Using geosocial search for urban air pollution monitoring

Using Geosocial Search for Urban Air Pollution Monitoring

Matteo Sammarco[a], Rita Tse[b], Giovanni Pau[a,*], Gustavo Marfia[c,*]

[a] *LIP6, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France*
[b] *Macao Polytechnic Institute, Rua de Luís Gonzaga Gomes, Macao*
[c] *Dipartimento di Scienze per la Qualità della Vita, Università di Bologna, 237 Corso D'Augusto, Rimini, Italy*

Abstract

While Twitter and other Online Social Networks (OSNs) or microblogs are considered as a source of information for breaking news or uproarious and unexpected events, they could also be exploited as a dense worldwide sensors network for physical measurements. The *corpus* of geotagged posts from OSNs includes people's feedbacks about a wide range of topics, with precise temporal and geographical metadata, that can be used as a support or an improvement to hardware sensors. For instance, if collocated people, independently and at the same time, write posts complaining about high temperatures, it could effectively denote a raise of heat in that place. In this paper, we explore the feasibility to use a geographical search on social networks, that is, a geosocial search, about air pollution related posts, as effective air impureness measurements. We evaluate our assumption in large cities over three continents of the planet, where a minimum increment about the number of air pollution related posts in a area, indeed corresponds to a raise of minimum pollution values in such area. Such a correlation can be exploited to integrate and extend existing air pollution monitoring networks. At the end of the manuscript we propose to further employ the time series of posts returned by the geosocial search to predict next pollution values.

Keywords: Air Pollution, Online Social Networks, Text Analysis, Pollution Forecast

1. Introduction

Photography in Interstellar, Christopher Nolan's last movie, well depicts the current landscape of some of the most populated and industrialized places in the world. During air pollution peaks, the usual blue sky becomes covered by a thick reddish coat due to the massive presence of pollutants, especially sulfur dioxide, particulate matter, and nitrogen oxides (Figure 1).

By now, this environmental issue is so diffused and evident that countries from all around the world are active to propose and ratify treaties such as the Gothenburg and Kyoto protocols. Such concern dates back in time, as over 40 years ago, in 1974, the United Nation Environment Programme (UNEP) and the World Health Organization (WHO) started the Global Environment Monitoring System for air pollution (GEMS/Air), which deployed air monitoring equipment in over 50 large cities throughout the world [1]. Mega-cities are indeed one the most affected places for the air pollution plague [2], mainly due to prolonged exposures to high road traffic levels and factory emissions.

Nowadays, although institutional agencies and independent entities have improved the monitoring networks in large cities with the deployment of more sensing stations, they are usually not sufficient. For instance, MonitorAr-Rio, the agency in charge of monitoring air quality in Rio de Janeiro, comprises only nine fixed stations covering a territory of 1,255 squared kilometers [3].

*Corresponding author
Email addresses: giovanni.pau@lip6.fr (Giovanni Pau), gustavo.marfia@unibo.it (Gustavo Marfia)

Figure 1: Paris landscape during an air pollution peak (left side) and during a day with under-threshold values (right side). The massive presence of pollutants in the air heavily affects the view.

Although the cost of pollution sensing platforms is progressively decreasing, a pervasive and widespread deployment of such technologies is still far away in time for its costs. Both satellite and terrestrial solutions are being studied for a fine grained pollution estimation, with different results in terms of sensitivity, resolution and accuracy [4, 5, 6]. However, the common denominator 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 automatic system has, instead, so far, employed feedbacks received from the public.

In this work we explore the possibility of integrating air pollution sensing networks with measurements based on air pollution related posts, spontaneously generated by users on Online Social Networks (OSNs), through geosocial search. A geosocial search is the operation of finding events and user activities advertised on OSNs, in a specific geographical area [7, 8]. OSNs include a huge *corpus* of geotagged posts related to various topics, which may be exploited to discover some knowledge pertaining a specific area. For example, if many co-located posts, written by different users, complain about car noise coming from the street, it is likely that a traffic jam is currently happening in that specific area. Similarly, if a geosocial search concerning high temperatures reveals many protests in a defined area, that place may be far from green spaces, thus more people suffer from heat. Also, geosocial searching for air allergies we may infer which are the most exposed places to air pollution or to spring pollen.

Resorting to geosocial search for urban air pollution monitoring brings several advantages. Mainly, in high density urban territories, post generation and upload is dense. Integrating information gathered from OSNs within traditional monitoring systems, may provide spatially more extensive and finer grained information. Such information could also be integrated in intelligent transport systems (ITS) or in in-car connected systems, as high pollution levels could indicate traffic congestion or particular weather conditions. Thus, traffic, for example, could be rerouted accordingly. In addition, people could decide on the sport where it is best to jog or to have a bike ride. In essence, city administrations and citizens at large could take advantage from the geosocial search for many different urban applications.

Apart from geographic coordinates, posts include sharp temporal information too, while many traditional monitoring networks, like MonitorAr-Rio, for example, only provide pollutant level values on a day-by-day granularity. Finally, gathering public posts from OSNs does not require particular costs as it totally avoids the installation, deployment ad maintenance costs that are typical of traditional stations.

Nevertheless, using geosocial search for air pollution monitoring is challenging. Socio-cultural differences, which are substantial when changing language and location, must be accounted for in order to implement an effective search application [9]. People, in addition, very often adopt local expressions or slang to express their feelings in short messages. Thus, selecting a relevant set, from the massive stream of all posts, is, indeed, a very hard task. However, even when a relevant set were found, it may not be sufficient to detect pollution, as it may be possible that different people react in different ways, as they exhibit different tolerance thresholds to air pollution. Moreover, information on OSNs could be biased or untrue [10, 11].

The contribution of this work is the design and implementation of a geosocial keyword-based search system to find where people complain most about bad air quality. The proposed system takes for geocultural

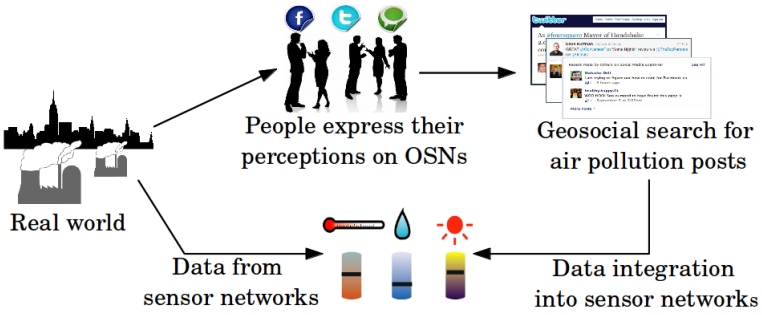


Figure 2: System proposal.

differences with specialized dictionaries of keywords. When the correlation between the time series of pollutant measurements and related posts is proved, we also provide an algorithm which exploits OSN posts to forecast the most likely pollution level in a given area.

The system that is here proposed is depicted in Figure 2. As people, independently from each other and from pollution sensors data, express their protests concerning bad air quality on OSNs, their air pollution related posts are selected from the unstructured mass of posts. Filtering is done either *a priori* resorting to a keyword-based searching scheme, or *a posteriori* through a post classification mechanism based on machine learning [12] and natural language processing (NLP) [13]. When many air pollution related posts are found in a given area, this likely indicates an actual raise of pollution in that area. If such information comes from an area where an existing correlation between pollutant levels and OSN posts has been previously verified, it can then be integrated with other pollution sensing network data, whenever available.

In a nutshell, we verify the feasibility of the proposed system as follows. We first show that air pollution related posts, with or without a per-user credibility assessment, can be put to good use as real air quality measurements. Such fact is corroborated by the large scale measurements that have been conducted in mega-cities over three continents. Secondly, we show that is it possible to leverage the posts time series to forecast pollution levels in the near future and give a measure of air pollution a under given probability guarantees.

The present manuscript is structured as follows. Section 2 introduces the proposed system and the dictionary constructed for an effective geosocial search. Section 3 describes the dataset including OSNs posts and air pollution measured values. Section 4 shows the feasibility evaluation of our proposal for the considered cities, while Section 5 extends the work with an original pollution level forecasting approach. Finally we present related works in Section 7 and our conclusions, together with future works and applications, in Section 9.

