Hindawi
Mobile Information Systems
Volume 2022, Article ID 5562317, 14 pages
https://doi.org/10.1155/2022/5562317

Research Article

A Geo-Social Characterization of Health Impact from Air Pollution in Mexico Valley

Roberto Zagal Flores,[1] **Christophe Claramunt**,[2] **Miguel Felix Mata Rivera**,[3] **Laura Ivoone Garay Jiménez**,[3] **Hugo Jiménez Hernández**,[4] **Ana Marcela Herrera Navarro**,[4] and **Amadeo José Argüelles Cruz**[5]

[1] *Instituto Politécnico Nacional, ESCOM-IPN, Ciudad de México 07320, Mexico*
[2] *Naval Academy Research Institute, Brest 29240, France*
[3] *Instituto Politécnico Nacional, UPIITA-IPN, Ciudad de México 07340, Mexico*
[4] *Universidad Autónoma de Querétaro, Fac., Informática, Juriquilla, Mexico*
[5] *Instituto Politécnico Nacional, CIC-IPN, Ciudad de México 07738, Mexico*

Correspondence should be addressed to Miguel Felix Mata Rivera; mmatar@ipn.mx

Received 12 February 2021; Revised 1 June 2022; Accepted 6 August 2022; Published 29 August 2022

Academic Editor: Mehrbakhsh Nilashi

The impact of the air pollution phenomenon has been long studied, but most often with a fragmented approach, without closely looking at the relationship between different components that characterize it, such as sensor-based data, health data from institutional databases, and data on how it is perceived by human beings in social media. The research developed in this study introduces an integrated methodological framework that analyses sensor data on air pollution distributed in space and time, combined with health data and social data narratives that reflect how different communities perceive this phenomenon in space and time; exploring how these different heterogeneous sources can be combined to better understand the impact of air pollution phenomena at the large-city level in the Valley of Mexico. We introduce a Spatio-temporal data integration and mining framework that aims to discover trends and insights regarding the distribution of the impact of an air pollution phenomenon in terms of human health and perception. The main peculiarity of our methodological framework is the integration of different large data sources by combining a series of methods: NLP (topic modeling), data mining (data cubes, unsupervised learning, and clustering), and GIS capabilities (spatial interpolation, choropleth maps) that together provide a better understanding of the quantitative and qualitative patterns emerging at a different spatial scale and temporal granularity. Overall, this shows how social data, when combined with quantitative data, can provide a better understanding of the impact of a given phenomenon, such as air pollution.

1. Introduction

Many environmental and urban phenomena have been long described by institutional, sensor-based databases or time-consuming surveys [1–4] (e.g., crime figures, air pollution, and noise levels). While these studies provide broad quantitative insights for analysis and understanding of a phenomenon of interest, as many open data initiatives also facilitated their availability, they do not provide instant and qualitative reports of what is happening in the city. This is the case, for example, when the aim is to study how inhabitants perceive the impact of crime patterns or pollution levels. In recent years, social networks have provided valuable alternatives to record citizens' opinions on the progress of some urban facts and events, and this is distributed across the city and at different levels of scale and over time [5, 6].

As expressed in social networks, narratives provide spontaneous and personal descriptions of geo-social and environmental phenomena that might provide complementary insights when combined with a repository of descriptive data. Indeed social media data are far from being as objective as sensor-based or statistical databases. However, they have the advantage of being regularly expressed, almost freely available, and generally much more qualitative, although they are surely heterogeneous, relatively imprecise, and not always complete. On the methodological side, the unstructured component of social data opens a new avenue of research as such narratives imply the development of Natural Language Processing (NLP) algorithms whose objective is to extract and reorganize the data to give a manipulable form to the whole data.

Therefore, the two questions that arise are as follows: first, how to reconcile these two data sources at different levels of description in space, time, and according to the embedded semantics? Second, can social data complement descriptive data available through open data sources, enhancing a knowledge discovery process and thus enriching the understanding of a phenomenon? This is the main methodological aim of this research, which explores how open social data associated with open data on a given phenomenon can lead to a better understanding of an urban or environmental phenomenon, and not only how it is distributed in the city over time but also how it is perceived by people acting in the city. The context of our research is the impact of air pollution as recorded in the city and in the region; and how it is perceived by human beings at the local level.

The reason for this choice is that air pollution provides a context with substantial quantitative data available in the city over time and a sensitive for city inhabitants, as it has important health impacts. In particular, it has been observed that humans are very reactive in social networks when the observed pollution levels worsen. Humans are also keen to describe some symptoms associated with respiratory diseases, this expressing a potential qualitative relationship with air pollution levels. We observe three fundamental problems in the observation of air pollution: (1) enriching the understanding of air pollution from social networks in combination with open data, (2) extracting narratives that describe health effects in social networks, (3) providing a workflow able to enrich social perceptions with trends obtained from environmental and health databases.

From a methodological point of view, our research develops a framework that applies a series of data mining, natural language processing (NLP), and geographic information systems (GIS) techniques. The goal is to extend the understanding of air pollution impact from a regional perspective to an individual granularity by providing a workflow capable of, first, enriching social insights with historical environmental and health trends from open data sources and secondly allowing data integration and exploration. The main steps of our methodological framework are as follows: (1) record the historical behavior of a phenomenon, representing regional trends over time using multidimensional data cubes combining open and social data; (2) search for phenomenon narratives and emerging themes or topics in space and time using; (3) characterize the emerging social perceptions using a Topic Modeling process (an NLP algorithm) aiming at discovering the semantic structures hidden in large text datasets. The application of this algorithm should highlight the main themes that could reveal embedded social-health patterns associated with air pollution phenomena and; (4) finally, present and highlight trends emerging using Spatio-temporal analysis tools, such as choropleth maps and kernel interpolations on the distribution and concentration of PM10, PM 2.5, and CO pollutants. It should provide a global analysis of the correlation between pollution patterns in space and time, respiratory diseases, and trends revealed by health data and their perception through social media. The peculiarity of our approach is that it combines sensor-based pollution data, health data, and human perceptions in an integrated framework that favors the analysis and discovery of the respective objective and subjective impact of air pollution and urban health phenomena.

The main advantage of our approach is that it analyses how trends and insights can be obtained through a robust integration of open data, health data, and people's perceptions as extracted from social media. It provides a relatively complete view of a given phenomenon, from quantitative descriptive data to qualitative patterns derived from social media narratives. However, the work might still be extended by first providing a stronger computational integration of the different NLP, data mining, and GIS capabilities. A further degree of flexibility might also be provided at the interface level by allowing health experts to "play with the data" to explore the patterns that emerge throughout all steps of data processing. Finally, additional visualization capabilities might be developed at different levels of spatial scale and temporal granularity.

The remainder of the paper is organized as follows. Section 2 introduces the related work, while Section 3 develops the methodological principles of our approach. Section 4 presents the results and experiments, while finally, Section 5 draws the conclusions and outlines future work.

2. Related Work

While many previous works have applied GIS, NLP, and data mining together to some extent to analyze urban patterns, our methodological framework goes further by providing an integrated framework that offers a robust approach that combines their potential in time and space, and by applying the whole approach to a relatively large dataset composed of sensor-based data, historical health data, and human narratives as expressed in a social media. The main peculiarity of our approach is that it supports a spatiotemporal correlation of quantitative and qualitative data that provides a better understanding of the real and perceived impact of air pollution phenomena in an urban environment.

2.1. Spatial Data Analysis Applied to Air Pollution and Health Studies.

In related work, hourly data of PM2.5 (fine particle matter with diameters of generally 2.5 micrometers), PM10 (particles with diameters of 10 micrometers and smaller), and CO pollutants (carbon monoxide) were collected from

336 Chinese cities for two years to uncover the geographic and time variations and influential factors of these pollutants [7]. The study showed that all the pollutants exhibited significant weekly and diurnal cycles. These results highlighted the impact of meteorology on air pollution in China, the geographical and temporal variations, and the role played by a series of additional factors.

