

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Sensing Pollution on Online Social Networks: A Transportation Perspective

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

R. Tse, Y. Xiao, G. Pau, S. Fdida, M. Roccetti, G. Marfia (2016). Sensing Pollution on Online Social Networks: A Transportation Perspective. *MOBILE NETWORKS AND APPLICATIONS*, 21(4), 688-707 [10.1007/s11036-016-0725-5].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/560984> since: 2022-02-28

*Published:*

DOI: <http://doi.org/10.1007/s11036-016-0725-5>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Tse, R., Xiao, Y., Pau, G. et al. Sensing Pollution on Online Social Networks: A Transportation Perspective. Mobile Netw Appl 21, 688–707 (2016). <https://doi.org/10.1007/s11036-016-0725-5>**

The final published version is available online at: <https://doi.org/10.1007/s11036-016-0725-5>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

# Sensing pollution on online social networks: a transportation perspective

Rita Tse · Yubin Xiao · Giovanni Pau ·  
Serge Fdida · Marco Roccetti · Gustavo  
Marfia

Received: date / Accepted: date

**Abstract** Transportation policy and planning strategies, as well as Intelligent Transportation Systems (ITS), can all play important roles in decreasing pollution levels and their negative effects. Interestingly, limited effort has been devoted to exploring the potential of social network analysis in such context. Social networks provide direct feedback from people and, hence, potentially valuable information. A post telling how a person feels about pollution at a given time at a given location, could be useful to policy-makers, planners or environmentally-aware ITS designers. This work verifies the feasibility of sensing air pollution from social networks and of integrating such information with real sensors feeds, unveiling how people advertise such phenomenon, acting themselves as smart objects, and how online posts relate to true pollution levels. This work explores a new dimension in pollution sensing for the benefit of environmental and transportation research in future smart cities, confronting over 1,500,000 posts and pollution readings obtained from governmental on-the-field sensors over a one-year span.

**Keywords** Smart objects · transportation · social networks · sensors · traffic · human perception

## 1 Introduction

Traffic amounts to a societal problem in many different countries of the world. Although high traffic levels are typically considered a sign of prosperity, and,

---

Rita Tse, Yubin Xiao  
Macao Polytechnic Institute, Macao, China

Giovanni Pau, Serge Fdida  
University Pierre et Marie Curie Sorbonne Universities, Paris, France

Marco Roccetti, Gustavo Marfia (corresponding author)  
University of Bologna, Bologna, Italy

as such, initially often welcomed, prolonged exposures to high traffic levels also  
 5 bring in a number of economic and health issues that include: loss of productivity, drivers' stress, and a plethora of diseases caused by the excess of chemical agents that are released by combustion engines (e.g., microscopic particles, carbon monoxide, etc., (Peters et al., 2004), (Townsend, 2002)). Clearly, many different countermeasures have been taken during the decades, which include,  
 10 but are not limited to, implementing better public transportation systems (Florian, 1977), gradually substituting petroleum-based combustion engines with cleaner propellers (e.g., electric, ethanol or methane (Reed and Lerner, 1973), (Agarwal, 2007)) and to the enforcement of congestion taxes (for those who drive through given city areas during pre-defined time frames) (Rotaris  
 15 et al., 2010). Such countermeasures are very often the result of the transportation policy and planning community research works and discussions, aiming at the development of cleaner and more sustainable cities. In essence, from the policy and planning perspective, the problem amounts to understand how transportation decision-making can better support public health objectives,  
 20 including reduced crashes and pollution emissions and increased physical activity. In the past, conventional transportation planning overlooked its negative (or, also, positive) health impacts, while now a general awareness exists that raising the priority of health objectives may support planning reforms which lead to more balanced transportation systems. In essence, it is today  
 25 widely understood that the implementation of sustainable and *green* transportation policies and plans, in the context of an integrated smart city, may be a cost-effective way of improving public health.

In the meantime, computerized systems capable of managing traffic resources (e.g., traffic light cycle times) and providing real-time traffic related  
 30 information to drivers (e.g., construction sites, traffic loaded roads) have flourished, with the aim of optimizing the use of the available road and signaling infrastructures with the development of Intelligent Transportation Systems (ITSs) (WEILAND and Purser, 2000). The ITS community has initially focused its attention on technical problems, in areas ranging from mathematical optimization (Ran and Boyce, 2012), (TANIGUCHI et al., 2001), (Betts,  
 35 1998), to synchronous and asynchronous distributed communications (Hartenstein and Laberteaux, 2008), (Ben Jaballah et al., 2014a), (Ben Jaballah et al., 2014b), queuing theory (Faouzi et al., 2011) and traffic congestion detection and forecasting (Marfia and Roccetti, 2011), just to mention a few. In fact,  
 40 the primary scope of ITSs was at large interpreted as that of measuring traffic states in order to facilitate, through centralized (e.g., enforcing traffic light timings) (Zhou et al., 2010) or decentralized means (e.g., providing traffic information to personal navigation systems) (Zhao, 2000), traffic flows (Wang, 2010). A fundamental role in ITSs has been played by traffic sampling systems (i.e., those systems that measure and report the amount of traffic that  
 45 is flowing through a given road) (Kong et al., 2013), systems which exploit different type of technologies (e.g., cellular, sensor and vehicular ad hoc network data) to estimate traffic flows in real-time. Now, with the progressive spread of a general environmental awareness and the development of novel

sustainable mobility paradigms, the reduction of pollution levels has moved from being considered one of the positive side effects produced by the use of ITSs to become one of their primary objectives (Boriboonsomsin et al., 2012), (Banister, 2008). As a consequence, researchers have started investigating how pollution information could be used to modify and influence traffic control algorithms, giving environmental measurements a clear, policy-sensitive role in future traffic management and control schemes. Pollution sensors have, hence, become important feeds of information, just as important as traffic flow sensors, to be considered in the design of ITSs (Blythe et al., 2008), (Costabile and Allegrini, 2008). Such evolution should clearly push urban areas to deploy pervasive pollution sensor systems at large, as their output information could then be put to good use to improve traffic management and emission containment operations.

However, although the policy makers and managers of many governmental, and non, industrial and transportation actors require the enforcement of laws and processes leading to a clean environment and although the cost of pollution sensing platforms is progressively decreasing, a pervasive and widespread deployment of such sensing technologies is still not financially sustainable (ranging from a few thousand dollars for light-scattering particulate matter monitors to a hundred thousand dollars for microbalance ones). Both satellite and terrestrial solutions are being studied for a fine grained pollution estimation and simulation, with different results in terms of sensitivity, resolution and accuracy (Tang and Wang, 2007), (Ma et al., 2008), (Honicky et al., 2008), (Ibarra-Berastegi et al., 2008), (Athanasiadis et al., 2009), (Hoff and Christopher, 2009), (Thatcher and Hurley, 2010), (Barzyk et al., 2015), (Fallah Shorshani et al., 2015). The common denominator, however, of all these technologies is the use of specialized hardware devices capable, by different means, of estimating the presence and concentration of specific pollutants. No system has instead involved so far any feedback received from the public. The reasons of such oversight may be mainly found in the following two points: (a) a rapid and large scale collection of information from the public has been difficult, so far, and, (b) no large scale study provides any proof regarding the reliability of such information (i.e., studies performed on limited groups of residents of given areas revealed that actual pollution situations may be very different from what people perceive) (Hyslop, 2009).

New lifestyles, habits and ways of communicating have, however, emerged, with the now extensive use of Online Social Networks (OSNs). With OSNs, personal communications are no more limited to one-to-one patterns of exchange of information (typical with phone calls and emails), but often (e.g., the case of a public post with multiple replies) follow a structure where information flows in a one-to-all fashion, allowing posts and comments to be read, answered and reposted by a multitude of users. In addition, posts can in principle touch upon any topics, as users can communicate anything with no censorship, in some cases acting as a sensor (Srivastava et al., 2012), (Fraternali et al., 2012). All this to express the following idea: the big data of OSN user posts, that typically exposes what a person thinks and how s/he feels

and behaves, could be an interesting source of pollution related information, actively contributing to the greater good (Ferretti et al., 2010), (Palazzi et al., 2010). In essence, a person equipped with a mobile phone, when posting on an OSN, can himself/herself act as a smart object, providing relevant information, i.e., pollution information, about the surrounding environment. The validity of such idea, clearly, requires a proof. In particular, it requires assessing whether and how frequently people touch upon pollution topics when moving within an urban setting. In addition, even if a sufficient number of posts were available, their value would also need to be checked: could the observation of a given number of posts complaining about air quality really represent a red flag indicating that pollution conditions are not contained within their required limits? The quantity and quality of user posts need, hence, to be evaluated in order to assess how researchers may effectively exploit such information.

The contribution of this paper is that of laying the foundation to better understand: (a) how OSN data could be integrated into transportation policy, planning and system operations, (b) utilizing pollution related posts provided by people moving into urban settings, (c) while assessing whether the submitted posts follow a pattern which reflects effective pollution conditions. To do so, we: (a) analyze how OSN data may fit, in practice, in transportation design and implementations, (b) verifying the feasibility of utilizing mobile users as smart objects while reporting pollution feedback. This is performed analyzing over 1,500,000 Sina Weibo (a very popular social networking platform in China) posts uploaded during a period of one year in five Chinese cities, Hong Kong, Guangzhou, Beijing, Chengdu and Shanghai.

This paper is organized as follows. After reviewing the approaches that fall closest to the one that has been presented in Section 2, a discussion regarding the implications of this work on transportation policy, planning and systems is carried out in Section 3. An analysis of how OSN data relates to pollution sensor data in Sections 4, 5 and 6. The paper concludes with Section 7.

## 2 Related Work

A wealth of research has been carried out on transportation policy, planning and systems (Rotemberg, 1985), (Kashani and Saridis, 1983), (Papageorgiou, 1984), (Leontiadis et al., 2011), (Marfia and Roccetti, 2011), (Franchi and Malpezzi, 2013), (Mahmassani et al., 1993), (Ben-Akiva et al., 1997), (Jayakrishnan et al., 1993), (Ben-Akiva et al., 1994), . Only recently, however, clearly due to their sudden popularity and widespread use, the opportunity of putting to good use any information that could be found on social networks has been considered. The following briefly analyzes a few representative examples.

In (Kaufman, 2012), the author recommends the adoption of social media policies for transportation providers seeking to inform, engage and motivate their customers. The author, in particular, considers the specific case of New York, where millions of commuters and tourists daily move. In essence, social media are selected as they may not only serve to spread traffic and mobility

information, but also to support existing communities (e.g., commuters, single parents taking kids to school, etc.), while providing a direct channel of communication which may solicit and encourage direct feedback from customers. Twitter and Facebook, and other microparticipation media, for example, are increasingly being exploited by transportation policy-makers and planners. The contribution of Evans-Cowley and Griffin moves along these lines examining more than 49,000 posts on Twitter to assess public engagement, sentiment and the impact of user posts on the decision-making process required by the design of a city wide strategic mobility plan (Evans-Cowley and Griffin, 2012). The conclusion of the authors is that microparticipation through social media is potentially effective, but nevertheless experiences substantial communication barriers to make it really effective for the support of decision making processes. In (Goodspeed, 2013) the author pushes forth a controversial perspective to the use of social media for a new theoretical understanding of social life in cities. In essence, the author emphasizes the limits of such data in different ways: posts are typically poor of content, hence, may be easily misinterpreted, while being hardly generalizable as they often report upon specific places at given times. Concluding, in his contribution Goodspeed urges the adoption of mixed methods and action research. The work presented in this paper moves into such direction, as it aims at finding and assessing any pollution-related information that may be available from geo-tagged posts on OSNs in order to expand the reach of conventional sensors.

