Computing
https://doi.org/10.1007/s00607-021-00944-8

SPECIAL ISSUE ARTICLE

Analysing environmental impact of large-scale events in public spaces with cross-domain multimodal data fusion

Suparna De[1] · Wei Wang[2] · Yuchao Zhou[3] · Charith Perera[4] · Klaus Moessner[5] · Mansour Naser Alraja[6]

Received: 4 August 2020 / Accepted: 24 March 2021

Abstract

In this study, we demonstrate how we can quantify environmental implications of large-scale events and traffic (e.g., human movement) in public spaces, and identify specific regions of a city that are impacted. We develop an innovative data fusion framework that synthesises the state-of-the-art techniques in extracting pollution episodes and detecting events from citizen-contributed, city-specific messages on social media platforms (Twitter). We further design a fusion pipeline for this cross-domain, multimodal data, which assesses the spatio-temporal impact of the extracted events on pollution levels within a city. Results of the analytics have great potential to benefit citizens and in particular, city authorities, who strive to optimise resources for better urban planning and traffic management.

Keywords Air pollution · Multimodal data fusion · Social event-pollution correlation · Social computing · Urban computing

Mathematics Subject Classification 62J10 · 62H20 · 91B76

This work was supported by the European Commission, Horizon 2020 Programme, TagItSmart! Project, under Contract 688061. M. N. Alraja's work was supported by The Research Council (TRC), Sultanate of Oman (Block Fund-Research Grant).

✉ Suparna De
s.de@ieee.org

[1] Departments of Digital Technologies, University of Winchester, Winchester, UK

[2] School of Advanced Technology, Xian Jiaotong Liverpool University, Suzhou 215123, China

[3] Institute for Communication Systems (ICS), University of Surrey, Guildford GU2 7XH, UK

[4] School of Computer Science and Informatics, Cardiff University, Cardiff, UK

[5] Faculty of Electrical Engineering and Information Technology, Chemnitz University of Technology, Chemnitz, Germany

[6] Department of Management Information Systems, College of Commerce and Business Administration, Dhofar University, Salalah, Oman

Published online: 12 April 2021

1 Introduction

With increasing migration of people to urban areas (55% of the global population lives in urban areas in 2018, expected to rise to 68% by 2050 [1]), sustainable urbanization has been identified as key to successful development [2]. This necessitates a successful management of urban growth, which largely depends on three dimensions: economic, social and environmental [2]. An insight into the interplay of these dimensions, e.g., by using urban computing techniques, can contribute to an understanding of the evolving needs for limited urban resources (e.g. roads, public transport and other shared public spaces). Existing initiatives in this regard, comprising mainly the environmental and/or economic dimensions, focus on the impact of citizens economic activities (manifest through human mobility and traffic) on the environment. Some representative studies include data-driven machine learning (ML)-based approaches for air quality characterization of cities [13,19,20,41,45], derivation of human mobility patterns based on activities [23,28,30] and prediction of carbon emissions from city transport [27].

Social events in cities, such as sports, cultural or staged demonstrations, involve much more large-scale human and traffic movement than rural areas [3]. The environmental cost (manifested through impact on pollution levels) of such events is largely unexplored, which can partially be attributed to the lack of real-time data sources. In recent years, online social networks (OSNs) have been flourishing and contain rich information about such events [21,42]. This is driven by the growing ubiquitous use of smartphones enabled with GPS tracking capabilities and recent progress in communication networks, which has led to the rise of people sharing city-related messages and mobility updates on OSNs such as Twitter and Foursquare [42]. Citizen sensing has been widely recognized as a complementary and corroborative information source for understanding a citys dynamics [42], with the massive amount of data generated at high frequency, which is representative of the natural, unconstrained human behaviour at very large scales [21]. There has been research on the use of open data from OSNs to detect city-specific events, e.g., large-scale people and traffic movements [21,23,30,42], traffic incidents [4,5] and natural disasters [7,10,11].

This work aims to address the question of evaluating the environmental impact of large city-wide social events, which usually involve lots of people and traffic movement. It combines social event detection and sensor data processing in order to analyse and quantify such impact. To meet this aim, we analyze spatio-temporal big urban data, i.e. datasets containing spatial, temporal and category information (s, t, c) [12,45] of cross-domain and multimodal nature, by exploiting social sources such as Twitter (for city events) and open sensor observations' datasets for pollutants. Existing approaches addressing fusion of sensor observations with OSN data, fall mainly into two categories: (1) terms mined from OSN messages are used as 'subjective context descriptors' for anomalies or trends observed in the physical sensor data (e.g. traffic or air quality patterns) [6,34,40], and (2) those that correlate OSN and physical sensor streams, where both are in the same modality, i.e. numerical [22,25,33,36]. Hence, the existing state of the art does not address fusion of multimodal data streams, i.e. OSN text messages and air pollutant numerical data.

1.1 Motivating scenario

To understand the needs and challenges behind the work, we present a motivating scenario. Consider that there is a football World Cup match planned to be held in a stadium located at the outskirts of the city in a couple of days' time. With prior experience of the environmental impact of sporting events at such scale, the city authorities could schedule an appropriate number of extra public transport means along the main transport routes. With fans arriving in the city a few days in advance of the match, there is suddenly a flash mob of thousands of people congregating at the city centre. The city authorities become aware of this in real-time from messages posted on Twitter, receiving also an approximation of the number of people involved. Simultaneously, air pollution peaks are observed at a number of monitoring sites around the city, showing appreciable correlations at the locations people are tweeting from. As a result, the pollution and congestion alerts in the city dashboard are updated and reflected in screens around the city: bus stops, car parks, shopping centres etc. High pollution alerts, together with locations affected, are pushed to the city apps that the citizens have installed on their smartphones. The coordinated public transport and emergency services are also prepared effectively and efficiently to respond to the situation.

Realization of the above scenario presents some needs: (1) to identify the relation between large social activities and environment impact at a more fine-grained level; (2) to better inform citizens about the environment to enable them to make better decisions for plans and activities; (3) to inform the city authorities about the possible problems and causes for better transportation and infrastructure planning. This scenario also introduces a number of challenges since it involves real-time analysis of textual and numerical data in conjunction, in order to derive meaningful information. The system needs to derive pollution anomalies in an unsupervised manner since different urban regions may have different spatio-temporal baselines (i.e. seasonal and locality variances) for pollution levels. This scenario can also be extended to include sensor networks providing a diversity of data types, i.e. multimedia and scalar types. Scalar data can be textual or numerical and multimedia data can contain audio, video, or image segments (e.g. from surveillance cameras, audio sensors). The system also needs to consider these heterogeneous data types which suffer from their own specific challenges, i.e. requiring handling of inconsistencies from, e.g. sensor anomalies/breakdowns, data transmission issues. Thus, there are at least two main challenges relevant to any enabling system, those resulting from a sensor network perspective and those related to converting textual OSN data into a form suitable for fusion analysis with sensor data.