Under similar principles, but this time with SO2, NO2, PM2.5, PM10, CO, and O3 pollutants recorded over one year, they have aggregated an air quality index in 338 Chinese cities [8]. The air quality index values showed remarkable spatiotemporal variations across the country. The main findings were that air quality index values generally remained high throughout the country. Spatially, high or low index values were discovered in cities located in the North or South of China, and high index values were observed in the West and East of China. It also appears that the concentrations of PM contribute significantly to the index of air quality. The study presents trends and spatial patterns of pollutants in cities, with clusters of high pollutants in the Southwest Xinjiang province and clusters of low pollutants in cities in southern and northeast China. However, despite the interest in these two approaches, no comparison with health data and people's perception of the health impact of these pollution patterns has been developed.

In related work, a different approach studied risk perception as an indicator of the public perception of air pollution [9]. The authors found significant differences in public risk perception and attitude toward air pollution amongst regions. They applied a hierarchical linear model to explore the effects of demographic, environmental, and economic factors on the trends that appear. They found that PM2.5 has a significant influence on perceived risk factors and a negative correlation between risk perception and user satisfaction. However, health data, as potentially available from governmental institutions, were not integrated into the framework. Another study investigated the effects of ambient air pollution on hospital admissions for cardiovascular and respiratory diseases in the city of Bangkok [10]. The study analyses daily air pollution concentration (O3, NO2, SO2, PM10, and CO) and meteorological variables from January 2006 to December 2014 and daily hospital admissions for cardiovascular and respiratory diseases. A time series analysis examined the effects of air pollution on hospital admissions and other potential confounders. The results showed a series of clear patterns and evidence to show the effects of air pollution (O3, NO2, SO2, PM10, and CO) on hospital admissions for cardiovascular and respiratory diseases. Here, no correlation was made with public perception of these health patterns.

Another study examined the variability of the impact of air pollution on life satisfaction in the cities of Beijing and Shanghai [11]. A robust negative impact of air pollution on subjective well-being was demonstrated. The authors applied a surface interpolation technique on pollution data as sensed from different monitoring sites to spatially estimate SO2, NO2, PM10, and PM2.5 pollution; the results showed that all pollutants have robust negative impacts. This work uses a similar interpolation technique.

Related work explores the particle matters (PM 2.5) and its relationship with lung cancer incidence in France [12], and the lung cancer burden attributable to PM2.5 exposure corresponded to 3.6% of all cases treated in 2015. The study uses a spatially refined nationwide chemistry-transport model with a spatial resolution of 2 km, neighborhood-scale population density data, and a relative risk from a published meta-analysis. However, the approach is purely quantitative and does not integrate additional qualitative data that could be drawn from people's perceptions of such health impacts.

Overall, these studies show the importance of detecting historical trends from open data sources about air pollution and health, as well as contextualizing such phenomena. Although air pollution impacts were either derived from health databases or, to a certain degree, social media, no sound cross-comparison and correlations were explored to evaluate the overall quantitative and qualitative impact of these health patterns over space and time.

2.2. Geo-Social Studies.

A general and conceptual framework has been developed to study specific situations of interest (e.g., epidemic outbreaks) using large-scale spatio-temporal multimedia streams [13]. Flu reports, as well as growth rates, were specifically extracted and aggregated from a case study. However, no explorations were reported to uncover trends emerging from new data in a sort of discovery knowledge process.

Another related work introduced a topic tracking system to identify, monitor, analyze, and visualize important local events posted on Twitter in urban environments [14]. The main idea was to obtain not only the spatial distribution of certain geo-topics but also to analyze the evolution of the patterns that emerged. However, the data were limited to social data and did not integrate additional data derived from additional resources.

A relevant example of the integration of social networks with sensor-based data to study the impact of air pollution has been developed in previous work [15]. The authors developed an implementation of a keyword-based geo-social search mechanism to look for spatial patterns in air quality complaints as revealed by Twitter data. The results showed a significant correlation over time in a series of cities in France, Brazil, and China. With the help of a dictionary of pollution-related terms, relevant posts are identified, classified, and then mapped to different urban neighborhoods and cross-compared with socio-cultural differences as they appeared from the city layout. From these historical patterns, some predictions are also generated. While sharing the methodological principles with this previous work, our methodological framework goes further by also analyzing the correlation with respiratory disease cases (e.g., "headache," "chest pain"). In addition, a series of space-time visualizations are generated, at different scales, achieving a better understanding of the patterns under review. Last but not least, our approach covers two years of tweets and 30 years of air quality measurements in a large urban region with over 20 million inhabitants.

3. Methodology

This section introduces the main principles of our methodology and geo-social framework. The large and complex incoming data comprise three components: (1) recorded and structured digital data from long or short periods [open and structured data], (2) time-stamped and geo-referenced public posts from social networks [unstructured data], and (3) geographical information associated to air pollution phenomena [structured data].

The first component, air quality data, contains the measurements per year of the following pollutants: PM10 (particles with diameters of 10 micrometers and smaller), PM2.5 (fine Particulate Matter with a diameter of generally 2.5 micrometers), and CO (Carbon Monoxide). These data are gathered through a network of 40 monitoring stations distributed in the Valley of Mexico administered by the Mexico City Government [16]. CO, PM10, and PM2.5 are the main parameters considered since these pollutants are produced by transport and industry in urban environments [7] and are related to respiratory diseases [10, 17, 18].

Health data contains information on the hospital unit and patient, diseases treated, date, and hospital postal address [19]. Population data describes the population profile in terms of locality, city, state, sex, and age (1990, 2000, and 2010). The open geospatial data set contains the administrative boundaries of Mexico. Finally, narratives are collected through a web extraction process with the public Twitter API. The following challenges were identified: (a) integration of complex structured and unstructured data obtained from social networks and open data, respectively, (b) analysis of geographic and temporal data with their representation, and finally; (c) integration of all data for pattern matching, mining, and exploration to gain additional insights; (d) search parameters on what, where and when to extract narratives from Twitter. The data sets used in this work (structured data and gathered tweets) are available in [20].

This geo-social methodology analyses an air pollution phenomenon from a social perspective in different places and at different times. The methodology looks for trends, insights, and patterns in space and time that have occurred in the last three decades. The approach considers the heterogeneous and complex nature of the data, where new open/social media datasets appear regularly. Figure 1 depicts the main principles behind this methodology.

Figure 1 summarizes the three stages of the methodological workflow. The first phase comprises the extraction of the geographic patterns of air pollution (i.e., how the pollutants are distributed in Mexico Valley over the last 30 years). The idea is to identify the main and significant historical trends of air pollution in the Mexico Valley and to provide a reference to guide the extraction of social data (where and when to extract data). Second, the search for emerging social issues or topics where social data is required to understand what mainly people are saying about air pollution, detecting topics related to different social contexts in space and time (e.g., where and when health and air pollution issues emerge). Then, unsupervised learning methods are applied to extract hidden topics (e.g., headache) in large text data sets that describe their semantics [21].

Finally, the last step of the framework applies geo-social patterns integration, where the goal is to look for correlations between air pollution data and open data such as health, diseases, and symptoms reflected by social media. For example, to look for a relationship between PM10, PM 2.5, and CO with the terms of diseases and respiratory symptoms (i.e., health impact expressed as headache) reported in social media.

3.1. Stage 1: Building Historical Trends from Open Data. Stage 1 explores 30 years of digital history of air pollution patterns in the Mexico Valley, as they emerge in space and time from open data sources (understanding phenomena from open data). It comprises three tasks: (i) construction of data cubes from open data (i.e., air pollution) to identify trends in space and time to apply inference mechanisms, (ii) cross-correlation of such open data patterns with Spatio-temporal health trends, and (iii) Spatio-temporal interpolation of pollutants.

3.1.1. Data Cube Principles. The main principle of a data cube architecture is to organize, blend and summarize a large database commonly derived from open data to derive a new structure (a data cube implementation phase). We have kept a spatial star schema introduced in a related work primarily oriented to high-dimensional data analysis [22].

Data cubes categorize information according to multiple hierarchies or semantic dimensions, allowing analysis of location, pollutants, and health in a multidimensional way. The goal is to detect emergent patterns in open data as they appear in space and time. We differentiate between fact tables (e.g., records of pollution measurements that are analyzed quantitatively) and dimension tables (e.g., tables that describe facts using categorical or numeric information, such as time or location).

