen SOM and Geo-SOM (k=2)

	SOM			Geo-SOM (k=2)		
	5x5	10x10	15x15	5x5	10x10	15x15
QE	0.424	0.211*	0.308	5.45	0.232*	0.312
TE	0.350	0.316*	0.370	0.326*	0.343	0.385
GE	-	-	-	1.796	0.794	0.623*

All maps have been trained 10,000 times with a subsequent fine tuning phase until convergence is reached, to determine most similar node weights with the smallest distance to the input vector (geospatial, temporal and disruption attributes), known as the Best-Matching Unit (BMU). Since the main goal is to observe local geospatial cluster patterns, the maximum geographical tolerance to search for every BMU is defined as $k=2$ (Bação et al., 2005). Furthermore, weight vectors for every computed Geo-SOM are bivariately correlated (Pearson's r) and also spatially (Local Moran's I) (Anselin, 1995) autocorrelated (threshold distance are all neighboring neurons) in order to assess the spatial interactions of clustering or dispersing BMUs, as well as the topological relationships of identified traffic disruption clusters across the different Geo-SOMs.

4. Case study and results

The analysis framework described in section 3 has been applied in our case study for official traffic disruptions and the Twitter dataset. This section summarizes the results of our analysis.

For our case study of Greater London, we analyzed and compared 129,651 real-time traffic disruptions from Transport for London with 63,407 georeferenced tweets for one month. For the Twitter data acquisition process, only georeferenced tweets within a given bounding box (see table 2 for further description) have been crawled, without any further keyword filtering. Furthermore, tweets covering traffic-relevant information were extracted by semantically analyzing each Twitter post using the Latent Dirichlet Allocation (Blei et al., 2003) topic model. The generative topic modeling approach is a computer linguistics technique which analyzes each tweet's content and assigns a probability-based topic indicator to each tweet, summarizing the frequency and distribution of how words appear and are semantically related to each other. By analyzing the frequency and distributions how words appear within tweets and how certain words relate to each other (e.g. "traffic" and "stuck" since users always mention them together), a list of "latent" topics describing each tweets

content can be retrieved. Only tweets showing the highest probability to be assigned to a given traffic topic were selected for the subsequent analysis (for a detailed explanation of the applied LDA[2] semantic methodology and the corresponding evaluation see (Steiger et al., 2015c)).

Table 2. Meta-information for the selected traffic disruptions and Twitter dataset.

Dataset	Greater London (UK)
Bounding Box (WGS 84)	-0.303, 51.238, 0.554, 51.731
Time span	01/04/2015-31/04/2015
Number of georeferenced tweets before LDA-processing	724,025
Number of georeferenced tweets after LDA-processing	63,407
Number of individual Twitter users after LDA-processing	27,193
Number of georeferenced traffic disruptions (TIMS)	179,651

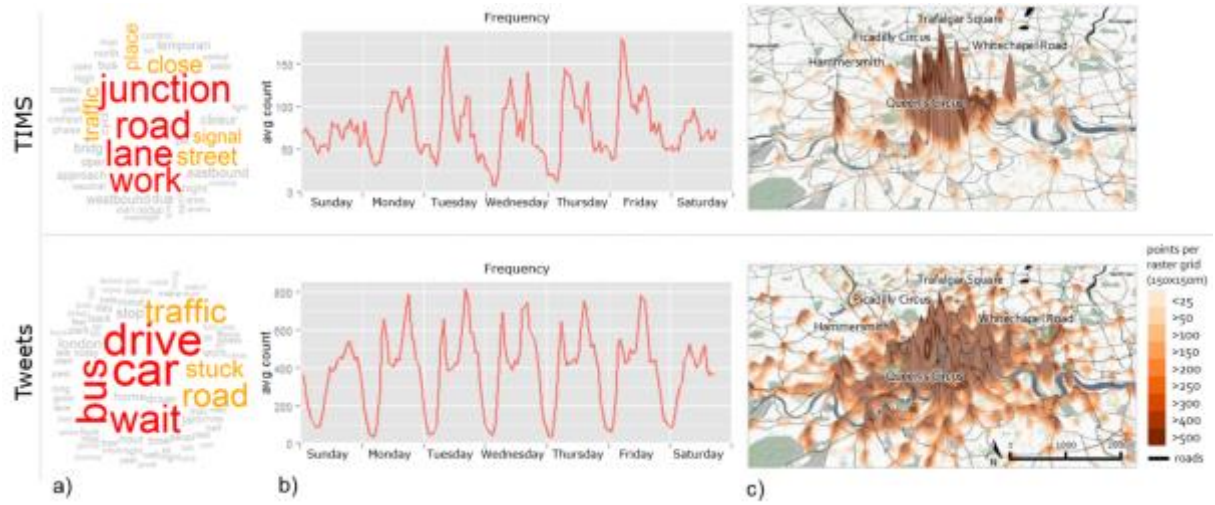
Figure 3 provides an initial comparison between semantic, temporal and geospatial information from official traffic disruptions and traffic-related georeferenced tweets. The observed, most frequently occurring words indicate the diverging semantic content; while, traffic disruption messages include more detailed contextual information about specific landmarks, the semantic information from tweets appears to be more general (3a). Regarding the temporal frequency of traffic-related topics (3b), one can observe a similar daily distribution with the characteristics of morning and evening peak and a decreasing amount of traffic-related posts on weekend periods. However, traffic-related tweets are more homogeneously distributed.

Analyzing the geospatial-semantic distribution of traffic-related posts for both datasets (3c), we can discern a geographical clustering of official traffic messages along central London mobility hubs (e.g., Trafalgar Square, Piccadilly Circus), whereas tweet point densities scatter more across the London road network. However, in order to further investigate similarities and to uncover more complex latent relationships between observed traffic disruption patterns and tweets, the following sub-section 4.1 presents the results of the conducted SOM analysis.

