The Correlation Between Ozone and Respiratory Illness using Project INDICATOR data for Urbana-Champaign

Elizabeth J Surbeck
University of Illinois, Urbana-Champaign
surbeck2@illinois.edu

ABSTRACT
The purpose of this analysis is to examine the correspondence between short-term ozone exposure and respiratory illness hospital reports, adopting from the methods used in “Meta-analysis of the Association between Short-Term Exposure to Ambient Ozone and Respiratory Hospital Admissions” by Meng Ji, Daniel S. Cohan, and Michelle L. Bell. This meta-analysis was published in 2011 on studies that have connected asthma cases in hospitals and ozone measurements in air quality reports.

Studies seem to suggest that there is a strong correlation between asthma and ozone level increases in the outdoor air. The point of exploring their research is to see if one could take real time collected data from an area such as Champaign-Urbana and prove or disprove the results of this meta-analysis. The datasets to attempt fulfilling this purpose have been drawn from Project Indicator, the Environmental Protection Agency’s (EPA) Air Quality System (AQS), and Midwestern Regional Climate Center’s (MRCC) Application Tools Environment (cli-MATE). The tools required to accomplish this analysis included Oracle’s Data Miner, MS Excel, and R and along with methods for analysis. These included linear regression to understand the interaction of the quantitative data from the two main data variables, respiratory illnesses and ground ozone and classification to explore in more detail other variables that interacts with respiratory illnesses and ground ozone. The results of the study ultimate suggested that ozone and respiratory illness correlate with multiple environmental variables though they do not significantly correlate with each other.

Categories and Subject Descriptors
H.2.8: Database Applications—I.1.2:Algorithms—I.5: PATTERN RECOGNITION

General Terms
Algorithms, Measurement, Human Factors, Verification

Keywords
Nonparametric analysis, asthma, syndromic surveillance, respiratory illness, ground ozone, environmental health, classification, naïve bayes, decision tree

1. INTRODUCTION
The purpose of this analysis is to examine the correspondence between short-term ozone exposure, environmental weather factors, and respiratory-related hospital reports, exploring the methods and observations in “Meta-analysis of the Association between Short-Term Exposure to Ambient Ozone and Respiratory Hospital Admissions” by Meng Ji, Daniel S. Cohan, and Michelle L. Bell. This meta-analysis was published in 2011 on studies that have connected asthma cases in hospitals and ozone measurements in air quality reports.

These studies suggest that there is a relationship between asthma and ozone level increases in the outdoor air. The purpose of this meta-analysis study was to attempt to find a way to bring multiple recent studies together and show a meaningful result to compare with this observation.

The point of adopting from their study and using it as inspiration to explore this relationship more is to see if one could take real time collected data from a town area like Champaign-Urbana and prove or disprove their observation within that regional context. The datasets used to attempt fulfilling this purpose have been drawn from Project Indicator, the Midwestern Regional Climate Center’s (MRCC) Application Tools Environment (cli-MATE), and the Environmental Protection Agency’s (EPA) Air Quality System (AQS).

Project Indicator currently offers a more or less complete public health data portrait of Champaign, Urbana, and the immediate populated area. Having teamed up with the Champaign-Urbana Public Health District (CUPHD), Project Indicator pulls data from records developed from Emergency Departments, convenient care ambulatory clinics, patient advisory nurses, and reportable diseases. In addition, they pull data from local school attendance, animal control, weather, the humane society, veterinary clinics, mosquito traps, pollen, and Twitter. In this particular case, I focused on the emergency department records in Project Indicator from Christie Clinic’s emergency department.

Project Indicator, however, does not show air quality or weather information reports, which would be essential in order to look at both asthma cases and ozone levels. The Environmental Protection Agency’s (EPA) Air Quality System (AQS) collects ambient air pollution data from national, state, local, and tribal agencies in the United States and became the data source of ozone measurements for this study. AQS data includes meteorological data and geographic and operator data of each monitoring station. This information is used for reports to Congress following the Clean Air Act. The data for ozone densities in air quality was pulled directly from the AQS site in a csv (comma-separated values) format using a hour sampling of every hour of every day from January 1, 2006 to December 31, 2012 within all measurement sites in Urbana-Champaign, IL.