Focusing the attention now on ITSs, the authors of VoiceTweet, for example, devised a system that leverages on Twitter feeds to communicate traffic information perceived by a driver (Sha et al., 2013). In practice, a driver, when stuck in traffic, can share his/her driving experience recording a voice tweet with his/her social navigator. The social navigator then: (a) tags the message with its timestamp and the vehicle's location, and, (b) sends the message to a server that, in turn, groups together all the messages that are received from nearby locations during a given time frame. The server periodically sends out tweet digests on social channels: when receiving such messages, social navigators prune off any unnecessary information (e.g., outdated or off route information) and compute a new route, or update the existing one. Clearly, VoiceTweet requires the creation of a platform and of a wide user base (but this may not be an issue, as the Waze experience has demonstrated (Jatowt et al., 2013)). Nevertheless, such solution only marginally exploits the enormous potential of OSNs, as Twitter is here employed as a communication channel conveying the information created by a new service customer set, rather than as an available source of spontaneous data. Social networking has also been exploited, within the ITS context, to alleviate traffic through the implementation of a dynamic ride sharing community service. A recent work proposes the use of an information grid system, which, leveraging on OSNs, implements a carpooling service (Fu et al., 2008). Pedestrians and drivers can utilize the proposed system uploading their schedule and locations of interest (i.e., origin and destination points) on OSNs. The system downloads such information from OSNs and seeks for the best time/location correspondences between pedestrians and

drivers. If feasible solutions are found, both parties are informed and conveniently left with the freedom of taking advantage of the proposed opportunity, or not. This work, just as the previous one, exploits OSNs as communication means, rather than as independent sources of information. Interestingly, both (Miller, 2013) and (Mahmassani, 2011) discuss the opportunities that could be created with the integration of cooperative technologies and ITSs. In particular, Miller envisions a virtual environment where data streams are fused, interpreted and made available with tools for human engagement and shared decision-making (Miller, 2013). Following a similar line of reasoning, Mahmassani acknowledges that ITSs have incredibly progressed with the advent of mobile platforms and apps, but also recognizes that there is a long way to go to take full advantage of the personalization/customization/socialization opportunities that they pose ahead of ITS planners and designers. While both works provide important intuitions, neither one, nor the other, practically explain how such ideas could be put to good use to boost existing ITS solutions.

A second stream of research relevant for this study includes all those works which analyze the human awareness and perception of air pollution. Several scientific contributions have shown how, for example, a direct exposure to pollution dramatically increases a population's awareness for this problem. Such finding is congruent with the belief that everyday personal experience is important in providing evidence of environmental risk (Bickerstaff and Walker, 2001). In addition, it has been observed that long-term residents tend to exhibit denial or exaggerated perceptions when compared to newcomers (Edelstein, 1988). Interestingly, the level of annoyance exhibited by the population for air pollution has grown to be considered an effective measure of overall environmental quality, to the point that some countries also record such quantity for national environmental monitoring purposes (Jacquemin et al., 2007). In essence, all of the reported studies conclude that a general worry for pollution exists and that such concern is acknowledged being important for policy and planning purposes. Nevertheless, no work has to this date considered assessing such risk resorting to the richest available source of human information, a.k.a. the Internet. This work opens such path, carrying out a large assessment based on over one 1,500,000 posts found on online social networks.

In essence, to the best of the authors' knowledge, this work is the first aiming at unveiling whether the people who move inside an area of interest may be exploited, through the lens of social media, as smart objects for the detection of pollution.

### 3 Pollution-aware transportation policy, planning and systems

A methodical integration of OSN sources of information in transportation policy, planning and systems requires an understanding of what such data may effectively provide and of how it may influence future mobility solutions. Before proceeding, a clarification is required regarding how OSNs may fit transportation needs. Agreeing with the caution recommended by many works, regarding



the reliability of public perception, in general, and the use of any information that may be found on OSNs, this contribution aims at advancing the state of the art regarding the integration an OSN source of pollution data with the feeds gathered from traditional sensors (Bickerstaff, 2004), (Goodspeed, 2013). The following subsections, hence, proceed providing an overview of how OSN based pollution information may fit existing transportation policy, planning and systems.

### 3.1 Policy, Planning and OSN-based Pollution

As reviewed in Section 2, there is a widespread concern for pollution: many countries constantly monitor pollutant levels in order to drive their transportation policy and planning (Bickerstaff, 2004), (Lindholm and Blinge, 2014). In addition, traditional sensors are not the only means by which pollution assessment is being conducted, there is growing interest for the annoyance suffered, as a consequence of pollution, by people (Jacquemin et al., 2007). Annoyance, in fact, has raised to be regarded as a useful measure of perceived ambient quality and a complementary tool for health surveillance and transportation management. In such scenario, it is important to understand the utility that any information gathered on OSNs may have.

The utility of OSN-based information may be assessed adopting two different perspectives. In particular, aiming at automatically estimating pollutant levels, problems such as perception, reliability and interpretability of OSN posts emerge. In essence, consider the simple case where a human being, say John, perceives a given pollutant. Even if John correctly perceives that pollutant, the reliability of what he says online is not guaranteed. In addition, say John perceived the pollutant correctly and is convinced he reported what he perceived online, his post may still be misunderstood by a computer algorithm, for a long list of reasons (e.g., jargon, ironic expression, cultural approach, etc.). Adopting such perspective, hence, the validity of the OSN information regarding a given location at a given time should be verified finding ways of confronting posts with pollutant levels measured by nearby traditional sensors (Figure 1).

Regardless of the accuracy of any OSN-based information, such information may still be interesting for policy and planning purposes. In fact, considering the matrix of possible information outcomes, shown in Figure 2, it is possible to analyze four possible cases:

1. Both OSN-posts and pollution sensors signal high pollution;
2. Both OSN-posts and pollution sensors signal low pollution;
3. The annoyance that emerges from OSNs is high, while pollution levels are low;
4. The annoyance that emerges from OSNs is low, while pollution levels are high.

While finding 1) and 2) would indeed be interesting in order to adopt OSN posts as sensors, also scenarios 3) and 4) are relevant for policy and planning.

In fact, both of such situations signal a mismatch between what is perceived by the population, according to OSNs, and the reality of things. Finding a high annoyance level and low pollution values can require responding to questions like: are the methodologies adopted to assess pollution levels sound? Why are people perceiving (and complaining) so harshly about pollution, when pollution appears contained below reasonable values? Similar questions may be asked in the opposite situation, while a lack of interest for the problem may pose an interesting policy question, regarding the degree of awareness and knowledge of the population concerning such problem.



Fig. 1: Information mix: extending sensor coverage using OSN posts.

### 3.2 ITSs and Pollution

Before describing which type of ITS architecture could effectively incorporate OSN-based sources of information, it is worthwhile summarizing how such complex systems are organized. Two main approaches have emerged with the aim of limiting the business and societal costs of vehicular congestion (Ben-Akiva et al., 1994), (Wootton et al., 1995), (Hellenga et al., 1995). The first

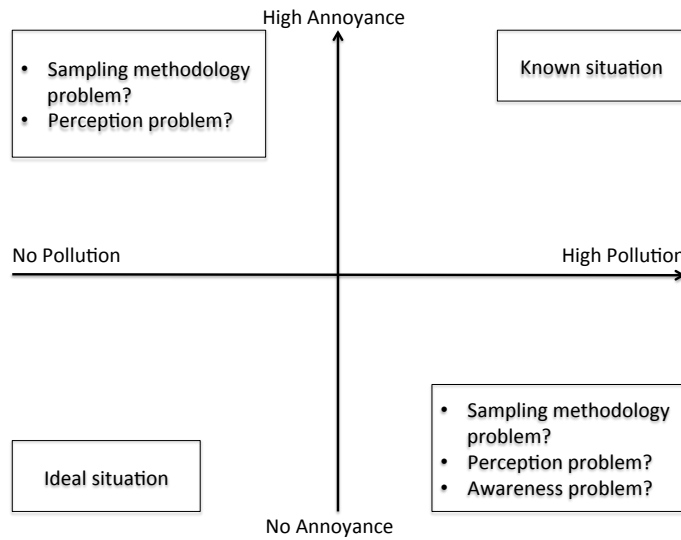


Fig. 2: Information alignment matrix: possible scenarios and implications for transportation policy and planning.

amounts to provide aggregate traffic information (e.g., intensity of traffic volumes estimated and lane occupancy rates estimated with video cameras and induction loops) to Advanced Traffic Management Systems (ATMSs) to control the road infrastructure (e.g., traffic light cycles or congestion charges) and to provide aggregate traffic information to drivers (e.g., with dynamic message signs and FM radios). The second approach is, instead, based on the idea of feeding road traversal times, sampled with probe vehicles, to Advanced Traveler's Information Systems (ATISs) which, in turn, supply drivers with a feedback on traffic and suggested routes (i.e., a driver's optimal route) to their destinations.

Both of such approaches deal with a plethora of traffic information sources (e.g., cellular networks, vehicular ad hoc networks, induction loops, video cameras, etc.) that have or are in the process of becoming available (?). The idea is to also integrate data derived from OSNs, just as sensor data, into such infrastructure (Figure 3). In order to do this, for the particular case of pollution data, it is important to remind that the two systems are operated by players with completely different points of view and needs: transportation authorities run ATMSs, while ATISs are marketed by private businesses. The authorities could implement on a short-term basis what has been discussed in the previous Sub-

section: require ATMSs to take online decisions depending on past, current or expected situations, discouraging traffic flows from traversing areas that were, are or are expected to be experiencing high pollution levels. ATISs could also be involved into such process (e.g., provide concerned drivers with sustainable routes, implement carpooling systems, indicate multi-modal transportation solutions that may jointly optimize travel time and sustainability-related variables) although with potentially many more difficulties, as the immediate goal of any driver is that of diminishing the time s/he spends on the road.

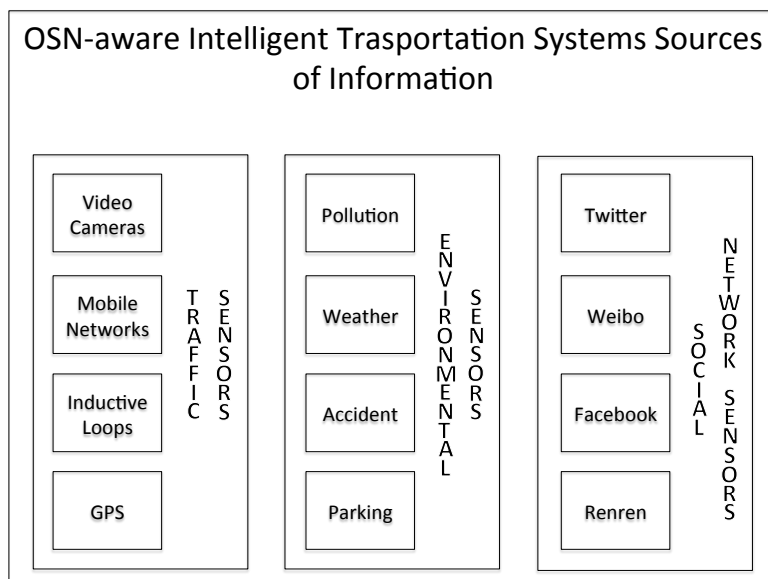


Fig. 3: Possible sources of information for transportation policy and planning.