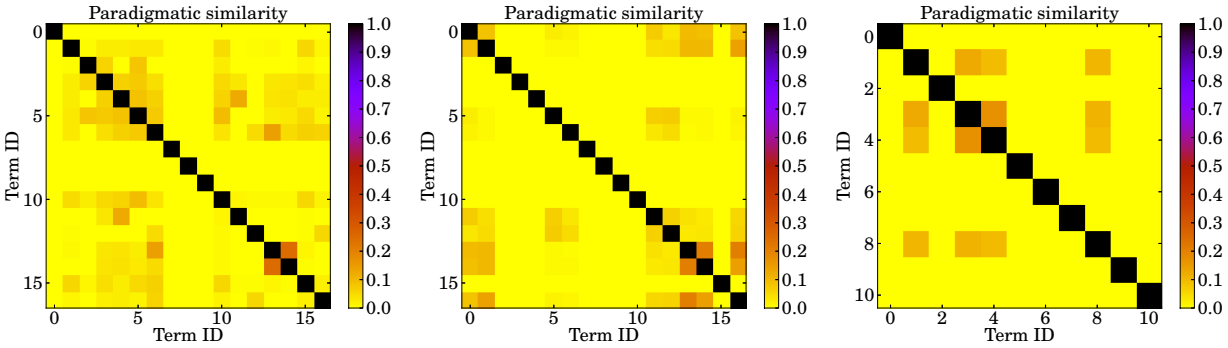
2. Rationale and system description

Every second, on average, 6 thousand tweets are posted in the world, which corresponds to over 360 thousand tweets sent per minute and 500 millions tweets per day [14]. Sina Weibo, Twitter's counterpart in China, presents values on the same scale, as other popular social networks and microblogs such as Instagram, Facebook, and Google+. Usually posts are associated to photos and they come with a large set of metadata including location, and user identifier. This feature makes posts gain value and interest to find places corresponding to certain terms. This data is thus exploitable for geosocial search.

Apart from geographical coordinates, the second important dimension to describe a territory is time. Providing instantaneous results is absolutely crucial to inform in real-time what is happening somewhere. For this reason we preliminary filter the stream of posts submitted to social networks in a given area, selecting only posts containing keywords contained in a previously-defined dictionary. In this way, we drastically

Table 1: Pollution-related terms dictionary for three languages.

Category	English translation for French terms (term ID)	English translation for Portuguese terms (term ID)	English translation for Chinese terms (term ID)
Pollution	bad air quality (0), pollution (1), pollutant (2), polluted (3), particulate (4), ozone (5), dioxin (6)	atmosphere (0), pollution (1), pollutant (2), carbon monoxide (3), nitrogen dioxide (4), ozone (5), dioxide (6), particulate (7)	pollution (0), poor air (1), gray sky (2), air quality (3)
Weather	fog (7), haze (8), gray sky (9)	air quality (8), greenhouse effect (9), ozone layer (10)	haze (4), fog (5), gray sky (6), bad weather (7)
Traffic	traffic jam (10), congestion (11)	traffic jam (11), ethanol (12)	traffic jam (8), congestion (9)
Health	asthma (12), conjunctivitis (13), rhinitis (14), breath (15), stifling (16)	asthma (13), bronchitis (14), stifling (15), pneumonia (16)	unable to breath (10)



(a) French (b) Portuguese (c) Chinese

Figure 3: Paradigmatic relations among words in the French, Portuguese and Chinese dictionary.

reduce the set of posts we work on, avoiding the need of *a posteriori* classification, using machine learning techniques [12], clustering algorithms, likelihood models [15] and the Mechanical Turk service [16]. Time constraints are even harder if the collected information must be instantly available for applications such as multi-modal transport systems.

2.1. Dictionary construction

Using geosocial search to identify worldwide famous places and entities is quite straightforward when they explicitly appear in the subject of the post and the geographic place is well known. For example, a post generated close to coordinates 48.8582° Nord, 2.2945° East and containing the term "Eiffel Tower", clearly refers to the French capital's symbol. More cumbersome is to identify posts referring to abstract or intangible entities.

Air pollution, for instance, could be detected by different people in different ways (with the sight or the smell sense), during special occasions (e.g., while jogging or bike riding in the city center) or could implicitly appear as the consequence of given health symptoms, but without directly referring to the air pollution cause.

Moreover, geosocial search is highly dependent on cultural differences and term usage from country to country, region to region [9]. For this reason, in order to build our dictionaries of air pollution related terms, we took into account the specificities of each country. For example, for the Brazilian dictionary we adopted

the word "ethanol" since that is the most common fuel for vehicles from that country. Also within the same country, considerable cultural differences are possible: people could prefer slang or local expressions to share a concept.

For all these reasons we created the dictionaries shown in Table 1 following the next three steps:

1. From an initial dataset of unlabeled posts we identify some categories of topics and select the most appropriate and related to air pollution (e.g., breathing health problems, weather, etc.). Identification is easily conducted with a supervised machine learning.
2. For each category, we detect a bag of most prevalent words. TF or TD-IDF metrics can be used at this step. We discovered some invariant terms (i.e., words frequently present in every language-related dictionary) like "pollution" or "traffic jam". Some other terms are added to account territory specificities.
3. Having a large bag of keywords clearly makes raise the amount of posts retrieved. At the same time it increases noise in the dataset with posts unrelated to air pollution or duplicate. Thus, we try to keep the dictionary dimension to a small size. We also want that terms found so far belong to different topics. Computing the paradigmatic similarity we found out that for our dictionaries at most 20 terms satisfied our conditions from the initial bag of terms. It is to note that this cut-off value is strongly related to the set of terms employed at the beginning of the process. Having a small bag words sharing a low paradigmatic similarity we cover a large set of discussions and topics with a minimal set of terms.

Dictionary construction is laborious and we need a special dictionary per geographical place, making the system not immediately scalable worldwide. On the other hand, having a customized dictionary leads to a better match between effects and causes (air pollution in our case). In addition, a geosocial search, for its nature, implies the use of specific social aspects in a geographical area. We also point out that the dictionary construction process is done only once at the beginning.

Categories found at the first step are the following:

- **Pollution.** Terms strictly related to pollution and air quality.
- **Weather.** Terms strictly related to those weather conditions that are highly affected by air pollution (e.g., "haze").
- **Traffic.** This set of words includes all those terms which reveal high levels of traffic, as vehicular emissions amount to one of the major causes for air pollution.
- **Health.** This category includes symptoms and effects of a prolonged air pollution exposure.

Clearly, depending on the country and language, terms in each category change as explained in the second step.

Terms used to create the dictionary in this manuscript have also been selected to have the minimum paradigmatic similarity among them. The paradigmatic relation is used in text mining to find semantic relations between terms in a *corpus* of documents [17]. Two terms, w_1 and w_2 have an high paradigmatic similarity if they can be naturally substituted one with the other, like "car" and "automobile", without altering the meaning of a phrase. On the contrary, a low paradigmatic value is observed when two words belong to different contexts. For the purpose of this work, the utilization of a dictionary of terms which share a low paradigmatic similarity amounts to cover a large set of discussions and topics with a minimal set of terms.

Let us consider the *corpus* of OSNs posts $\mathcal{C} = \{p_1, p_2, ..., p_n\}$ and each post as a set of terms $p_i = \{w_1, w_2, ..., w_i\}$. We model posts as a directed graph $G(\mathcal{V}, \mathcal{E})$ where each distinct term w_j is represented by a node and a directed edge (w_i, w_j) exists if term w_i appears before w_j in a post p_i.

We compute the paradigmatic similarity as follows:

$$ParSim(w_1, w_2) = \frac{Sim(Left(w_1), Left(w_2)) + Sim(Right(w_1), Right(w_2))}{2}, \quad (1)$$

where

$$Left(w_i) = \{w \in \mathcal{V} : \exists w \rightsquigarrow w_i \in \mathcal{E}\}, \tag{2}$$

and

$$Right(w_i) = \{w \in \mathcal{V} : \exists w_i \rightsquigarrow w \in \mathcal{E}\}. \tag{3}$$

$Sim(\cdot)$ amounts to the Jaccard similarity function between sets: $Sim(A, B) = \frac{A \cap B}{A \cup B}$ returns a value of 0 for completely unrelated terms, a value of 1 when two terms can be naturally substituted one with the other. Figures 3 show the paradigmatic relations between the terms that have been selected for all the dictionaries used in this manuscript. In general, values are very low for each pair of terms.

2.2. Integration with traditional monitoring systems

One of the goals of the proposed system is the integration with traditional air pollution monitoring systems, that is systems employing fixed stations deployed in a urban context. Despite the sophistications of such stations and the capability to monitor many pollutants, they have same drawbacks like purchase, installation, and handling costs. Thus, the proposed system can provide an extended and denser coverage on the territory.

Due to the considerations made in Section 2.1, we foreseen the construction of a dictionary of terms for each city. Although the use of a dictionary filters out a large amount of posts, we do not make any assumption on the amount of posts retrieved per minute which might be massive in large cities in presence of pollution peaks. For this reason we recommend to use a distributed system able to analyze streams od data in near real-time like Samza or Spark Streaming.

Given a tessellation of the city under monitoring, a sampling time window and a relative threshold as number of detected posts, an air pollution warning event in a cell is triggered by the excess of number of air pollution related posts for more than n time windows. The responsible pollutant might be identified by the effect described on the posts or by pollutant levels measured by the closest fixed traditional stations.

All the data related to posts is stored and iteratively analyzed to refine predictions as described in Section 5.

Traditional monitoring system and the system proposed in this work are complementary. Fixed stations are in turn used to supervise the events triggered by the proposed system, and to control users' credibility like a base-truth.