[2] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

This is the “Accepted Version” of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

Figure 3. Comparison of a) overall most frequent associated, occurring words between analyzed tweets and official traffic disruptions in word cloud, b) temporal word frequencies of tweets and traffic messages during weekdays are shown, and c) the geospatial-semantic point densities of tweets and TIMS messages, vertically extruded as spikes and aggregated. (base map: Stamen Design CC BY 3.0, Data by OpenStreetMap CC BY SA, TfL TIMS data: TfL Publicly-available data licensed under the Open Government License v.2.0)



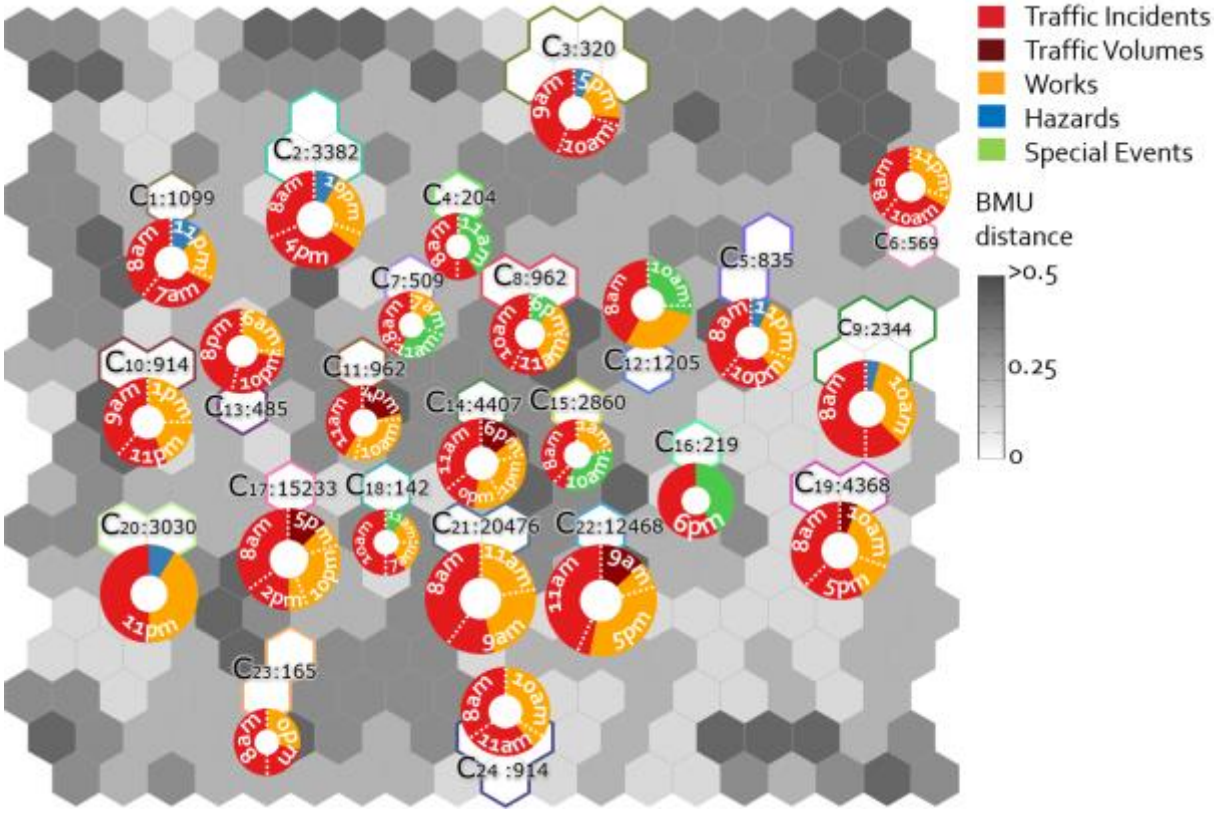
4.1 Results SOM traffic disruption patterns (Figure 4a/4b)

This section presents the resulting traffic disruption clusters after analysis of the geospatial, temporal and attribute information. SOM U-matrices provide the first visual insight into the data characteristics and topological structures by plotting the Euclidean distance between the derived codebook vectors on a two-dimensional hexagonal output space. Darker colors (>0.65) represent distant codebook vectors and dissimilar input attributes of traffic disruptions, whereas light colors denote close adjacent neurons (0.10-0.65) and indicate the presence of distinct traffic disruption clusters with similar geospatial-temporal characteristics.

On the applied standard SOM U-matrix (Figure 4a), the codebook vector distances between the map units are only representing the topological neighborhood relationships of non-geospatial attributes (in our case the traffic disruptions and their temporal characteristics see Figure 3).

This is the “Accepted Version” of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

Figure 4a) U-matrix results of the 10x10 SOM analysis showing inferred cluster patterns with close BMU weight vectors (<0.2), indicating similar input characteristics. Pie chart sizes correspond to the amount of associated disruptions for each cluster.



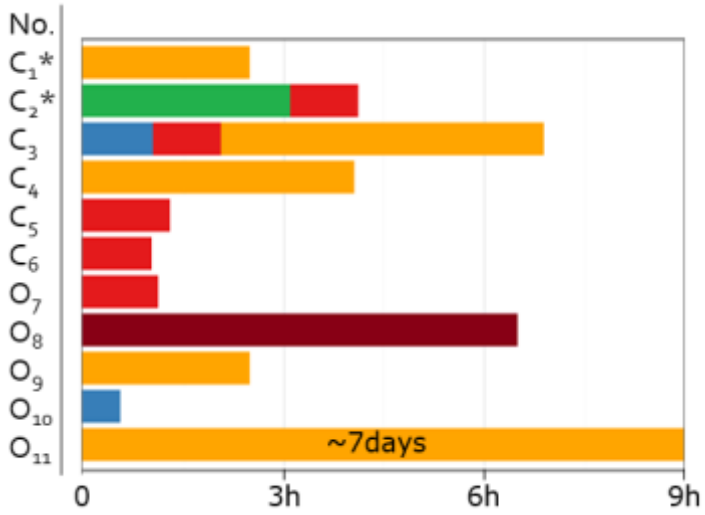
In Figure 4a, the most notable insight of all emerged patterns is that the attributes of traffic events within identical disruption categories strongly differ, depending on the temporal characteristics and appear as separate cluster patterns (dissimilar and distant BMU between C_3&C_4 and C_5&C_6) on the U-matrix. The most associated work disruptions occur during weekdays between 9-11pm (C_3) with a low severity (3-4) and show slightly differing characteristics than work disruptions, which occur in the morning (8-11am) (C_4) with medium severity (2-3). The average disruption length of both clusters is similar (~3 hours). Traffic incidents occurring between Mondays and Wednesdays (C_6) peak between 8-10am and 5-6pm, reflecting typical commuting activities similar to the Geo-SOM results (see Figure 5c). In contrast, traffic incidents between Thursday-Friday (C_5) are assigned to a separate cluster pattern occurring later in the morning (10-12am) and they peak earlier (mostly at 1pm and 5pm) with a longer average disruption length (Figure 4b). Both clusters share some similarities between their input components (e.g., same disruption category). In contrast, work disruptions and traffic incidents on weekends show similar temporal distributions with no

differentiation between morning and evening times, suggesting a stronger similarity between input components, other than during weekdays.

