Visual Analysis for Exploring the Relation between Air Pollution, Environmental Factors and Respiratory Diseases

Ning Yu[1], Mingwen Zheng[2], Ximena Andrade[3], Ron Patane[4]
{nyu@brockport.edu[1], mingwen1986@gmail.com[2], andradepatricia2017@outlook.com[3], rpatane@uscupstate.edu[4]}

State University of New York College at Brockport, 350 New Campus Drive, Brockport, NY, USA 14420[1], Resurgent Capital Services, 55 Beattie Place, Greenville, SC USA 29601[2], Greenville Health System, 701 Grove Rd, Greenville, SC USA29605[3], University of South Carolina Upstate, 800 University Way, Spartanburg, SC USA 29303[4]

Abstract. Using visualization tool for data analytics makes the complex data more understandable so that some inherent patterns and relationships can be revealed more efficiently and clearly. This paper provides a such paradigm for visual analysis on public health issues. By exploiting the advanced visualization tool based on cloud and artificial intelligence, we collect and analyze several data sets on PM2.5 air pollution, environments and public health in Beijing in order to discover the intrisic relations between those factors. The visual analytics illustrates that the incidences of respiratory diseases increase as the concentration of PM2.5 in the air increases and that a strong correlation exists to temperature. PM 2.5 concentrations are higher in cold winter months than in hot summer months. Meanwhile, an unexpected discovery from the investigation was the pattern found for air concentrations of PM2.5 throughout a 24-hour period. The levels of PM2.5 reach peak at midnight and drop to their lowest level at 2pm in the daytime. Most importantly, there exists a direct positive correlation between the number of emergency room visits for respiratory diseases and high levels of PM2.5 in the atmosphere. These interesting findings can provide direct supports for strategy/decision making in public health and emergency administration.

Keywords: Visualization Tools, Visual Analysis, PM2.5, Air Pollution, Temperature, Respiratory Diseases

1 Introduction

China's burgeoning economy is fueled by worldwide demand for manufactured goods. It became the world's top exporter in 2009 [1]. The United States is the top consumer of China's manufactured goods in the world. In 2015, the U.S. imported 18% of China's total exports [2]. These economic circumstances have environmental implications for both nations. Just as China's economy has grown, so have its environmental problems. One of the major areas of public concern is air quality. High levels of pollutants in the atmosphere are the result of industrialization and urbanization [3]. Urbanization increases air pollution as a result of an increased number of vehicles in transit and the use of fossil fuels for residential heating. Most of the pollutants, however, come from industries, power plants and biomass burning [4]. All of these sources of air pollution release particulate matter, making it the major pollutant in China's

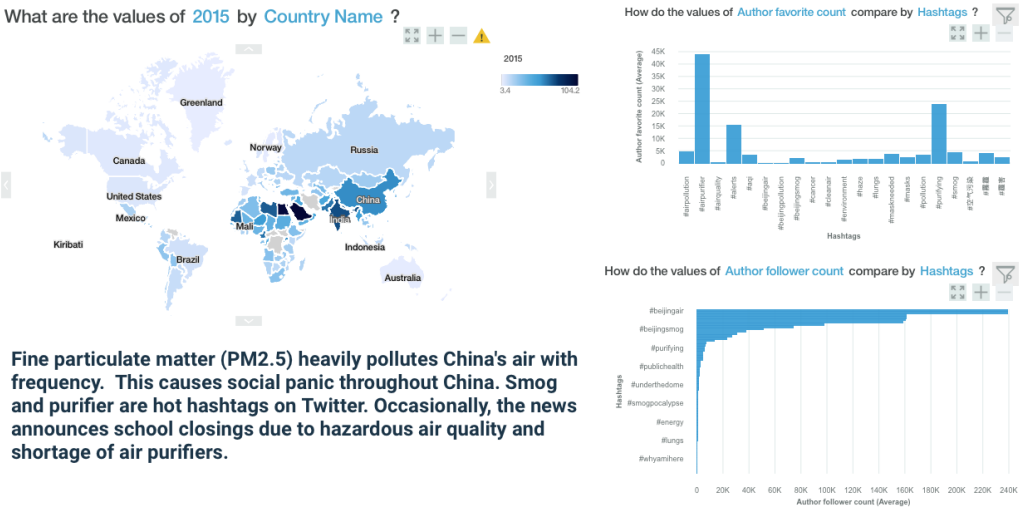
MOBIMEDIA 2018, June 21-22, Qingdao, People's Republic of China