3. Air pollution measures and posts collection

Our proposal is based on geosocial search, which must take into account socio-cultural differences from place to place. Thus we want to apply our approach on considerably far and different places in the world. Finding correlations between real air pollution levels and OSNs pollution related posts, implies the presence of an air pollution monitoring network already deployed on the territory to use as a base-truth. In addition, places under monitoring must constantly or sporadically experience air pollution peaks phenomena. Usually they are the most crowded and economically advanced cities in the world. Emissions from industrial activities, biomass combustions, and, above all, vehicles emissions, are the main causes for air pollution due to the introduction of aerosols into the atmosphere.

Finally, we need a certain amount of OSNs posts to generate a continuous correlation with air pollution values. Posts must be generated and geotagged as close as possible to monitoring stations. Therefore, places under monitor must be cities with a dense presence of OSNs user-generated posts too.

3.1. Air pollution measurements

For the reasons introduced in this Section, we apply and evaluate our approach on some major cities over three continents:

- **Europe**. Paris, France, and the Grand Paris agglomeration area.
- **South America**. São Paulo, Brazil and the São Paulo agglomeration area (MASP).

Table 2: Paris AirParif air pollution monitoring stations used in this manuscript and relative measurement capabilities.

	Station						
	S_1	S_2	S_3	S_4	S_5	S_6	S_7
Name	Central	Gennevilliers	Ivry-sur-Seine	Neuilly-sur-Seine	Place Victor Basch	Rue Bonaparte	Aubervilliers
Latitude	48.856614	48.929810	48.817917	48.880746	48.832781	48.856340	48.903686
Longitude	2.352222	2.294333	2.394060	2.277453	2.330581	2.334031	2.384688
$PM_{2.5}$	✓	✓					
PM_{10}	✓	✓			✓		
NO_2	✓	✓	✓	✓	✓	✓	✓

Table 3: São Paulo CETESB air pollution monitoring stations used in this manuscript and relative measurement capabilities.

	Station											
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
Name	Osasco	Capão Redondo	S.André Capuava	Mauá	Parelheiros	Congonhas	Guarulhos	Santos-Ponta da Praia	Sorocaba	Parque D.Pedro II	S.Bernardo Paulicéia	Taboão da Serra
Latitude	-23.525838	-23.666545	-23.645681	-23.669424	-23.776595	-23.616040	-23.456214	-23.980514	-23.502606	-23.544763	-23.670962	-23.609054
Longitude	-46.801009	-46.783021	-46.493339	-46.465158	-46.696296	-46.663296	-46.518289	-46.300167	-47.477611	-46.631060	-46.585041	-46.755939
$PM_{2.5}$					✓	✓		✓				
PM_{10}	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NO_2	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓

- **Asia**. Major Chinese cities: Beijing, Guangzhou Shanghai, Chengdu, and Hong Kong.

These cities are equipped with a more or less developed network of air pollution monitoring stations. Each sensing station is configured or is able to sense one or more pollutants. In this manuscript we rely our model analysis mainly on the following pollutants:

- Respirable Suspended Particles ($PM_{2.5}$), particulate matter with a diameter of 10 μm or less. They contribute to haze in urban contexts.
- Fine Suspended Particles (PM_{10}), particulate matter with a diameter of 2.5 μm or less.
- Nitrogen Dioxide (NO_2), a reddish-brownish gas with a pungent odor. It is an important component of city smog.

Paris. With more than two millions habitants and about seven thousands private vehicles, the French capital is one of the biggest European cities which sometimes experiences air pollution peaks. The most struck period is the spring season with high concentration of fine particulate matter as shown in Figure 1.

Institutional countermeasures include vehicular traffic restriction and free public transports during peaks, and the progressive reduction of the most pollutant vehicles, diesel engine equipped cars above all.

Air quality monitoring is France is guaranteed by independent associations. AirParif is one of them [18] maintaining a network of 70 stations in the whole Paris region.

Table 2 indicates geographical coordinates of stations employed as base-truth of pollution measurements, as long as their capability to monitor $PM_{2.5}$, PM_{10} or NO_2 concentrations. Three of the stations are placed in the city center and the last four are placed around the neighborhoods. All of them are able to monitor NO_2 levels, while only stations S_1 and S_2 can measure $PM_{2.5}$, and PM_{10} levels too.

Beijing, Guangzhou, Shanghai, Chengdu, and Hong Kong. These Chinese cities have impressive numbers both as industrial development and population density and also as bad air quality conditions all along the year. Even if belonging to the same country, these cities are very far from each other and present different morphological features (presence of sea, altitude, etc.).

We rely on air quality measurements taken by U.S. Consulates or Embassies in the considered cities, one per city [19]. All of them are able to measure $PM_{2.5}$ concentrations. Pollution values for Hong Kong are taken from the relative Environmental Protection Department [20].

São Paulo. The Metropolitan Area of São Paulo (MASP) is the most industrialized area in Brazil, including 19 million people and seven million vehicles. São Paulo is located at 800 meters above sea level and fine particulate matter concentrations are inversely correlated with precipitations [21].

Table 4: Chinese air pollution measurement stations.

	Station				
	S_1	S_2	S_3	S_4	S_5
Name	Beijing	Guangzhou	Shanghai	Chengdu	Hong Kong
Latitude	39.939690	23.119406	31.208801	30.572269	22.285167
Longitude	116.456041	113.321231	121.447220	104.066541	114.156833
$PM_{2.5}$	✓	✓	✓	✓	✓
PM_{10}					
NO_2					

(a) Paris (b) MASP

Figure 4: Tweets heatmap and stations location for Paris and MASP.

The other side of the coin of São Paulo development and industrialization regards various environmental problems associated with the growth of its population in urban areas. Air quality is a major concern, because the reported concentrations of certain regulated pollutants, typically ozone and fine particulate, have often exceeded Brazilian national standards.

Air quality in MASP is monitored by the CETESB company (standing for Companhia Ambiental do Estado de São Paulo) [22] counting a network of 63 stations. Monitoring a larger area, we employ more stations than Paris. Table 3 indicates stations names, coordinates, and pollutant measurement capabilities. All the stations are able to measure PM_{10} levels, most of them NO_2, while only few of them can measure $PM_{2.5}$ levels too.

3.2. Air pollution posts

We have conducted a geosocial search on Paris and São Paulo during the time period coming from March 5^{th} until April 30^{th}. These two months are a critical period for both cities from the air pollution point of view. In the last few years (2014 and 2015), Paris has experienced high pollution peaks during the spring period (Figure 1). In São Paulo, humidity, wind speed and precipitation values are lower during winter months (from April to September), resulting increasing $PM_{2.5}$ concentrations [23].

The geosocial search through keywords reported in Table 1, allows to filter the huge mole of user generated posts upstream. We further filter the posts dataset, considering only posts generated between 6 a.m. and 12 p.m. and geolocalized within a certain distance from the city center. Different kinds of locations are available in social network posts. We consider only latitude/longitude coordinates coming from GPS enable mobile devices. We have included the time constraint to restrict the analysis when cities are fully industrious.

For Paris the covering circle has a radius of 8 Km. Such a distance includes the whole city of Paris, plus the dual-carriageway ring main road around and the suburbs that surround the French capital. This area is also known as Grand Paris. For São Paulo, we use a much larger radius, so to include the large MASP area.

Dataset is composed only by genuine user generated posts. No retweets are admitted nor Twitterbot's posts. Almost the totality of users have published no more than 4 posts as shown in Figure 5. We manually checked that the rest of the dataset was not automatically generated. Following this rationale we have found among Sina Weibo posts two users who have published 384 and 257 posts respectively. Being automatic generated posts, they have been excluded from the analysis.

Figures 4(a) and 4(b) show the distribution of posts over Paris and São Paulo respectively. In the former case, posts are mainly concentrated in three areas: the city center and the two main train stations. These points are covered by AirParif stations S_1, S_5, and S_6. The ring highway around Paris as long as some suburbs are covered by the other four stations.

São Paulo is more densely populated than Paris. Three areas present a high concentration of posts: the industrialized littoral zone covered by station S_8, the western Sorocaba's industrial park covered by station S_9, and all the São Paulo city area is jeopardized and covered by the rest of 10 CETESB stations.

Since air pollution in Chinese cities is present for longer periods, we collect posts for a longer period too, starting from October 3^{rd} 2012 till April 3^{rd} 2014. The geosocial search is set to return Sina Weibo posts generated within a radius of 5 Km from the sensing stations. At the end of the filtering pipeline we got about 19,000 posts.

4. Feasibility evaluation

In this Section we evaluate the feasibility of using a geosocial search to monitor air pollution levels in an urban area.

Differently from standard sensors, which periodically or continuously monitor physical quantities, using crowdsourcing measurements based on published OSNs posts involve many complications and uncertainties. Primarily, people are mainly binary sensors having different thresholds affecting their claims credibility and they express their perception occasionally. Then, people experiencing difficulties to breath due to air pollution, could differ their notification for many reasons: they could want to immediately find a safer place, they could not be online or they could be doing something more impelling. Thus, a delay is expected between the event occurrence and the posting time, typically following an homogeneous Poisson process[24].