#### 4 Posts-Pollutants Interconnection Model

Although quantifying the annoyance level of the population online, as a result of experienced pollution conditions, may always be useful for transportation purposes, as discussed in the previous Section, this work aims at revealing those OSN signals which may interpreted as symptoms of effective pollution, if any exist. Finding patterns that bind pollution-related OSN posts to chemical agent levels, hence, requires providing answers to two questions: (a) which are the pollution emissions and levels (if any) that push OSN users to write given types of posts, and, (b) if such posts exists, do they strictly touch upon pollution, or do they include, instead, references to other topics (e.g., weather,

congestion, ability to breath, etc.)? An answer to both of such questions can be given proceeding in the following order: the candidate sentinel topics and the pollutants of interest are identified in this Section, whereas the search for any existing patterns between such quantities is postponed to the on-the-field data analysis conducted as part of this work.

Now, the peculiarity of an OSN source (i.e., where the vast majority of information is provided by human beings), when compared to hardware sensors, is that a certain amount of interpretation is required. For example, a person complaining about haze (i.e., haze in general may be due to humidity and have nothing to do with pollution) could in reality be experiencing a highly polluted situation. The same may be said for other variables other than pollution, weather-related and not. It is, hence, important to individuate a set of pollution-related areas that may provide reliable pollution-sentinel variables. A non-exhaustive list of such areas is given by the following: (a) traffic situation (e.g., congestion, parking, accidents, etc.), (b) weather, and, (c) health, just to mention the most evident ones (an exemplar semantic map is provided in Figure 4).

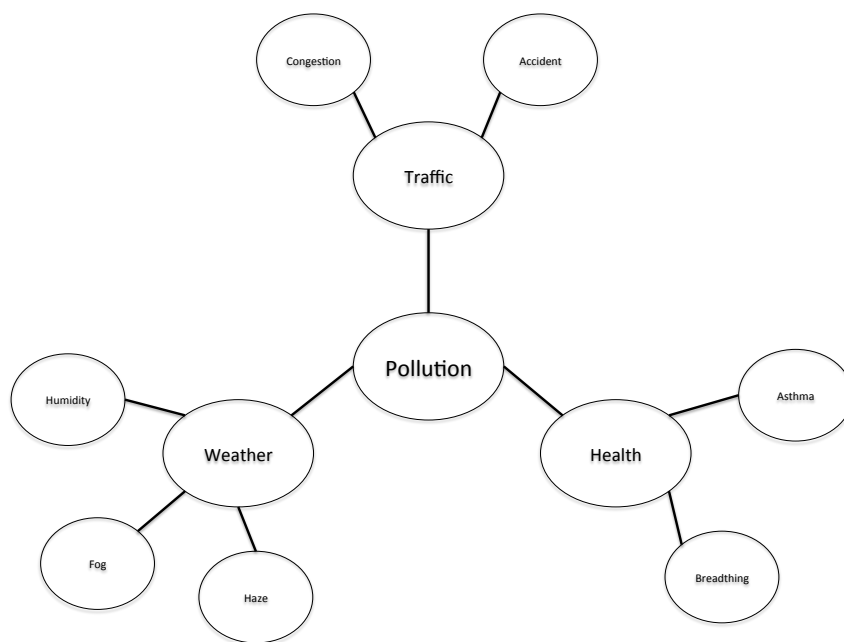


Fig. 4: Exemplar semantic map for pollution.

Traffic is widely recognized as the main causes of air pollution in urban areas. In some areas the contaminants introduced by traffic may become so disruptive to provoke heavy modifications to normal weather conditions (e.g., haze, fog, etc.) (Künzli et al., 2000), (Huang et al., 2014), (Wang et al., 2006). For this reason posts indicating high congestion levels (or also the occurrence of accidents which will eventually lead to congestion) and the presence of haze and foggy conditions could be, in reality, respectively indicating that pollution levels will be soon climbing or have already exceeded acceptable limits. While it may be reasonable to consider posts complaining about traffic congestion or particular weather conditions, for example, as indicators of high pollution levels, the use of health related posts requires particular care. Specific health problems (e.g., asthma, etc.) are not taken here into account, as the subject is so critical and multifaceted (i.e., understanding the causal relationship at the root of given health impairments can require years, while ITS decisions are taken within seconds to at most hours from an event) that it deserves a stream of research on its own. In addition, evidence exists from previous works that who chronically suffers from respiratory problems tends to exacerbate his/her perception of pollution (Bickerstaff, 2004). This work will consequently limit to consider only those posts that explicitly signal the experiencing of respiratory problems due to poor air conditions (which are still related to the health of a person, but on a much shorter time scale).

It is now possible to define which pollutants have been considered for the purposes of this work. Although many different elements are classified as pollutants, as possible causes of diseases and physical impairments, the ones that may be directly related to vehicular traffic include chemical agents such as (Beckerman et al., 2008):

- Sulfur Dioxide ( $SO_2$ ), characterized by an irritating odor, combined with other elements can contribute to the production of haze and reduced visibility;
- Ozone ( $O_3$ ), bluish color, in high concentrations its smell is sharp, resembling the smell of electrical equipment;
- Carbon Monoxide (CO), colorless, odorless and non-toxic;
- Nitrogen Oxides ( $NO_x$ ), which include Nitric Oxide (NO), a colorless, odorless and non-toxic gas, and Nitrogen Dioxide ( $NO_2$ ), a reddish-brownish gas with a pungent odor, an important component of city smog;

and particulate matter, which can also contribute to haze in urban contexts, such as:

- Respirable Suspended Particles ( $PM_{10}$ ), particulate matter with a diameter of 10  $\mu\text{m}$  or less;
- Fine Suspended Particles ( $PM_{2.5}$ ), particulate matter with a diameter of 2.5  $\mu\text{m}$  or less.

Despite the fact a causal relationship exists between the listed pollutants and what a person could feel (e.g., pungent odor, hazy weather, etc.) inhaling them, searching for relationships between what is written online and pollution

is clearly different than running a controlled experiment in a lab. In essence, confronting an OSN-based approach to a controlled experiment study, the OSN one lacks of any additional information which may help interpreting posts (e.g., gender, age, census, etc.). Nonetheless, this work explores two dimensions that cannot be dealt with in a controlled environment: (a) exploiting a high number of *participants*, as well as, (b) any existing spatial and time correlations between what people report online and the sensed levels of the chemical agents of interest.

## 5 Data Sources and Classification

Searching for spatiotemporal relations between pollution levels and OSN posts requires, as a first step, finding reliable sources of information, amounting to: (a) a large-scale source of geo-located posts, and, (b) pollution sensor data. This, however, only amounts to the first step: even if both pollution-related posts and sensor data were found for the same areas and in the same time frames, the design of an algorithm that may automatically and reliably select pollution-related posts from the enormous corpus that is daily published in any of the most popular OSNs is not trivial. The following explains how such problem has been approached.

### 5.1 Source selection

The sources of data selected for this study are: (a) an OSN with an abundant number of posts published nearby accessible pollution sensors, (b) at least two different cities notable for their pollution emissions (i.e., with, possibly, appreciable differences in terms of average pollution levels). For the initial purposes of this study, both of such conditions have been met by the cities of Hong Kong and Guangzhou, in China. Hong Kong and Guangzhou are both very populated (7.2 and 8.5 million of inhabitants, respectively), and, both can exhibit high pollution levels. Due to a number of factors (i.e., distance from the ocean, etc.), Hong Kong average pollution levels are typically lower than those experienced in Guangzhou (who, 2014).

Hong Kong and Guangzhou also share another interesting property: they are among the cities where registered users of China's most popular microblogging system, aka Sina Weibo, are most active (Guo et al., 2012). Sina Weibo, a Twitter and Facebook hybrid, features a penetration rate above the 30% of Chinese Internet users. In December 2012, Sina Weibo had 503 million registered users, with about 100 million daily posted messages. More than 70% of Sina Weibo users daily utilize such service from mobile, thus, not only sharing their posts but also their positions. Sina Weibo is, hence, an ideal source of information, as in the majority of cases it is possible to verify whether a post that has been shared has been written anywhere nearby one of the pollution sensors taken here into consideration.

In addition, for Hong Kong it has been possible to access multiple sources of pollution information:  $SO_2$ ,  $CO$ ,  $O_3$ ,  $NO_x$ ,  $NO_2$ ,  $PM_{10}$  and  $PM_{2.5}$  sensors from multiple sites. For the purpose of this work, the Central, Causeway Bay and Mongkok sites have been chosen, as they cover the downtown areas of the city. In Guangzhou, instead, a downtown sensor located inside and operated by the American Embassy provides  $PM_{2.5}$  hourly information.

## 5.2 Post classification

Starting on October 3rd 2012, throughout April 3rd 2014, approximately 640 thousand Sina Weibo posts written by 228,684 users and 910 thousand posts published by 505,033, have been recorded, respectively, in the areas of Hong Kong and Guangzhou. Of all this corpus of data, only those posts that have been written within a radius of 5 km from any of the accessible pollution sensors (i.e., a value chosen as a tradeoff between finding a sufficient number of posts and observing pollution conditions consistent with the reference sensors) have been considered. Now the final step required to be able to support a comparison between pollution levels and pollution-related posts was the selection of the latter. Pollution-related post classification has been conducted according to the strategy indicated in Section 4, i.e., searching for any posts that were related to air pollution in any of the following four categories: (a) pollution, (b) weather, (c) traffic, or, (d) health. To do so in an automatic way, three different paths have been followed.

The first path has been that of utilizing a well-accepted methodology in text classification, which is a Naïve Bayesian classifier (McCallum and Nigam, 1998). This has been performed resorting to the PyMining Naïve Bayes implementation, which supported the construction of a database of relevant features concerning the pollution, weather, traffic and health. In order to test the soundness of the classifications operated with this algorithm, a total number of 1000 posts, randomly taken from those that have been classified as related to pollution, have been divided into 10 subsets of 100 posts each and checked by 10 independent human classifiers, respectively. Unfortunately, this check revealed that the average performance of this method was as low as 61%.

The second path has, hence, involved utilized a second popular classification mechanism, namely a Support Vector Machine (SVM) (Hearst et al., 1998). In order to classify the Chinese text, this strategy required walking through two successive steps, training and testing. During the training part, a training document containing relevant terms related to pollution was manually created and its words were segmented to formulate a model from the computed eigenvalues and term-frequency inverse-document-frequency matrix. The application of such model improved the classification performance, compared to the use of Naïve Bayes. The same test that was utilized for Naïve Bayes, in this case revealed an average performance of 88% (i.e., 88% of the posts classified as pollution-related where effectively pollution-related).