1.2 Contributions and outline

The proposed novel cross-domain data fusion method combines both textual OSN data and environment pollution data to detect and identify correlations between events and pollution in public spaces. The contributions can be summarised as follows:

1. fusion technique for cross-domain data in different modalities (numeric and text) and scales of measure (nominal and ratio).
2. unsupervised pollution episode (anomaly) detection method that does not require an offline mining step and is location-agnostic (thus, being applicable to different urban regions).
3. quantification of the social impact of the events, through the region of influence that identifies specific areas of the city most affected in terms of the identified pollution episodes.

The rest of the paper is organised as follows: Sect. 2 reviews the state of the art in data-centric urban computing approaches that analyze physical and social network data together. Section 3 presents the study and the statistics of the collected data; followed by Sect. 4 which presents the architecture of the developed data fusion system. Section 5 explains the method of extracting events from tweets, together with an estimation of the involved population and location tagging. Section 6 presents the data analytics methods for detecting pollution episodes from the environmental data, followed by the data fusion methods for correlating environmental episodes with events. Section 7 evaluates the findings in terms of a well-known effect size analysis metric. Section 8 concludes the paper and discusses the planning implications of the findings for future urban development.

2 Literature review

There has been great interest in applying urban computing techniques to identify patterns from urban big data and infer unknown knowledge. While there exist many studies that have explored OSN data as alternative data sources of urban data [29], only the research works that perform analysis of observation data from physical sensors (as numeric or time series data) in conjunction with data generated by citizens on social networking platforms (as numeric or text data streams) to build models for data analytics, are relevant to this work, and are reviewed in the sub-sections below.

2.1 Social network data as context descriptor for sensor observations

Early research in fusing physical sensor observations with OSN data has primarily focused on statistical analysis of the sensor data in isolation and then using the social data to provide a semantic context descriptor to the patterns derived from sensor data. Examples of this class of methods include the traffic anomaly detection work in [34,40] that use GPS traces from sensors mounted in taxis. The detected traffic anomaly is then described by mining terms from social media. These works fusing OSN data with sensor measurements are similar to our approach. However, in contrast to the approach in [34] that requires an offline mining step for anomaly detection (the offline mining derives the normal behavior as a pre-cursor to detecting anomalies), our fusion framework proposes an unsupervised anomaly detection method that infers the normal data pattern as part of the online anomaly processing. A recent citywide deployment proposal [6] takes a similar approach to the existing state of the art, where OSN data trends/analytics, citizen questionnaires and mobile phone sensor crowdsourcing methods are proposed to provide a subjective perception from citizens, as a complementary sensing method to the distributed static and mobile air quality and noise sensor deployments. The daily frequency and negative sentiment expressed in OSN messages is also employed as an indicator of public satisfaction with perceived air quality in [38]. The authors in this study also found that there is a marked association between a higher AQI in a city and the frequency of OSN messages discussing air pollution topics in that city.

However, these existing methods only utilise the mined social media terms to describe the detected anomaly or explain the causes, whereas our approach goes beyond this to convert the mined event topics into a numerical representation that can be conveniently correlated with the detected anomalies.

There are also studies that process physical data of different types simultaneously that is in the same scale of measure and mode, e.g., numerical data in the ratio scale. Such studies mainly utilize location-based services, such as the work by Komninos et al. [25] analyzing Foursquare check-in data and its correlation with diurnal pollutant levels and traffic volume in Patras, Greece, and that by Jara et al. [22] to correlate traffic behavior with temperature in the city of Santander. Similar location-based analysis is done in [33], where the authors mine machine-learning features from cellular base station data and apply check-in patterns from the Foursquare OSN, as semantic annotations for the areas near to base transceiver station (BTS) cells in terms of urban activities such as park, travel, food, shop etc.

2.2 Urban informatics analyzing different sensor streams with open datasets

Considerable research has applied data mining and machine learning-based methods to analyze sensor data from multiple sources together with open datasets pertaining to cities to determine trends in city dynamics or to perform prediction and classification analysis. For instance, the work in [27] combines taxi GPS data with datasets describing the road network, points of interest (POI) and meteorological data to predict transportation carbon emissions within different city grids in Zhuhai, China. The study in [14] determines the correlation of transport density on the road network and weather changes by analyzing sensor data of taxi trajectories and regional weather data, and open datasets such as the road network data and regional information, which includes social factors such as house age, number of neighbors, number and characteristics of POIs etc.

Urban models to predict air quality in city districts without installed monitoring stations have been proposed in [17,45], by considering a range of spatio-temporal urban big data sources such as meteorology, vehicular traffic and POI. The authors in these studies predict the causality between these urban sources and Air Quality Index (AQI) and apply this to find the most influential data for air quality estimations. While the authors in [45] use Granger causality measures to determine the causality associations between AQI and urban sources, Ge et al. [17] construct a region similarity matrix and construct a deep learning framework to fuse the air quality data with regional spatial metadata. The research reported in [36] investigates environmental impacts on socio-economic development (measured via GDP) by applying a sophisticated multi-criteria decision-making enhanced TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) model to the following air pollutants: SO2, NO2, PM10, AQI, as well as pH and dustfall. The authors calculate a joint indicator between air pollution and GDP by determining the relative distance between sample points and the optimal/worst sample. The results show that GDP has been rising steadily for the 20 years that the measurements have been considered (1996–2015), while air pollution peaked in 2010 and then dropped. Bermudez-Edo et al. [8] propose a sliding time window pipeline to detect correlations between two traffic streams over distinct road segments. They employ both Pearson correlation and mutual information methods to investigate their effectiveness at detecting spatio-temporal correlations, with one of the findings being, that if the correlation displays a pattern, then a temporal series can be used to analyze if certain data changes can predict other values over a period of time.