Drill-down operations for data exploration are performed considering the various dimensions and granularity by applying statistical operations such as *mean*, *min,* and *maximum values* in quantitative columns. For example, to identify the periods and regions with the highest pollution levels, the “*average*” measure is applied in the <“[*pollutantMeasure*]”> column and includes location information, such as city limits or geographic coordinates of the monitoring station by performing drill-across operations. This average or summary is used individually for each pollutant.

The first type of data cube is oriented to air pollution data using the time and location dimensions: *{[idPollutant], [idMayoralty], [PolygonCity], [Time]}, {[Average(pollutant Measure)]}* where *[PolygonCity]* is the geometry of each city, extracted and transformed from the geospatial open data collected from INEGI (Mexican government agency of Statistics and Geography). The dimension *[Time]* contains four granularity levels: *[Year], [Month], [Day]* and *[Hour]*. The {<[pollutionMeasure]>} dimension is computed depending on the required level of analysis in time and/or space (e.g., <year> or/and <city>). Open datasets are integrated with information from 69 air monitoring stations, containing *<ID>, <location>, <altitude>, <operation status>*, etc. and attributes on sensed air pollutants such as:*<ID>, <name of pollutant>, <unit of measurement>*, etc.

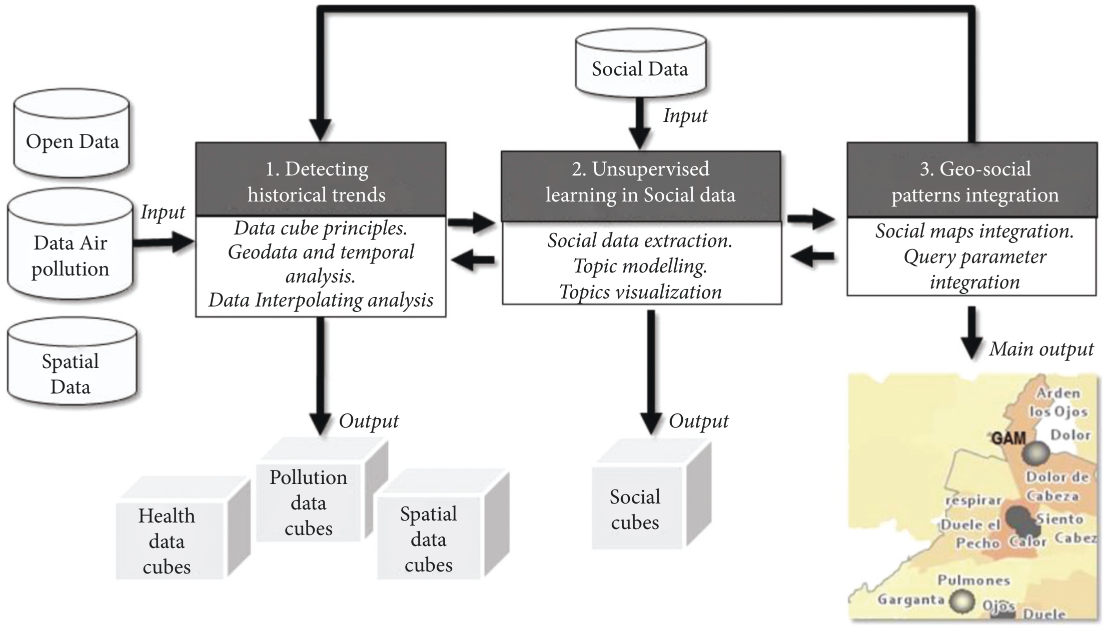


Figure 1: The geo-social methodology.

The second type of data cube (Spatial cube) is geared to respiratory diseases, with the following dimensions: *{< [ID_ICD-10]>, <[ID_Hospital]>, <[Year]>, <[Month]>, Count(∗)}*. Each respiratory disease describes where and when it was treated. The column <*[ID_CIE10]*> is a key that identifies the information about the respiratory disease (International Statistical Classification of Diseases(ICD)). While <*[ID_Hospital]*> gives the spatial criteria as it stores the hospital postal address. For instance, to identify areas with the highest number of diseases for a given period, the data cube is defined as follows: *{[ID_CIE10], [idCity], [PolygonCity], [Year], [Month]}, {[Count(∗)]}*. For example, the max number of "respiratory diseases" can aggregate records by "year" and "city."

The third type of data cube (Social Cube) is focused to social media; derived from collected tweets with the following attributes: <[idTweet]> <[Text]> <[Date-Time]>. When a tweet is published, additional information is progressively associated as followers can express some opinions about it: < [Likes]>, <[Shares]>, <[latitude], [longitude]>. The latitude/longitude coordinates give the location from where a tweet was fetched; this location is available when the set of tweets is extracted. This social cube will be complemented by additional new topics derived from the next stage.

Figure 2 shows the average values of PM10 and PM2.5 from 2003 to 2019 (i.e., the whole monitoring period) and shows the most polluted months in the Valley of Mexico. The PM values range from 0 to 70 micrograms per cubic meter (μg/m^3). According to the Mexico City government, from 0 to 50 μg/m^3 the air quality is good; from 51 to 100 μg/m^3, the quality is regular, and it is unhealthy in values higher than 100 μg/m^3. The historical trends show an increase from September to December (PM10: 60 μg/m^3; PM2.5: 30 μg/m^3) and a decrease from January to March (PM10: 50 μg/m^3; PM2,5: 25 μg/m^3). For the same period, the historic max values were in October (PM2.5:800 μg/m^3) and December (PM10: 1000 μg/m^3); there is another increment of PM10 levels in April (62 μg/m^3), May for PM2.5 (30 μg/m^3), the max values were in March (PM10: 1600 μg/m^3) and April (1000 μg/m^3). There are four months with good air quality concentrations in the Valley of Mexico.

In contrast to PM2.5 and PM10 trends, the highest average CO pollution levels raised from September (1.45 PPM (Particles per million)) to December (2.2 PPM) and then decreased from January (2.2 PPM) to June (1.45 PPM). The historic max values were in January (35 PPM), May (36.9 PPM), and December (35.40 PPM). The average values reflected a tolerable air quality that established a limit for the concentration in ambient air of 11 PPM for an average of 8 hours. Accordingly, September-December and January-May are critical periods to guide the social data extraction process from Twitter to search for a better impact understanding of air pollution. The main idea is to identify the relevant months to extract social data at stage 2 (i.e., unsupervised learning from social data).

3.1.2. Geographical Data and Temporal Analysis. The purpose is to visualize the pollution trends that emerge by considering health and social data integration with pollutant measures as they appear in space and time. A specific focus is made on a multi-criteria analysis of pollutant data CO, PM10, and PM2.5 in conjunction with respiratory diseases in the Valley of Mexico, such as "chronic obstructive pulmonary disease (COPD)," "acute lower respiratory illness (ALRI)," "cerebrovascular disease (CEV)," "ischemic heart disease (IHD)," "COPD and lung cancer (LC)." The aim is to explore their possible association, using data cubes to summarize (i.e., drill-up operation) 30 years of pollutant records.