The finest temporal granularity given by air pollution measurement networks is one hour. We adopt a two hour time window sampling, with averaged pollution values, and we resample the time series of number of generated posts accordingly, summing the amount of post published in that time window.

Thus, for each pollution-sensed sample, we count the number of posts published, result of the geosocial search, in the following time window and within a radius of 5 Km away from the sensing station. We group the amount of posts in four sets:

- 0, the result of the geosocial search is empty in that time window,

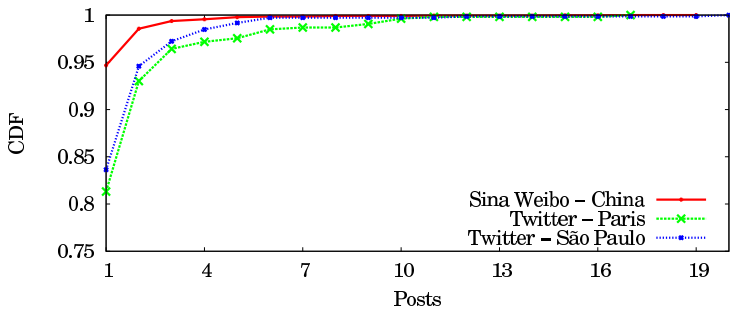
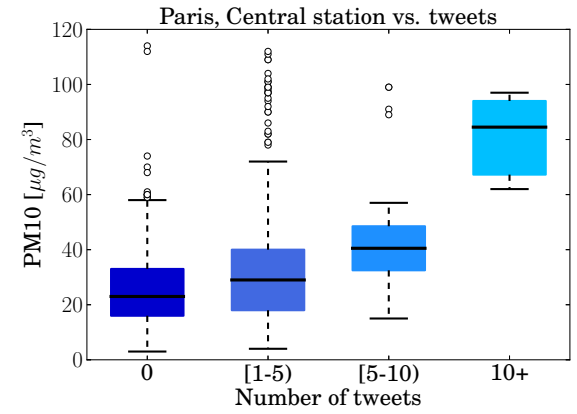
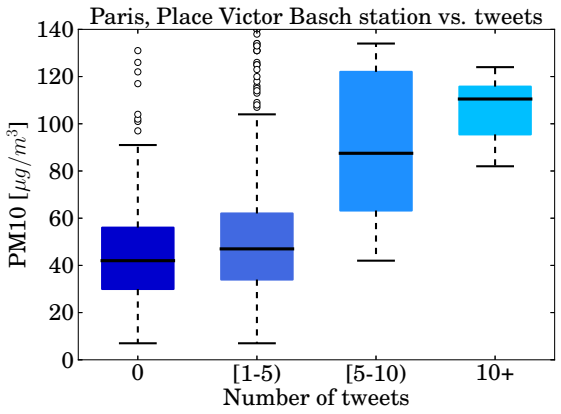


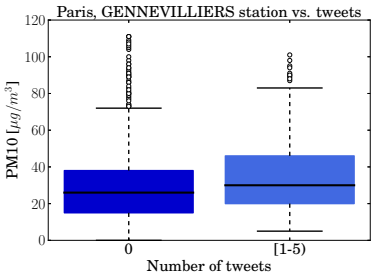
Figure 5: CDF of posts per users.



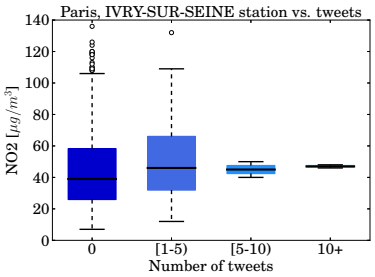
(a) S_1, PM_{10}



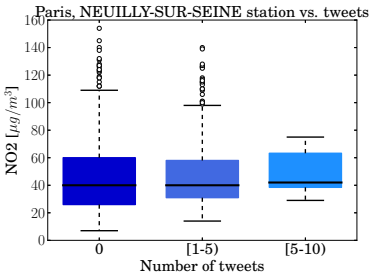
(b) S_5, PM_{10}



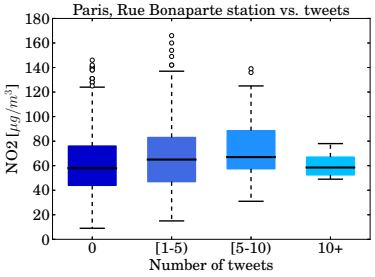
(c) S_2, PM_{10}



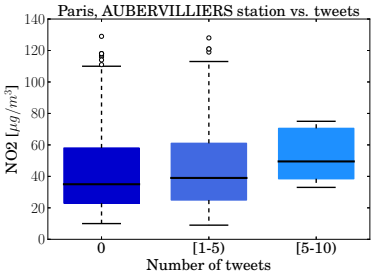
(d) S_3, NO_2



(e) S_4, NO_2



(f) S_6, NO_2



(g) S_7, NO_2

Figure 6: OSNs posts vs. measured PM_{10} and NO_2 concentration values in Paris.

- $[1-5)$, one to four posts are present on the OSNs in the considered time window
- $[5-9)$, the geosocial search returns five to nine posts,
- 10+, ten or more posts have been published in the time window.

It is worth to note that the chosen grouping limits depend on the sampling time window. With this configuration, also in very populated cities like Beijing or Paris we did not get more than 20 posts per sampling window. In case of a larger sampling window, it is worth to stretch the sets of posts, vice-versa shrinking the sampling window. Given the distribution of number of posts per sampling window, we get

a percentile value of 0.3698539, 0.5256751, 0.8818061, and 0.9852811 for 0, 1, 5, and 10 number of posts respectively.

Feasibility of the proposed model must show a raising level of pollution as the number of published posts increases, or a pollution level under the alert threshold for the first group, and an increasing dirtiness of air condition for the others.

Then, in Section 5, we describe the model to forecast the amount of posts generated in the next time window and find the most probable pollution level.

4.1. Paris measurements

Figures 6 show the key statistics of a given pollutant concentration (minimum, first quartile, median, third quartile and outliers), in according to the number of published posts, for measurement stations in Paris. For stations which are able to measure particulate matters (S_1, S_2, and S_5)we consider PM_{10} values; $PM_{2.5}$ typically constitutes 60% of the PM_{10} mass [25]. For the other station we rely on NO_2 values.

The European directive 1999/30/EC states that concentration of particles with a diameter of 10 μm or less should not exceed the hourly mean of 50 $\mu g/m^3$, while 200 $\mu g/m^3$ for Nitrogen Dioxide.

For each station and for every pollutant, minimum pollution values raise as the number of posts increases.

In particular, Figures 6(a) and 6(b) show PM_{10} values measured by stations S_1 and S_5. They cover large part of the city center, thus even more than 10 posts are present during the considered time window. Until four detected posts, pollution is around acceptable values. From five posts up, instead, air pollution begins to gain more and more importance, achieving 100 $\mu g/m^3$ for station S_1 and 120 $\mu g/m^3$ for station S_5.

Even if NO_2 measured values are under the acceptance threshold, Figures 6(d)–6(g) show that a larger set of posts returned by the geosocial search indicates a higher minimum value of this pollutant, although under the legal limits. This result indicate also that people effectively sense this pollutant.

4.2. São Paulo measurements

All the 12 São Paulo monitoring stations measure PM_{10} concentrations, so Figures 7(a)–7(l) show PM_{10} measured values key statistics related to posts generated within a radius of 5 Km away from the stations. CETESB scores air quality with the following indices:

- $[0 - 40]$, good quality,
- $[41 - 80]$, moderate quality,
- $[81 - 120]$, bad quality,
- $[121 - 200]$, very bad quality.

Although during the considered period air quality is at mostly moderate falling in the range $[41 - 80]$ and no more than five posts have been generated during any sampling window, the minimum pollution values of PM_{10} raises as soon as a post is published.

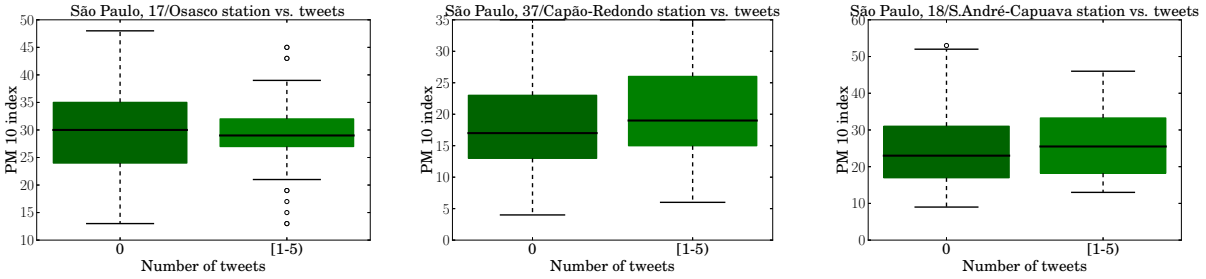
For this kind of results where the correlation is not as clear as in the other examples, the simplest corrective is to stretch the sampling window. Figures 8 show the correlation between OSNs posts and measured PM_{10} values for the last three stations considered in Figure 7, with a double sampling window. A longer sampling window often aggregates more posts, thus the correlation becomes more clear.