The third and final path entailed utilizing a dictionary constructed utilizing the keywords most frequently found in pollution-related posts. Table 1 summarizes the keywords that have been used to individuate posts falling into each of the four categories. Interestingly, the use of such terms provided an average performance of 91%, higher than any of the machine learning algorithms employed so far. We hence opted for utilizing such methodology for post classification in the rest of this work.

Table 1: Pollution-related terms dictionary.

Category	Terms (English translation)
Pollution	Pollution, poor air, gray sky, air quality
Weather	Haze, fog, gray sky, bad weather
Traffic	Traffic jams, congestion
Health	Unable to breath

As further and final checks, additional tests have been performed on the final sets of data (ca. 15,000 posts in Hong Kong and almost 4,000 posts in Guangzhou), in order to ensure that the obtained posts were effectively pollution-related ones and avoid jeopardizing the rest of this work. Figures 5 and 6 provide a visual representation of where the pollution sensors are and where the majority of pollution-related posts appeared in Hong Kong and Guangzhou, respectively. Now, the proposed post-pollutants model is evaluated in the next Section in order to verify whether any relation emerges between such quantities from the analyzed sources of data.

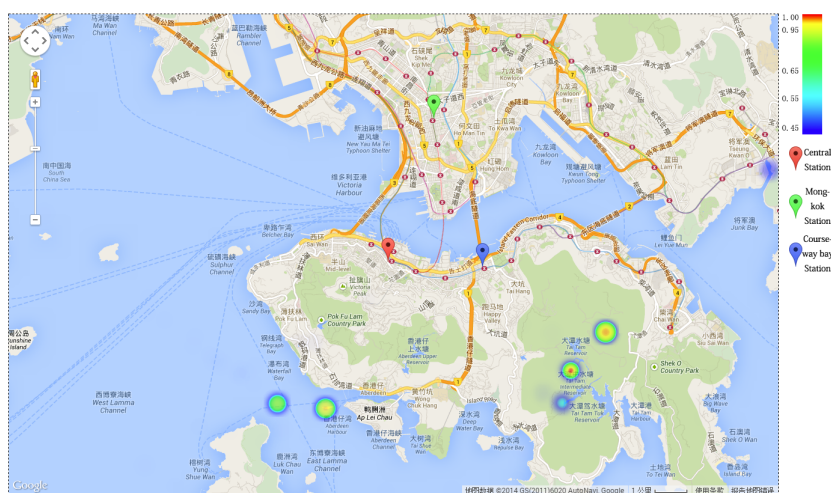


Fig. 5: Pollution-related posts and sensor positions in Hong Kong.

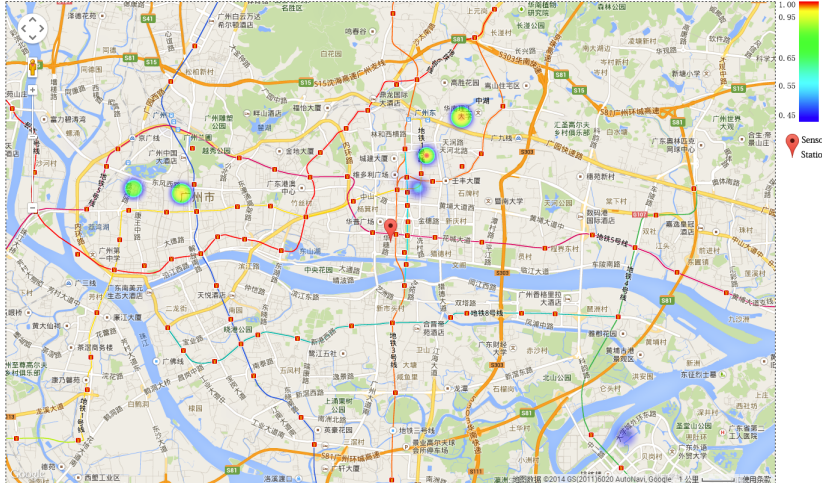


Fig. 6: Pollution-related posts and sensor positions in Guangzhou.

## 6 Model Evaluation

An evaluation of the proposed post-pollution interconnection model requires finding any patterns connecting such two quantities. To do so, a hypothesis regarding how a user behaves after being exposed to elevated pollution levels is needed. The hypothesis that is adopted in this work is: when exposed to high pollution, a person will, with a high chance, first of all try to find a way of getting out of such distressing situation. Once in a safe area (e.g., indoors), that person may, with some probability, then advertise the occurrence of such situation.

Now, one of the problems that derive from such hypothesis is that the longer anyone waits to signal his/her distress due to experienced pollution conditions, the lower the value of such information for any system which operates relying on real-time data (e.g., ITS).

This justifies the use of the following evaluation strategy: for each pollution-sensed value, count the number of pollution-related posts that have been published during the T hours that follow, where T has been set to 2 hours.

In the following hence, it is shown how pollution values vary, as the minimum number of posts published in the next 2 hours, within a range of 5 km from the sensing station, increases. Anticipating here one of the results of this work, an insufficient number of traffic and health posts were found, hence the remainder of this Section will concentrate on the most significant results obtained with the posts that contained pollution and weather key terms.

## 6.1 Term Categories vs. Pollution Levels

When utilizing pollution keys, the key statistics of given pollutant concentrations varied as the minimum number of posts published increased. This, in particular, happened with Carbon Monoxide in downtown Hong Kong at the Central pollution station (as shown in Figure 7) and with Fine Suspended Particulates in downtown Guangzhou (Figure 8), where minimum, first quartile, median, third quartile statistics and any outliers (i.e., empty dots) are plotted for a given minimum number of posts. When, at least 5 pollution posts were published in Central Hong Kong, CO values always exceeded a value just below  $500 \mu g/m^3$ . Such value drops dramatically (i.e., almost  $0 \mu g/m^3$ ), when the post number threshold is set to 1. Similarly, in Guangzhou the presence of at least 5 pollution posts yields  $PM_{2.5}$  values which exceed  $100 \mu g/m^3$ , as opposed to all those situations where at least 1 relevant post is published, returning a minimum  $PM_{2.5}$  value close to  $0 \mu g/m^3$ .

Resortin to the US Environmental Protection Agency (EPA) it is possible to gain a practical understanding of these values. Average CO concentrations should not exceed the hourly average value of  $40 \mu g/m^3$  and average  $PM_{2.5}$  concentrations should not exceed  $35 \mu g/m^3$  over a 24 hours time interval (i.e., for the sake of completeness, such value is typically lower than the maximum value admitted over an hour averaging) (epa, 2014). The trends shown in Figures 7 and 8 reveal that CO values in Hong Kong are well below the EPA recommended limits, while the  $PM_{2.5}$  in Guangzhou are not.

Almost all weather-related posts found during the period of interest were related to haze. Disturbing haze conditions, as anticipated in Section 4, are consistent with the presence of high levels of given pollutants (e.g.,  $NO_2$ ,  $NO_x$ ,  $SO_2$  and particulate matter). We hence here concentrate on the relation between pollutants and haze related posts.

Nitrogen Dioxide may be a cause of haze in urban areas, this is indicated by the fact that more posts are recorded as its minimum value increases (Table 2). However, considering that EPA recommends an  $NO_2$  concentration limit equal to 100 ppb ( $188 \mu g/m^3$ ) over an hour time frame, it is possible to observe that an increased number of posts do not substantially indicate that such value has been exceeded.

Table 2: Number of haze posts vs. minimum  $NO_2$  values.

Minimum $NO_2$ value		Pollution sensor location
no. of posts $\geq 1$	no. of posts $\geq 5$	
11 $\mu g/m^3$	39 $\mu g/m^3$	Causeway Bay, HK
10 $\mu g/m^3$	56 $\mu g/m^3$	Central, HK
1 $\mu g/m^3$	48 $\mu g/m^3$	Mongkok, HK

$PM_{2.5}$  levels exhibit the most interesting relation with haze posts. This does not apparently result from the post-pollution values obtained for Hong Kong and reported in Table 3. An explanation to this may simply be the fact

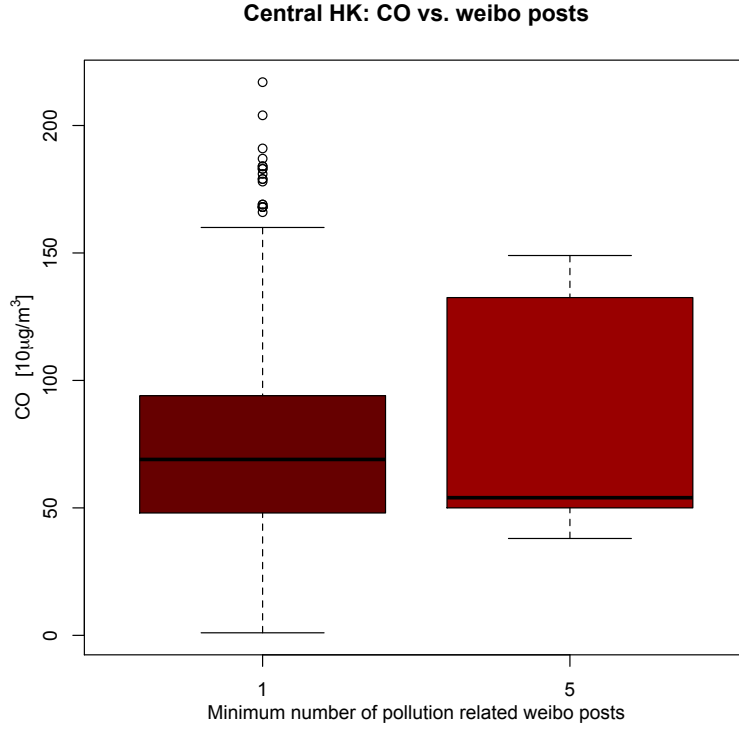


Fig. 7: Pollution posts vs. CO values in Central, Hong Kong.

that average values in Hong Kong, for  $PM_{2.5}$ , fall in the 30-40  $\mu g/m^3$  range, in essence approximately equal to the EPA 35  $\mu g/m^3$  limit. Hazy conditions are reported in literature when  $PM_{2.5}$  values exceed those obtained in Hong Kong (Sun et al., 2006). The situation in Guangzhou appears different, according to the visual relation between haze-related posts and  $PM_{2.5}$  levels returned by Figure 9: more posts (i.e., at least 15 haze related posts) are published as the minimum value of  $PM_{2.5}$  falls above 100  $\mu g/m^3$ , hence, well beyond acceptable values (Zhang et al., 2014). A closer look at such data reveals that the days when Sina Weibo users in Guangzhou noticed higher haze levels, publishing at least 15 posts, are the 13th of December 2013, the 8th and the 25th of January, the 17th 18th and 26th of February and the 3rd and 12th of March 2014, all part of a period where high haze and pollution have been recorded for Guangzhou.

It was possible to establish a dependence also between the values of CO concentration and the number of weather-related posts published by Sina Weibo users. Table 4 shows this trend, which is more evident in Central than in other locations in Hong Kong.