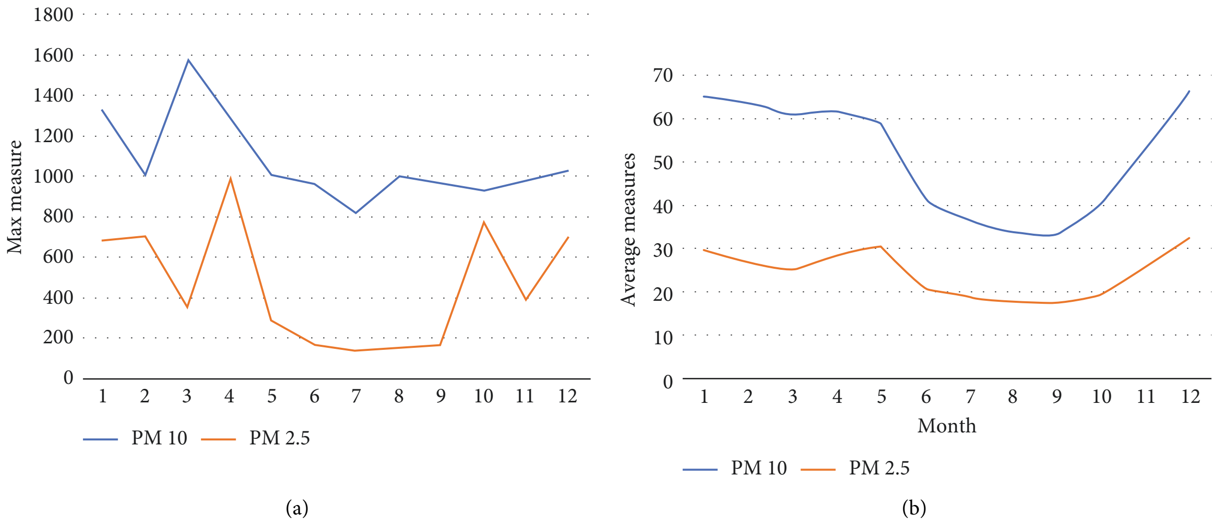


Figure 2: Monthly average and max values of PM10 and PM 2.5 from 2003 to 2019.

Spatial analysis is done by transforming tabular data into map layers [23]. The respiratory disease data, geographical boundaries, and pollutants values stored in data cubes are transformed into layers. Then, they are overlaid to observe some emerging patterns according to different time granularities over the last three decades (respiratory disease data is available over days from 2005 to 2015, CO pollutants data from 1990 to 2019, PM 10 from 1995 to 2019, and PM 2.5 from 2003 to 2019, social data is available from 2018 to 2019).

The process starts by averaging pollutant values by decade and at the mayoralty level, the data cube is as follows: *{[idPollutant], [PolygonCity], [Year]}, {[Average(pollutantMeasure)]}*.

The obtained data shows that during the 90 s decade, the maximum average values of CO remained stable in the municipalities or counties of the central-South region. In the 2000s-decade, average levels increased in the central region and some counties in the north of the Valley of Mexico, and in the north of the State of Mexico. As denoted by Figure 3(a), from 2010 to 2019, the maximum average values of CO (1.44 to 1.77 PPM) are in the central-north region. In contrast, the South shows intermediate and low values averages of CO (0.26 to 0.49 PPM). Likewise, from 1995 to 1999, maximum PM10 values were concentrated in the northeast of the Valley of Mexico. However, in the 2000s, this trend also appeared in additional counties located to the South and North. We found that PM10 concentrations have progressively increased in the east and center of Mexico Valley from 2010 to 2019, while CO pollutant has extended from the center to the North of Mexico Valley. The analysis is similar to the pollutant PM2.5 that was monitored from 2003 to 2019, although some mayoralties are not included in the monitoring air pollution network (Figures 3(c) 3(b)).

Pollution trends remain constant in the municipalities and mayoralties of the north-central region of Mexico Valley and match high-density population areas and the highest number of respiratory disease reports. It is worth mentioning that in Mexico, people are generally treated in hospitals far from their houses because the hospital designation criteria are made according to their ailment and not on the proximity of the hospital to their house. Therefore, large patient mobilities are generated (e.g., people who, from their house to the hospital, make a journey of up to two hours on average).

3.1.3. Pollutant Distribution Analysis. Spatial interpolation has been applied to estimate pollutant values at locations where no data appeared as monitoring stations were available. The purpose is to estimate and derive a surface pattern of the highest pollutant distribution concentrations using all pollutant values measured per date-time (yy/mm/dd, hours) over the last 30 years. We applied a kernel interpolation [24] that allows us to present the geographical and temporal distribution of these pollutants. It is worth mentioning that kernel interpolation has already been applied to interpolate environmental phenomena such as air pollution modeling [25]. The parameters used in the kernel interpolation are as follows:

(i) Input feature layer: Geolocated monitoring stations data layer (<"[pollutant measurement]">)

(ii) Value point: Pollutants measured per hour (<"[pollutantMeasure]">): PM10 (5.7 million records), PM2.5 (2.9 million rows), and CO (10.4 million records). For example, {Pollutant: PM10; Measure: 30; Date:2008-07-1; hour: 11}

(iii) Kernel function: Fifth-order polynomial function.

The results of kernel interpolation identify homogeneous surface patterns that denote a continuous intensity of air pollution concentration. Figure 4 shows the spatial distribution of pollutant values of PM10, PM2.5, and CO as derived from hourly pollutant values.

The patterns reflected by PM10 and PM2.5 are relatively similar, extending from the northeast and northwest toward the center region, while the minimum values appear in the center. There is an interesting pattern of CO values: the highest values are increasing from the center to the North of the Valley of Mexico, possibly due to the high population

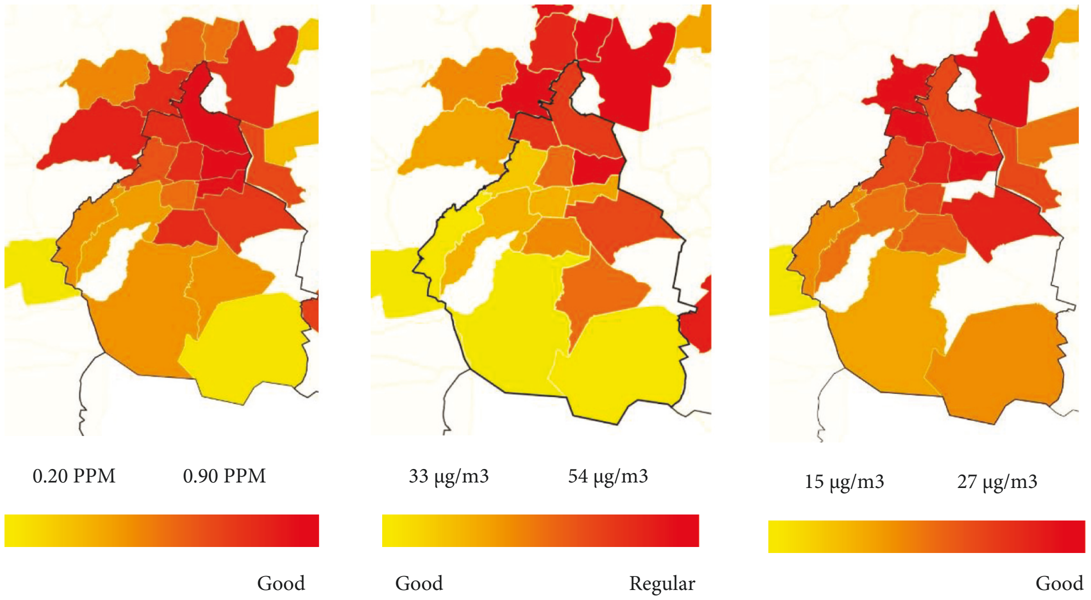


Figure 3: CO, PM10, and PM2.5 average levels per mayoralty and municipality. (a) 2010–2019 (CO), (b) 2010–2019 (PM10), (c) 2010– 2019 (PM2.5).