Table 1 provides a comparison of the current state-of-the-art methods according to a number of metrics that are illustrative of the characteristics of studies looking at fusing physical sensors' observations with OSN data. The metrics related to the data sources, modality of data sources and fusion aspects, bring out the challenges relevant to this work.

Our investigation of the literature work shows that data analytics from both social media data and physical sensor data simultaneously remains a challenging problem due to the large semantic gap among the different data types. Most of the existing approaches mainly focus on correlating numerical physical and social sensing data, or use text OSN data as a annotation element, whereas our proposed system is able to leverage cross-space multimodal data (textual data from Twitter and numerical time-series pollutant data).

3 Case study and dataset characteristics

This work uses the 2012 London Olympics as a validating use-case. Taking this large-scale event as a case study mitigates the risk of the data sparsity challenge [44], where in addition to missing values for the relevant pollution sensor measurements, there may be no pollutant measurements available for the same spatial and temporal range as the events detected from the Twitter social platform.

3.1 Air quality monitoring sites

As the NO_2, pollutant is highly related and sensitive to traffic and mobility of urban residents, we chose it as the target pollutant. The dataset used in this work was recorded by retrieving data from LondonAir [26], the London Air Quality Network (LAQN) website, which provides the data from the large-scale deployment of air pollution monitoring sites across London. It contains data for the duration of the Olympic Games (26 July 12 August 2012). NO_2, concentrations were retrieved at 15-min intervals. The monitoring sites were carefully chosen, including those situated in the vicinity of

Table 1 Comparison of related work

Ref.	Data sources	Modality	Volume	Duration	Fusion aspects	Application	City
[40]	Taxicab, bike rental, POI, road network	Numeric	165M taxi, 8M bike, 862 roads, 14 POIs	Jan 2014–Jan 2015	Anomalies detected across road segments using taxi and bike trip in/out flows	Anomaly event detection	NY, USA
[34]	GPS taxi trip traces, Twitter messages	Numeric, text	19M taxi trips, Twitter messages	Mar–May 2011	Detected traffic anomaly described by mining Twitter constrained to road sub-graph	Traffic anomaly detection	Beijing, China
[38]	AQI, PM2.5, Sina Weibo, Tecent Weibo messages	Numeric, text	–	Oct–Nov 2015	Sentiment classification of OSN messages, association of negative sentiment in air pollution messages in OSN with measured AQI	Citizens sentiment correlation with AQI	Beijing, China
[25]	NOx and CO, traffic volume, Foursquare check-ins	Numeric	1889,043 Foursquare check-in data	Mar–Apr/Jul Sep 2009/12	Correlations (Pearson) found between diurnal Foursquare check-ins and traffic volume, NOx and pollutants	Urban dynamics model	Patras, Greece

Table 1 continued

Ref.	Data sources	Modality	Volume	Duration	Fusion aspects	Application	City
[22]	Traffic volume, temperature	Numeric	97 temperature sensors, 38 traffic sensors	Apr Dec 2013	Correlations between mean temperature and traffic on 2 highways	Traffic interpolation	Santander, Spain
[14]	Taxi, weather, POI, community, road	Numeric	4,529 taxi trajectories	Jan 2006–Nov 2007	Weather based prediction, key factor analysis for POI, community feature weighting	weather-traffic index (WTI) system	Shanghai, China
[17]	Air quality, meteorology, weather forecast, POI, road networks	Numeric	Hourly pollutants (36 sites), meteorological data (16 districts)	May 2014–May 2015	Deep spatial-temporal fusion network for urban heterogeneous data and capturing the affecting factors	Air quality prediction	Beijing, China
[27]	Taxi GPS, carbon emission, road, POIs, meteorology	Numeric	38M taxi, 2208 carbon, 38,815 POIs, 30,227 road segments	Aug–Oct 2015	Three-layer perceptron neural network to learn the characteristics of collected data and infer the transportation carbon emission	Transportation carbon emission prediction	Zhuhai, China

Table 1 continued

Ref.	Data sources	Modality	Volume	Duration	Fusion aspects	Application	City
[8]	Vehicles count for different road sectors	Numeric	2016 observations per week	–	Pearson correlation and mutual information for dependency detection	Traffic dependency detection	Aarhus, Denmark
[33]	Call data and Foursquare check-ins	Numeric	2.5M Foursquare Venues, 20M calls and text messages	Sep 2010	Combining call/Check-in data for activity prediction	Urban characterisation by user activity	Madrid Barcelona, Spain
[36]	Air pollutants, meteorological data, GDP	Numeric	20 yearly average for all data)	1996–2015	Entropy-based weight-training combining air quality indicators with GDP	Pollution impact on sustainability	Wuhan, China
[6]	Pollutants, ambient sound, app data and questionnaire on OSN	Numeric, Likert-scale	Deployment planned	–	Daily exposure of citizens to air pollution and noise, correlated with their subjective observations captured through questionnaires	Impact of air quality and noise on users daily habits	Patras, Greece

Table 1 continued

Ref.	Data sources	Modality	Volume	Duration	Fusion aspects	Application	City
This work	Air pollutant, Twitter messages	Numeric, text	NO2 from 13 monitoring sites at 15-min intervals, 1.6M tweets	Jul–Aug 2012	Correlation between pollutant anomalies and large-scale social event in city	Environmental impact of large-scale city event on public	London, UK

the main Olympic stadiums, and also in central London to capture the main transport directions towards the stadiums. By focusing on this subset of monitoring sites, our work aims to detect local pollution episodes that occur for a limited temporal duration and may be distinct from larger episodes attributed to weather phenomena or seasonal variations. To capture the effects of events on pollution, Roadside (situated 2–10m away from the road) and Urban Background (>10m away from main roads and >30m from busy roads) sites were chosen to account for both pollutant generation and dispersal. Table 2 shows the sites information, including the site type and site code. Figure 1 maps the monitoring site locations, together with that of the Queen Elizabeth (QE) Olympic Stadium.

3.2 Twitter data

The tweets corresponding to the time span of the validating use case [using time constraints: since: 26-07-2012 until: 13-08-2012 (excluded)] and with place keywords (i.e. London) were retrieved from the Twitter search API. A total number of 1,625,508 tweets were collected. Figure 2 shows the distribution of tweets retrieved over the event time span, with two major peaks in tweet activity corresponding to the opening (27 July) and closing (12 August) ceremonies of the London Olympics Games, during which 234,912 and 164,194 tweets were collected, respectively. The numbers of tweets on other days vary from 20,000 to 100,000.