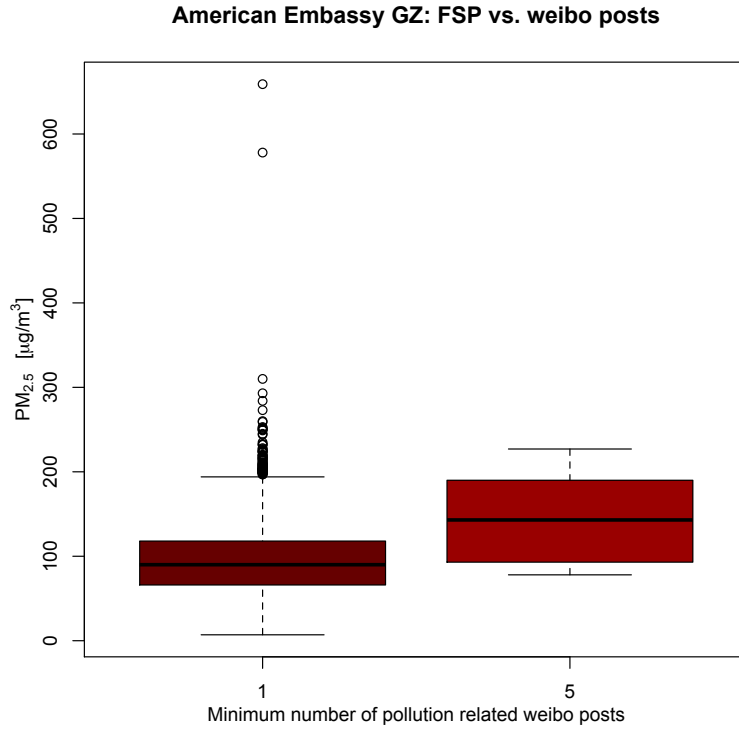


Fig. 8: Pollution posts vs.  $PM_{2.5}$  values at the American Embassy, Guangzhou.

Table 3: Number of haze posts vs. minimum  $PM_{2.5}$  values.

Minimum $PM_{2.5}$ value		Pollution sensor location
no. of posts $\geq 1$	no. of posts $\geq 5$	
2 $\mu g/m^3$	7 $\mu g/m^3$	Causeway Bay, HK
3 $\mu g/m^3$	8 $\mu g/m^3$	Central, HK
1 $\mu g/m^3$	5 $\mu g/m^3$	Mongkok, HK

## 6.2 Extending the Evaluation

The results described in Subsection 6.1 strongly indicate the existence of a trend between: (a) haze, a topic often mentioned on OSNs by the citizens of Hong Kong and Guangzhou and, (b) high  $PM_{2.5}$  values. To corroborate such result Guangzhou data is further validated and the same approach is extended to Beijing, Shanghai and Chengdu.

A further check on Guangzhou data entailed first establishing whether persons or machines had published the posts considered in Figure 9 concerning haze in Guangzhou. The entire data set has been reviewed. Of almost 4000

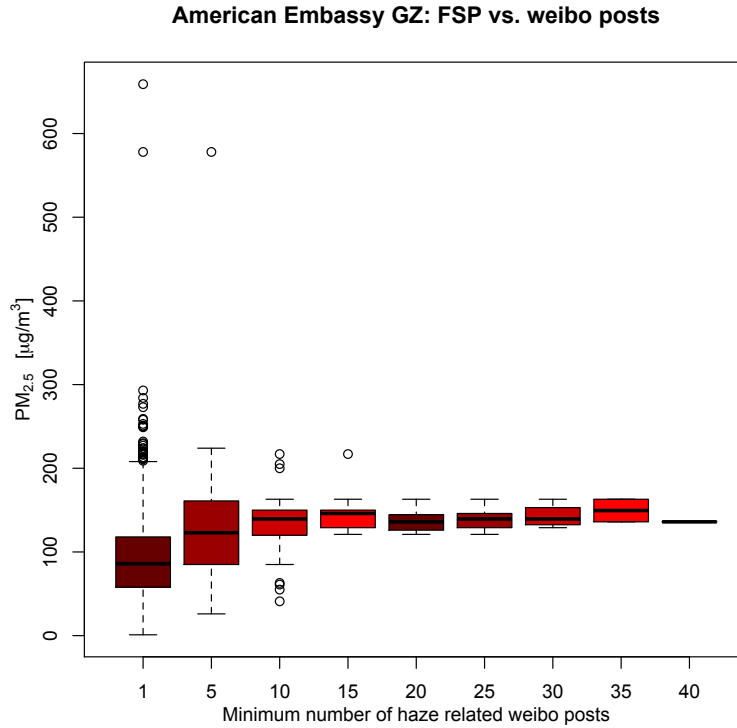


Fig. 9: Weather posts vs.  $PM_{2.5}$  values at the American Embassy, Guangzhou.

Table 4: Number of haze posts vs. minimum  $CO$  values.

Minimum $CO$ value		Pollution sensor location
no. of posts $\geq 1$	no. of posts $\geq 5$	
180 $\mu g/m^3$	410 $\mu g/m^3$	Causeway Bay, HK
10 $\mu g/m^3$	370 $\mu g/m^3$	Central, HK
330 $\mu g/m^3$	510 $\mu g/m^3$	Mongkok, HK

posts, approximately 300 (ca. 8%) were automatically produced by a meteorological site, yielding 3,726 posts written by 3,493 human users between October 18th 2013 and April 17th 2014. After removing such unwanted data, we further operated as follows. We also included into our analysis the sensor readings with no following haze-related posts. We excluded from such data set the sensor readings produced between 12AM and 6AM, as vehicular traffic is less intense at night, and confronted such distribution with the pollutant-level distributions.  $PM_{2.5}$  distributions have been obtained for  $[1, 5)$ ,  $[5, 10)$ ,  $[10, 15)$ ,  $[15, \infty)$  number of messages subsequently posted on OSNs. The trend

emerging in Figure 10a confirms that minimum  $PM_{2.5}$  levels increase as the number of haze-related published posts increase.

Now, for the four sites that have been so far considered, the average  $PM_{2.5}$  values for 2013 were  $33 \mu g/m^3$  in Hong Kong Mongkok, 34 in Hong Kong Central, 45 in Hong Kong Causeway Bay and 52.2 in Guangzhou, respectively. The same average values obtained by the sensor stations taken under consideration in Beijing, Shanghai and Chengdu were 90.1, 60.7 and  $86.3 \mu g/m^3$ , respectively. The observation period for the three new cities taken under consideration is October 21th 2014–December 17th 2014.

A total number of 292,590, 283,179 and 248,939 posts were published from 135,870, 127,320 and 111,641 accounts, respectively, in Beijing, Shanghai and Chengdu. Processing such posts, it has been possible to observe that 1,492, 426 and 416 haze-related posts were published by respectively 1421, 415 and 379 human users in the three cities of interest (Beijing, Shanghai and Chengdu, in this order). Of these posts, 1312, 248 and 228 were published within a 5 km radius from a source of  $PM_{2.5}$  data. Figures 10b, 10c and 10d confirm what was first observed in Guangzhou: as the number of haze-related posts increase, the minimum value of  $PM_{2.5}$  emissions increase, generally indicating haze-related posts as a symptom of high  $PM_{2.5}$  values.

### 6.3 Model Analysis

The result inferred in Subsection 6.2, concentrations of  $PM_{2.5}$  exceeding given values are often signalled by multiple posts concerning haze appearing on Weibo, does not leverage on the fact that given Weibo users may be more reliable than others. If this were true, monitoring selected users may result convenient in order to better track  $PM_{2.5}$  concentrations. The frequency graphs shown in Figure 11 indicates such approach would not work in the observed cities, as no user posted more than 5 messages concerning haze, while the great majority posted at most 2. Such scenario does not support the design of algorithms that may leverage on the past history of users, at least on the time intervals considered in this paper.

Now, to better quantify the relation between posts and pollution, the Spearman's rank correlation coefficient is adopted. In essence, such coefficient provides a non-parametric measure of statistical dependence between two variables assessing how well the relationship between two variables can be described using a monotonic function. A Spearman correlation value of 1 indicates the variables are a perfect monotone function one of the other. Table 5 provides the Spearman correlation coefficients and their associated statistical significance values obtained when comparing two variables: number of posts vs. the  $k$  smallest pollution values recorded in correspondence of the given number of posts. In essence, coefficients close to 1 prove what already observed in Subsection 6.2, a higher number of haze-related posts signals an increased minimum concentration value of  $PM_{2.5}$ .

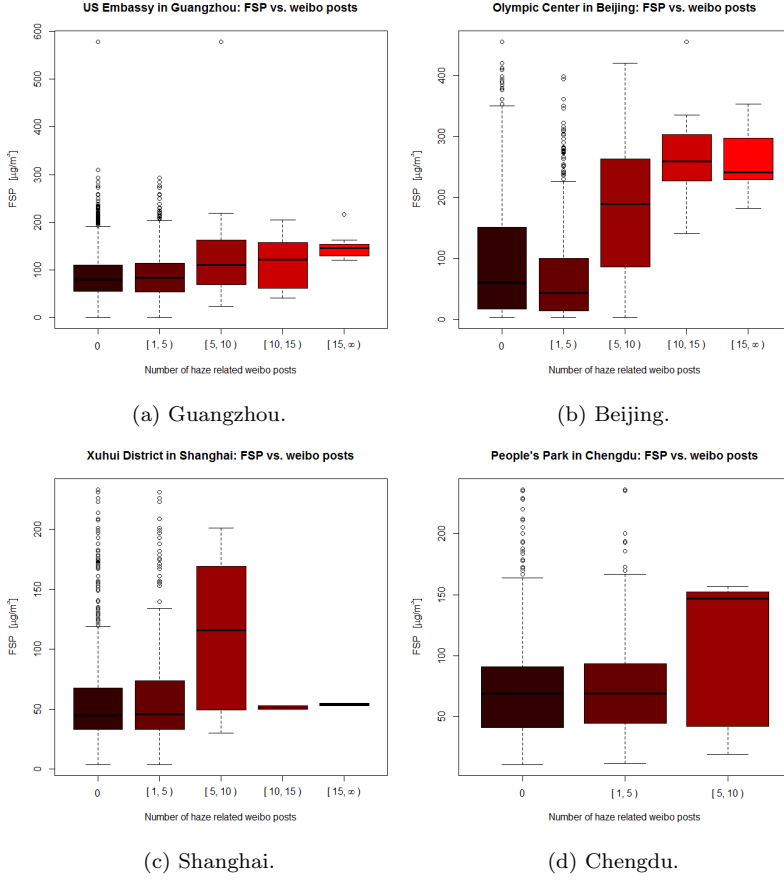


Fig. 10: Weather posts (w/o automatically generated posts) vs.  $PM_{2.5}$  values.