Moreover, the SOM approach is also useful for detecting cluster outliers (O_8/O_9/O_{10}/O_{11} with distant adjacent neurons), which share only a few similarities in comparison to the entire dataset. These observations form a distinctive abnormal traffic pattern with only a small number of associated traffic disruptions. Thus, it would be worthwhile to further investigate the causes of disruption by analyzing the cluster attribute characteristics. All outliers have been labeled with a high severity (1-2) regarding the current or expected impact of the disruption on traffic, requiring a high level of operational attention. O_8 and O_9 are exceptional severe and long lasting traffic volumes and construction works which always occur on Mondays, differing from the other weekday disruption intensities.

Furthermore, cluster outlier O_{10} represents unplanned hazards, which therefore occur at unusual times (most frequently at night between 1-4am) with an average disruption length of less than 32 minutes. Cluster outlier O_{11} consists of long-term construction sites with an average disruption length of 7 days, which mainly concentrate in the inner city.

Figure 4b). Average disruption length per disruption category within observed SOM clusters and outliers. Note that *marked clusters only have distinctive characteristics on weekend periods and are thus not observed during weekdays.



4.2 Results Geo-SOM traffic disruption patterns (Figure 5a/5b/5c)

On the Geo-SOM U-Matrix in Figure 5a, one can generally observe several notable disruption patterns. Overall, the highest number of associated, similar traffic disruption patterns cluster

along major arterial roads and public squares (C_{21}: Westway with 20,476 disruption messages, C_{17}: Victoria Square with 15,233 messages, C_{22}: Elephant & Castle with 12,468 messages, C_{14}: Trafalgar Square with 4,407 messages). Special and planned events, such as concerts, demonstrations, marches, sporting events, etc., mostly take place in the inner city (C_7/C_{12}/C_{15}), whereas most of the Geo-SOM clusters with hazard disruptions (obstructions, damages, fire etc.) are concentrated along major arterial roads (C_1/C_2/C_5). The longest average disruption length (~2 hours) within all traffic message categories is detected along major motorways and ring roads (C_1/C_2/C_3/C_5/C_9/C_{19}) (see Figure 5b).

Focusing on the temporal characteristics of the dataset, a strong concentration of traffic incidents (accidents, breakdowns, etc.) with similar geospatial-temporal attributes forms strong clusters with close BMU along major arterial roads (C_1/C_2/C_3/C_6/C_{10}/C_{19}/C_9/C_{24}). These clusters most frequently occur during weekdays in the morning (8am-11am) and in the afternoon at 4-5pm, reflecting typical commuting peak hours (Figure 5c). Major traffic volume disruption patterns, due to the "sheer weight of traffic" (predetermined category of disruption as specified in TIMS feed), occur predominantly on weekdays within the inner city around 4-6pm, causing an average disruption length of longer than 40 minutes (C_{11}/C_{14}).

Furthermore, a clear temporal dependence of traffic disruptions, reflecting an inbound and outbound traffic flow, can be seen along the outer London Ringways (C_5/C_6/C_9), where traffic incidents most frequently cluster on weekdays at 8am. These roads merge into the London Inner Ring Road (strong Geo-SOM cluster C_8 with close neurons), where all occurring traffic incidents generally peak later on weekdays at 10am. A similar cluster pattern exists between the Rochester Way C_{19} (traffic incidents peak weekday 8am) and major square Elephant & Castle C_{22} (most traffic incidents occur on weekdays at 11am).

Within the work disruption category, the planned element of these disturbances is also reflected within the temporal dimension, since main utility works are carried out during weekdays within low traffic periods at night (11pm-1am). Work disruptions during the day (C_9/C_{19} construction works at 10am), result in a notable increase of traffic incidents and traffic volume disruptions since they have a similar temporal evolution. C_4/C_{16} only occur on weekends and are in the proximity of large stadiums/public squares, consisting of a notable amount of traffic disruptions due to special events (10-11am), along with traffic accidents (happening earlier at 8am with an average disruption length of 25-28 minutes).

This is the “Accepted Version” of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

Figure 5a). U-matrix results of the 10x10 Geo-SOM (k=2) analysis showing the inferred cluster pattern with close BMU weight vectors (<0.2), indicating similar input characteristics. The sizes of all pie charts denote the amount of associated disruptions for each cluster.