DOI 10.4108/eai.21-6-2018.2276626

metropoles and contributing to some of the worst air quality in the world (Figure 1). Beijing is one such city with air quality that has been heavily impacted by the proximity of industries and urbanization.

Those affected by China's air pollution problems, however, are not limited to residents of China. A study conducted in 2013 concluded that the United States not only receives manufactured goods from China, but also their pollution. Polluted air originating in China is carried by westerly winds to the west coast of the United States and travels as far its eastern states [5]. This foreign pollution reduces the effectiveness of many environmental efforts in the U.S.

This investigation sought to study the correlation between emergency room visits for respiratory diseases in Beijing and levels of PM2.5, and how environmental factors such as temperature affect them. Also, we study the social media data to explore the sentiment and the concerns of the residents in Beijing regarding the quality of the air.



Fine particulate matter (PM2.5) heavily pollutes China's air with frequency. This causes social panic throughout China. Smog and purifier are hot hashtags on Twitter. Occasionally, the news announces school closings due to hazardous air quality and shortage of air purifiers.

Figure 1. Introductory dashboard. Here is a geospatial visualization allowing the exploration of 2015 PM2.5 levels worldwide. It is easy to see which nations had the highest and lowest levels of PM2.5 contamination. The dashboard also provides visualizations of the number of hashtags related to air pollution.

2 Method

We collected datasets pertaining to Beijing for three primary categories, air pollution (PM2.5 concentration), environmental factors (temperature and season change), and respiratory diseases (emergency room visits- total and disease specific). Additionally, worldwide exposure data for PM2.5 was obtained. Dataset details and source are as follows: (1) The World Bank [6]: Dataset contains world PM2.5 air pollution, mean annual exposure in the measurement of micrograms per cubic meter (μg/m3). (2) U.S. Department of State Air Quality Monitoring Program - Mission China [7]: Dataset contains PM2.5 concentration (μg/m3) for Beijing by year, month, and hour. The Air Quality Index (AQI) (Figure 2) was consulted to interpret the PM2.5 (μg/m3) values from this dataset. This website yielded data for the years 2013, 2014,

2015, and 2016. (3) PLOS ONE research [8]: Data set contains air pollution levels (PM2.5, SO2, CO2, NO2, O3), meteorological data (temperature and relative humidity) and emergency room visits data for Beijing for the year 2013. This dataset also contains the number of emergency room visits (ERV) for upper respiratory tract infections (URTI), lower respiratory tract infections (LRTI- which refers to bronchitis or pneumonia), asthma, influenza, and acute exacerbations of chronic obstructive pulmonary disease (AECOPD). (4) Social media data sets [9]: IBM Watson provides the capability to import data from various social media platforms. We selected to search for hash tags related to air pollution in Beijing. Twitter data from 1/1/2015 to 12/31/2016 with the #Beijing, #smog, and #purifier was obtained and analyzed. Twitter data was unavailable for earlier years.

The World Bank, Mission China and PLOS ONE datasets were preprocessed for Watson analytics. Data cleaning and preprocessing included removal of formatting, invalid data, redundant data, and integration of yearly data by joining tables. All files were loaded into Watson for analysis.

Watson Analytics is an advanced data analysis and visualization tool that we use in this study with guided data discovery and cognitive abilities based on the state-of-the-art techniques in artificial intelligence and cloud computing. It is able to provide powerful features on automated predictive analytics and visualization over a variety of data representations. Moreover, Watson Analytics can simplify and accelerate the procedures of data preprocessing and preparation.