The coefficient values provided in Table 5 establish the existence of a monotonic relationship between the number of posts and the minimum concentration of  $PM_{2.5}$ . Such relation is further assessed computing another figure of merit:  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$ , where  $\rho_{PM_{2.5}}$  amounts to the concentration of  $PM_{2.5}$ ,  $\rho_{thr}$  to a constant (i.e., concentration threshold value of interest) and  $n_{posts}$  to the number of haze related posts. In simple words,  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$  represents the probability that a given  $PM_{2.5}$  concentration value is exceeded, conditioned to the number of posts that appeared on the OSN. Along with the values of  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$ , for  $\rho_{thr} = 35$  and 70 in Figures 12 and 13 and  $\rho_{thr} = 35$  in Figures 14 and 15, also the respective logistic regression curves are provided. In essence, higher values of  $n_{post}$  correspond to a higher probability of observing pollution samples which exceed  $\rho_{thr}$ . In particular, a value of  $n_{posts} > 5$  (obtained maximizing the true positive occurrence rate throughout the four scenarios of interest) returns high



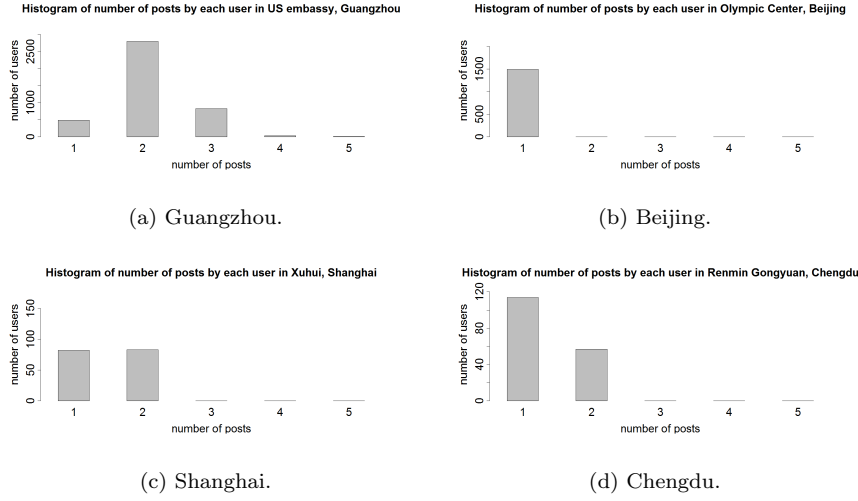


Fig. 11: Post number per user frequency plot.

probability values in all cases. As expected when  $n_{posts} \leq 5$  nothing can be really said, especially in Shanghai and Chengdu (Figures 14 and 15). This fact well emerges from the data reported in Table 6, where the number of true negative, false negative, true positive and false positive number of samples are reported taking adopting the threshold values of  $\rho_{thr} = 35$  and  $n_{posts} > 5$ . Now, although  $\rho_{thr} = 35 \mu g/m^3$  represents the  $PM_{2.5}$  threshold recommended by EPA, such value is well exceeded by the  $PM_{2.5}$  average concentrations observed in the cities that have been taken under consideration. For this reason a further analysis is performed on the two cities where more posts are available, Guangzhou and Beijing. Increasing the value of  $n_{posts}$  until a 90%-10% ratio is maintained between those samples which fall in the True Positive set and those that instead fall in the False Positive one, the values of 6 and 10 are found for Guangzhou and Beijing, respectively (Table 7).

## 7 Conclusion

The results shown in paper show a trend exists between OSN posts and  $PM_{2.5}$  levels. Unlike what one could possibly expect (i.e., the existence of an observable trend between posts complaining about traffic or pollution and pollutant recorded levels), the strongest ties that have been found in this work are between pollution levels and haze. This is a key finding, as it means that, at least for the cities taken into consideration, the concern for a particular weather condition like haze is high and that such concern can be associated with the observation of high pollution levels.

Table 5: Minimum  $k$  recorded  $PM_{2.5}$  values compared with the number of posts.

Spearman's coefficient and significance		
	Correlation Coefficient	$p$
<b>Guangzhou</b>		
$k = 1$	0.86	0.0001
$k = 2$	0.89	0.0001
$k = 3$	0.86	0.0001
$k = 4$	0.86	0.0001
$k = 5$	0.88	0.0001
<b>Beijing</b>		
$k = 1$	0.88	0.0001
$k = 2$	0.82	0.0001
$k = 3$	0.83	0.0001
$k = 4$	0.86	0.0001
$k = 5$	0.86	0.0001
<b>Shanghai</b>		
$k = 1$	0.90	0.0374
$k = 2$	0.84	0.0088
$k = 3$	0.75	0.0079
$k = 4$	0.67	0.0082
$k = 5$	0.62	0.0085
<b>Chengdu</b>		
$k = 1$	1.00	0.0001
$k = 2$	0.96	0.0028
$k = 3$	0.90	0.0011
$k = 4$	0.88	0.0004
$k = 5$	0.87	0.0001

Table 6: Number of pollution concentration samples which fall beyond  $\rho_{thr} = 35 \mu g/m^3$  (true positive) and which not (false positive) for  $n_{posts} > 5$ .

Number of pollution samples				Location
True Neg.	False Neg.	True Pos.	False Pos.	
243	2495	42	0	Guangzhou
342	572	94	10	Beijing
426	619	10	1	Shanghai
208	896	4	0	Chengdu

Table 7: Number of pollution concentration samples which fall beyond  $\rho_{thr} = 70 \mu g/m^3$  (true positive) and which not (false positive).

Number of pollution samples				$n_{posts}$	Location
True Neg.	False Neg.	True Pos.	False Pos.		
1127	1630	21	2	10	Guangzhou
527	415	72	4	6	Beijing

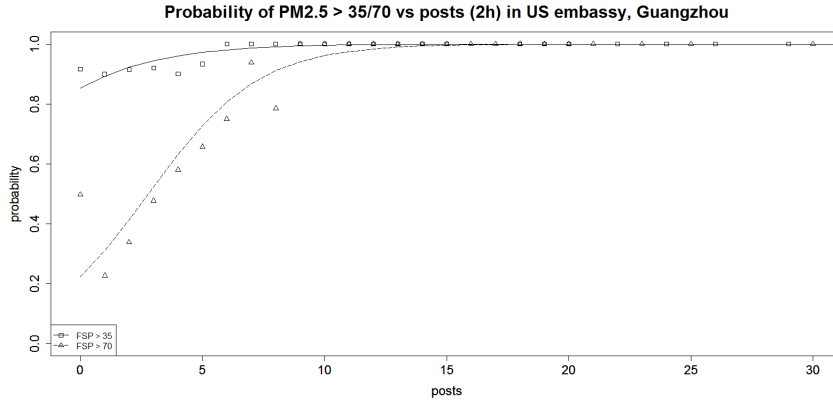


Fig. 12:  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$  values and logistic regression curves for  $\rho_{thr} = 35 \mu g/m^3$  and  $\rho_{thr} = 70 \mu g/m^3$  in Guangzhou.

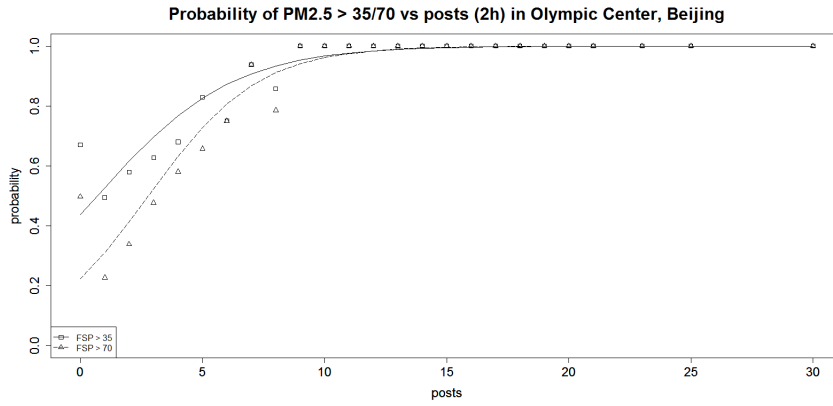


Fig. 13:  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$  values and logistic regression curves for  $\rho_{thr} = 35 \mu g/m^3$  and  $\rho_{thr} = 70 \mu g/m^3$  in Beijing.

This work leaves a number of open challenges and unanswered questions. A first challenge derives from the limitation of using humans (clearly, not designed to sense or advertise such type of events) as pollution sensors: high pollutant levels may not be detected in many cases (e.g., the maximum value of  $PM_{2.5}$ ,  $659 \mu g/m^3$ , is observed within a 2-hour interval where only a single post was published). How OSN information could be best balanced with pollution sensor information, to improve the quality of pollutant level estimates in those areas where no specific sensors are available, remains hence an open question.

A second open challenge is of cultural and sociological nature: how do people from different areas of the world react to pollution on OSNs? Is it

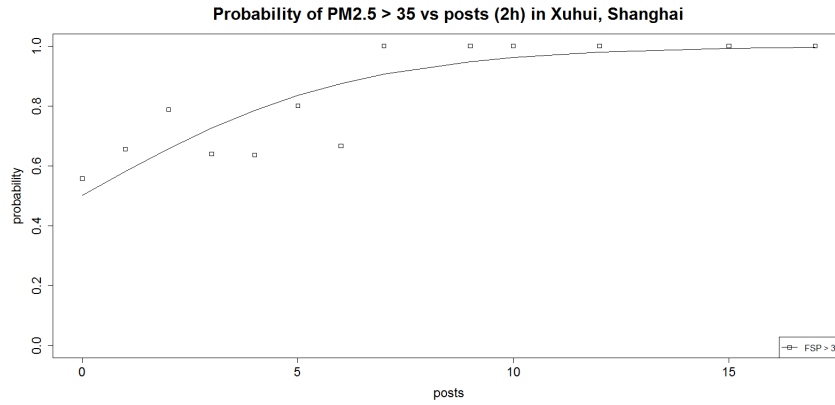


Fig. 14:  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$  values and logistic regression curve for  $\rho_{thr} = 35 \mu g/m^3$  in Shanghai.

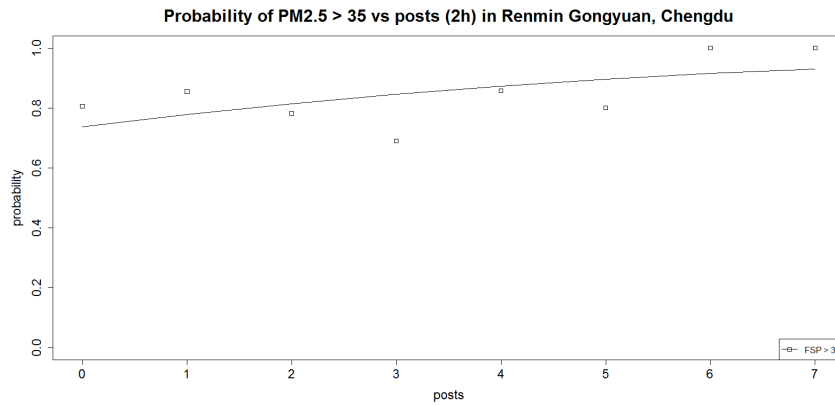


Fig. 15:  $P(\rho_{PM_{2.5}} > \rho_{thr} | n_{posts})$  values and logistic regression curve for  $\rho_{thr} = 35 \mu g/m^3$  in Chengdu.

possible to extend what we have found for China to other areas of the world?  
 Or should these results only be contextualized to the Chinese case or to given  
 pollution configurations?

The strongest indication that this work provides, at this point, in the field  
 of transportation, is that significance of OSN posts goes well beyond that of  
 a simple indicator of annoyance, in many cases such information can reveal  
 effective trends in pollutant concentrations. Human beings, hence, through  
 the use of OSNs, could integrate with the information they provide other  
 environmental sensors. Although a very interesting opportunity, this should  
 however be accomplished considering the fact that human sensors are subject  
 to randomness and noise and, as such, any pollution information they provide

should be periodically corroborated with other means of assessing pollution levels.