The Midwestern Regional Climate Center’s (MRCC) Application Tools Environment (cli-MATE) offers hourly, daily, monthly, and yearly weather information from 1997 to the present day. The MRCC is a project jointly run by the National Climatic Data Center (NCDC) and the Illinois State Water Survey in Champaign, IL. It also partners with other regional climate centers and National Oceanic and Atmospheric Administration (NOAA),
which is a management entity of NCDC. MRCC holds a mission to facilitate discovery of evidence-based explanations for Midwestern climate changes and their impact for the sake of issues such as human health, climate change, agriculture, and other related topics.

2. Related Work

Project Indicator is comparable to other, similar disease monitoring systems across the county. A couple of these include the UNC Department of Emergency Medicine Carolina Center for Health Informatics and the New York City Department of Health and Mental Hygiene.

The North Carolina Disease Event Tracking and Epidemiologic Collection Tool, also known as NC DETECT gives regular data on public health surveillance. They report on statewide emergency department data and other pertinent information for public health safety and knowledge. Like Indicator, NC DETECT has syndromic surveillance capabilities, using datasets drawn from emergency departments, emergency medical services, urgent care, and the Carolinas Poison Center. There data has successfully been used to track statewide epidemics such as the path of influenza outbreaks.

The New York City Department of Health and Mental Hygiene is another comparable organization that established their syndromic surveillance system for the purpose of monitoring emergency department activity to detect disease trends and outbreaks. The system evaluates daily chief complaint information and looks for key symptoms related to severe illnesses. Past use of the system has accurately pointed to influenza, norovirus, and rotavirus outbreaks.

For the purposes of this study, which focuses on Champaign-Urbana, these data resources could not be used. Other data resources though provided some relatable information to the purposes of this analysis. These data resources, however, did not ultimately fit the analysis’s needs. Ozone and asthma-related events occur over short periods of time so the need for day-to-day sets data for an entire year was needed. Additionally, because no medical professionals were a part of the analysis, specific ICD-9 codes were required to pinpoint the exact nature of each medical case report for the most reliable results. Nevertheless, these datasets are worth noting as they do ultimately relate to the analysis’s concerns and can show other ways public health data is collected for Illinois and the greater United States of America.

The Illinois Department of Public Health’s Hospital Discharge Database uses data collected by the Illinois Department of Public Hospital Association and provides information of hospitalizations within the state of Illinois. The data collected includes patient demographics, reasons for hospital admission, care received at the facility, and the outcome of the hospitalization. Some of the databases limitations include variations in medical labeling and the limited locations of the Illinois hospitals providing this data. Hospitals that contributed to this database are members of the IHA and are located typically in the region of Cook County rather than being spread out throughout the state, so seems to be too limited in showing a statewide data picture. The data is retrieved through a query on a webpage interface and cannot be downloaded. Any raw data would likely have to be obtained by contacting the Illinois Department of Public Health and making a specific request.

The Integrated Health Interview Series dataset resource, which is maintained by the University of Minnesota, provides documentation of public use files from the U.S. National Health Interview Survey, otherwise known as NHIS. The data is designed so users can extract data from multiple variables, which include medical, geographic, and temporal factors. Thought it might not be able to show Champaign area-specific data, it could show asthma cases in larger, related geographic areas and using IHIS’s yearly data samples.

Asthma Call-back Survey data acts as a part of the Centers for Disease Controls’ National Asthma Control Program with the purpose of tracking the state of asthma on the state and national levels. The survey is given two weeks after the Behavioral Risk Factor Surveillance Survey to respondents who report any past or current asthma diagnosis. Whether children are included in the survey depends on the state that gave the Behavioral Risk Factor Surveillance Survey. Some limitations with this data source includes that the data is publicly available only in html-formed tables on web pages in yearly samples on a state level from 2006 to 2010. Asthma Call-back Survey data are the data from the National Health Interview Survey, which comes in annually to represent national estimates of the citizen population’s health status. This data source specifically draws from cases of lifetime and current asthma and asthma attacks of those suffering from current asthma. The data is only given in html-formed tables and limited years that would not include the most recent ones.

These sources may be useful as anecdotal information, though they do not help the analysis come to its point in an efficient way. They could be used to look at monthly and yearly samples of data instead of daily and could be used to watch for any trends in ozone and asthma over years instead of months and on a national or regional level instead of a local one. Local and daily information was required for this particular analysis, however. Data resources comparable to AQS data do not seem as accessible as that which the AQS and EPA can offer.