4 System overview

Figure 3 shows the system architecture for data fusion and analytics, which consists of three parts: event detection, pollution episode extraction and data fusion.
Event Detection: as a precursor to analyzing the pollution data, messages from social media, i.e., the tweets, are retrieved using the time span of the chosen scenario and tagged with the broad location name (e.g. city or city region). Following pre-processing with tokenization and stop-word removal, events are determined from the cleaned tweets by extracting latent topics by using the Twitter-LDA method. More precise event location is also determined by deriving the location terms contained in the tweets.
Pollution episode extraction: as shown in the middle of Fig. 3, the pollution episode extraction step takes in the pollutant data from the various monitoring sites of a city. A pollution episode represents points of inflection at which the sensed data may show sharp and sudden changes. Pollution episodes reported by the majority of the monitoring sites covering the entire or large regions of a city may be attributed to weather phenomena or seasonal variations. To capture the effects of events on pollution episodes manifested through increased human and traffic flows, our work focuses on local anomalies that are detected only by a subset of monitoring sites located close to each other for a limited temporal duration. The collected datasets can be in different data formats (e.g., CSV or JSON) and scale; in the pre-processing step, the

Table 2 Information of monitoring sites

Local authority	Site code	Site name	Site type
City of London	CT6	City of London Walbrook Wharf	Roadside
Hackney	HK6	Hackney Old Street	Roadside
Islington	IS2	Islington Holloway Road	Roadside
Newham	NM2	Newham Cam Road	Roadside
Redbridge	RB4	Redbridge Gardner Close	Roadside
Tower Hamlets	TH2	Tower Hamlets Mile End Road	Roadside
Tower Hamlets	TH4	Tower Hamlets- Blackwall	Roadside
City of London	CT1	City of London Senator House	Urban Background
City of London	CT3	City of London Sir John Cass School	Urban Background
Islington	IS6	Islington Arsenal	Urban Background
Newham	NM3	Newham Wren Close	Urban Background
Tower Hamlets	TH1	Tower Hamlets Poplar	Urban Background
Tower Hamlets	TH5	Tower Hamlets Victoria Park	Urban Background

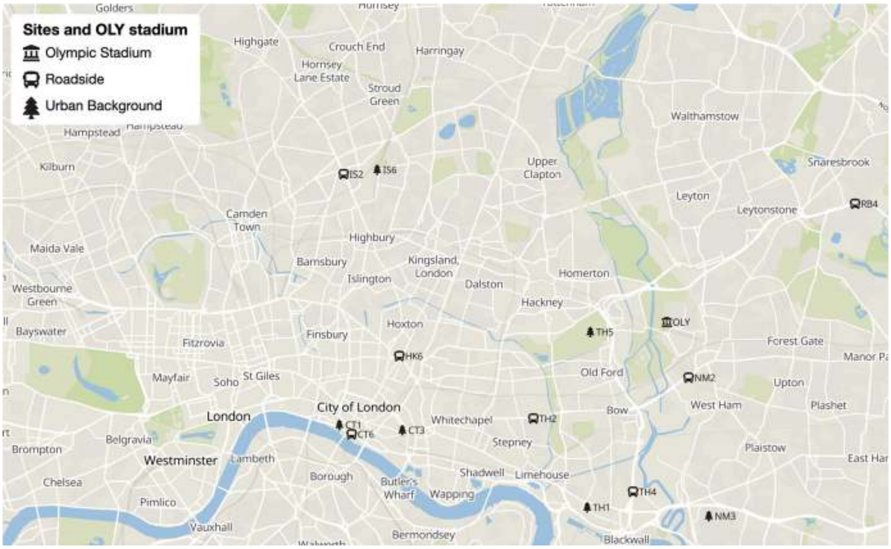
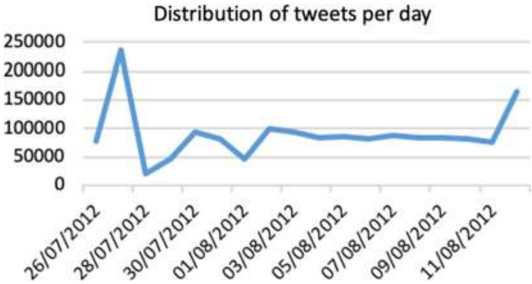


Fig. 1 Pollution monitoring sites and event (QE Olympic Park) locations

Fig. 2 Distribution of tweets during the 2012 London Olympics duration



raw data needs to be cleaned and then integrated. Finally, the Kolmogorov Complexity (KC) [24] score analysis is applied to generate the anomalous graphs.

Data fusion: the steps depicted in the right column of Fig. 3 aim to analyze and explain the pollution episodes in order to derive knowledge about the relations between the episodes and events. This is achieved by finding the presence of significant correlations between the anomalous graphs and the representative event topics. The analysis also includes determining the scale and impact of the event on the deviation of the pollutant levels.

5 City event detection from Twitter

This section applies the method from our previous work on event extraction from Twitter [42] and briefly explains the Twitter-LDA and Gibbs sampling algorithms for topic selection and classification.

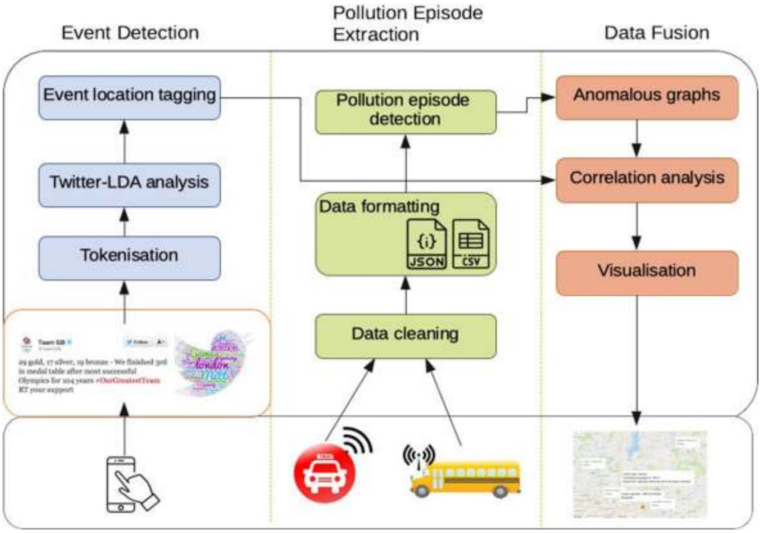


Fig. 3 System architecture

5.1 Tweets pre-processing

The place names and temporal information (date parameters d_i) are used as input to retrieve relevant tweets from the Twitter search API. All the tweets posted on a day are considered as a document, which is subjected to tokenisation, stop words and noisy words removal (expressions such as ya and ha). URL links and unreadable codes are also removed. Given the nature of the tweets and lack of unlabeled data, we adopt an unsupervised approach to infer topics discussed in tweets as well as their relevant keywords and distributions of topics over a day. To better interpret the meanings of these inferred, unlabeled topics, we map them to the broad event categories developed by Ritter et al. in [35]. The categories identified include: [Traffic | Culture | Sports | Air Quality | Weather | Disaster | Non-event].

