

Predicting The Ideal Time To Purchase Allergy Medication

For The Albuquerque Area

Matthew Letter

University of New Mexico: Department Of Computer Science
Big Data
Albuquerque, New Mexico
Mletter1@unm.edu

Abstract

Medication waste is a growing problem in the United States. This paper addresses ways to stem allergy medication waste, for the Albuquerque area, through predicting what months have high pollen counts and what factors contribute to pollen levels. The author of this research believes outlining stimuli related to pollen levels can help with medication purchase determination. Data coinciding with pollen levels and correlating factors are used to make this determination. If we follow the methods outlined in this paper for choosing when to purchase allergy medication we may very well avoid a large amount of medication waste and improve our ability to predict bad pollen years for the Albuquerque area.

The data yields that the highest contributing factor to pollen levels in the Albuquerque area is rainfall precipitation. Drought years (years with low rainfall) show a drastic drop in pollen level implying that pollen levels are directly related to rainfall totals. In this was a prediction that a drought year yields low pollen counts has been made using tableau. Furthermore, on average, the start of the allergy season in the Albuquerque area is February with the elevation of pollen levels such as juniper. The peak of allergy season is April with Mulberry being the strong overarching pollen producer over all area plants tested. By November the pollen levels have bottomed out therefor it is recommended that this is the best month for purchasing allergy medication, as allergy medications expire in 12-month increments (such as 1 year, 2 year...). Furthermore, for years with droughts it is recommended that you wait to buy your medications and see if you allergies are bad enough to warrant purchasing medication, as the level of pollen in the air is on the order of magnitude 10 times lower than years not in drought.

Keywords; Pollen, Albuquerque, Purchasing, Allergy, Medications, Calendars, Weather.

I. INTRODUCTION

People spend billions of dollars a year trying to control their allergies [2]. Unfortunately a large amount of allergy medication are thrown away because people buy their allergy medication at the wrong point of time during the year, which leads to expiration before they can be used. The goal of this study is to give the people of the Albuquerque area a way of determining when to buy their allergy medication, so as to not let them go to waste. This paper also explores pollen

trends in Albuquerque to help people with specific allergies predict when to buy their medications.

Items that are looked at include: how do pollen levels stack up based on the plant type, what months of the year do certain plants pollinate, what weather factors contribute to pollen levels, and what insight can be gleaned from the weather. These factors were all used to come up with a set of conclusions that help determine the proper time of the year to buy medications based off of insights from the collected data sets. These insights give us the ability to predict what years may have high levels of pollen relative to the norm.

II. RELATED WORK

2.1 Allergenic Pollen and Pollen Allergy in Europe

Allergies have been impacting people all over the world and weather conditions appear to play a significant role in the ability of plants produce pollen [1]. Data obtain from aerobiological studies and allergological investigation; have contributed to the creation of pollen calendars with the approximate flowering period of plants in a given region. It has been noted that pollen data correlates with weather conditions for given regions. Allergens such as pollen have contributed to degrading the quality of life of people suffering from the many respiratory conditions, such as asthma.

2.2 A Revised Nomenclature for Allergy: An EAACI Position Statement from the EAACI Nomenclature Task Force

The European Academy of Allergy and Clinical Immunology (EAACI), an association promoting basic and clinical research, reports a revised nomenclature for allergic reactions based on the present knowledge of the allergy mechanisms. Allergic reactions can express themselves in many different bodily organs through many different mechanisms. According to [2], an allergy, a specific type of hypersensitivity, is a clinical reaction in which an immunologic mechanism is proven or strongly implicated. An allergy can be antibody or cell mediated. In an effort to avoid misunderstandings between patients, physicians, and their colleagues, a very clear designation and nomenclature of the allergic disorders must be adhered to.

2.3 A Holocene pollen record of persistent droughts from Pyramid Lake, Nevada, USA

Pollen and algae preserved in the sediments from Pyramid Lake, Nevada, provide evidence for periods of persistent drought. Pollen levels increase when the ground has sufficient saturation for plant reproduction. The level of pollen for a wet year will be increased from that of a dry year. This fact was used to analyze core samples from Pyramid Lake, Nevada, showing that 7600 to 6300 cal yr B.P., was the driest period of the Holocene age [3].

2.4 Superiority of an Intranasal Corticosteroid Compared With an Oral Antihistamine in the As-Needed Treatment of Seasonal Allergic Rhinitis

Daily use of nasal spray corticosteroids or histamine receptor antagonists have proven to be effective in the treatments of allergic rhinitis. One thing that is noted is that the use of histamine receptor antagonists should be taken as needed. Therefor it is important to keep track of when to buy your allergy medications. This study shows that the as-needed intranasal corticosteroid outperformed antihistamines when controlling allergic rhinitis [4]. Both of these medications are good for the treatment and control of yearly allergy symptoms.