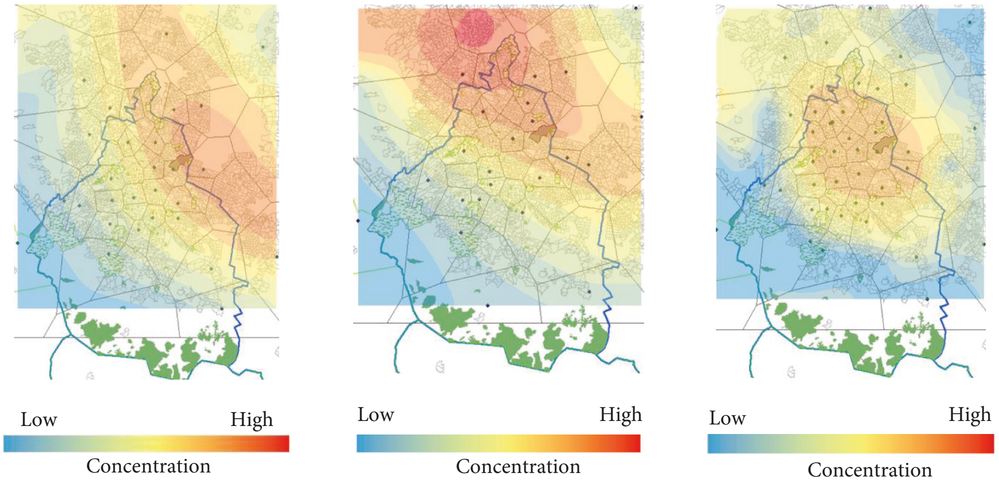


Figure 4: Spatial distributions of PM10, PM 2.5, and CO. (a) PM10 (1995–2019). (b) PM2.5 (2003–2019). (c) CO (1990–2019).

density in this area and heavy transportation infrastructure. A considerable concentration of respiratory diseases could be related to this. In addition, the patterns are helpful in identifying where to extract data from social networks. This is a sort of pattern that needs to be enriched by social discussions related to this region; to understand in a better way why the pollutants are dispersed in these directions and how people react to it (visualization in an integrated way of social-health insights and affected regions by one or more pollutants).

3.2. Stage 2: Unsupervised Learning from Social Data. The goal is to detect relevant Spatio-temporal topics from Twitter social narratives. An unsupervised learning approach identifies which relevant topics are reflected in the tweets. The first task is to identify how many tweets belong to each identified topic (e.g., "headache" topic groups 3% of the tweet dataset). For instance, a dataset of thousands of tweets can be classified into several topics and a high number of categories. This reflects one of the main challenges of our study: to find out what and how many topics categories could be contained in the whole dataset. This stage is divided into three tasks: data extraction, discovery, and topic visualization.

3.2.1. Data Extraction. The data extraction process applied to the Twitter network uses historical trends data, as obtained in the previous phase, as search parameters (where and when) to refine the temporal period and spatial extent to explore (e.g., September-December, north-center). These search parameters provide a starting reference for orientating the social data extraction. Pollutant data is monitored monthly over September-December and January-May as concentrations are higher during these periods.

The first extraction challenge is to define a suitable search parameter (what to extract) that can identify many tweets related to causes or effects of air pollutants (e.g., concentration of traffics, excessive and illegal use of fireworks, health and environmental effects). We collected more than 38,000 geo-referenced tweets from 2018 to 2019. The extraction process runs on a semi-automatic web process that collects and monitors narratives about air pollution. The extraction search parameters are keywords, hashtags, terms, location, and an extraction buffer (i.e., 5 to 10 kilometers to cover the largest area of each municipality or city). The extraction process still requires to be refined to extract additional tweets that could be directly or indirectly associated with air pollution. The next section addresses this.

3.2.2. Discovering Topics. This task attempts to discover the underlying semantic structure in text datasets to identify recurring patterns; these are called topics. The patterns may or may not correspond to our intuitive notion of a topic. However, they are useful for analyzing relationships between concepts contained in a set of tweets.

From the tweet datasets extracted from the previous step, the next step is to discover relevant topics from narratives. For instance, "headache" is a term or topic that can be associated with a group of tweets that collect people's opinions about the health impact of exposure to high air pollutant concentrations. Furthermore, these topics might be associated with additional words (e.g., bi-gram or tri-gram terms) that can reflect some instances of pollution levels.

We retained a Topic Modeling technique previously applied to discover relevant topics introduced in related work [26]. This approach is a common mechanism in NLP, and it is based on statistics and linguistics inference. The main input is a set of tweets, and the output is a list of terms or topics where each topic represents a group of tweets. Let us describe the whole process that combines NLP techniques like bigram and trigram extraction, the K-means algorithm for textual clustering, and Latent Dirichlet Allocation (LDA) algorithm for Topic Modeling. The design process has been inspired by a series of related works where K-means and LDA algorithms have been mixed successfully in similar text mining scenarios [27–29].

Our process architecture is illustrated in Listing 1. The output includes a list of topics (e.g., a name identified during the LDA process), cluster size, and tweet IDs. The overall sequence is as follows: (1) setting hyperparameters; for example, the "K" parameter is the optimal number of topics that are defined after trial and error of many LDA models executions evaluating the performance (setting a "K" value to avoid repeating the same terms in different topics that offer meaningful and interpretable topics [30]); (2) defining stop words, necessary terms and symbols to be eliminated during analyses at each execution step (e.g., bad words, repeated terms, key-terms for data extraction); (3) running K-means; (4) running LDA; (5) finally, the discovered topics are used to create social data cubes for further visualization.

In order to visualize the topics in space and time, the next step extends the structure of the social cube (mentioned in section 4.1) by the topic modeling results. This is done by the dimensions " <*[TopicName]*>" that identifies the topic name, and <*[ClusterSize]*> that denotes the number of tweets grouped by topic. The "location" and "date" of the tweets that belong to the topic are included (i.e., information obtained when the tweet is extracted). The principle behind this is to cross-related the data to connect each cluster with the tweet's information. The final structure considers the following dimensions: [*Topic*], *[Date-Time], [Latitude], [Longitude], {Max(ClusterSize)}*.

This social cube organizes the topics by date and/or location. For example, to compute what is the max size of the tweets with the topic "headache" grouped for a specific location, or at what times the term "headache" appeared per month in 2019. This can be done for each category identified as a key term. In general, this highlights the main topics that appear in social data, when pollutant concentrations increase or decrease, and what other events are involved during this temporal frame.

From the results obtained, and as a sort of recursive process, the search process can be refined at the data extraction level introduced in the previous section. This might help the extraction process using better search parameters. For instance, while the initial extraction process was executed in 2018, we obtained the terms "air pollution" and "chest pain " which significantly appeared in the topic extraction process. This term, in fact, has been used by people as a local and popular expression to describe a symptom caused by exposure to high levels of air pollutants. In 2019, the search process was improved by including the term "chest pain" as a search parameter. The results of this data extraction generated a more specialized set of tweets related to air pollution events.

3.2.3. Topics Visualization. The purpose of this task is to visualize the discovered topics. A common technique that helps to do this is the word cloud technique [31]. We found that the topic "pollution" ("contaminación" in Spanish) is the larger one: 27% of the tweets collected represent some textual narratives associated with air pollution. While the term "traffic" ("trafico" in Spanish) corresponds to 11% of all gathered tweets, and "health" ("salud" in Spanish) represents 20% of the gathered tweets.

The "rain" term (in Spanish Lluvia") represents 4% of gathered tweets. This reflects the fact that in May 2019 in Mexico City, there was an environmental contingency (i.e., the highest concentration of pollutants). This forced the Mexican government to restrict the circulation of most cars and industry activities, coupled with the fact that the weather at this time is the hottest of the year. But two days of rain caused people to express that rain helps reduce the concentration of pollutants. Over these two days, this rain topic was the major one discussed and no longer the high index of pollutant concentration.

There are other categories of minor topics, such as "headache" ("dolor de Cabeza"), "chest pain" ("dolor de pecho"), "burning eyes" ("ardor ojos"), "headache" (dolor de

```
Input: tweets_dataset_input
   /* Setting parameters */
   K_clusters = 20
   Stop_words = {“http: //”, “https://”,”@”,”#”. . .}
   N-Gram_range = {1,4}
   LDA_passes = 10
Begin
   input = remove_stopwords(filtered_input)
   Corpus = gensim.CreatingN-Grams(input,N-Gram_range)
   Clusters = K-Means (Corpus, K_clusters)
   SocialTopics = LDA_process_gensim(Clusters, K_clusters,LDA_passes)
   Output:SocialTopics_Set
Call:
   CreatingSocialDataCubes(SocialTopics_Set)
```

Listing 1: Discovering social topics - algorithm.

Cabeza), “eyes” (“Ojos”) among others. Together, they represent about 7% of recollected tweets and describe tweet views related to physical terms such as the effects of air pollution.