5.2 Twitter-LDA analysis

The Twitter-LDA model [39] is a customisation of the original Latent Dirichlet Allocation (LDA) model, which is suitable for processing short text such as tweets. It assumes that a tweet only discusses one topic and contains a small number of background words that do not contribute to any topic. For more technical detail about the generative process of Twitter-LDA, readers are referred to [39].

Gibbs Sampling is used to infer the latent topic and keyword distributions in the Twitter-LDA model and is described in Algorithm 1. In the beginning, topics are randomly assigned to each word. The words are also attached with a decision of whether it is a topical or background word. The algorithm then iteratively samples a topic for the document and makes a decision over each word on whether it is background or topical, through the posterior distribution calculated from the previous iterations.

```
Algorithm 1: Gibbs sampling on Twitter-LDA
Input: Preprocessed Tweets collection on several dates $d_1, d_2, d_3 \ldots$
Output: Distribution of Topics over dates, $Distribution$
Initialise topics matrix T, background/topic decision matrix Bt
for iteration $i$ = 1,2,3... do
    for each dates tweets collection $d_j = tw_1, tw_2, tw_3 \ldots$ do
        for each tweet $tw_k = w_1, w_2, w_3 \ldots$ do
            $T[j][k] = Sample\_T()$            // Sample a topic for the tweet
            for each word $w_l$ do
                $Bt[j][k][l] = Sample\_Bt()$   // Sample a decision for the word
/* Generate distribution of topics over each day */
$Distribution = compute\_distribution()$;
return $Distribution$
```

The distribution is then updated accordingly. After a certain number of iterations, the distribution of topics over each day starts to converge. The output of the inference includes the topics with the list of top related keywords and the number of tweets for each topic. Since Twitter-LDA is an unsupervised method, the extracted topics are unlabeled. To better understand the meanings of the latent topics, they can be linked to type of events as defined in [35]. Each specified event type defines a list of related keywords, which are used to match to the top keywords for each latent topic inferred from Twitter-LDA. A topic will be classified as an event type if most of its top keywords match the keyword collection of the corresponding event type. If a topic cannot match any keywords, it will be classified as a non-event.

5.3 Event scale estimation and location tagging

The scale of a detected event is defined as the size of the population involved in that event. It can be estimated based on the frequency of the tweets relating to the event and the population of a city. As can be seen shortly, the estimated values are fairly close to the true values. The final step is to determine the precise location of the events detected from the tweets. An aggregation and rank-based location entity detection approach is developed which extracts the location entities in the relevant tweets using location named entity recognition model based on OpenNLP [16]. The detected location entities are aggregated and ranked by their occurrences, with the two ranked at the top considered representative of the event location. The associated geolocation coordinate information is determined by formulating a query to the Google Maps Geocoding API [18].

Table 3 introduces the top 10 detected topics, their determined category, and their top 5 related keywords. Topic T49 (london back love day london!) is matched to a Culture event according to its keywords. T52 (2012 medal olympic olympics gold), T63 (2012 olympics london olympic games), T12 (2012 gold bolt olympics usain), and T78 (2012 basketball team men's olympics) indicate particular Olympics sport events. T77 (lol lauren love im girl) is matched to a non-event topic. T69 (london, fashion august show tickets) is a culture event about a show and performance in London. T36

Table 3 Top-5 keywords for the top-10 frequent topics

Rank	Topic	Event type	Top-5 keywords
1	T49	Culture	London back love day london!
2	T52	Sport	2012 medal olympic olympics gold
3	T63	Sport	2012 olympics london olympic games
4	T77	Non-event	Lol lauren love im girl
5	T69	Culture	London, fashion august show tickets
6	T12	Sport	2012 gold bolt olympics usain
7	T36	Culture	Opening ceremony 2012 olympics Olympic
8	T78	Sport	2012 basketball team men's Olympics
9	T47	Culture	2012 closing olympics london ceremony
10	T97	Non-event	United #job #jobs london, kingdom

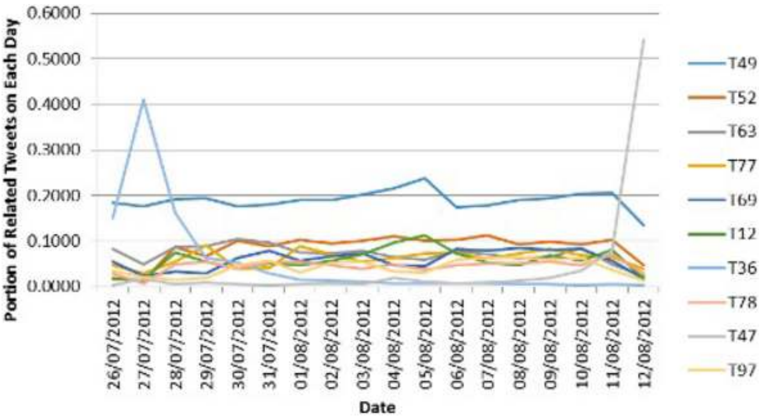


Fig. 4 Distribution of the top-10 frequent topics

(opening ceremony 2012 olympics olympic) and T47 (2012 closing olympics london ceremony) indicate the opening and closing ceremonies of the London Olympics Games. They are classified as culture events, since they are celebratory ceremonies. Figure 4 gives the distributions of the top 10 frequent topics. Two obvious spikes in the figure indicate topics discussing the opening ceremony (27 July—T36) and closing ceremony (12 Aug—T47). T97 (united #job #jobs london, kingdom) indicates a non-event topic for jobs.

The event scale estimation provides a figure of 70,317 members of the population being involved in the Olympics event, which is close to the 80,000 capacity of the Olympics stadium.