3. Method

Project Indicator’s data came in the form of a SQL file and an Excel file, which included information from 2006 to 2012 in the SQL files from Christie Clinic and the Excel file from Carle’s Emergency Room Department. Queries were run on MySQL and within Excel to look specifically at all ICD-9 codes that describe respiratory illnesses. The focus on these codes for this analysis was eventually narrowed to codes that reflected all respiratory illnesses, asthma-related illnesses, and non-asthma-related illnesses.

In order to refine the dataset selection of the respiratory case numbers, multiple discussions during the fall of 2013 with the provider of the information, Dr. Ian Brooks of Project Indicator, about the data revealed originally that for the purposes of this project, Christie’s convenient care clinic data was the only dataset that could provide reliable records of ICD-9 codes. Later, however, a reliable record with ICD-9 codes from Carle’s Emergency Room data the following spring of 2014 in the form of the Excel file so both major immediate care facilities in Champaign-Urbana were included.

The reason ICD-9 codes are so important to this analysis is because of the lack of medical diagnosis expertise of those
involved in this analysis. ICD-9 is a health code system supported by the World Health Organization of the United Nations. It is used for management, clinical, and research purposes within the medical community. It is used in American hospitals and related medical institutions for billing purposes.

Though six years worth of data was collected initially, there were still holes in the analysis that caused the years ultimately evaluated to be 2006 to 2009. One issue was the Christie’s data stops suddenly at December 17. That data is currently not immediately available. Another issue came from AQS. Large periods of time do not provide any information about the ozone in Champaign-Urbana in 2010 and 2011. The reason for this is not clear. The regression lines and classification results created from the data as a result cover a four-year span in recent Champaign-Urbana history.

The data from the SQL query results and Excel files from Christie and Carle were placed into csv files to be manually evaluated in R and MS Excel to organize the data of the data and total number of asthma cases, total numbers of other respiratory cases, and total numbers of both together specifically in R.

The data for daily ozone measurements in air quality was pulled directly from the AQS site in a csv file using a one-hour maximum sampling of every day from January 1, 2006 to December 31, 2009 within all measurement sites in Urbana-Champaign. The data was opened first in Excel for manual review and then R to calculate the mean measurements of each day over a 24-hour period. The results then were put into the main csv file for the final analysis procedures.

Data for daily weather information was pulled from MRCC’s Climate website from January 1, 2006 to December 31, 2009 from Champaign-Urbana in a csv file, also using a one-hour maximum sampling of every day. The data was also opened first in Excel for manual review and then R to calculate the mean measurements of each day over a 24-hour period. The variables used for this were air temperature in Fahrenheit, web bulb temperature, dew point temperature, relative humidity, station level air pressure, sea level air pressure, and wind speed. Because this analysis also desired the heat index temperature and wind chill temperature, wind speed, air temperature, and relative humidity were used and passed through a javascript that would input and output the information in a csv file.

All datasets were brought into a single master csv file to contain consistently formatted information that includes every day from 2006 to 2009. In the cases of no available data, data was treated slightly differently. Empty spaces were labeled as 0 for regression line variables, which was deemed safe in this case because these instances in both data sets were rare with the exceptions of obvious data holes. This large time spans of missing data were not included in the final analysis. Classification variables’ missing values were treated as null values as they would be sorted away from variables with meaning in the algorithm.

For the initial regression analysis, ozone air measurements and asthma report numbers and ozone air measurements and all respiratory illness report numbers were loaded in a main file with their corresponding dates in R to be evaluated by respiratory case scenarios. These numbers were generated into graphs and the linear regression formula was applied to them (see Figure 1, Figure 2, and Figure 3). This found that any positive correlations would have to be found within other variables.