Once more, people reveal themselves highly perceivable and the geosocial search effective.

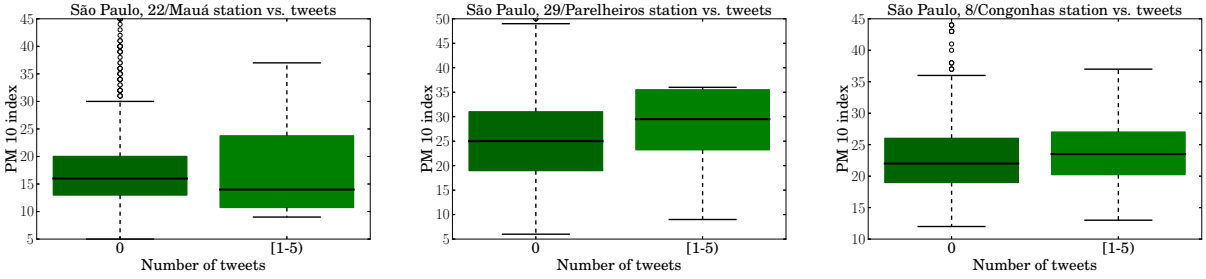
4.3. China measurements

Beyond a further confirmation of our feasibility assumption, dataset of Chinese posts also reveals the interesting aspect that we have highlighted in Section 2: the importance to have a dictionary of terms customized on the geographic point of interest for an effective geosocial search.

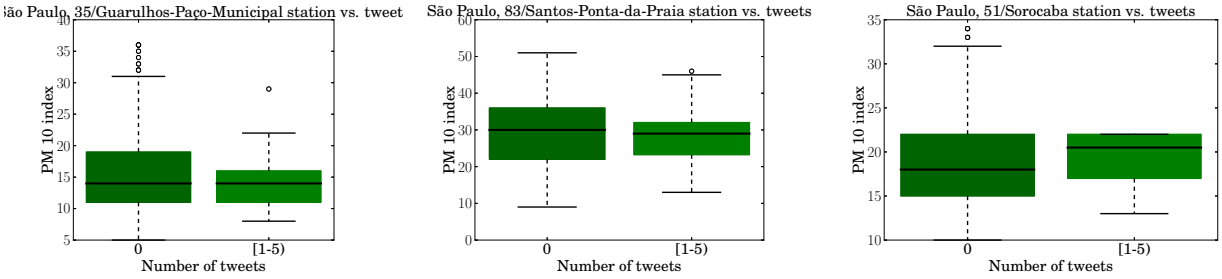
Almost all the weather-related posts found during the examined period were related to the “haze” term and these posts are the relative majority. French dictionary includes the same term but it does not have the same effect in that case. Disturbing haze conditions are consistent with the presence of high levels of given



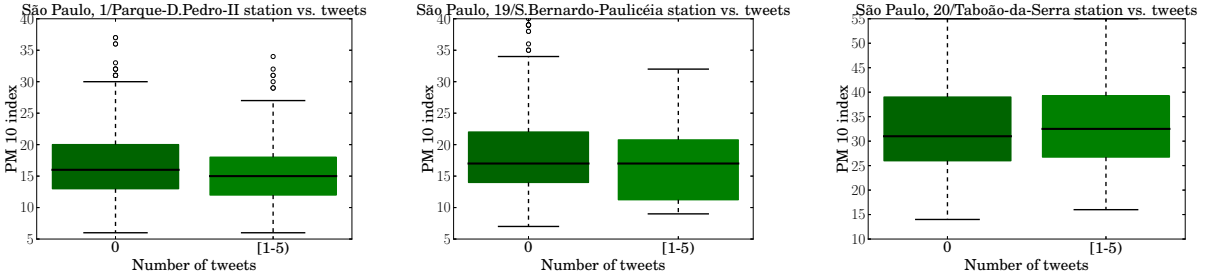
(a) Osasco (b) Capao Redondo (c) S. André Capuava



(d) Mauá (e) Parelheiros (f) Congonhas



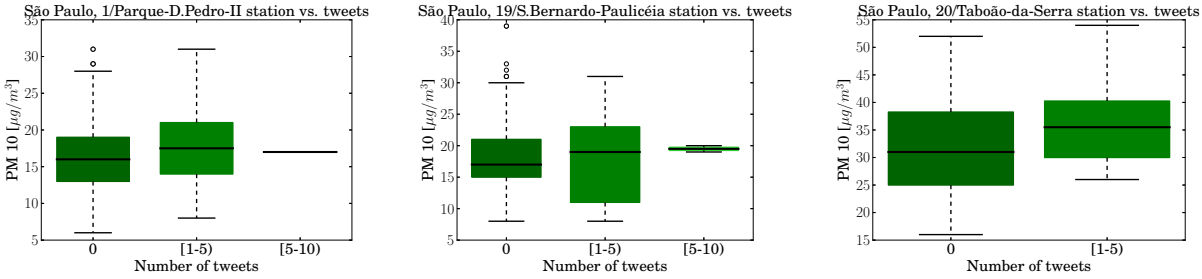
(g) Guarulhos (h) Santos Ponta da Praia (i) Sorocaba



(j) Parque Pedro II (k) S. Bernardo Paulicia (l) Taboão

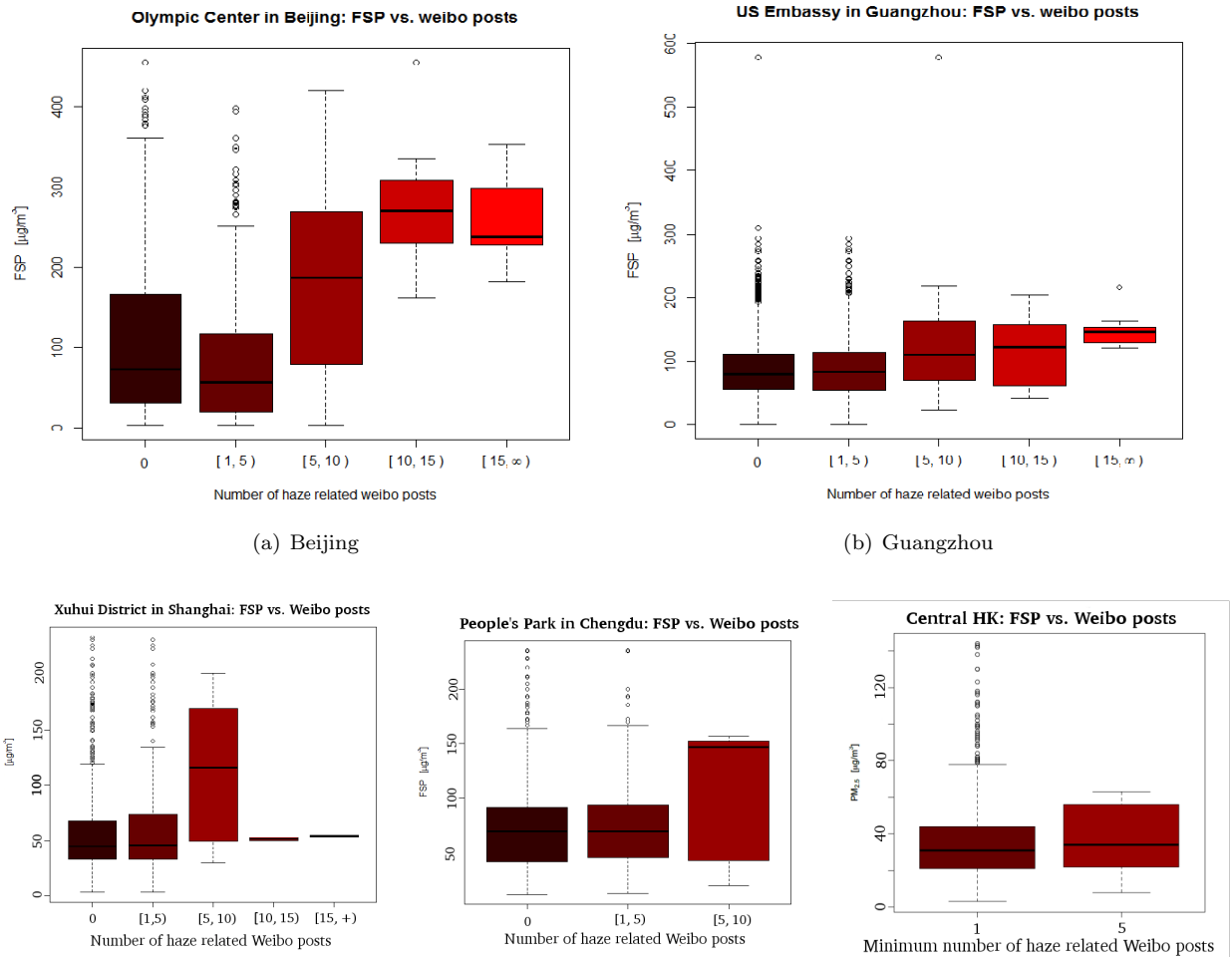
Figure 7: OSNs posts vs. measured PM_{10} values in São Paulo for all the CETESB stations.

pollutants, NO_2 and particulate matter among them. Figures from 9(a) to 9(e) prove what already observed in Sections 4.1 and 4.2. Moreover, in this case, a higher number of haze-related posts signals an increased minimum concentration of $PM_{2.5}$ values for all the considered sites. Despite the previous evaluations, the



(a) Parque Pedro II (b) S. Bernardo Paulicia (c) Taboão

Figure 8: OSNs posts vs. measured PM_{10} values in São Paulo, four hours time window sampling.