Watson Analytics offers a wide range of analytic tools and predictive capabilities. For example, the Data tab shows the imported datasets and provides a score to determine the quality of the data. Within the Data tab, further refinement of the data prior to analysis is possible. Watson's refinement capability was utilized to categorize the yearly PM2.5 data based on the Air Quality Index (AQI). The ability to use Watson's refining tool simplified interpretation of the data in relation to the AQI. The tools within the Discover tab provide the ability to formulate questions and select graphical representations. The wide selection of graphs made it possible to select the best visualizations for the insights acquired. The Display tab provided the ability for organizing the discoveries into dashboards. Five dashboards were built in this study for that the visualizations best represent the insights extracted from the data. Watson Analytics is a perfect tool for making comparisons and predictive analysis, finding trends and relationships, and geospatial mapping.

3 Analytical Results

The analysis produced by advanced visualization tools such as Watson Analytics demonstrated a strong correlation between the three parameters: PM2.5 and other pollutants, environment factors such as season and temperature, and emergency room visits (ERV) related to respiratory diseases. The highest number of respiratory diseases occur when the temperature is lowest and the air quality is most hazardous (Figure 2). We found from the analysis that January, February, March and especially the months of November and December, experience increased numbers of ERVs for respiratory diseases. Similarly, the concentrations of PM2.5 and CO, NO2, and SO2 are also highest during these months. It is important to note that ozone is different. Ozone reaches its highest levels during summer months. This is likely due to the catalysis by summer heat and sunlight that increase ozone forming reactions [10]. Research of the literature indicated that fossil fuels, which release PM2.5, are used for residential heating.

Increased use of fossil fuels for heating during winter months helps explain the rise of PM2.5 and other air pollutants as temperature decreases. The possibility that additional factors may influence the levels of airborne pollutants during certain times of the year cannot be dismissed.

Visualization tools analyzed the data by different measures of time by year, month, and hour. The results show that the majority of days in 2013 fell at or above the AQI categories of "Unhealthy for All". The analysis of PM2.5 levels by year show u-shaped curves (Figure 2). The hourly analysis shows an increase in PM2.5 levels between 6pm and 3am (Figure 2). The graph, displayed in Figure 3, shows rising levels of other airborne pollutants from January to March and November to December. The predictive model in Figure 4 denotes a positive correlation between CO and PM2.5. The last analysis, shown in Figure 4, shows a downward trend from 2013 to 2016 in the levels of PM2.5.

This set of analysis pertain to ERVs for respiratory diseases. The data was analyzed to show how the incidence of five respiratory diseases are reflected in the number of ERVs over a year. In Figure 3, it shows that incidences of URTI, LRTI, and AECOPD increase and decrease at the same times of the year. Asthma's incidence is the highest for most of the year, but its levels peak during the summer months. In Figures 4, the dashboard shows predictor models for drivers of respiratory diseases. The predictions show that females and two-year olds have the highest possibilities of visiting the ER in relation to temperature and PM2.5 levels.

Watson's social media analytics provided insights into the sentiment and concerns of Beijing's citizens. Air purifier, air and alerts are the most tweeted hashtags in Beijing from a list of hashtags related to air quality. Figure 5 shows the graph analyzing the number of tweets with the hashtags Beijing and smog. It reveals that the most tweets with those hashtags occur in November and December. Likewise, negative sentiment tweets are sent in November and December.

The visualizations that best communicated the insights found in the data were organized into dashboards. The introductory dashboard (Figure 1) exhibited the insights into worldwide PM2.5 levels for comparison with China. It provided visualization of hashtags pertaining to air quality. Figure 2 was of the Beijing PM2.5 visualizations. The graphics showed u-shaped curves for levels of PM2.5 with different measures of time. Figure 4 dashboard's visualizations were of the predictive model and the relationships between PM2.5 and other variables such as temperature. Figure 3 explored the correlations between time and temperature, PM2.5 and ERVs for Respiratory Diseases. Figure 5 was the Twitter dashboard with the resulting visualizations of the Twitter data analysis.