## References

EPA Air Criteria, 2014. URL <http://www.epa.gov/air/criteria.html>.  
Outdoor Air Database, World Health Organization, 2014.

A. K. Agarwal. Biofuels (alcohols and biodiesel) applications as fuels for internal combustion engines. *Progress in Energy and Combustion Science*, 33(3):233–271, jun 2007. ISSN 03601285. doi: 10.1016/j.pecs.2006.08.003. URL <http://www.sciencedirect.com/science/article/pii/S0360128506000384>.

I. N. Athanasiadis, A. E. Rizzoli, P. A. Mitkas, and J. M. Gómez, editors. *Information Technologies in Environmental Engineering*. Environmental Science and Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-88350-0. doi: 10.1007/978-3-540-88351-7. URL <http://www.springerlink.com/index/10.1007/978-3-540-88351-7>.

D. Banister. The sustainable mobility paradigm. *Transport Policy*, 15(2):73–80, mar 2008. ISSN 0967070X. doi: 10.1016/j.tranpol.2007.10.005. URL <http://www.sciencedirect.com/science/article/pii/S0967070X07000820>.

T. M. Barzyk, V. Isakov, S. Arunachalam, A. Venkatram, R. Cook, and B. Naess. A near-road modeling system for community-scale assessments of traffic-related air pollution in the United States. *Environmental Modelling & Software*, 66:46–56, apr 2015. ISSN 13648152. doi: 10.1016/j.envsoft.2014.12.004. URL <http://www.sciencedirect.com/science/article/pii/S1364815214003594>.

B. Beckerman, M. Jerrett, J. R. Brook, D. K. Verma, M. A. Arain, and M. M. Finkelstein. Correlation of nitrogen dioxide with other traffic pollutants near a major expressway. *Atmospheric Environment*, 42(2):275–290, jan 2008. ISSN 13522310. doi: 10.1016/j.atmosenv.2007.09.042. URL <http://www.sciencedirect.com/science/article/pii/S1352231007008412>.

M. Ben-Akiva, H. N. Koutsopoulos, and A. Mukundan. A DYNAMIC TRAFFIC MODEL SYSTEM FOR ATMS/ATIS OPERATIONS. *I V H S Journal*, 2(1):1–19, jan 1994. ISSN 1065-5123. doi: 10.1080/10248079408903812. URL <http://www.tandfonline.com/doi/abs/10.1080/10248079408903812#.VY6A3M4y1NM>.

M. Ben-Akiva, M. Bierlaire, J. Bottom, H. Koutsopoulos, and R. Mishalani. DEVELOPMENT OF A ROUTE GUIDANCE GENERATION SYSTEM FOR REAL-TIME APPLICATION. In *Transportation Systems 1997*, 1997. ISBN 0080429319. URL <http://trid.trb.org/view.aspx?id=505713>.

W. Ben Jaballah, M. Conti, M. Mosbah, and C. E. Palazzi. Fast and Secure Multihop Broadcast Solutions for Intervehicular Communication. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):433–450,

feb 2014a. ISSN 1524-9050. doi: 10.1109/TITS.2013.2277890. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6595027>.

W. Ben Jaballah, M. Conti, M. Mosbah, and C. E. Palazzi. A secure alert messaging system for safe driving. *Computer Communications*, 46:29–42, jun 2014b. ISSN 01403664. doi: 10.1016/j.comcom.2014.03.010. URL <http://www.sciencedirect.com/science/article/pii/S0140366414000966>.

J. T. Betts. Survey of Numerical Methods for Trajectory Optimization. *Journal of Guidance, Control, and Dynamics*, 21(2):193–207, mar 1998. ISSN 0731-5090. doi: 10.2514/2.4231. URL <http://arc.aiaa.org/doi/abs/10.2514/2.4231>.

K. Bickerstaff. Risk perception research: Socio-cultural perspectives on the public experience of air pollution. *Environment International*, 30(6):827–840, 2004. ISSN 01604120. doi: 10.1016/j.envint.2003.12.001.

K. Bickerstaff and G. Walker. Public understandings of air pollution: the ‘localisation’ of environmental risk. *Global Environmental Change*, 11(2): 133–145, 2001. URL <http://www.sciencedirect.com/science/article/B6VFV-430WWRB-4/2/a7d1de642bdee7ade8f2542793c2697c>.

P. Blythe, H. Bryan, P. Watson, B. Sharif, J. Neasham, S. Edwards, J. Wagner, M. Bell, and V. Suresh. An environmental sensor system for pervasively monitoring road networks. In *IET Road Transport Information and Control Conference and the ITS United Kingdom Members’ Conference (RTIC 2008)*, pages 91–91. Institution of Engineering and Technology, jan 2008. ISBN 978-0-86341-920-1. doi: 10.1049/ic.2008.0805. URL <http://digital-library.theiet.org/content/conferences/10.1049/ic.2008.0805>.

K. Boriboonsomsin, M. J. Barth, W. Zhu, and A. Vu. Eco-Routing Navigation System Based on Multisource Historical and Real-Time Traffic Information. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1694–1704, dec 2012. ISSN 1524-9050. doi: 10.1109/TITS.2012.2204051. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6236175>.

F. Costabile and I. Allegrini. A new approach to link transport emissions and air quality: An intelligent transport system based on the control of traffic air pollution. *Environmental Modelling & Software*, 23(3):258–267, mar 2008. ISSN 13648152. doi: 10.1016/j.envsoft.2007.03.001. URL <http://www.sciencedirect.com/science/article/pii/S1364815207000369>.

M. R. Edelstein. *Contaminated communities: The social and psychological impacts of residential toxic exposure*. 1988.

J. S. Evans-Cowley and G. Griffin. Microparticipation with Social Media for Community Engagement in Transportation Planning. *Transportation Research Record: Journal of the Transportation Research Board*, 2307(-1):90–98, 2012. ISSN 0361-1981. doi: 10.3141/2307-10. URL <http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2307-10>.

M. Fallah Shorshani, M. André, C. Bonhomme, and C. Seigneur. Modelling chain for the effect of road traffic on air and water quality: Techniques, current status and future prospects. *Environmental Modelling &*

*Software*, 64:102–123, feb 2015. ISSN 13648152. doi: 10.1016/j.envsoft.2014.11.020. URL <http://www.sciencedirect.com/science/article/pii/S1364815214003466>.

N.-E. E. Faouzi, H. Leung, and A. Kurian. Data fusion in intelligent transportation systems: Progress and challenges A survey. *Information Fusion*, 12(1):4–10, jan 2011. ISSN 15662535. doi: 10.1016/j.inffus.2010.06.001. URL <http://www.sciencedirect.com/science/article/pii/S1566253510000643>.

S. Ferretti, M. Furini, C. E. Palazzi, M. Roccetti, and P. Salomoni. WWW recycling for a better world. *Communications of the ACM*, 53(4):139, apr 2010. ISSN 00010782. doi: 10.1145/1721654.1721692. URL <http://dl.acm.org/ft{ }gateway.cfm?id=1721692{&}type=html>.

M. Florian. A Traffic Equilibrium Model of Travel by Car and Public Transit Modes. *Transportation Science*, may 1977. URL <http://pubsonline.informs.org/doi/abs/10.1287/trsc.11.2.166>.

F. Franchi and S. Malpezzi. Infomobility: an Integrated Framework, 2013. ISSN 2147-5369. URL <http://www.world-education-center.org/index.php/P-ITCS/article/view/1918>.

P. Fraternali, A. Castelletti, R. Soncini-Sessa, C. Vaca Ruiz, and A. Rizoli. Putting humans in the loop: Social computing for Water Resources Management. *Environmental Modelling & Software*, 37:68–77, nov 2012. ISSN 13648152. doi: 10.1016/j.envsoft.2012.03.002. URL <http://www.sciencedirect.com/science/article/pii/S1364815212000849>.

Y. Fu, Y. Fang, C. Jiang, and J. Cheng. Dynamic ride sharing community service on traffic information grid. *Proceedings - International Conference on Intelligent Computation Technology and Automation, ICICTA 2008*, 2: 348–352, 2008. doi: 10.1109/ICICTA.2008.399.

R. Goodspeed. The Limited Usefulness of Social Media and Digital Trace Data for Urban Social Research. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media Understanding*, pages 2–4, 2013. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6178/6290>.

Z. Guo, Z. Li, H. Tu, and L. Li. Characterizing User Behavior in Weibo. In *2012 Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing*, pages 60–65. IEEE, jun 2012. ISBN 978-1-4673-1956-0. doi: 10.1109/MUSIC.2012.18. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6305825>.

H. Hartenstein and K. Laberteaux. A tutorial survey on vehicular ad hoc networks. *IEEE Communications Magazine*, 46(6):164–171, jun 2008. ISSN 0163-6804. doi: 10.1109/MCOM.2008.4539481. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4539481>.

M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, jul 1998. ISSN 1094-7167. doi: 10.1109/5254.708428. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=708428>.

- B. Hellinga, M. Baker, M. Carter, and M. V. Aerde. Linking ATIS/ATMS and environmental plume dispersion models. *Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future*, pages 251–258, 1995. doi: 10.1109/VNIS.1995.518847.
- R. M. Hoff and S. A. Christopher. Remote Sensing of Particulate Pollution from Space: Have We Reached the Promised Land? *Journal of the Air & Waste Management Association*, 59(6):645–675, jun 2009. ISSN 1096-2247. doi: 10.3155/1047-3289.59.6.645. URL <http://www.tandfonline.com/doi/abs/10.3155/1047-3289.59.6.645>.
- R. Honicky, E. A. Brewer, E. Paulos, and R. White. N-smarts. In *Proceedings of the second ACM SIGCOMM workshop on Networked systems for developing regions - NSDR '08*, page 25, New York, New York, USA, aug 2008. ACM Press. ISBN 9781605581804. doi: 10.1145/1397705.1397713. URL <http://dl.acm.org/citation.cfm?id=1397705.1397713>.
- K. Huang, G. Zhuang, Q. Wang, J. S. Fu, Y. Lin, T. Liu, L. Han, and C. Deng. Extreme haze pollution in Beijing during January 2013: chemical characteristics, formation mechanism and role of fog processing. *Atmospheric Chemistry and Physics Discussions*, 14(6):7517–7556, mar 2014. ISSN 1680-7375. doi: 10.5194/acpd-14-7517-2014. URL <http://www.atmos-chem-phys-discuss.net/14/7517/2014/acpd-14-7517-2014.html>.
- N. P. Hyslop. Impaired visibility: the air pollution people see. *Atmospheric Environment*, 43(1):182–195, jan 2009. ISSN 13522310. doi: 10.1016/j.atmosenv.2008.09.067. URL <http://www.sciencedirect.com/science/article/pii/S1352231008009217>.
- G. Ibarra-Berastegi, A. Elias, A. Barona, J. Saenz, A. Ezcurra, and J. Diaz de Argandoña. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. *Environmental Modelling & Software*, 23(5):622–637, may 2008. ISSN 13648152. doi: 10.1016/j.envsoft.2007.09.003. URL <http://www.sciencedirect.com/science/article/pii/S1364815207001740>.
- B. Jacquemin, J. Sunyer, B. Forsberg, T. Götschi, L. Bayer-Oglesby, U. Ackermann-Liebrich, R. de Marco, J. Heinrich, D. Jarvis, K. Torén, and N. Künzli. Annoyance due to air pollution in Europe. *International Journal of Epidemiology*, 36(4):809–820, 2007. ISSN 03005771. doi: 10.1093/ije/dym042.
- A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagan, and B. T. Dai, editors. *Social Informatics*, volume 8238 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2013. ISBN 978-3-319-03259-7. doi: 10.1007/978-3-319-03260-3. URL <http://link.springer.com/10.1007/978-3-319-03260-3>.
- R. Jayakrishnan, H. S. Mahmassani, and U. Rathi. User-Friendly Simulation Model for Traffic Networks with ATIS/ATMS. In *Computing in Civil and Building Engineering (1993)*, pages 833–840. ASCE, 1993. URL <http://cedb.asce.org/cgi/WWWdisplay.cgi?82191>.