Finally, a significant trend that appears is that the importance and extent of the topics related to air pollution constantly evolve, especially when some specific events related to new terms appear (e.g., rain or Popocatepetl volcano). Discussions on air pollution are taken into account and are considered as different degrees of “*active*” or “*inactive*” when this is more or less the case. Figure 5 shows to which degree of magnitude emerged the social topics associated with air pollution in 2019.

For instance, it appears that 11% of the tweets collected belong to the group of the topic “traffic,” a topic is described by different events and contexts, and that emerged in 2019. “traffic” and “pollution” topics emerged with magnitude differences. The topics “environmental contingency” and “pollution” emerged from February to September, and that coincided with an increase in PM10, PM2.5, and CO values from high to low. The topic “rain” emerged from February to May. It reappears in September as “acid rain” from tweets that describe poor water quality and “heavy rain” this matches with the rainy season in Mexico City.

Summing up, we discovered at least three relevant social topics with the “air pollution” phenomenon. They are: “health,” “pyrotechnics,” and “mass Traffic” that can be refined when performing social narrative extractions. In order to highlight the topic trends from a geographical perspective, another social cube is materialized by location. It then appears that topics with large cluster sizes are concentrated in the north-center of Mexico Valle, where there are higher pollutant concentrations and the highest index of the population. The emerging topics found are “pollution,” “headache,” “pyrotechnics,” “traffic,” and other respiratory diseases terms.

3.3. Stage 3: Social Patterns in Space and Time. The purpose is to derive insights that reflect geographic, temporal, and social patterns. This is done by analyzing trends and patterns discovered in the previous stages; through word clouds, social maps, and overall by applying GIS overlays. The “Social map” concept is introduced to overlay geographical topic distributions with other data layers (e.g., demography, respiratory diseases, air pollution surfaces concentration). The overlay operation can be applied to other events that have occurred in the same location and at different times. For example, the layers of respiratory symptoms topics are overlaid with “respiratory disease cases” and “pollution interpolation” layers. The overall process is similarly applied to other data layers. In order to interpret geo-social patterns, the topics and trends discovered in phase 1 can be used as input query parameters for cross-analyzing tweets with open data.

4. Experiments and Results

The objective of the experiments is to provide a geo-social characterization of the health impact of air pollution phenomenon using open data sources and social narratives. The main discovered insight is that social discourses reflect the air pollution phenomenon vary not only in function of the pollutant concentrations but also encompass a more general behavioral pattern that combines a series of additional dimensions and factors. The emerging trends reflect the roles of the population distribution, major social events, governmental policies, and additional environmental factors.

The approach is oriented toward the three most important general categories of topics that appear: “health,” “pyrotechnics,” and “traffic.” These topics highly correlate with concentrations of CO, PM10, and PM2.5 pollutants, which are distributed around the northeast of the Valley of Mexico. The main topics that emerge in the north of the Valley of Mexico from July to September are: “health (breathe, throat),” “stress,” and “traffic.” Also, “traffic” from April to June, “stress and traffic” from July to September, and finally, from October to December, “health topics” (e.g., breathing, throat), “pyrotechnics,” and “traffic” topics appeared.

The pollutant concentrations are compared with open data vs. social data patterns. The topics illustrated in Figure 6 show regular air quality values, but the highest values were in

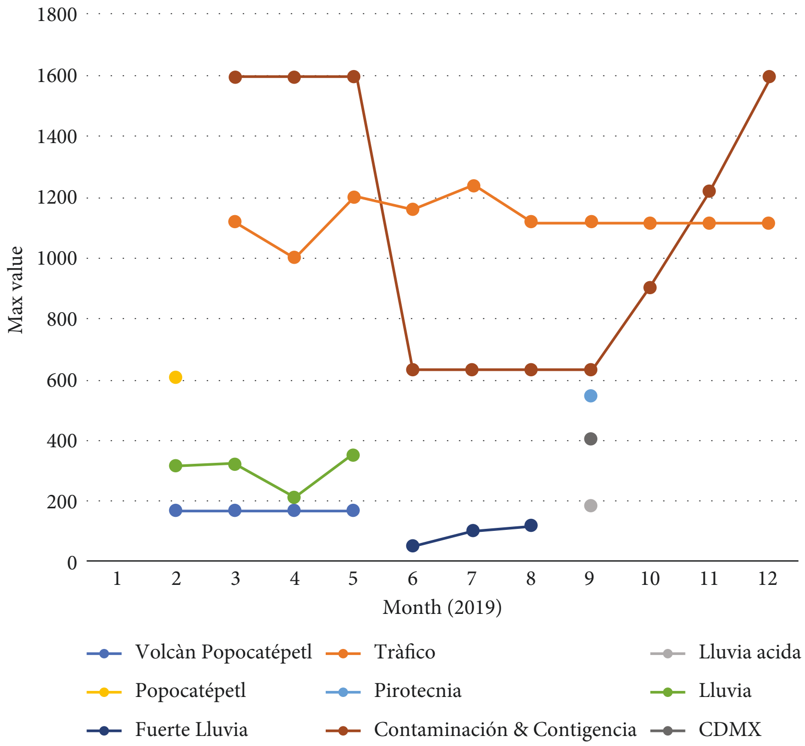
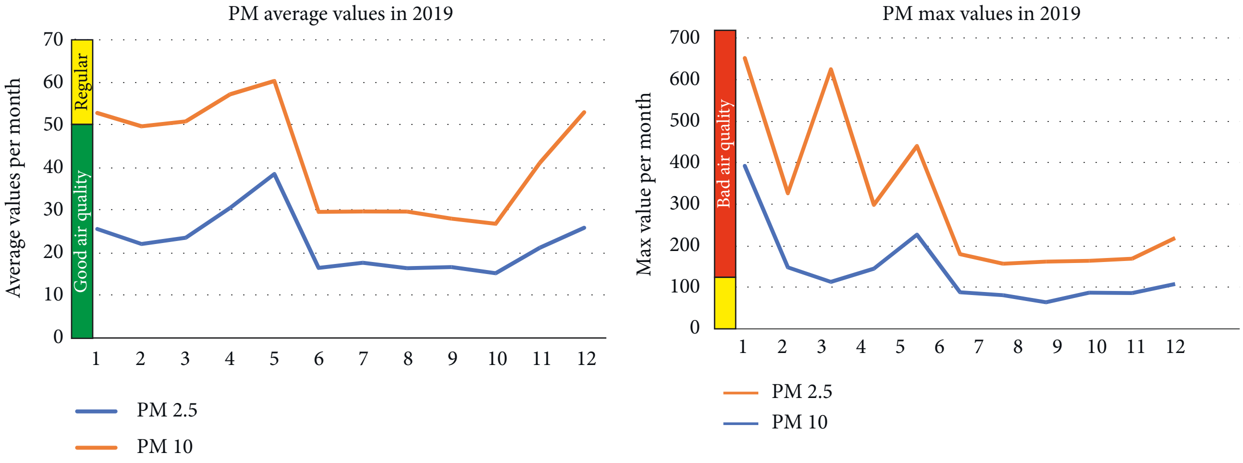


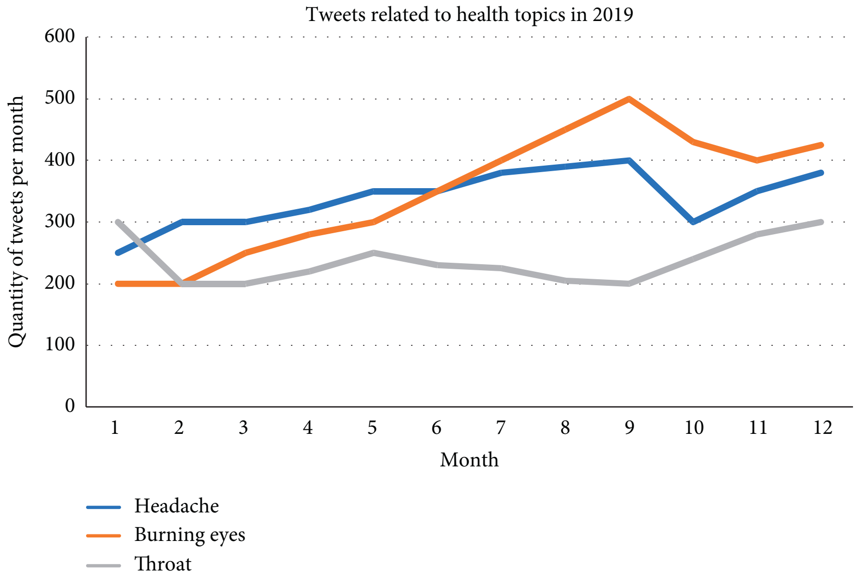
Figure 5: Social topics associated with air pollution in 2019.