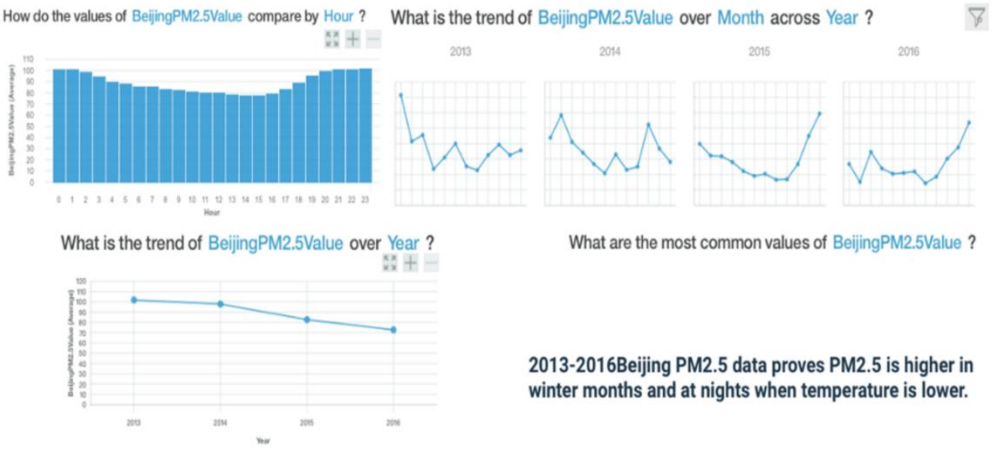


Figure 2. Beijing PM2.5 Dashboard. This dashboard is composed of graphs demonstrating the relationship between PM2.5 and season. It permits the user to see the levels of PM2.5 in the air over hour, month, and year.

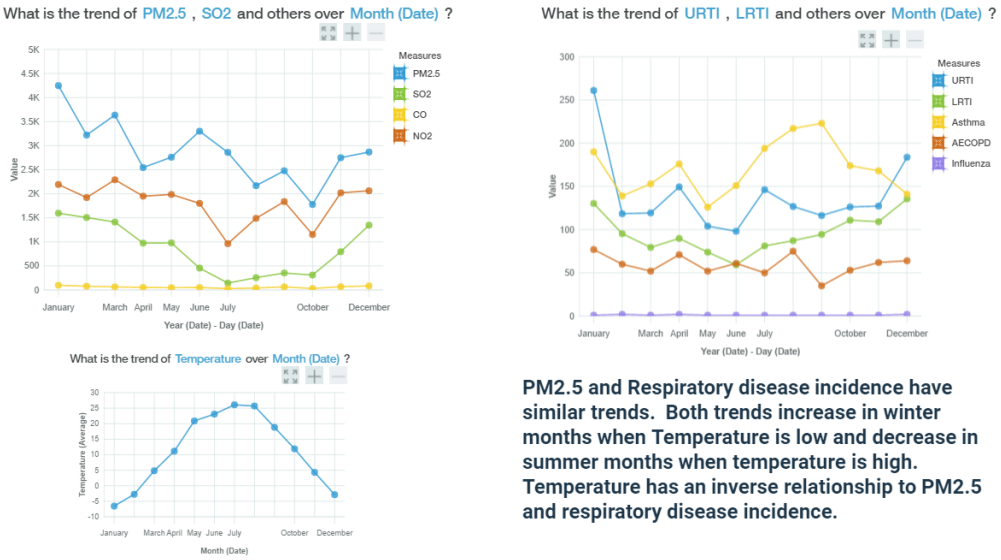


Figure 3. U shape and inverse shape Dashboard. The dashboard above was assembled to bring together the disease specific visualizations and to show the inverse and direct relationships amongst the variables.