6 Composite effects of detected events on pollutant levels

6.1 Pollution episode extraction

As the collected sensor data might contain duplicates, missing or incorrect values due to noise, equipment maintenance, recalibration or communication faults [17,44], pre-processing for such data is also needed. It consists of 2 steps: in the data cleaning step, the sensor observation data is collected and stored; then, each data point is checked and all negative values (below 0) are marked as invalid and removed; in the data formatting step, the cleaned data is formatted into time series by applying JSON or CSV scripts depending upon the retrieved data. Pollution episodes, where the data exhibits a pronounced departure from the normal behavior are extracted by calculating the information theoretic Kolmogorov Complexity (KC) [24] score, which is computed through the Kolmogorov-Smirnov test [15]. For each data point, it computes the Euclidean distance to all other data points in a random sampled sequence; this forms a sequence A. Next, a new sequence is sampled randomly: for the data points in this new sequence, their distances to data points of the first random sampled sequence are computed and form new distance sequences $B_1, B_2 \ldots B_n$, where n is the number of sampled data points. The Kolmogorov-Smirnov test is applied to A and each of B. The mean value of Kolmogorov-Smirnov test values is the KC score.

6.2 Correlation analysis and region of influence

Based on the geographical and the temporal constraints, the associations between detected pollution episodes and the events can be derived. In particular, the impact of the events extracted from social media on the pollution (e.g., the identified NO_2, patterns as well as the pollution episodes) is derived through correlation analysis.

To calculate the correlations, both anomalies in pollution and events need to be quantified as arrays of numbers. Also, the spatial and temporal information of the pollution and events should match, i.e., locations and time of the sensing data and events are close to each other. The pollutant representations consist of arrays of KC-score values of the sensing data across a number of different days, while events are represented by the ratio of tweets involved in the particular event topics to the total number of collected tweets on a day.

Since the impact of the events could involve a broader geographic scope than the exact event location, e.g., through traffic flows, the pollutant data from the nearby sensing locations is also considered in the impact analysis. The relevant monitoring sites are selected based on proximity (within a defined radius of the monitoring site where the pollution episode is detected and also the event location), using the geospatial search algorithm proposed in [43]. The search algorithm implements a distance query by taking as inputs the Geohash representations of the event location and returns the names of the monitoring sites within the required radius. The resultant sites are a subset of those presented in Table 1 and include the following 11: CT6, HK6, IS2, NM2, RB4, TH2, TH4, NM3, TH1, TH5 and CT3.

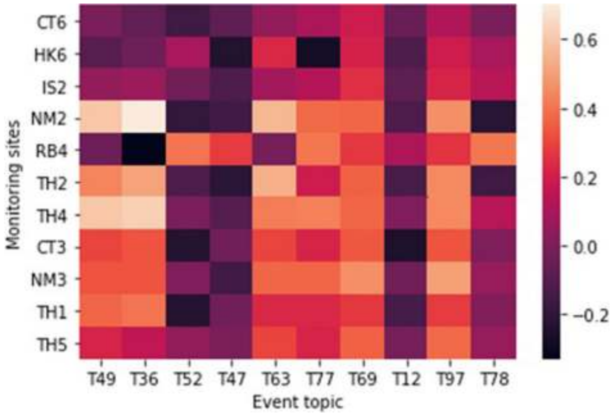


Fig. 5 Monitoring site pollutant detected event correlation

The implementation has been done using Matlab and Java, and is available as a Docker image hosted on the Docker hub repository,[1] together with a description of commands for pulling down the Docker images and their execution. The implemented component takes as input Twitter tweets and environmental pollution data, and computes the Pearson correlation between city events impacting people and/or traffic flows (identified from the tweets) with the measured pollution levels. It uses as input two files, one containing the measured pollution, e.g. NO_2, levels at a given place for a day, and the other collected tweets for the same time period and place. The pollution measurements file is a csv file that contains sensing data collected from different sensing sites, with the measurement dateTime (dd/mm/yyyy hh:mm) in the first column and the NO_2, values for the different sensing sites in the following columns. A default data.csv file is provided which contains pollution values from different sites in London. The second input file is a text file that stores the text of collected tweets for each day, with one tweet in each row. A default tweetslist.txt file is included with tweets from London for the Olympic event days.

Figure 5 shows the Pearson correlation results between the sensed NO_2, data and the detected event topics. As a complement of correlation analysis, p values, as shown in Table 4, are also calculated to determine whether it can reject the null hypothesis of no correlation between anomalies and events. A p value of < 0.05 denotes significant correlation, showing that a strong impact of events on pollution level reasonably exists.

In Fig. 5, roadside sites such as Newham-Cam road (NM2), Tower Hamlets Mile End (TH2) and Blackwall (TH4) show a strong correlation with T36 (opening ceremony 2012 olympics olympic), in terms of KC-scores. The Pearson correlation values are 0.70, 0.51 and 0.62, respectively. High values are also observed with the overall Games topic (T63: 2012 olympics london olympic games). As shown in Table 4, the corresponding p values (0.0, 0.03 and 0.01) for T36 confirm the strong relationship between the opening ceremony event and the detected anomaly in the pollution pattern at these sites. The sites of NM2 and TH2 display similar association with the overall

[1] https://hub.docker.com/r/ikaas/anomaly-detection.

Table 4 *p* values for event-pollutant episode correlation

Site code	T49	T36	T52	T47	T63	T77	T69	T12	T97	T78
CT6	0.95	0.77	0.52	0.75	0.84	0.67	0.46	0.85	0.64	0.96
HK6	0.72	0.87	0.69	0.35	0.36	0.27	0.40	0.65	0.46	0.69
IS2	0.85	0.79	0.91	0.64	0.72	0.59	0.32	0.75	0.39	0.58
NM2	0.01	0.00	0.46	0.57	0.02	0.12	0.13	0.62	0.05	0.39
RB4	0.88	0.18	0.10	0.25	0.94	0.09	0.28	0.66	0.29	0.09
TH2	0.07	0.03	0.64	0.41	0.02	0.45	0.14	0.59	0.06	0.53
TH4	0.01	0.01	0.96	0.70	0.08	0.08	0.14	0.99	0.06	0.59
CT3	0.13	0.08	0.56	0.55	0.03	0.49	0.13	0.50	0.09	0.72
NM3	0.10	0.10	0.68	0.98	0.18	0.13	0.14	0.68	0.08	0.97
TH1	0.07	0.07	0.76	0.65	0.10	0.16	0.07	0.78	0.04	0.95
TH5	0.37	0.53	0.84	0.98	0.23	0.38	0.14	0.94	0.12	0.82