(a) Beijing (b) Guangzhou

(c) Shanghai (d) Chengdu (e) Hong Kong

Figure 9: Sina Weibo haze-related posts vs. measured $PM_{2.5}$ values in Beijing, Guangzhou, Shanghai, Chengdu, and Hong Kong air pollution monitoring stations.

Table 5: Spearman's coefficient and significance, minimum recorded $PM_{2.5}$ values compared with the number of posts.

	Paris, S_1	**São Paulo,** S_{10}	**Guangzhou**	**Beijing**	**Shanghai**	**Chengdu**
Coefficient	0.95	0.78	0.86	0.88	0.90	1.00
ρ	0.0001	0.0446	0.0001	0.0001	0.0374	0.0001

minimum number of posts to detect an effective raise of air impurity level is higher: at least 10 posts for Beijing and at least 5 for the other sites. This can be explained as a habituation effect after a long period of exposure to bad air quality.

4.4. Statistical significance

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 smallest pollution values recorded in correspondence of the given number of posts. We present results for Chinese cities, and central stations in Paris and São Paulo.

In essence, coefficients close to 1 prove what already observed in this Section: a higher number of posts signals an increased minimum concentration value of $PM_{2.5}$.

The coefficient values provided in Table 5 establish the existence of a monotonic relationship between the number of posts and the minimum concentration of $PM_{2.5}$.

5. Air Pollution level forecast

We have shown in Section 4 and 4.2 how the pollution level monitored by a station, and in particular the minimum value, raises as the minimum number of posts published in the next sampling time window, within a range of 5 km from that sensing station, increases. In this section we show how it is possible to predict the next amounts of posts so as to define a certain level of air pollution.

The series of number of user generated posts sampled each 2 hours is a non-seasonal, stationary, time series process. Any trend, autocorrelation (ACF) frequency, or partial autocorrelation (PACF) frequency is evident during the inspection of our dataset.

Therefore, we use an Autoregressive Moving Average (ARMA) model to forecast the number of posts that will be likely generated in the next two hours [26] or more. A time series y_t is represented as an ARMA(p, q) model as follows:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t, \tag{4}$$

Parameters p and q are the autoregressive and the moving average order respectively. The stationary series y_t can be expressed as a linear combination of p past values and q past forecast errors with e_t as white noise. We take advantage of the R package "forecast" which contains routines to create ARMA/ARIMA models and forecast values in time series.

Once the number of posts in the next time window is predicted, the pollution level is estimated through a binomial logit regression which gives the $P(\rho_{PM_x} > \rho_{PM_{thr}} | n_{posts})$, where ρ_{PM_x} is the concentration of $PM_{2.5}$ or PM_{10}, $\rho_{PM_{thr}}$ is the relative threshold defined by the US Environmental Protection Agency (EPA) [27], and n_{posts} is the number of posts generated in a two hours window. For sake of clarity, Figure 10 shows the $P(\rho_{PM_{2.5}} > \rho_{PM_{thr}} | n_{posts})$ in according to pollution values time series of station S_1 in Paris and the time series of posts generated within 5 Km away from S_1. Both time series use a two hours sampling. Probability overcoming the threshold of $35 \mu g/m^3$ is 0.5 when the geosocial search returns 5 posts, 0.75 with 8 posts and approaching to 1 with more than 11 posts.

Table 6: One step ahead forecasting MASE error for Paris AirParif air pollution monitoring stations.

	Station						
	S_1	S_2	S_3	S_4	S_5	S_6	S_7
ARMA(2,0)	0.5369787	0.1907348	0.3973642	0.4190386	0.5247731	0.5201859	0.3667977
ARMA(1,1)	0.528387	0.1907348	0.3874301	0.332341	0.529819	0.5078005	0.3680847
ARMA(1,0)	0.554162	0.1907348	0.4470347	0.4190386	0.5348649	0.5366997	0.3667977

Table 7: One step ahead forecasting MASE error. China.

	Station	
	S_1	S_2
ARMA(2,0)	1.098646	1.72159
ARMA(1,1)	1.088443	1.704242
ARMA(1,0)	1.100714	1.85774

Table 8: One step ahead forecasting MASE error for São Paulo CETESB air pollution monitoring stations.

	Station											
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
ARMA(2,0)	0.28257	0.572204	1.033147	0.5563099	0.4325632	0.3814696	0.4238552	0.21353	0.3136	0.2472488	1.271565	1.059638
ARMA(1,1)	0.28257	0.572204	1.033147	0.5563099	0.4325632	0.3814696	0.4238552	0.21353	0.3136	0.2472488	1.351038	1.059638
ARMA(1,0)	0.374	0.60814	0.6357827	0.4087175	0.4423132	0.5364417	0.53797	0.22443	0.3245	0.7224804	0.9536741	0.978127

6. Forecasting accuracy

Since the pollution level estimation is based on the one step ahead forecast of the number of generated posts, we evaluate the ARMA model forecasting accuracy on our dataset.

We compare the accuracy of three models, namely ARMA(2,0), ARMA(1,1), and ARMA(1,0), for one step ahead forecasting. Initially, models are fitted with the first 200 samples from the time series, thus we compare real and forecast values from the 201^{st} sample on. When forecasting for the i^{th} sample, we tune the model with the first $i-1^{th}$ samples. Values are rounded to the closest integer.

As accuracy metric, we do not adopt neither the mean absolute error, which would be scale-dependent, nor the relative error that is not feasible since both real and forecast values could be equal to zero. For these reasons we compute the forecast accuracy with the mean absolute scaled error (MASE) [28].

$$\text{MASE} = \frac{1}{K}\sum_{k=1}^{K}|y_{T+k} - \hat{y}_{T+k}|/Q, \tag{5}$$

where

$$Q = \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|. \tag{6}$$

We denote with y_i the i^{th} observation and with $\hat{y}_i$ the forecast of y_i. T is the number of observations (initially, $T = 200$), while K are the number of forecast values. Q is indeed a stable measure of the scale of the time series.

Tables 6 and 8 show the one step ahead forecasting accuracy for the time series depicting the number of posts generated within a radius of 5 Km away from each station. In every case, the prediction error is low, especially using the ARMA(1,1) model. In particular, MASE error is only in few cases greater than one, indicating a better prediction than the one-step ahead in-sample random walk forecast. In some cases the accuracy is equal even changing the model, showing a very low variance when differencing such time series. one-step ahead in-sample random walk forecast

Table 9: Forecasting MASE error for different time windows.

	2H	**4H**	**8H**	**16H**
Paris, S_1	0.528387	0.55235	0.598234	0.818211
Sao Paulo, S_{10}	0.2472488	0.42319	0.49721	0.71358
China, S_1	1.088443	1.335301	1.324451	1.73151
China, S_2	1.704242	1.724646	1.708153	1.90146

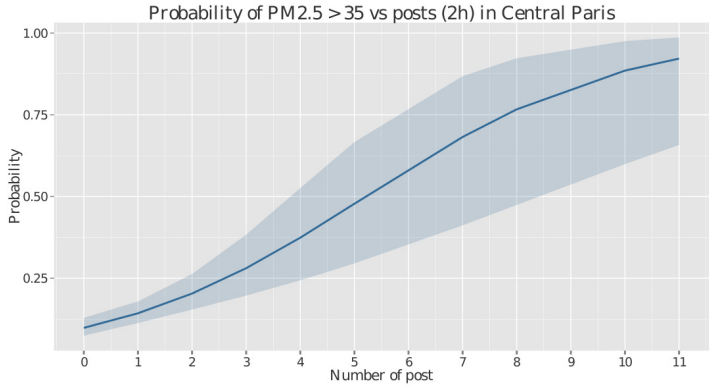


Figure 10: $P(\rho_{PM_{2.5}} > \rho_{PM_{thr}}|n_{posts})$ logit regression curve, for $\rho_{PM_{thr}} > 35\mu g/m^3$, for station S_1 in Paris

In Table 9 we compute the ARMA(1,1) forecasting MASE error for different time windows in order to check the forecasting horizon. Only for stations downtown Paris and Sao Paulo the error remains less than 1 up to 16 hours previsions. Thus, the prediction is better than one-step ahead in-sample random walk forecast for almost one day. It is to be noted that the error does not grows linearly and fluctuations are presents for the considered stations in China.