4 Discussion and Conclusion

Analysis of the data with Watson Analytics provided various insights. Upon analyzing the PM2.5 data it was revealed that in 2013 the majority of daily PM2.5 averages were in the unhealthy for all, very unhealthy, or hazardous categories. Monthly averages for PM2.5 from 2013 to 2016 (Figure 3) show a tendency for high levels of the pollutant in the months of January, February, and December. A surprising discovery emerged from the graphing of the levels of PM2.5 by hour (Figure 2). It showed that Beijing experiences the highest levels of PM2.5 during the night hours. The rise of PM2.5 begins at 6pm and continues until 3am. Further analysis uncovered a negative correlation between PM2.5 and temperature (Figure 4). Additionally, an analysis of other airborne pollutants was conducted (Figure 3). In general, the analysis showed that sulfur dioxide, carbon monoxide, and nitrogen dioxide concentrations are higher from January to March and in November and December. Ozone is the exception as it has higher concentrations in May, June and August. The predictive model (Figure 4) in this set of analysis determined that CO is a predictor of PM2.5 with 54% certainty. The final discovery was the decreasing trend of PM2.5 levels in Beijing (Figure 2) over the last four years.

The analysis of ERVs (Figure 3) found a general trend for URTI, LRTI, and AECOPD. The graph showed the highest numbers of ERVs for respiratory diseases occur in January, February, November and December. Asthma had the highest number of ERVs in July, August and September. Predictive models (Figure 4) in this set of discoveries determined that females are the most affected by levels of PM2.5 and that infants age 2 are the most affected by temperature.

Analysis of Twitter hashtags in 2015 and 2016 unveiled the concerns of Beijing's citizens who were greatly interested in air purifiers and paying attention to alerts. Coal as a reason and cancer as a consequence were all tagged (Figure 1). November and December had the highest number of tweets with the hashtags #Beijing and #smog. Those two months also had the highest numbers of tweets with negative sentiment. They were all consistent with the stunning rise of PM2.5 value in the year 2015 and 2016 (Figure 2).

The higher concentrations of airborne pollutants probably caused a surge of emergency room visits for respiratory diseases. The highest number of ERVs for URTI, LRTI, and AECOP occurred at the beginning and end of the year. Asthma sufferers, however, suffered the most. Visits to the ER for asthma were relatively high throughout the year. The highest numbers of asthma ERVs were strongly correlated to high ozone concentrations.

In this study, the analytics and visualization tool had illustrated a powerful capability in data analytics, visualization, and automated predictive analytics. It was also encouraging to see a downward trend of PM2.5 levels over the last four years (Figure 2). Prompted by the hosting of the 2008 Olympics in Beijing the Chinese government began efforts to reduce air pollution [11]. The downward trend demonstrated by the data analysis denotes changing environmental regulations. The Twitter analysis, however, revealed that there was still much to be done to clean up the air in Beijing. Current Tweets showed that smog in Beijing continues to be of great concern to its citizens.

Limited by the availability of data set, we analyzed only Beijing's data from 2013 to 2016. Another limitation of this study was that we investigated only the outdoor air quality data rather than the indoor data that probably also contributed to the respiratory diseases. In addition, more data sources such as Weibo may be included to further unveil other hidden patterns or relationship on environment and health issues in the future studies.