Games topic T63, with identical *p* values of 0.02. The results are in fact coinciding with what is anticipated. As these sites are roadside monitoring stations, the corresponding sensed pollution data is highly influenced by the traffic through the roads. In addition, the selected measurement, NO_2, is largely impacted by traffic. As shown in Fig. 1, Mile End Road (TH2) and Newham-Cam Road (NM2) are on a major navigation route from central London to the Olympic Park site (along the A11 primary road). The increased traffic on the opening day of the Olympics does impact the pollution levels. Similarly, T36 has a significant correlation with data from the sensing site at Blackwall (TH4), which is on a road (primary road A12) from the south of London to the Olympic Park. These results indicate the most preferred route taken by people travelling to the Olympic Games stadium from central London. The results show that the NO_2, pollution levels, mainly caused by traffic, are indeed highly influenced by the opening ceremony hosted in the Olympic Park. In contrast, urban background sites show a low correlation with the relevant Games topics, even for sites such as TH5, which is located close to the event venue. The closing ceremony of the London Olympics Games (topic T47) does not have such an evident relationship according to the calculated correlation results; since it was on the 12th Aug (Fig. 4), while the detected pollution data anomaly from the Mile End Road sensing site according to its calculated KC-score is on 10th Aug. Taken together, these results point to the region of influence of the Olympic Games event as well as the specific sub-event, e.g. the opening ceremony, on distinct geographical regions in terms of the recorded pollution anomalies.

7 Evaluation

Inspired by effect size quantification in education theory [9], causality measures are proposed in this section, in order to evaluate by how much the events statistically influence the pollutant anomaly values. This forms a key-factor analysis of the environmental impact of the events through statistical causality measures such as ANOVA η^2 [32] for effect size quantification. The η^2 measure estimates the magnitude of the effect of the independent variable (detected event in our case) and has quantified measures to categorize effect size (e.g., low, moderate or high). It is calculated as:

$$\eta^2 = SS_A/SS_T \tag{1}$$

where SS^T is the total sum of squares component, which can be partitioned into between-group sum of squares (SS_A) and the within-group or the error sum of squares ($SS_{s/A}$), representing the variation due to the independent variable and variation due to individual differences in the score, respectively:

$$SS_T = SS_A + SS_{s/A} \tag{2}$$

The sum of squares between groups (SS_A) examines the differences among the group means by calculating the variation of each mean ($\bar{Y}_{.j}$) around the grand mean ($\bar{Y}_{..}$), as shown in (3) below:

$$SS_A = n\sum_j(\bar{Y}_{.j} - \bar{Y}_{..})^2 \tag{3}$$

where n is the number of observations in each group, $\bar{Y}_{..}$ is the mean of the full sample and is calculated across all individuals and all groups.

A single score is represented by Y_{ij}, indicating the score is for an individual i, within a particular group, j. The "." refers to computing across that element, either individuals or groups. Then, $\bar{Y}_{.j}$ is the mean of a particular group, j; the "." is used in place of the i because the mean is calculated using all the 'i's for a particular group.

$SS_{s/A}$ is the variation of individual scores around each group mean and is calculated by (4):

$$SS_A = \sum_j\sum_i(Y_{ij} - \bar{Y}_{.j})^2 \tag{4}$$

To show the effect size of the independent variable, the sites are separated according to the site type (roadside and urban background) and η^2 values are calculated for all of the days (event and non-event days). Doing this enables us to determine the effect over the short period of the detected event in comparison to a long period when other events may affect the results. The resulting values show that the days without the events have smaller values than the period including the event days, for the roadside sites NM2, TH2 and TH4; other sites do not show any statistically significant effect. Moreover, the calculated values show a moderate effect for NM2 (0.06) and a lower effect for TH2 and TH4 (0.027 and 0.0068, respectively), according to the general rule of thumb for η^2 given by Miles and Shelvin [31].

These results corroborate the correlation analysis performed in the previous section, with the same monitoring sites showing statistically significant causality of the events on the monitored pollutant levels. With the events showing a statistical significant influence on the same monitoring sites as derived in the previous section, this also supports the region of influence inference as well.

8 Conclusions

This work attempts to quantify the environmental impact of social and cultural events which involve large-scale traffic and human movement in public spaces within a city. The developed novel cross-domain data fusion techniques combine social network data with environmental sensor data to detect and identify correlations between events and pollution levels. The results have great potential to contribute to an increased understanding for public space management and pollution (or air quality), which could improve the quality of life for citizens. The correlation of air quality with city events, which is ultimately influenced by mobility of people and vehicles, has important social implications for a connected city, e.g., for planning of public spaces and infrastructures such as hospitals and schools. This insight allows to adopt a data-enabled collaborative approach to plan and build responsive urban areas that helps inform peoples decision making and enables urban authorities to plan for the best possible use of limited city resources. This is in line with the observations from the pioneering urbanist William Whyte, who stressed on careful observation and collection of data to answer questions on building psychologically healthy urban spaces [37].

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. World population prospects: the 2017 revision (2017). https://www.un.org/development/desa/publications/world-population-prospects-the-2017-revision.html
2. World urbanization prospects: key facts (the 2018 revision) (2018). https://population.un.org/wup/Publications/Files/WUP2018-KeyFacts.pdf
3. Alberti M (2017) Grand challenges in urban science. Front Built Environ 3:6
4. Anantharam P, Barnaghi P, Thirunarayan K, Sheth A (2015) Extracting city traffic events from social streams. ACM Trans Intell Syst Technol 6(4):1–27
5. Balduini M, Valle ED, Dell'Aglio D, Tsytsarau M, Palpanas T, Confalonieri C (2013) Social listening of city scale events using the streaming linked data framework. In: Advanced Information Systems Engineering, pp. 1–16. Springer Berlin Heidelberg
6. Bardoutsos A, Filios G, Katsidimas I, Krousarlis T, Nikoletseas S, Tzamalis P (2020) A multidimensional human-centric framework for environmental intelligence: air pollution and noise in smart cities. In: 2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 155–164
7. Basu M, Shandilya A, Khosla P, Ghosh K, Ghosh S (2019) Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations. IEEE Trans Comput Soc Syst 6(3):604–618
8. Bermudez-Edo M, Barnaghi P, Moessner K (2018) Analysing real world data streams with spatio-temporal correlations: Entropy vs. pearson correlation. Autom Constr 88:87–100
9. Broadbent J (2017) Comparing online and blended learner's self-regulated learning strategies and academic performance. Int High Educ 33:24–32
10. Chen Q, Wang W, Huang K, De S, Coenen F (2020) Adversarial domain adaptation for crisis data classification on social media. In: 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)