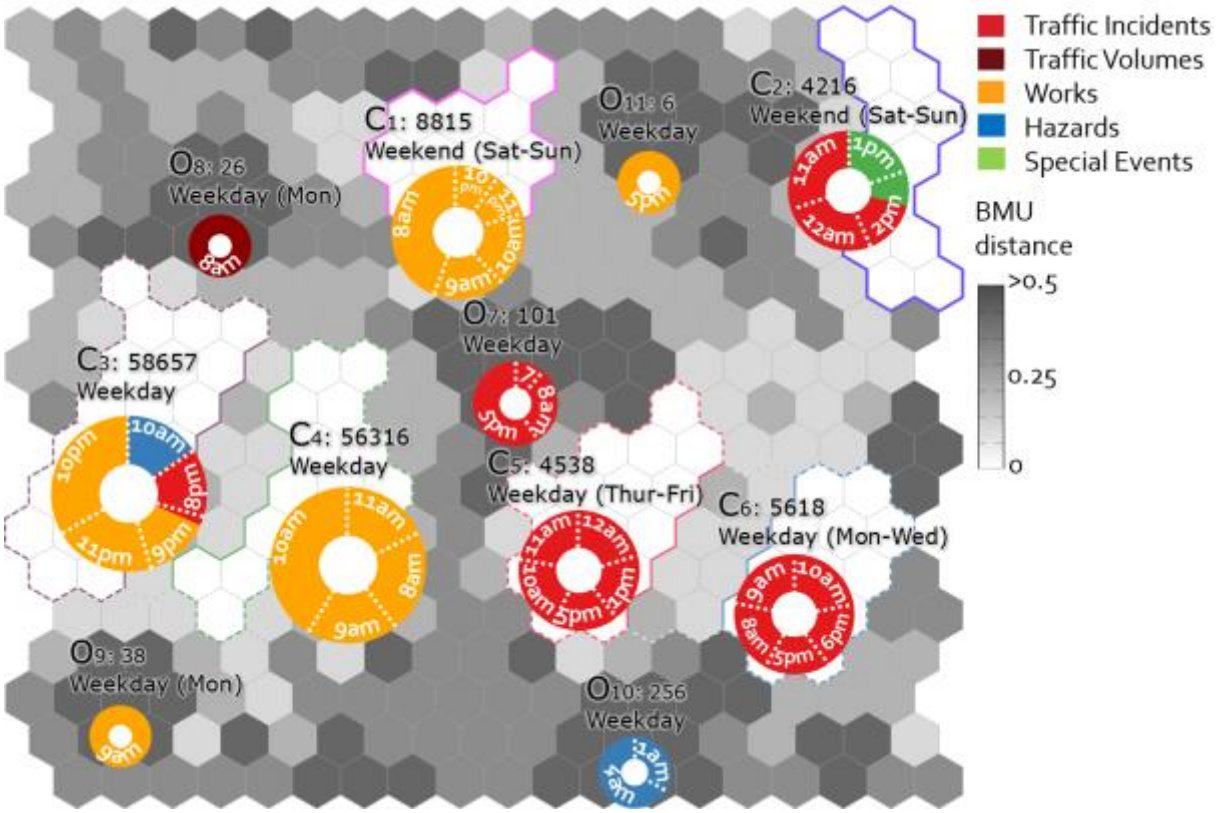
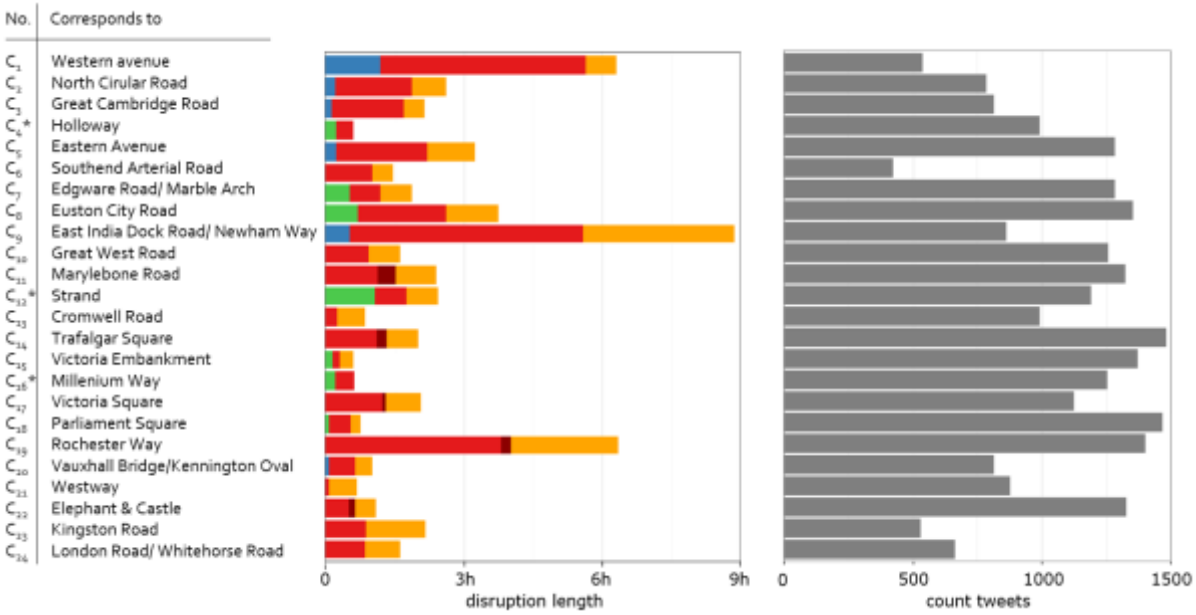


Figure 5b). Average disruption length per disruption category within observed Geo-SOM clusters and comparison with number of georeferenced tweets within these Geo-SOM disruption clusters. The different colors indicate the category of disruption and correspond to Figure 5a. Note that *marked clusters only have distinctive characteristics on weekend periods and are thus not observed during weekdays.