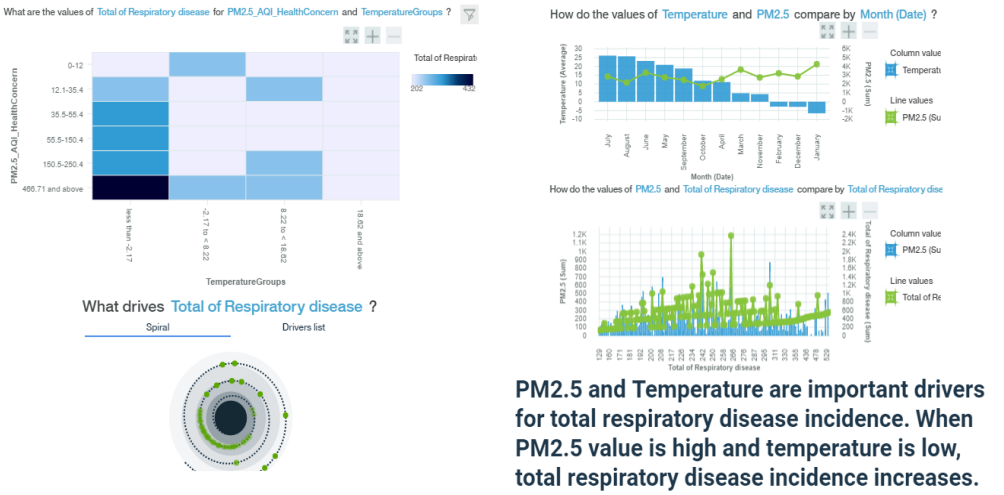


Figure 4. Relationship and Predictive Model Dashboard. This dashboard displays the analysis of the respiratory disease data. It allows the user to see the relationships between ERVs for respiratory diseases and other variables.

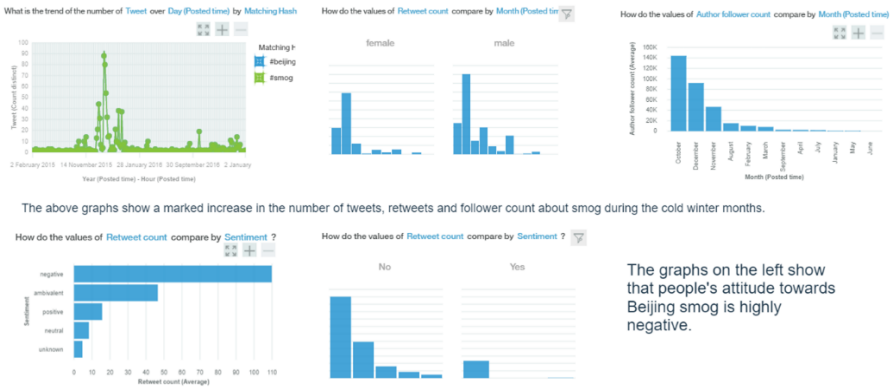


Figure 5. Twitter analysis dashboard. The final dashboard was constructed to bring together Twitter analysis and the visualizations made possible by Watson analytics. Watson provided plenty of data to explore the sentiment of the residents of Beijing regarding the quality of the air.

References

[1] Monaghan, A.: China surpasses US as world's largest trading nation. The Guardian, 2014. (2014)

[2] Workman, D.: China's top trading partners. World's Top Exports, 2017. (2017)

[3] Nielsen, C., and Ho, M.: Air pollution and health damages in China: An introduction and review. Clearing the Air: The Health and Economic Damages of Air Pollution in China. (2007)

[4] Rohde, R. & Muller, R.: Air pollution in China: Mapping of concentrations and sources. PLOS one. vol. 10 no. 8 (2015)

[5] Lin, J., Pan, D., and Davis, S., et al.: China's international trade and air pollution in the United States. Proceedings of the National Academy of Sciences of the United States of America. vol. 111 no. 5, pp. 1736-1741 (2013)

[6] Brauer, M. et al.: Dataset contains world PM2.5 air pollution, mean annual exposure in the measurement of micrograms per cubic meter (μg/m3). The World Bank. (2016)

[7] United States Department of State, Air Quality Monitoring Program. Retrieved from http://www.stateair.net/web/historical/1/1.html (2017)

[8] Xu, Q., et. al.: Fine Particulate Air Pollution and Hospital Emergency Room Visits for Respiratory Disease in Urban Areas in Beijing, China, in 2013. PLOS one. vol. 11 no. 4 (2016)

[9] Twitter: Twitter Data retrieved from IBM Watson Analytics with both Hash tags #Beijing and #smog, data ranging from January 1, 2015 to December 31, 2016 (2017)

[10] European Environment Agency: Hot summer weather exacerbating ozone pollution. (2013)

[11] Ramzy, A.: Beijing's Olympic war on smog. Time, 2008. (2008)