7. Related works

The ubiquitous accessibility to OSNs and the proliferation of user generated posts has recently motivated the use of geosocial search, leading to the creation of general frameworks for geo-social queries [29, 30], its usage to urban planning [31, 32] and activities detection [33, 34].

Although several studies regard OSNs, their utilization as a source of world measurements is a relatively original approach. Authors in [24] adopt a probabilistic spatio-temporal model to use tweets as earthquake sensors and give alerts.

Nevertheless, Goodspeed points out drawbacks and limitation of such approach [35]. One of the occurring problems is the tweets information credibility. This aspect has been studied in previous works [36, 37]. Techniques to infer credibility involve supervised machine learning [38], maximization likelihood model [39, 40], a combination of both of them [11], and observations from the physical world [15]. Authors in [41] use a managed participatory sensing system to improve traditional environmental monitoring systems constituted by static measurement stations.

Sentiment analysis of posts text, instead, is employed to infer contributors genuinity [13]. An extensive analysis of social differences in tweets is conducted by Poblete et al. [9].

While a large literature is dedicated to air pollution forecast, to the best of our knowledge, relying of social network monitoring is an original approach.

Large urban scenarios are a very complexity system of meteorological conditions and air pollution concentrations, making very difficult to provide fast and accurate pollution level forecasts, with often large over- or under-prediction errors [42, 43]. For this reason, statistical models on historical time series have been adopted to forecast pollutant concentrations. ARIMA has been previously used as a forecasting model to predict maximum daily surface O_3 concentration at Brunei Darussalam [44]. Chelani et al. propose a hybrid methodology combining the autoregressive integrated moving average model and nonlinear dynamical model [45], while Kim and Kumar apply an autoregressive model with threshold for ozone monitoring.

Hassan and Li's approach relies, as many other works, on AI to predict specific pollutant concentrations in restricted areas [46]. These works include forecasting methods based on artificial neural networks (ANN) [47, 48, 49], support vector machine (SVM) [50, 51], fuzzy logic [52], and a combination of fuzzy logic and Hidden Markov Model (HMM) [53]. A comparison between statistical methods and classification algorithms

is provided by Athanasiadis et al. [54], while Díaz-Robles et al. propose a hybrid ARIMA-ANN approach to improve forecasting of PM_{10} in Temuco, Chile [55].

Slini et al. use classification and regression trees in according to neural networks to predict PM_{10} levels in the Thessaloniki urban area, Greece [56], while Feng et al. use a combination of SVM and genetic algorithms to monitor and predict ozone concentrations [57]. NEMO is a NO_2 concentration predictor evaluated in Athens, Greece, based on a case-based-reasoning approach combining heuristic and statistical techniques [58].

PrevAir is an online service forecasting 3 days ahead air pollution levels in Europe, combining meteorological and chemistry models [59]. Authors in [60] also combine atmospheric and chemistry models to forecast pollutant concentrations in southern Brazil. A survey about $PM_{2.5}$ mass concentrations in six Brazilian cities is provided by De Miranda et al. [23].

8. Discussion and Challenges

Before leading the reader to the conclusions of this work, we wish to recall her attention on some keen points.

We have mentioned that people in China may have gotten used to pollution and for this reason they tweet less often about it. Thus the model should adapt to the behavior/characteristics of people in different regions around the world. This adaptation is one of the central point since the beginning of the manuscript, especially for the dictionary construction where differences are more evident: different terms are employed not only considering the translation but also territory peculiarities.

Another point to highlight is the possibility of colluding users maliciously reporting data from a given region: for example, a competitor may generate fake pollution tweets from a region around a factory. Although it is not the center of this work, credibility is a real issue dealing with social networks and crowdsourcing systems. The issue is so important that a large literature is produced about this topic [35, 36, 37, 38, 39, 40, 11]. Moreover, we have added some proposals to manage this issue: credibility of a user might be inferred controlling the pollution levels measured by the closest fixed station, by the claims of the other users, or by attached pictures in posts.

9. Conclusions

This manuscript describes the use of the geosocial search as a source for urban air pollution measurements. The aim of our work is to integrate collected measures to existing air pollution monitoring networks which often are not enough extended to cover large cities and their suburbs.

After the construction of a dictionary of terms to conduct an effective geosocial search, we show how the air pollution level measured is proportional to the number of posts returned by the geosocial search. The feasibility of our approach is proved in large cities with different morphological and cultural aspects over three continents.

Finally, despite the existing literature which uses pollution values and meteorological time series to predict future pollution values, we propose a forecasting approach relying on the time series of number of posts.

Once we have proved the feasibility, our methodology can be extended in many directions. Credibility of users' claims about air pollution could be corroborated by attached photos or by contemporary measures from the closest fixed traditional air monitoring stations.

Conducting sentiment analysis on post texts could reveal different topics to monitor apart from air pollution, such as sea bathing water quality or unsafe zones.

Cross regression between the time series of number of posts and pollution values is a promising technique to further improve previsions.

Acknowledgments

The authors of this work have been partially supported by the Macao Polytechnic Institute-Bridging Urban Sensing and Social Networks (RP/ESAP- 02/2014), the Italian Ministry of University and Research (project: CagliariPort2020, SCN_00281, FFO) and the ATOS/Renault Chair of Excellence research fund at the University Pierre and Marie Curie.

References

[1] M. D. Gwynne, The global environment monitoring system (gems) of unep, Environmental Conservation 9 (1982) 35–41. doi:10.1017/S0376892900019469.
URL http://journals.cambridge.org/article_S0376892900019469

[2] D. Mage, G. Ozolins, P. Peterson, A. Webster, R. Orthofer, V. Vandeweerd, M. Gwynne, Urban air pollution in megacities of the world, Atmospheric Environment 30 (5) (1996) 681 – 686, supercities: Environment Quality and Sustainable Development. doi:http://dx.doi.org/10.1016/1352-2310(95)00219-7.
URL http://www.sciencedirect.com/science/article/pii/1352231095002197

[3] Monitorar-rio - programa de monitoramento da qualidade do ar [online].

[4] R. M. Hoff, 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) (2009) 645–675.

[5] R. Honicky, E. A. Brewer, E. Paulos, R. White, N-smarts: networked suite of mobile atmospheric real-time sensors, in: Proceedings of the second ACM SIGCOMM workshop on Networked systems for developing regions, ACM, 2008, pp. 25–30.

[6] N. Maisonneuve, M. Stevens, M. E. Niessen, L. Steels, Noisetube: Measuring and mapping noise pollution with mobile phones, in: Information Technologies in Environmental Engineering, Springer, 2009, pp. 215–228.

[7] B. Pat, Y. Kanza, M. Naaman, Geosocial search: Finding places based on geotagged social-media posts, in: Proceedings of the 24th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2015, pp. 231–234.

[8] J. Caverlee, Z. Cheng, D. Z. Sui, K. Y. Kamath, Towards geo-social intelligence: Mining, analyzing, and leveraging geospatial footprints in social media., IEEE Data Eng. Bull. 36 (3) (2013) 33–41.

[9] B. Poblete, R. Garcia, M. Mendoza, A. Jaimes, Do all birds tweet the same?: characterizing twitter around the world, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 1025–1030.

[10] M. T. Al Amin, T. Abdelzaher, D. Wang, B. Szymanski, Crowd-sensing with polarized sources, in: Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on, IEEE, 2014, pp. 67–74.

[11] S. Sikdar, S. Adali, M. T. A. Amin, T. F. Abdelzaher, K. Chan, J. Cho, B. Kang, J. O'Donovan, Finding true and credible information on twitter, in: 17th International Conference on Information FusionFusion, FUSION 2014, Salamanca, Spain, July 7-10, 2014, 2014, pp. 1–8.
URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6915989

[12] D. Lary, T. Lary, B. Sattler, Using machine learning to estimate global pm2. 5 for environmental health studies, Environmental health insights 9 (Suppl 1) (2015) 41.

[13] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining., in: LREC, Vol. 10, 2010, pp. 1320–1326.

[14] Twitter statistics [online].

[15] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, H. Liu, Exploitation of physical constraints for reliable social sensing, in: Real-Time Systems Symposium (RTSS), 2013 IEEE 34th, 2013, pp. 212–223. doi:10.1109/RTSS.2013.29.

[16] Amazon mechanical turk [online].

[17] H. Kozima, T. Furugori, Similarity between words computed by spreading activation on an english dictionary, in: Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1993, pp. 232–239.

[18] Airparif - association de surveillance de la qualité de l'air [online].

[19] Us department of state air quality monitoring program [online].

[20] Aqhi - hong kong air quality health index [online].

[21] J. Seinfeld, S. Pandis, Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 1998.

[22] Cetesb - companhia ambiental do estado de são paulo [online].