(a)

(b)

Figure 6: Continued.



(c)

Figure 6: Health topics in comparison with PM trends in 2019.

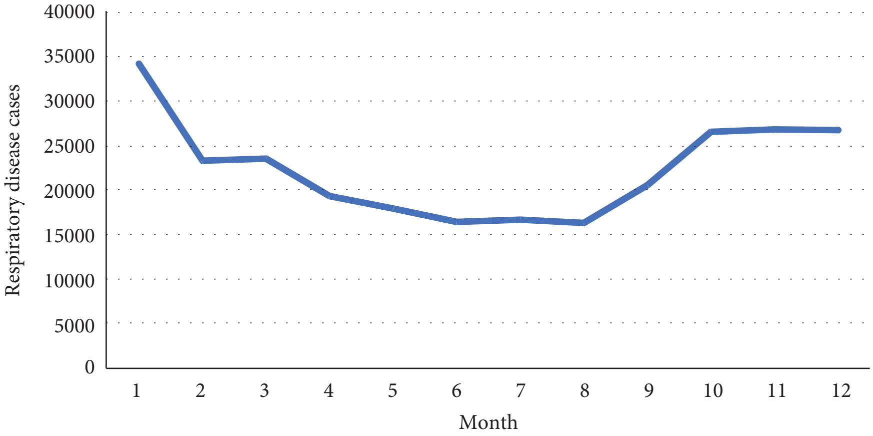


Figure 7: Respiratory disease cases per month (2005–2015).

January (PM10: 653 μg/m^3, PM2.5: 393 μg/m^3), May (PM10: 626 μg/m^3), and December (PM10: 359 μg/m^3, PM2.5: 303 μg/m^3). This could explain why people suffer physical discomfort due to bad air quality during this period. That is, between regular and bad air quality levels. Figure 6 shows some health impacts of air pollution such as "headache," "burning eyes," and "throat" topics as they emerged when PM10 and PM2.5 values increased from October to December (there is also a significant increase when pyrotechnics also emerged as a topic). Overall, "burning eyes" is the most popular impact of air pollution as reflected by social data patterns. The mentioned topics also emerged in the central, north, and east of the Valley of Mexico.

The next experiment considered health-related topics associated with pollution narratives: "nose," "eye," "head," "chest," "throat," "chest pain," "watery eyes," "headache" and "burning eyes." These human body parts or respiratory symptoms are often associated with physical discomfort caused by air pollution. These topics emerged in locations where there are large reports of respiratory diseases and during the most polluted months (January, February, March, May, October, and December). In addition, Figure 7 shows a historical increase in respiratory disease cases from October to December, considering only 10 years of available data (2005–2015) provided by the Mexican government.

Figure 8 (a) overlays respiratory diseases and health topics. The topic location of each topic is given by the tweet locations, while the topic color intensity denotes the topic magnitude. According to the last 5 years of respiratory diseases data (2005–2015), there is a relevant number in the northern cities of the Valley of Mexico. For example that in the East in "Ciudad Nezahualcóyotl" 3738, the counties like "Cuauhtémoc" 14250 and "Gustavo A. Madero (GAM)" 11835 cases of respiratory diseases converged with respiratory symptoms topics such as "headache" ("dolor de cabeza"), "eyes" ("ojos") and "throat" ("garganta").

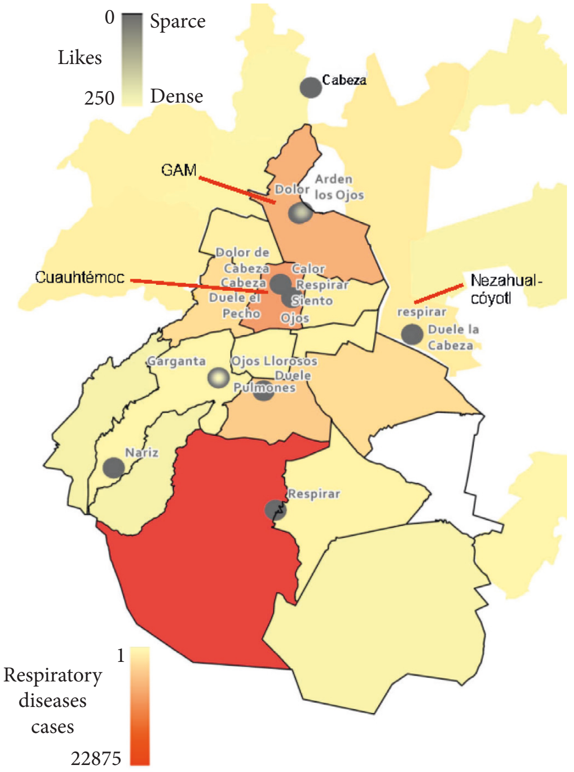


Figure 8: Health topics and respiratory diseases (2010–2015).

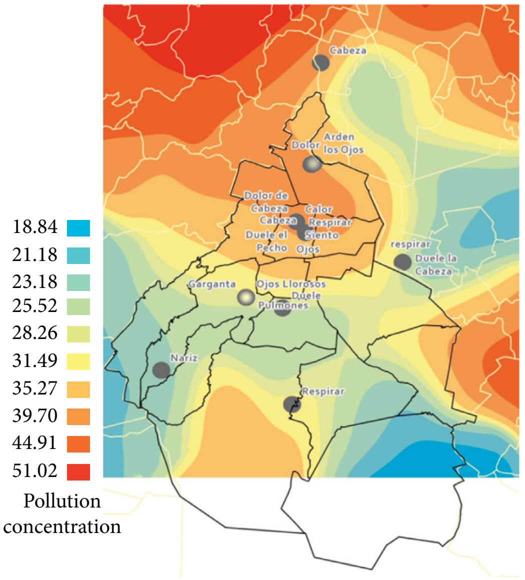


Figure 9: Health topics and pollutants concentration in 2019 (CO, PM 10, and PM 2.5).

Table 1: List of acronyms.

Full form	Acronym
Acute lower respiratory illness	ALRI
Application programming interface	API
Carbon monoxide	CO
Cerebrovascular disease	CEV
Chronic obstructive pulmonary disease	COPD
Geographic information system	GIS
Gustavo A. Madero (municipality name)	GAM
International statistical classification of diseases	ICD
Ischemic heart disease	IHD
Latent dirichlet allocation (topic modeling algorithm)	LDA
Lung cancer	LC
Mexican government agency of statistics and geography	INEGI
Natural language processing	NLP
Nitrogen dioxide	NO2
Ozone	O3
Particulate matter	PM

Figure 9 shows an overlay between health symptoms topics and interpolation surface of PM 10, PM 2.5, and CO pollutants concentration in 2019. The source describes the concentration of these pollutants in the north of Mexico. The locations of respiratory symptoms: “breathe” (“respirar”), “headache” (“dolor de cabeza”), “chest pain” (“duele el pecho”), “headache” (cabeza) and “burning eyes” (“arden los ojos”) are in regions where the air quality is regular or bad (orange and red color). These topics represent groups of citizen narratives that describe respiratory symptoms and high pollution levels.

Summing up, the most recurrent social discourses associated with air pollution emerged when related to the following facts: the health impact of the air pollution phenomenon.

(1) Health” under the combination with a high index of the population in the area, a high index of pollutants concentration, rainy or non-rainy periods (South and North).

(2) Health patterns appear, not surprisingly, as strongly correlated with the distribution of population and pollutants and under the influence of air pollution conditions.

(3) Respiratory symptoms emerge in social networks when respiratory disease cases increase in hospitals.