![Figure 1](image1.png)

**Figure 1**

Asthma and Ozone Regression Analysis in Champaign Urbana, 2006-2009

![Figure 2](image2.png)

**Figure 2**

Respiratory Illness and Ozone Regression Analysis in Champaign Urbana, 2006-2009

![Figure 3](image3.png)

**Figure 3**

Respiratory Illness that is not asthma and Ozone Regression Analysis in Champaign Urbana, 2006-2009

The classification analysis for naïve bayes and decision tree took the numbers from asthma cases, non-asthma respiratory cases, and both asthma and non-asthma respiratory in order to evaluate them with ozone measurement and the weather information variables
mentioned earlier in this report. This was done within Oracle Data Miner. Also, within Oracle Data Minor, the numbers were binned according to their summarization information in R with some specific exceptions. Respiratory illness case scenarios were divided based on the average number of cases in order to find a yes-no scenario. Yes was above average numbers of cases. No was at or below average number of cases. Variables such as temperature were divided based on effect on the state of water such as whether it would be freezing. Variables that revealed a potential relationship with any of the three respiratory case scenarios or ground ozone measurements were then also evaluated for a second time within R for regression.

4. Results

For the initial regression line analysis, there does not seem to be a correlation between ozone and any of the respiratory case scenarios. The regression coefficient for ozone and just the asthma cases for all of 2006 to 2009 was -23.039 while the regression coefficient for respiratory cases that weren't asthma as -562.1. All possible respiratory cases and ozone had a regression coefficient of -585.13. The regression lines in the months that showed the strongest positive signals of correlations from 2006 to 2009 were March, July, and December. Though March and July could arguably have rising temperatures, December does not. What this suggested, if anything, was that there were other variables that were yet uncertain that played a part in ozone's interaction with respiratory illnesses. During the classification phase, however, a more complex though clear picture took shape.

In the case of looking exclusively at asthma cases, the decision tree only broke apart between above and at and below average at one variable in the data, which was month. At or below the average number of cases for asthma showed to have a propensity to occur during December, January, February, April, May, June, July, and August. The performance matrix can be found in Table 1. An above average number of cases had a propensity to occur during September, October, November, and March. The results of the Naïve Bayes algorithm agreed with these results. The performance matrix can be found in Table 2.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Above Average</th>
<th>At or Below Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above Average</td>
<td>226</td>
<td>60</td>
</tr>
<tr>
<td>At or Below Average</td>
<td>100</td>
<td>186</td>
</tr>
<tr>
<td>Total</td>
<td>326</td>
<td>246</td>
</tr>
<tr>
<td>Correct %</td>
<td>69.3252</td>
<td>75.6098</td>
</tr>
</tbody>
</table>

Table 4

<table>
<thead>
<tr>
<th></th>
<th>Above Average</th>
<th>At or Below Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above Average</td>
<td>206</td>
<td>80</td>
</tr>
<tr>
<td>At or Below Average</td>
<td>102</td>
<td>184</td>
</tr>
<tr>
<td>Total</td>
<td>308</td>
<td>264</td>
</tr>
<tr>
<td>Correct %</td>
<td>66.8831</td>
<td>69.697</td>
</tr>
</tbody>
</table>

A regression analysis for these seemingly significant variables showed the following coefficients. Sea level pressure had a 26.2 coefficient. Air temperature had a -0.0657 coefficient. Wet bulb temperature had a -0.7198 coefficient. Wind Speed had a 1.59 coefficient.

In the case of looking at respiratory cases that were not related to asthma the decision tree results showed a propensity to the fall and winter months with high sea level pressure ratings and various wind speed ratings. The performance matrix can be found in Table 3. Naïve Bayes results for above average cases with 50% propensity or above showed at and below freezing and cold and above freezing wet bulb temperatures, low and somewhat low ozone ratings, at and below freezing and cold and above freezing air temperatures, high and highest possible sea pressure for the area, and high and highest wind speed for the area all occurring during November, December, January, February, and March. The performance matrix can be found in Table 4.
A regression analysis for these seemingly significant variables showed the following coefficients. Wind speed had a 1.603 coefficient. Dew point temperature had a -0.7032 coefficient. Wet bulb temperature had a -0.7194 coefficient. Sea level pressure had a 25.75 coefficient. Air temperature had a +0.6578 coefficient.

Ozone itself was finally explored within the Oracle Data Miner, using the already available variables. It found in the results of the decision tree algorithm that ozone could be high at any particular time in the year. However, it rates higher during the spring and summer and relative humidity is moderate while temperature is pleasant. The performance matrix can be found in Table 7. The higher levels of ozone within the naïve bays algorithm results for above average cases with 50% propensity or above showed a low to medium station level air pressure, a warm to pleasant air temperature, a warm or pleasant dew point, a warm to pleasant wet bulb temperature, a low to medium to high relative humidity, and a high to low to medium sea level air pressure measurement. The performance matrix can be found in Table 8.