[23] R. de Miranda, M. de Fatima Andrade, A. Fornaro, R. Astolfo, P. de Andre, P. Saldiva, Urban air pollution: a representative survey of pm2.5 mass concentrations in six brazilian cities, Air Quality, Atmosphere and Health 5 (1) (2012) 63–77. doi:10.1007/s11869-010-0124-1.
URL http://dx.doi.org/10.1007/s11869-010-0124-1

[24] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 851–860. doi:10.1145/1772690.1772777.
URL http://doi.acm.org/10.1145/1772690.1772777

[25] W. H. Organization, Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, Global update 2005 Summary of risk assessment, WHO, 2006.

[26] R. Shumway, D. Stoffer, Arima models, in: Time Series Analysis and Its Applications, Springer Texts in Statistics, Springer New York, 2011, pp. 83–171. doi:10.1007/978-1-4419-7865-3_3.
URL http://dx.doi.org/10.1007/978-1-4419-7865-3_3

[27] Us environmental protection agency [online].

[28] R. Hyndman, A. B. Koehler, Another look at measures of forecast accuracy, International Journal of Forecasting 22 (4) (2006) 679–688.
URL http://EconPapers.repec.org/RePEc:eee:intfor:v:22:y:2006:i:4:p:679-688

[29] N. Armenatzoglou, S. Papadopoulos, D. Papadias, A general framework for geo-social query processing, Proc. VLDB Endow. 6 (10) (2013) 913–924. doi:10.14778/2536206.2536218.
URL http://dx.doi.org/10.14778/2536206.2536218

[30] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, J. Sperling, Twitterstand: news in tweets, in: Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems, ACM, 2009, pp. 42–51.

[31] V. Frias-Martinez, V. Soto, H. Hohwald, E. Frias-Martinez, Characterizing urban landscapes using geolocated tweets, in: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), IEEE, 2012, pp. 239–248.

[32] B. Pat, Y. Kanza, M. Naaman, Geosocial search: Finding places based on geotagged social-media posts, in: Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume, 2015, pp. 231–234. doi:10.1145/2740908.2742847.
URL http://doi.acm.org/10.1145/2740908.2742847

[33] F. Kling, A. Pozdnoukhov, When a city tells a story: Urban topic analysis, in: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12, ACM, New York, NY, USA, 2012, pp. 482–485. doi:10.1145/2424321.2424395.
URL http://doi.acm.org/10.1145/2424321.2424395

[34] N. Malleson, M. Birkin, New insights into individual activity spaces using crowd-sourced big data.

[35] R. Goodspeed, The limited usefulness of social media and digital trace data for urban social research, in: Seventh International AAAI Conference on Weblogs and Social Media, 2013.

[36] A. Gupta, P. Kumaraguru, Credibility ranking of tweets during high impact events, in: Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, PSOSM '12, ACM, New York, NY, USA, 2012, pp. 2:2–2:8. doi:10.1145/2185354.2185356.
URL http://doi.acm.org/10.1145/2185354.2185356

[37] B. Kang, J. O'Donovan, T. Höllerer, Modeling topic specific credibility on twitter, in: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, IUI '12, ACM, New York, NY, USA, 2012, pp. 179–188. doi:10.1145/2166966.2166998.
URL http://doi.acm.org/10.1145/2166966.2166998

[38] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, WWW '11, ACM, New York, NY, USA, 2011, pp. 675–684. doi:10.1145/1963405.1963500.
URL http://doi.acm.org/10.1145/1963405.1963500

[39] D. Wang, M. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, H. Le, Using humans as sensors: An estimation-theoretic perspective, in: Information Processing in Sensor Networks, IPSN-14 Proceedings of the 13th International Symposium on, 2014, pp. 35–46. doi:10.1109/IPSN.2014.6846739.

[40] D. Wang, L. Kaplan, H. Le, T. Abdelzaher, On truth discovery in social sensing: A maximum likelihood estimation approach, in: Proceedings of the 11th International Conference on Information Processing in Sensor Networks, IPSN '12, ACM, New York, NY, USA, 2012, pp. 233–244. doi:10.1145/2185677.2185737.

[41] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, L. Thiele, Deriving high-resolution urban air pollution maps using mobile sensor nodes, Pervasive and Mobile Computing 16, Part B (2015) 268 – 285, selected Papers from the Twelfth Annual {IEEE} International Conference on Pervasive Computing and Communications (PerCom 2014). doi:http://dx.doi.org/10.1016/j.pmcj.2014.11.008.
URL http://www.sciencedirect.com/science/article/pii/S1574119214001928

[42] J. H. Seinfeld, S. N. Pandis, Atmospheric chemistry and physics : from air pollution to climate change, Hoboken, N.J. J. Wiley, 2006.
URL http://opac.inria.fr/record=b1128402

[43] A. A. Argiriou, Use of neural networks for tropospheric ozone time series approximation and forecasting ? a review, Atmospheric Chemistry and Physics Discussions 7 (2) (2007) 5739–5767.
URL https://hal.archives-ouvertes.fr/hal-00302742

[44] Y. A. S. M. H. H. J. V. Kumar, K., Forecasting daily maximum surface ozone concentrations in brunei darussalam–an arima modeling approach, Journal of the Air and Waste Management Association 54.

[45] A. B. Chelani, S. Devotta, Air quality forecasting using a hybrid autoregressive and nonlinear model, Atmospheric Environment 40 (10) (2006) 1774 – 1780. doi:http://dx.doi.org/10.1016/j.atmosenv.2005.11.019.
URL http://www.sciencedirect.com/science/article/pii/S135223100501071X

[46] R. Hassan, M. LI, Urban Air Pollution Forecasting Using Artificial Intelligence-Based Tools, INTECH Open Access Publisher, 2010.
URL https://books.google.fr/books?id=p7qGoAEACAAJ

[47] H. Niska, T. Hiltunen, A. Karppinen, J. Ruuskanen, M. Kolehmainen, Evolving the neural network model for forecasting air pollution time series, Engineering Applications of Artificial Intelligence 17 (2) (2004) 159–167.

[48] M. Cai, Y. Yin, M. Xie, Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach, Transportation Research Part D: Transport and Environment 14 (1) (2009) 32 – 41. doi:http://dx.doi.org/10.1016/j.trd.2008.10.004.

[49] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, J. Wang, Artificial neural networks forecasting of pm2.5 pollution using air mass trajectory based geographic model and wavelet transformation, Atmospheric Environment 107 (0) (2015) 118 – 128. doi:http://dx.doi.org/10.1016/j.atmosenv.2015.02.030.

[50] W.-Z. Lu, W.-J. Wang, Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends, Chemosphere 59 (5) (2005) 693–701.

[51] E. G.-H. C. P.-O. J. R.-A. A. Sotomayor-Olmedo, M. Aceves-Fernández, J. Vargas-Soto, Forecast urban air pollution in mexico city by using support vector machines: A kernel performance approach, International Journal of Intelligence Science 3 (3) (2013) 126 – 135. doi:10.4236/ijis.2013.33014.

[52] F. C. Morabito, M. Versaci, Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data, Neural Networks 16 (3) (2003) 493–506.

[53] M. M. Hossain, M. R. Hassan, M. Kirley, Forecasting urban air pollution using hmm-fuzzy model, in: Advances in Knowledge Discovery and Data Mining, Springer, 2008, pp. 572–581.

[54] I. Athanasiadis, K. Karatzas, P. Mitkas, Contemporary air quality forecasting methods: a comparative analysis between classification algorithms and statistical methods, in: Fifth international conference on urban air quality measurement, modelling and management, Valencia, Spain, 2005.

[55] L. A. Díaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, J. A. Moncada-Herrera, A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of temuco, chile, Atmospheric Environment 42 (35) (2008) 8331–8340. doi:10.1016/j.atmosenv.2008.07.020.

[56] T. Slini, A. Kaprara, K. Karatzas, N. Moussiopoulos, Pm10 forecasting for thessaloniki, greece, Environ. Model. Softw. 21 (4) (2006) 559–565. doi:10.1016/j.envsoft.2004.06.011.
URL http://dx.doi.org/10.1016/j.envsoft.2004.06.011

[57] Y. Feng, W. Zhang, D. Sun, L. Zhang, Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification, Atmospheric Environment 45 (11) (2011) 1979–1985.

[58] E. Kalapanidas, N. Avouris, Short-term air quality prediction using a case-based classifier, Environmental Modelling and Software 16 (3) (2001) 263 – 272. doi:http://dx.doi.org/10.1016/S1364-8152(00)00072-4.
URL http://www.sciencedirect.com/science/article/pii/S1364815200000724

[59] Prevair - air quality forecasts and observations in france and europe [online].

[60] M. D. F. Andrade, E. D. Freitas, R. Y. Ynoue, E. Todesco, A. V. Vara, S. Ibarra, L. D. Martins, J. Martins, V. S. Carvalho, Air quality forecasting system for southeastern brazil, Frontiers in Environmental Science 3 (9). doi:10.3389/fenvs.2015.00009.
URL http://www.frontiersin.org/interdisciplinary_climate_studies/10.3389/fenvs.2015.00009/abstract