Full forms and acronyms mentioned in the study are shown in Table 1.

5. Conclusions and Future Work

The methodology and computational approach presented in this study combine sensor-based data with social media narratives in order to describe and study the negative effects of air pollution in the Valley of Mexico. By combining social networks and open data, the study revealed a series of health-geo-social patterns associated with the impact of the air pollution phenomenon. The social media narratives related to air pollution reveal the main topics that emerge when high levels of PM10, PM2.5, and CO arise. In general, the approach can also be used as a monitoring and predictive mechanism to anticipate some pollution patterns and thus enable some countermeasures.

The fundamental principle behind our approach is to extract social narrative patterns that reflect peoples' perception of an air pollution phenomenon at the scale of a large city and over time. For example, by considering the historically most polluted months and regions obtained by mining 30 years of data, search terms are refined by observing the behavior narratives that emerged from the tweets gathered in a 2-year period. Our approach shows how social and open data complement each other to describe the impact of an environmental phenomenon, from the regional level to the personal level, and when combined with government and open data on respiratory disease cases to contrast health findings as they merged from social networks.

The geo-social characterization of air pollution is derived from an implementable workflow framework by applying data mining and GIS methods whenever necessary and NLP techniques. We introduced a social characterization of an air pollution phenomenon that describes the Spatio-temporal dynamics of social narratives related to air quality impacts. An additional contribution is a data exploration process by topics and trends, which are further used as smart parameters to design data exploration mechanisms that support Spatio-temporal cross-analysis of tweets and open data. Another interest of the approach is that a given topic emerges not only when some associated patterns arise but also when some triggering precursor conditions or circumstances are likely to activate it.

Our framework combines traditional techniques from Natural Language Processing (Topic Modeling), Unsupervised Learning (Clustering), Geographic Information Systems (spatial interpolation and choropleth maps), and Data Mining (data cubes modeling) to integrate the understanding of air pollution impact knowledge from regional to individual granularity. Our work can be improved by considering new methods for Spatio-temporal topic tracking, sentiment analysis, deep learning, semantic modeling with ontologies to improve the extraction of social data, and new big data technologies.

The whole methodological workflow framework that combines three complementary resources (i.e., text analysis, spatial data mining, topic modeling) has the advantage of being reproducible and applied to other phenomena at different spatial and temporal scales. Future work will be focused on integrating deep learning to analyze tweet patterns and study the health and social impacts of the COVID-19 pandemic associated with air pollution patterns.

Data Availability

The data sets used in this work (structured data and gattered tweets) are available on http://antacom.org.mx/opendata/airpollutioncdmx.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors want to thank God, CONACYT (PROJECT 7051), SIP (SIP-20220557), IPN, COFAA, UPIITA-IPN, and ESCOM-IPN for their support. Dedicated to the memory of Roberto Zagal (father). The authors want to thank CONACYT (PROJECT 7051), SIP (SIP-20220557), IPN, COFAA, UAQ, CIC-IPN, UPIITA-IPN, and ESCOM-IPN for their support.

References

[1] G. Chen, S. Li, Y. Zhang et al., "Effects of ambient PM 1 air pollution on daily emergency hospital visits in China: an epidemiological study," *The Lancet Planetary Health*, vol. 1, no. 6, pp. e221–e229, 2017.

[2] J. Yang, B. Shi, Y. Zheng, Y. Shi, and G. Xia, "Urban form and air pollution disperse: key indexes and mitigation strategies," *Sustainable Cities and Society*, vol. 57, Article ID 101955, 2020.

[3] J. Burkhardt, J. Bayham, A. Wilson et al., "The effect of pollution on crime: evidence from data on particulate matter and ozone," *Journal of Environmental Economics and Management*, vol. 98, Article ID 102267, 2019.

[4] D. Sui and M. Goodchild, "The convergence of GIS and social media: challenges for GIScience," *International Journal of Geographical Information Science*, vol. 25, no. 11, pp. 1737–1748, 2011.

[5] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*, Springer Science & Business Media, vol. 12, 2012.

[6] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 1–55, 2014.

[7] R. Li, Z. Wang, L. Cui et al., "Air pollution characteristics in China during 2015-2016: spatiotemporal variations and key meteorological factors," *Science of the Total Environment*, vol. 648, pp. 902–915, 2019.

[8] W. F. Ye, Z. Y. Ma, X. Z. Ha, H. C. Yang, and Z. X. Weng, "Spatiotemporal patterns and spatial clustering characteristics of air quality in China: a city level analysis," *Ecological Indicators*, vol. 91, pp. 523–530, 2018.

[9] S. Pu, Z. Shao, M. Fang et al., "Spatial distribution of the public's risk perception for air pollution: a nationwide study in China," *Science of the Total Environment*, vol. 655, pp. 454–462, 2019.

[10] A. Phosri, K. Ueda, V. L. H. Phung, B. Tawatsupa, A. Honda, and H. Takano, "Effects of ambient air pollution on daily hospital admissions for respiratory and cardiovascular diseases in Bangkok, Thailand," *Science of the Total Environment*, vol. 651, pp. 1144–1153, 2019.

[11] G. Du, K. J. Shin, and S. Managi, "Variability in impact of air pollution on subjective well-being," *Atmospheric Environment*, vol. 183, pp. 175–208, 2018.

[12] I. Kulhánová, X. Morelli, A. Le Tertre et al., "The fraction of lung cancer incidence attributable to fine particulate air pollution in France: impact of spatial resolution of air pollution models," *Environment International*, vol. 121, pp. 1079–1086, 2018.

[13] V. K. Singh, M. Gao, and R. Jain, "Situation recognition: an evolving problem for heterogeneous dynamic big multimedia

data," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1209–1218, 2012, October.

[14] F. Yao and Y. Wang, "Tracking urban geo-topics based on dynamic topic model," *Computers, Environment and Urban Systems*, vol. 79, Article ID 101419, 2020.

[15] M. Sammarco, R. Tse, G. Pau, and G. Marfia, "Using geosocial search for urban air pollution monitoring," *Pervasive and Mobile Computing*, vol. 35, pp. 15–31, 2017.

[16] Air Quality in Mexico City, http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBh%27, 2022.

[17] D. Vallero, *Fundamentals of Air Pollution*, Academic Press, 2014.

[18] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, no. 7569, pp. 367–371, 2015.

[19] General Directorate of Health Information in Mexico, http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html, 2022.

[20] R. Zagal, F. Mata, and C Claramunt, *Recollected Tweets and Open Data Integrated Records: A Data Sample for Mexico City*, http://antacom.org.mx/opendata/airpollutioncdmx.html, 2020.

[21] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining Text Data*, pp. 77–128, Springer, Boston, MA, 2012.

[22] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[23] N. Andrienko and G. Andrienko, "Exploratory analysis of spatial and temporal data: a systematic approach," *Springer Science & Business Media*, 2006.

[24] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications*, Routledge, 2018.

[25] M. Goutham, S. Jayalakshmi, and R. Samundeeswari, "A study on comparison of interpolation techniques for air pollution modelling," *Indian Journal of Scientific Research*, pp. 2250–0138, ISSN, 2018.

[26] J. Kawash, N. Agarwal, and T. Özyer, *Prediction and Inference from Social Networks and Social Media*, Springer International Publishing, 2017.

[27] A. N. Srivastava and M. Sahami, *Text Mining: Classification, Clustering, and Applications*, CRC Press, 2009.

[28] X. Chen, X. Zhou, T. Sellis, and X. Li, "Social event detection with retweeting behavior correlation," *Expert Systems with Applications*, vol. 114, pp. 516–523, 2018.

[29] J. L. P. Barker and C. J. Macleod, "Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities," *Environmental Modelling & Software*, vol. 115, pp. 213–227, 2019.

[30] M. Hofmann and A. Chisholm, *Text Mining and Visualization: Case Studies Using Open-Source Tools*, CRC Press, vol. 40, 2016.

[31] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski, and S. Franconeri, "An evaluation of semantically grouped word cloud designs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 9, pp. 2748–2761, 2020.