5. Conclusion

While there are patterns in the behaviors of asthma illnesses, non-asthma respiratory illnesses, all respiratory illnesses, and ground ozone, these patterns do not appear to relate completely with each other. Respiratory illnesses and ground ozone simply do not appear under similar conditions and during even the same time of year in the Champaign-Urbana area.

Asthma appears to be highly seasonal, which was already suggested in the initial regression analysis results. Other respiratory illnesses with and without asthma included showed a strong relationship with higher sea level pressures and wind speeds while, like just asthma alone, occurred primary in the colder months. Respiratory illnesses without asthma seemed to be most prevalent when the sea level pressure was higher, the air and wet bulb temperatures were a little older, and the wind speed was a little higher. Respiratory illnesses with asthma in the numbers as well showed also a presence particularly during higher wind speeds, lower dew point, wet bulb, and air temperatures, and higher sea level pressure measurements.

Ground ozone, however, seemed to have the highest density during warmer months with higher dew point, wet bulb, and air temperatures with lower station and sea level pressures and humidity. Unlike respiratory illnesses, ground ozone seemed to thrive best in the middle of warmer, drier environments when the weather is perhaps more cloudy, given the lower air pressure.

It is possible that higher ground ozone density measurements and higher respiratory case reports have some relationship under very specific conditions. The data results from this analysis suggest, however, that such incidents are more coincidental rather than correlated in the Champaign-Urbana area.

6. Discussion

Though this analysis states that the patterns found with asthma, non-asthma respiratory cases, all respiratory cases, and ground ozone do not appear to correlate with each other, this does not necessarily mean that any studies looking at respiratory health and ground ozone are wrong in showing relationships. Some of these
factors are limiting and may suggest further study in other directions.

This analysis only looked at emergency department and convenient care reports for asthma and other respiratory-related report numbers. If a case occurred and it was either recorded in a different department or did not merit a hospital trip in some way, it could not have been accounted for in these methods. As a result, these numbers only account for serious cases.

In the future, it would be informative to look at ICD-9 codes from other departments and other medical facilities in the Champaign-Urbana area. Still, this would not account for home incidents, which would be incredibly difficult to report. Monitoring Champaign-Urbana area residents through social media outlets such as Facebook and Twitter where users tend to mention their state of health voluntarily might accomplish this. This latter collection method though would also likely possess flaws.

Within the medical data itself, a variable that is currently unavailable but would offer useful information would be the age of patients. Children are more prone to suffer from asthma and having an age variable may offer a more detailed image of how ozone interacts specifically with asthma sufferers. The likeliness of obtaining information of this nature from Champaign-Urbana medical institutions right now, however, is low. Another flaw in having such basic patient information is the chance of number overlap. These asthma reports do not have unique identifiers for patients for privacy reasons. A lack of overlap cannot be guaranteed.

Though finding another good resource for ozone measurements outside of what the EPA provides is unlikely, other air quality variables such as small particle matter and pollen may be informative if included in the study. Nitrogen data for this area was not available from the EPA. Currently it is also difficult to find a current and available pollen database on this area even though these measurements are reported on a daily basis in mainstream weather outlets.

The physical site for analysis, Champaign-Urbana, may be more useful as a comparison site to an industrial area, which likely possesses higher rates of ozone and potentially also more asthma cases, such as the East St Louis area and industrial zones in the Chicago area. The Champaign-Urbana area is not heavily industrialized and is not close to any industrial zones, which likely also affects its data. It has a north temperate climate, which may also possibly make its atmospheric events slightly different than other potential study sites. Its population demographics as a city area with a large university may make a difference as well. With more factors such as those listed here put into consideration, more defined answers to this analysis’s concerns may be obtained.

7. ACKNOWLEDGMENTS

Thanks to Dr. Ian Brooks, GSLIS Faculty and Director of the Health Sciences Group at the National Center for Supercomputing Applications for making Project INDICATOR data available. Thanks also to Dr. Catherine Blake for her academic guidance and fellow classmate, Craig Evans, for his guidance on Java programming.

8. REFERENCES