- H. Kashani and G. Saridis. Intelligent control for urban traffic systems. *Automatica*, 19(2):191–197, mar 1983. ISSN 00051098. doi: 10.1016/0005-1098(83)90091-2. URL <http://www.sciencedirect.com/science/article/pii/0005109883900912>.
- S. M. Kaufman. How Social Media Moves New York:. (1), 2012. doi: <http://wagner.nyu.edu/rudincenter/publications/how-social-media-moves-new-york.pdf>. URL <http://wagner.nyu.edu/rudincenter/>.
- Q.-J. Kong, Q. Zhao, C. Wei, and Y. Liu. Efficient Traffic State Estimation for Large-Scale Urban Road Networks. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):398–407, mar 2013. ISSN 1524-9050. doi: 10.1109/TITS.2012.2218237. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6353219>.
- N. Künzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filliger, M. Herry, F. Horak, V. Puybonnieux-Texier, P. Quénel, J. Schneider, R. Seethaler, J. C. Vergnaud, and H. Sommer. Public-health impact of outdoor and traffic-related air pollution: a European assessment. *Lancet*, 356(9232):795–801, sep 2000. ISSN 0140-6736. doi: 10.1016/S0140-6736(00)02653-2. URL <http://www.sciencedirect.com/science/article/pii/S0140673600026532>.
- I. Leontiadis, G. Marfia, D. Mack, C. Mascolo, G. Pau, and M. Gerla. An Opportunistic Traffic Management System for Vehicular Networks. *Performance Evaluation*, 12(4):1537–1548, 2011.
- M. E. Lindholm and M. Blinge. Assessing knowledge and awareness of the sustainable urban freight transport among Swedish local authority policy planners. *Transport Policy*, 32:124–131, 2014. ISSN 0967070X. doi: 10.1016/j.tranpol.2014.01.004. URL <http://dx.doi.org/10.1016/j.tranpol.2014.01.004>.
- Y. Ma, M. Richards, M. Ghanem, Y. Guo, and J. Hassard. Air Pollution Monitoring and Mining Based on Sensor Grid in London, jun 2008. URL <http://www.mdpi.com/1424-8220/8/6/3601/htm>.
- H. S. Mahmassani. Impact of information on traveler decision. In *Transportation Research Board 90th annual meeting*, Washington, DC, 2011.
- H. S. Mahmassani, S. Peeta, T.-Y. Hu, and A. Ziliaskopoulos. DYNAMIC TRAFFIC ASSIGNMENT WITH MULTIPLE USER CLASSES FOR REAL-TIME ATIS/ATMS APPLICATIONS. In *Large Urban Systems. Proceedings of the Advanced Traffic Management Conference*, 1993. URL <http://trid.trb.org/view.aspx?id=406594>.
- G. Marfia and M. Rocchetti. Vehicular congestion detection and short-term forecasting: A new model with results. *IEEE Transactions on Vehicular Technology*, 60(7):2936–2948, 2011. ISSN 00189545. doi: 10.1109/TVT.2011.2158866.
- A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*, pages 41–48, 1998. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.9324>.

- H. J. Miller. Beyond sharing: Cultivating cooperative transportation systems through geographic information science. *Journal of Transport Geography*, 31:296–308, 2013. ISSN 09666923. doi: 10.1016/j.jtrangeo.2013.04.007.
- C. E. Palazzi, M. Rocchetti, and G. Marfia. Realizing the unexploited potential of games on serious challenges. *Computers in Entertainment*, 8(4):1–4, dec 2010. ISSN 15443574. doi: 10.1145/1921141.1921143. URL [http://dl.acm.org/ft\\_gateway.cfm?id=1921143&type=html](http://dl.acm.org/ft_gateway.cfm?id=1921143&type=html).
- M. Papageorgiou. Multilayer control system design applied to freeway traffic. *IEEE Transactions on Automatic Control*, 29(6):482–490, jun 1984. ISSN 0018-9286. doi: 10.1109/TAC.1984.1103573. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1103573>.
- A. Peters, S. von Klot, M. Heier, I. Trentinaglia, A. Hörmann, H. E. Wichmann, and H. Löwel. Exposure to traffic and the onset of myocardial infarction. *The New England journal of medicine*, 351(17):1721–1730, 2004. ISSN 0028-4793. doi: 10.1056/NEJMoa040203.
- B. Ran and D. Boyce. *Dynamic Urban Transportation Network Models: Theory and Implications for Intelligent Vehicle-Highway Systems*. Springer Science & Business Media, 2012. ISBN 3662007738. URL <https://books.google.com/books?hl=en&lr=&id=VLb0CAAAQBAJ&pgis=1>.
- T. B. Reed and R. M. Lerner. Methanol: A Versatile Fuel for Immediate Use: Methanol can be made from gas, coal, or wood. It is stored and used in existing equipment. *Science*, 182(4119):1299–1304, dec 1973. ISSN 0036-8075. doi: 10.1126/science.182.4119.1299. URL <http://www.sciencemag.org/content/182/4119/1299.short>.
- L. Rotaris, R. Danielis, E. Marcucci, and J. Massiani. The urban road pricing scheme to curb pollution in Milan, Italy: Description, impacts and preliminary costbenefit analysis assessment. *Transportation Research Part A: Policy and Practice*, 44(5):359–375, jun 2010. ISSN 09658564. doi: 10.1016/j.tra.2010.03.008. URL <http://www.sciencedirect.com/science/article/pii/S0965856410000479>.
- J. J. Rotemberg. The efficiency of equilibrium traffic flows. *Journal of Public Economics*, 26(2):191–205, mar 1985. ISSN 00472727. doi: 10.1016/0047-2727(85)90004-0. URL <http://www.sciencedirect.com/science/article/pii/0047272785900040>.
- W. Sha, D. Kwak, B. Nath, and L. Iftode. Social vehicle navigation: integrating shared driving experience into vehicle navigation. *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, pages 1–6, 2013. doi: 10.1145/2444776.2444798.
- M. Srivastava, T. Abdelzaher, and B. Szymanski. Human-centric sensing. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1958):176–97, jan 2012. ISSN 1364-503X. doi: 10.1098/rsta.2011.0244. URL <http://rsta.royalsocietypublishing.org/content/370/1958/176.short>.
- Y. Sun, G. Zhuang, A. Tang, Y. Wang, and Z. An. Chemical Characteristics of PM 2.5 and PM 10 in HazeFog Episodes in Beijing. *Environmental Science & Technology*, 40(10):3148–3155, may 2006. ISSN 0013-936X. doi: 10.1021/

es051533g. URL <http://dx.doi.org/10.1021/es051533g>.

U. Tang and Z. Wang. Influences of urban forms on traffic-induced noise and air pollution: Results from a modelling system. *Environmental Modelling & Software*, 22(12):1750–1764, dec 2007. ISSN 13648152. doi: 10.1016/j.envsoft.2007.02.003. URL <http://www.sciencedirect.com/science/article/pii/S136481520700028X>.

E. TANIGUCHI, R. G. Thompson, T. YAMADA, and R. van Duin. CITY LOGISTICS. NETWORK MODELLING AND INTELLIGENT TRANSPORT SYSTEMS. 2001. URL <http://trid.trb.org/view.aspx?id=673352>.

M. Thatcher and P. Hurley. A customisable downscaling approach for local-scale meteorological and air pollution forecasting: Performance evaluation for a year of urban meteorological forecasts. *Environmental Modelling & Software*, 25(1):82–92, jan 2010. ISSN 13648152. doi: 10.1016/j.envsoft.2009.07.014. URL <http://www.sciencedirect.com/science/article/pii/S1364815209001765>.

C. L. Townsend. Effects on health of prolonged exposure to low concentrations of carbon monoxide. *Occupational and Environmental Medicine*, 59(10):708–711, oct 2002. ISSN 13510711. doi: 10.1136/oem.59.10.708. URL <http://oem.bmj.com/content/59/10/708.short>.

F.-Y. Wang. Parallel Control and Management for Intelligent Transportation Systems: Concepts, Architectures, and Applications. *IEEE Transactions on Intelligent Transportation Systems*, 11(3):630–638, sep 2010. ISSN 1524-9050. doi: 10.1109/TITS.2010.2060218. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5549912>.

Y. Wang, G. Zhuang, Y. Sun, and Z. An. The variation of characteristics and formation mechanisms of aerosols in dust, haze, and clear days in Beijing. *Atmospheric Environment*, 40(34):6579–6591, nov 2006. ISSN 13522310. doi: 10.1016/j.atmosenv.2006.05.066. URL <http://www.sciencedirect.com/science/article/pii/S1352231006005437>.

R. J. WEILAND and L. B. Purser. INTELLIGENT TRANSPORTATION SYSTEMS. *Transportation in the New Millennium*, 2000. URL <http://trid.trb.org/view.aspx?id=639268>.

J. Wootton, a. García-Ortiz, and S. Amin. Intelligent transportation systems: A global perspective. *Mathematical and Computer Modelling*, 22(4-7):259–268, 1995. ISSN 08957177. doi: 10.1016/0895-7177(95)00137-Q.

Z. Zhang, J. Wang, L. Chen, X. Chen, G. Sun, N. Zhong, H. Kan, and W. Lu. Impact of haze and air pollution-related hazards on hospital admissions in Guangzhou, China. *Environmental science and pollution research international*, 21(6):4236–44, mar 2014. ISSN 1614-7499. doi: 10.1007/s11356-013-2374-6. URL <http://link.springer.com/10.1007/s11356-013-2374-6>.

Y. Zhao. Mobile phone location determination and its impact on intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 1(1):55–64, mar 2000. ISSN 15249050. doi: 10.1109/6979.869021. URL <http://ieeexplore.ieee.org/articleDetails.jsp?>

arnumber=869021.

- B. Zhou, J. Cao, X. Zeng, and H. Wu. Adaptive Traffic Light Control in Wireless Sensor Network-Based Intelligent Transportation System. In *2010 IEEE 72nd Vehicular Technology Conference - Fall*, pages 1–5. IEEE, sep 2010. ISBN 978-1-4244-3573-9. doi: 10.1109/VETECF.2010.5594435. URL <http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5594435>.