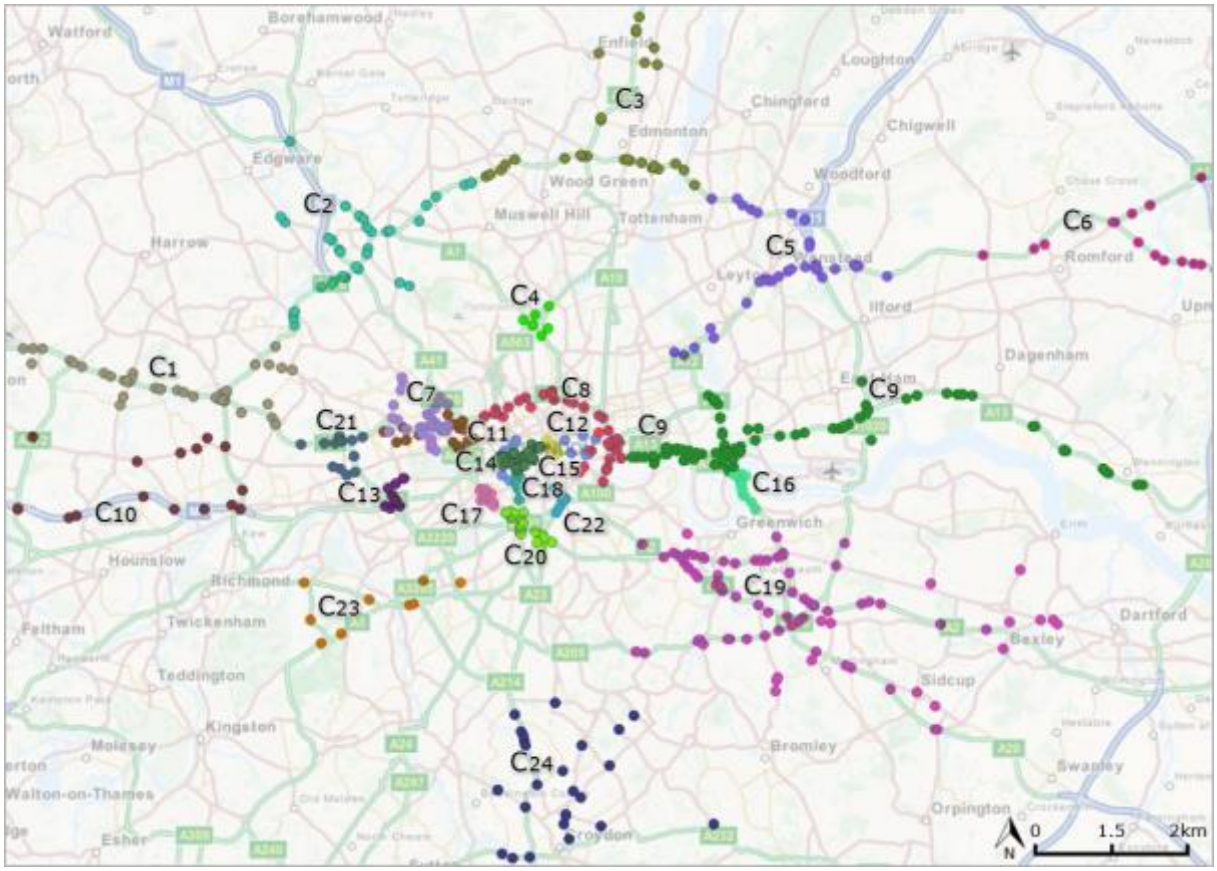


4.3 Result correlation of derived Geo-SOMs (Figure 6)

We applied the Geo-SOM approach, described in section 3, to official traffic disruption data (TIMS) as a reference dataset by using geospatial, temporal and disruption attribute information. Next, the Geo-SOM is also utilized to cluster semantically classified and pre-filtered, georeferenced tweets covering traffic-related events by using their geospatial and temporal information (see Figure 2 for the overall analysis framework). Subsequently, the resulting BMU weight vectors from both Geo-SOMs are autocorrelated to measure the degree of spatial association for all BMUs and their neighboring neurons, in order to statistically quantify non-random local indicator of spatial association (LISA) (Anselin, 1995). Afterwards, we applied the bivariate Pearson correlation to assess the mutual statistical dependence of each BMU weight vector for all Geo-SOMs. This way, latent similar structures of inferred mobility disruption patterns, divided into five traffic disruption categories (hazard, special events, traffic volume, traffic incident, construction works), are correlated and compared to the observed characteristics from tweets to analyze whether the official traffic disruptions are reflected within a similar tweet post behavior. Figure 6 visualizes the

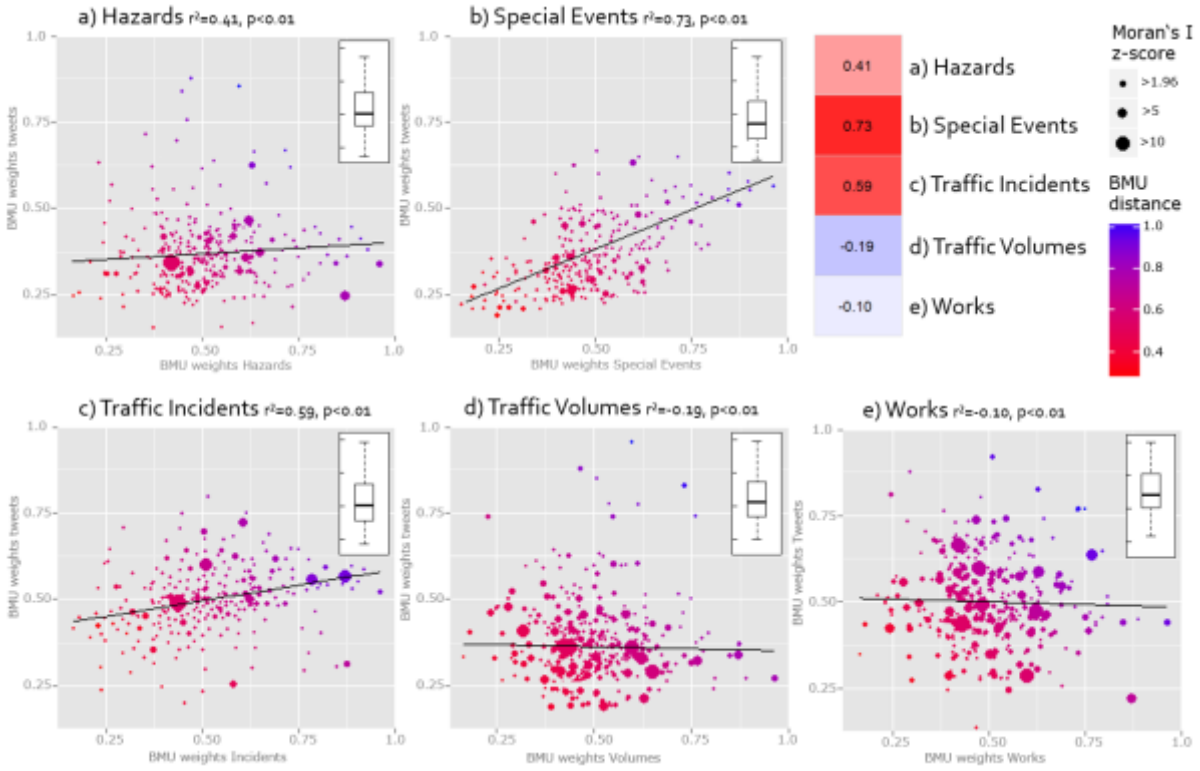
coherence between the computed neuron weights for each traffic disruption category and respective tweets. The bivariate correlation of BMU distances indicates similar (BMU weights for mutually compared Geo-SOM <0.5) and dissimilar (BMU weights for mutually compared Geo-SOM >0.5-1) input data characteristics for tweets and each traffic disruption category. A positive spatial autocorrelation ($-1.96<I_i>1.96$), denoted by different point sizes, implies the presence of distinctive, non-random spatial pattern for the given neuron weight and its neighboring neuron weights.

Figure 5c). Derived clusters of the Geo-SOM mapped to the geographic space (base map: Tiles Courtesy of MapQuest, data from OpenStreetMap contributors, TfL TIMS data: TfL Publicly-available data licensed under the Open Government License v.2.0).



This is the "Accepted Version" of the paper published as Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010.

Figure 6. Results of correlated Geo-SOM BMU distance weights for every disruption category with traffic-related tweets.



From the correlations depicted in Figure 6, we can infer the following:

- Geo-SOM BMU weight vectors between **tweets and special event disruptions** show a **strong correlation** (r=0.73). BMU weight vectors are very close (boxplot: middle quartile<0.4, lower whisker with the first quartile starting from 0.1) and the surrounding neurons' weights also spatially autocorrelate, which demonstrates the overall similarity of patterns (Figure 6b).
- Geo-SOM BMU weight vectors between **tweets and traffic incident** disruptions show a **moderate correlation** (r=0.59). BMU weight vectors are close (boxplot: middle quartile=0.5, lower whisker with the first quartile starting from 0.2) and cluster patterns have a similar spatial autocorrelation (Figure 6c).
- Geo-SOM BMU weight vectors between **tweets and hazard disruptions** show a **weak correlation** (r=0.41). BMU weight vectors have a medium distance (boxplot: middle quartile=0.5, lower whisker with the first quartile starting from 0.25) and cluster patterns have a similar spatial autocorrelation (Figure 6a).

- Geo-SOM BMU weight vectors between **tweets and traffic volume** (r=-0.19) **and tweets and construction works** (r=-0.10) show **no correlation,** since the observed BMU weight vectors have a medium to far mutual distance (boxplot: middle quartile>0.6, lower whisker with the first quartile beginning from 0.26) and geospatial, temporal attribute patterns show no association with a very dissimilar spatial autocorrelation (Figure 6d and 6e).