11. Chen Q, Wang W, Huang K, De S, Coenen F (2020) Multi-modal adversarial training for crisis-related data classification on social media. In: 2020 IEEE International Conference on Smart Computing (SMARTCOMP)
12. De S, Christophe B, Moessner K (2014) Semantic enablers for dynamic digitalphysical object associations in a federated node architecture for the internet of things. Ad Hoc Netw 18:102–120
13. De S, Jassat U, Wang W, Perera C and Moessner K (2021) Inferring latent patterns in air quality from urban big data. IEEE Internet Things Mag 4(1):20–27. https://doi.org/10.1109/IOTM.0011.20000
14. Ding Y, Li Y, Deng K, Tan H, Yuan M, Ni LM (2017) Detecting and analyzing urban regions with high impact of weather change on transport. IEEE Trans Big Data 3(2):126–139
15. Emmert-Streib F (2010) Statistic complexity: combining kolmogorov complexity with an ensemble approach. PLoS ONE 5(8):e12256
16. Foundation, T.A.S.: Opennlp (2017). https://opennlp.apache.org
17. Ge L, Zhou A, Li H, Liu J (2019) Deep spatial-temporal fusion network for fine-grained air quality prediction. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 536–543
18. Google: Google maps geocoding api (2020). https://developers.google.com/maps/documentation/geocoding/intro
19. Hu K, Rahman A, Bhrugubanda H, Sivaraman V (2017) HazeEst: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors. IEEE Sens J 17(11):3517–3525
20. Hu K, Sivaraman V, Bhrugubanda H, Kang S, Rahman A (2016) SVR based dense air pollution estimation model using static and wireless sensor network. In: 2016 IEEE SENSORS. IEEE
21. Hu T, Bigelow E, Luo J, Kautz H (2017) Tales of two cities: using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. IEEE Trans Big Data 3(1):55–66
22. Jara AJ, Genoud D, Bocchi Y (2014) Big data for smart cities with KNIME a real experience in the SmartSantander testbed. Softw Pract Exp 45(8):1145–1160
23. Jiang S, Ferreira J, Gonzalez MC (2017) Activity-based human mobility patterns inferred from mobile phone data: a case study of singapore. IEEE Trans Big Data 3(2):208–219
24. Kim M (2012) Anomaly detection. http://uk.mathworks.com/matlabcentral/fileexchange/39593-anomaly-detection/content/kse_test_matlab/kse_test.m
25. Komninos A, Stefanis V, Plessas A, Besharat J (2013) Capturing urban dynamics with scarce check-in data. IEEE Pervasive Comput 12(4):20–28
26. LondonAir: London air quality network (laqn) (2020). https://www.londonair.org.uk/LondonAir/Default.aspx
27. Lu X, Ota K, Dong M, Yu C, Jin H (2017) Predicting transportation carbon emission with urban big data. IEEE Trans Sustain Comput 2(4):333–344
28. Luo X, Yuan Y, Li Z, Zhu M, Xu Y, Chang L, Sun X, Ding Z (2019) FBVA: a flow-based visual analytics approach for citywide crowd mobility. IEEE Trans Comput Soc Syst 6(2):277–288
29. Machado KLS, Boukerche A, Cerqueira EC, Loureiro A (2019) A data-centric approach for social and spatiotemporal sensing in smart cities. IEEE Internet Comput 23(1):9–18
30. Marakkalage SH, Sarica S, Lau BPL, Viswanath SK, Balasubramaniam T, Yuen C, Yuen B, Luo J, Nayak R (2019) Understanding the lifestyle of older population: mobile crowdsensing approach. IEEE Trans Comput Soc Syst 6(1):82–95
31. Miles J, Shevlin M (2000) Applying Regression and Correlation. SAGE Publications Inc. https://www.ebook.de/de/product/3768816/jeremy_miles_mark_shevlin_applying_regression_and_correlation.html
32. Myers JL, Well AD, Robert F, Lorch J (2010) Research Design and Statistical Analysis. Taylor & Francis Inc
33. Noulas A, Mascolo C, Frias-Martinez E (2013) Exploiting foursquare and cellular data to infer user activity in urban environments. In: 2013 IEEE 14th International Conference on Mobile Data Management, vol. 1, pp. 167–176
34. Pan B, Zheng Y, Wilkie D, Shahabi C (2013) Crowd sensing of traffic anomalies based on human mobility and social media. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13
35. Ritter A, Mausam Etzioni O, Clark S (2012) Open domain event extraction from twitter. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'12

36. Wang Q, Dai HN, Wang H (2017) A smart MCDM framework to evaluate the impact of air pollution on city sustainability: a case study from china. Sustainability 9(6):911
37. Whyte W (2001) The social life of small urban spaces. Project for Public Spaces, New York
38. Zhang X, He M, Shao B, Ren C (2016) Physical-social fusion to assist public services in the war against air pollution in china. In: 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), pp. 916–920
39. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: Lecture Notes in Computer Science, pp. 338–349. Springer Berlin Heidelberg
40. Zheng Y, Zhang H, Yu Y (2015) Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS'15
41. Zhou Y, De S, Ewa G, Perera C, Moessner K (2018) Data-driven air quality characterization for urban environments: a case study. IEEE Access 6:77996–78006
42. Zhou Y, De S, Moessner K (2016) Real world city event extraction from twitter data streams. Procedia Comput Sci 98:443–448
43. Zhou Y, De S, Wang W, Moessner K (2014) Enabling query of frequently updated data from mobile sensing sources. In: 2014 IEEE 17th International Conference on Computational Science and Engineering
44. Zhou Y, De S, Wang W, Wang R, Moessner K (2018) Missing data estimation in mobile sensing environments. IEEE Access 6:69869–69882
45. Zhu JY, Sun C, Li VOK (2017) An extended spatio-temporal granger causality model for air quality estimation with heterogeneous urban big data. IEEE Trans Big Data 3(3):307–319