Furthermore, the severity and duration of inferred traffic disruption clusters (Figure 5a) are compared to the number of assigned traffic-related tweets for each neuron (Figure 5b). Generally, one can observe a higher amount of traffic-related tweets in the vicinity of major public squares and within the inner city (C_7/C_8/C_{14}/C_{17}/C_{18}) reflecting the varying distribution and population frame of Twitter posts. Over all categories, there is no association between the average durations and severities of the disruption clusters and generated tweets. However, there is a weak correlation (r=0.27) between traffic volume disruptions, their average duration as well as severity and the amount of tweets assigned to the same neuron. Geo-SOM clusters with a high amount of special events (C_8/C_{12}) and traffic volume disruptions (C_{11}/C_{14}/C_{19}) have, compared to all the other Geo-SOM clusters, in average 2% more assigned tweets.

5. Discussion of results and applied methods

The SOM result within sub-section 4.1 revealed varying durations of observed disruption patterns depending on the time of their occurrence. Detected incidents and corresponding types of disruptions differ in severity and duration through the day time during weekdays and weekends, reflecting the bimodal (peak hour) distribution of human mobility (Wang et al., 2014). The SOM assists to explore these temporal variations of input attributes by topologically grouping dissimilar and similar disruptions together. Hence, work and traffic incident related disruption categories with a dissimilar duration and severity between mornings and evenings on weekdays are more distant to each other and appear as separate structures on the neuronal map (see Figure 4a C_3&C_4 and C_5&C_6). In contrast, work and traffic disruptions frequently occurring on weekends (C1 and C2) have similar daily and are thus mapped to closer relating neurons. As a result, we can confirm that in the SOM, official TIMS messages reflect typical well-known human mobility patterns and are therefore a reliable and trustworthy reference dataset for the comparison with traffic-related human sensor observations (tweets). Distant neurons on the SOM U-matrix (Outliers $O_{8\text{-}11}$) indicate

distinctive abnormal traffic events with differing characteristics (e.g. exceptional duration) from the majority of disruptions.

When incorporating the geospatial components using a Geo-SOM (sub-section 4.2), one can identify frequently repeating, daily patterns with similar time-dependent disruption characteristics along major arterial (ring) roads, being a proxy indication of a specific collective inflow and outflow mobility behavior. The U-matrix (see Figure 5a) and the detected topological cluster structures, show that disruption messages and their attributes are allocated to similar neurons and form distinct clusters along road segments, when linked back to geographic space (see Figure 5c). With the Geo-SOM, consistent cluster characteristics of disruptions show similar temporal patterns at certain geographic locations and help to uncover complex topological structures of the London street network. Geo-SOM structures of disruption messages form distinct clusters along road segments and public squares. Outer peripheral roads appear as similar cluster patterns on the Geo-SOM and have typical evening and morning rush hour peaks with a characteristic distribution of disruption categories (predominantly traffic incidents). These patterns are different from inner city clusters along major hubs and central ring roads, where the types of traffic disruptions occurring are more diverse. Thus, the geographical clusters (see Figure 5c) of frequently repeating disruption patterns, reflect the daily intra-urban movement of traffic flow and the underlying hierarchical road system of arterial roads leading into the city, which typically intersect at major transportation hubs.

The correlation results of the computed Geo-SOMs (see sub-section 4.3) between official traffic messages, as a reference dataset, and traffic-related tweets, have shown that observations from Twitter data and their geospatial-temporal characteristics are statistically associated and share similarities with certain traffic events, depending on the type of disruption event. Special and planned events, together with traffic incidents and hazards, are very well reflected in tweets since these observations are mapped to similar neurons (see Figure 6), indicating similar geospatial-temporal patterns on each computed Geo-SOM with a positive spatial autocorrelation. This result also demonstrates the heterogeneity of the given social media observations (as stated in sub-section 2.1), because other types of traffic disruptions (e.g., construction works and traffic volumes) show no correlation. Geo-SOM cluster patterns with a high amount of special events and traffic volume disruptions also show a slightly higher amount of associated tweets compared to, for instance, work disruptions, suggesting that these events trigger more tweets. In general, the presented framework

demonstrated its usefulness for the exploration of complex mobility disruption patterns. The assessment of geospatial and non-geospatial input components with the SOM approach revealed characteristic temporally repeating incident and disruption patterns depending on the type of disruption event. It is also worth noting that the approach enables the detection of distinctive abnormal traffic events with a high severity and increasing duration length. The analysis of geospatial components uncovered time-dependent disruption characteristics along the London street network as a proxy for the underlying mobility flows Further, the proposed workflow is suited for analyzing Twitter data that are characterized by uncertainty in geographic and semantic dimensions.

5.1 Limitations

Several limitations of the conducted analysis need to be addressed. First, the user generated, textual content of tweets is noisy, making it challenging to apply natural language processing (NLP) techniques to identify meaningful information (Ling et al., 2012). Thus, the application of the LDA model for the selection of traffic-relevant tweets may result in misclassification of latent semantic topics. However, this is a common limitation of studies that involve text mining methods and we apply commonly known semantic analysis techniques, which reflect the current state of research within computational linguistics (Aggarwal and Zhai, 2012).

Furthermore, in our analysis, we assume that the semantic content of tweets reflects an observation at a certain location and time they have been posted. However, there might be a temporal latency between real-world events and their appearance in social media. Nevertheless, existing social media studies (Sakaki et al., 2010) have shown that events are actually broadcasted even earlier in social media than in official data sources, making it potentially promising also for the application of real time traffic event detection. Therefore traffic related tweets are also of value for the real-time traffic detection of events, since these observations are reflected earlier in social media than within the official traffic data reports. Further, Twitter data itself is sparse and has a heterogeneous geospatial-temporal distribution. Therefore, the study needs to be reproduced in other areas to further compare how human sensor observations from Twitter data are related to traffic event characteristics in other urban environments. The use of additional demographic and economic variables could further explain the influence of socioeconomic factors on the detected traffic patterns.

Regarding the application of SOMs and Geo-SOMs one main issue arises. In general, the inference of traffic patterns from official data by using SOMs has been proven by several

existing studies and our previous research also confirms the validity and suitability of SOMs for the exploration of geospatial-temporal and semantic clusters from Twitter data. As a dimensionality reduction tool, SOMs allow the analysis and topology preservation of the input properties from high-dimensional attributes in a combined manner. Prior to the comparison of each computed Geo-SOM, QE, TE and GE between every input unit and the mapped output neuron after each Geo-SOM training iteration have been quantified. The results showed a stable output of neuronal map structures with constant distance errors (see error computation results sub-section 3.1).

6. Conclusion and future work

This paper presents the results of a combined SOM/Geo-SOM analysis framework for the detection of distinctive mobility disruption patterns from official TIMS messages and the comparison between traffic-relevant, georeferenced Twitter messages.

We have chosen a SOM (sub-section 4.1) and a Geo-SOM (sub-section 4.2), in order to assess non-geospatial components and the combination with geospatial components separately.

First, we uncovered latent temporal relationships of traffic disruption properties and their temporal variations on a non-geospatial, standard Kohonen SOM. This approach only preserves the input dataset characteristics of non-geospatial components, regardless of geographic space and is not restricted to any geographical neighborhood. The results (4.1) showed for construction work messages and traffic incidents a characteristic difference of duration and severity between commuting peak hours (morning and evening rush hour) and between weekdays and weekends.

Second, we assessed the influence of location on a geographically enforced Geo-SOM (4.2). The Geo-SOM enables the comparison and correlation of several Geo-SOM results (with same training parameters), because features similar in attribute space and geographic space are also mapped onto similar output Geo-SOM locations after the dimensionality reduction. Therefore, they represent a fraction of the relative, original geographic properties of the input data. The resulting geographic clusters of linked road segments with similar temporal traffic disruption patterns facilitate the characterization of a certain mobility inflow and outflow behavior and revealed previously unexpected connections.

Answering RQ1, our findings have shown strong correlations between inferred spatiotemporal clusters from tweets with proximity to special events, traffic incidents and hazard reports. Therefore, georeferenced tweets are helpful for the real-time detection of

event-related traffic disruptions (concert, demonstrations, sport events etc.) and are reflected in collective human mobility patterns. Thus tweets can add information in order to estimate the effect (e.g. severity/ intensity) on the infrastructure when no official traffic data source are available, especially for unplanned events. In addition, official real time traffic information served as a reference source providing ground-truth for the reliability and plausibility of these observed mobility patterns.

Moreover, in association with a high severity and long duration of traffic volume disruptions, there is statistical evidence that the amount of generated Twitter posts with similar geospatial, temporal characteristics are increasing accordingly.

With regard to RQ2, we have shown the similarities of mobility phenomena from tweets and their connection with traffic disruptions, making social sensor observations a reliable proxy for some traffic-related events. This opens up the possibility to utilize the more heterogeneous and wider geospatial-temporal distribution (see Figure 3c) of social media messages to harness further information regarding the detection of near real-time flow conditions inside road networks, especially during the development of an actual disruption event.

Our results suggest that particularly special events, such as concerts, demonstrations, sports events, etc. are well reflected within Twitter and provide complementary information about possible collective movements, since people talk about the event beforehand and follow similar mobility patterns (Steiger et al., 2015a). These complex events especially, are hard to forecast from classic detectors and therefore social media can be used to enrich existing information. This newly gained knowledge may support decision-makers during traffic events in a way that social media and official authorities complement each other.

Therefore the results answer how and when tweets should be used for extracting mobility behavior: Answering “How” the presented SOM framework analyzes the temporal, spatial and textual dimension of each tweet in a combined manner. Furthermore the results can be easily compared with official data to underline the significance of social media for human mobility analysis. Answering “When” the results show which traffic disruption categories are reflected in social media (special events, traffic incidents and hazards), demonstrating in what traffic analysis scenarios social media can be used for as an additional source of information. In opposition to the geospatial-temporal distribution, the textual information from tweets can only be used marginally to semantically enrich traffic disruption information, due to the detailed resolution of traffic conditions (see comparison of the most

frequent terms in Figure 3a). Nonetheless, the results demonstrated the effectiveness of the proposed methodology to uncover similar characteristics and latent disruption patterns from official data and georeferenced tweets (see Figure 5a). This implies the practical use of tweets to detected real-time traffic events which can add information when no official traffic data sources are available, especially for unplanned events such as demonstrations. Tweets here help to detect these events but also to extract the underlying mobility pattern in order to estimate the effect (severity/intensity) on the infrastructure.

As for future research, the proposed framework will be expanded and applied across other geographic areas (e.g. cities) in order to further compare inferred traffic event characteristics. Additionally, characteristics of input variables could be assessed and tracked over time by applying a SOM/Geo-SOM for every resulting disruption cluster in order to investigate principal, temporally changing components. Another promising direction is the prediction of certain human behavior and disruption patterns before an actual event, based on the exploration of existing messages and their cluster characteristics. Finally, we are planning to use additional demographic and economic variables within the proposed SOM/Geo-SOM approach to observe how certain socioeconomic neighborhood characteristics (e.g., car ownership, age, etc.) might influence and explain observed traffic patterns.

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. Furthermore, we also thank Transport for London for providing free available real time transportation data licensed under the Open Government License v.2.0.

References

Agarwal, P., Skupin, A., 2008. Self-Organising Maps: Applications in Geographic Information Science. John Wiley & Sons.

Aggarwal, C., Zhai, C., 2012. Mining text data. Springer Science & Business Media.

Anselin, L., 1995. Local Indicators of Spatial Association-LISA. Geogr. Anal. 27, 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Asamer, J., Din, K., Werner, T., 2007. Self organizing maps for traffic prediction, in: Artificial Intelligence and Applications. pp. 24–29.

Asif, M.T., Dauwels, J., Goh, C.Y., Oran, A., Fathi, E., Xu, M., Dhanya, M.M., Mitrovic, N., Jaillet, P., 2014. Spatiotemporal patterns in large-scale traffic speed prediction. IEEE

Trans. Intell. Transp. Syst. 15, 794–804. doi:10.1109/TITS.2013.2290285

Bação, F., Lobo, V., Painho, M., 2005. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. Comput. Geosci. 31, 155–163. doi:10.1016/j.cageo.2004.06.013

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022. doi:10.1145/950000/944937/3-993

Boulet, R., Jouve, B., Rossi, F., Villa, N., 2008. Batch kernel SOM and related laplacian methods for social network analysis. Neurocomputing 71, 1257–1273. doi:10.1016/j.neucom.2007.12.026

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’11. ACM, New York, NY, pp. 1082–1090. doi:10.1145/2020408.2020579

Couronne, T., Beuscart, J., Chamayou, C., 2013. Self-Organizing Map and social networks: Unfolding online social popularity. arXiv Prepr. arXiv1301.6574.

Cranshaw, J., Schwartz, R., Hong, J., Sadeh, N., 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City., in: Sixth International AAAI Conference on Weblogs and Social Media - ICWSM. AAAI.

Feng, C.-C., Wang, Y.-C., Chen, C.-Y., 2014. Combining Geo-SOM and Hierarchical Clustering to Explore Geospatial Data. Trans. GIS 18, 125–146. doi:10.1111/tgis.12025

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

Fotheringham, A., Wong, D., 1991. The modifiable areal unit problem in multivariate statistical analysis. Environ. Plan. A 23, 1025–1044. doi:10.1068/a231025

Gao, S., 2014. Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age. Spat. Cogn. Comput. 15, 86–114. doi:10.1080/13875868.2014.984300

Gonzalez, M., Hidalgo, C., Barabasi, A., 2008. Understanding individual human mobility patterns. Nature 453, 779–782. doi:http://dx.doi.org/10.1038/nature07850

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. doi:10.1007/s10708-007-9111-y

Gorricha, J., Lobo, V., Costa, A., 2013. A Framework for Exploratory Analysis of Extreme Weather Events Using Geostatistical Procedures and 3D Self-Organizing Maps. Int. J. Adv. Intell. Syst. 6.

Hagenauer, J., Helbich, M., 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: Proceedings of the 25th Conference of the International Cartographic Association.

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, in: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp ’13. ACM Press, ACM, New York, NY, p. 1. doi: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. Cartogr. Geogr. Inf. Sci. 41, 260–271. doi:10.1080/15230406.2014.890072

Helbich, M., Hagenauer, J., Leitner, M., Edwards, R., 2013. Exploration of unstructured narrative crime reports : an unsupervised neural network and point pattern analysis approach. Cartogr. Geogr. Inf. Sci. 40, 326–336. doi:10.1080/15230406.2013.779780

Jiang, B., Harrie, L., 2004. Selection of Streets from a Network Using Self-Organizing Maps. Trans. GIS 8, 335–350. doi:10.1111/j.1467-9671.2004.00186.x

Jiang, B., Yin, J., Zhao, S., 2009. Characterizing the human mobility pattern in a large street network. Phys. Rev. E 80.

Jurdak, R., Zhao, K., Liu, J., 2015. Understanding Human Mobility from Twitter. PLoS One 10. doi:10.1371/journal.pone.0131469

Kangas, J., 1992. Temporal knowledge in locations of activations in a self-organizing map, in: I. Aleksander and J. Taylor (Ed.), Artificial Neural Networks 2. Amsterdam, pp. 117–120.

Kauko, T., 2005. Using the self-organising map to identify regularities across country-specific housing-market contexts. Environ. Plan. B 32.

Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78, 1464–1480.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69.

Krumm, J., Caruana, R., Counts, S., 2011. Learning Likely Locations, in: User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, pp. 64–76. doi: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. doi:10.1371/journal.pone.0096180

Lenormand, M., Tugores, A., Colet, P., Ramasco, J.J., 2014. Tweets on the road. PLoS One 9, e105407. doi:10.1371/journal.pone.0105407

Li, L., Goodchild, M.F., Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartogr. Geogr. Inf. Sci. 40, 61–77. doi:10.1080/15230406.2013.777139

Liu, Y., Kang, C., Gao, S., Xiao, Y., Tian, Y., 2012. Understanding intra-urban trip patterns from taxi trajectory data. J. Geogr. Syst. 14, 463–483. doi:10.1007/s10109-012-0166-z

Liu, Y., Sui, Z., Kang, C., Gao, Y., 2014. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS One 9. doi:10.1371/journal.pone.0086026

Miller, H., Han, J., 2009. Geographic data mining and knowledge discovery, in: Handbook of Geographic Information Science. pp. 352–366. doi:10.4324/9780203468029_chapter_1

Miller, H.J., Goodchild, M.F., 2014. Data-driven geography. GeoJournal 1–13. doi:10.1007/s10708-014-9602-6

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. A tale of many cities: universal patterns in human urban mobility. PLoS One 7. doi:10.1371/journal.pone.0037027

Openshaw, S., Blake, M., Wymer, C., 1995. Using neurocomputing methods to classify Britain’s residential areas. Innov. GIS 2, 97–111.

Peng, C., Jin, X., Wong, K., Shi, M., Liò, P., 2012. Collective human mobility pattern from taxi trips in urban area. PLoS One 7. doi:10.1371/journal.pone.0034487

Sagl, G., Delmelle, E., Delmelle, E., 2014. Mapping collective human activity in an urban environment based on mobile phone data. Cartogr. Geogr. Inf. Sci. 41, 272–285. doi:10.1080/15230406.2014.888958

Sakaki, T., Okazaki, M., 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, ACM. ACM, New York, NY, pp. 851–860. doi:10.1145/1772690.1772777

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

Song, C., Qu, Z., Blumm, N., Barabási, A., 2010. Limits of predictability in human mobility. Science (80-.). 327, 1018–1021. doi:10.1126/science.1177170

Spielman, S., Thill, J., 2008. Social area analysis, data mining, and GIS. Comput. Environ. Urban Syst. 32, 110–122. doi:10.1016/j.compenvurbsys.2007.11.004

Steiger, E., Ellersiek, T., Resch, B., Zipf, A., 2015a. Uncovering latent mobility patterns from Twitter during mass events. GI Forum J. 1–2015, 525–534. doi:10.1553/giscience2015s525

Steiger, E., Resch, B., Zipf, A., 2015b. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. Int. J. Geogr. Inf. Sci. 30, 1694–1716. doi:10.1080/13658816.2015.1099658

Steiger, E., Westerholt, R., Resch, B., Zipf, A., 2015c. Twitter as an indicator for whereabouts of people ? Correlating Twitter with UK census data. Comput. Environ. Urban Syst. 54, 255–265.

Transport for London, 2007. Data Feed Specification for Developers, Surface Science.

Ultsch, A., Vetter, C., 1995. Self-Organizing-Feature-Maps versus statistical clustering methods: a benchmark. Research Report.

Uriarte, E., Martín, F., 2005. Topology preservation in SOM. Int. J. Appl. Math. Comput. Sci. 1, 19–22.

Wakamiya, S., Lee, R., Sumiya, K., 2011. Crowd-based Urban Characterization: Extracting Crowd Behavioral Patterns in Urban Areas from Twitter, in: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM, ACM, New York, NY, pp. 77–84. doi:10.1145/2063212.2063225

Wang, X.-W., Han, X.-P., Wang, B.-H., 2014. Correlations and scaling laws in human mobility. PLoS One 9, e84954. doi:10.1371/journal.pone.0084954

Watts, M., Worner, S., 2009. Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. Ecol. Modell. 821–829. doi:10.1016/j.ecolmodel.2008.12.016