III. DESIGN AND APPROACH

Aproach and Design

Three languages were considered Python, Java, and C++. Python has many libraries for parsing, machine learning, and connecting to databases. C++ has many of the same feature capabilities as python but you have to worry about memory management. Finally java was looked at and chosen for the language to implement the parsers. Java was chosen manly because of the SAXPARSER library. The sax parser is designed for parsing XML files.

Albuquerque data warehouse, which is the location of the data set for that Albuquerque area pollen data, presented the data in the form of XML. The sax parser took in the XML pollen data and cleaned the attributes subsequently putting them into an SQL database [5]. The second parser was for parsing the CSV data files obtained from Wunderground that contained all the weather data for the Albuquerque area over the past 100 years [6]. This parser took in the CSV file and cleaned the data for proper representation to the form of a relational database. Java was chosen to implement this because it was already being used for the Sax Parser. After the data was in the MySQL database it was ready for analysis Using Ubiq and Tableau.

Once the data was in the MySQL database Ubiq was used for initial discovery and insight. The MySQL database was connected to Ubiq through an ssh tunnel. Once Ubiq had the connection to the database manipulation of the data was done. The first goal was to obtain a representation of the average pollen counts for each month, and a weather data trend line comparison later on. Once average data plots are produced it is necessary to investigate other aspects of the pollen data set. One aspect of the data that immediately sticks out is that the pollen readings were taken at two different locations in Albuquerque. The two locations are categorized under the names Eastside and Westside referring to their respective portions of the Albuquerque area. Insight into how location of the data collected affected the resulting graphs and plots was taken into consideration. Once the aspects of the pollen data are exhaustively examined weather data correlations are looked at.

The weather data and pollen data tables are inner joined on the date attribute yielding a large data set. Once the data is in this form correlation between the two sets can be examined by comparing the data on the joined timeline attribute. Graphs and charts are made with trend lines and bar charts for predictive purposes. The overarching goal of this step in the approach is to find correlating features between the pollen data and the weather data. Finally with all this information in hand predictions are made about when to purchase allergy medication. All of the charts and graphs and any other insight from the data will be use to draw these conclusions.

IV. RESULTS

The Pollen Data Set

First the data was virtualized to the form of average pollen per month by tree type. Figure 2 shows the results of the mapping of the pollen averages by tree type. November and December are the months with the lowest pollen counts. March and April are the months with the highest pollen levels. Pollen levels start to rise in the month of January. Figure 2 shows the SQL statement that generated the data for figure 1. Each tree was separated by tree type and an average aggregate call was run on the pollen level of the tree over all years of the data. The data for all years for each tree was averaged by month for easy referencing. Next other aspects of the pollen data are investigate

A study on the location attribute, of the pollen data, was conducted. The location was separated into two categories Eastside and Westside, referencing their respective location of the Albuquerque area. It was then combined with a count on the number of readings taken and pollen levels. The last dimension added into the mix was a time dimension. This dimensionality reduction of the data yielded figure 3, which has 4 graphs, associated with it. From top to bottom the graphs will be reference alphabetically starting from (a).

Figure 3, graph (a) shows the pollen level per day from 2004 to the end of 2015 for the Eastside of town. Notice how the pollen levels begin to drop in 2009; a correlation will be

```

select date_format('pollen_data','submission_date') as 'submission_date',avg(if('pollen_data'.'tree_type'='Acacia',pollen_data.'level',null)) as 'level_Acacia',avg(if('pollen_data'.'tree_type'='Alder',pollen_data.'level',null)) as 'level_Alder',avg(if('pollen_data'.'tree_type'='Ash',pollen_data.'level',null)) as 'level_Ash',avg(if('pollen_data'.'tree_type'='Beech',pollen_data.'level',null)) as 'level_Beech',avg(if('pollen_data'.'tree_type'='Birch',pollen_data.'level',null)) as 'level_Birch',avg(if('pollen_data'.'tree_type'='Chenopodiaceae',pollen_data.'level',null)) as 'level_Chenopodiaceae',avg(if('pollen_data'.'tree_type'='Cottonwood',pollen_data.'level',null)) as 'level_Cottonwood',avg(if('pollen_data'.'tree_type'='Elm',pollen_data.'level',null)) as 'level_Elm',avg(if('pollen_data'.'tree_type'='End of Season',pollen_data.'level',null)) as 'level_End of Season',avg(if('pollen_data'.'tree_type'='Ephedra',pollen_data.'level',null)) as 'level_Ephedra',avg(if('pollen_data'.'tree_type'='Equipment Error',pollen_data.'level',null)) as 'level_Equipment Error',avg(if('pollen_data'.'tree_type'='Goldenrod',pollen_data.'level',null)) as 'level_Goldenrod',avg(if('pollen_data'.'tree_type'='Grass',pollen_data.'level',null)) as 'level_Grass',avg(if('pollen_data'.'tree_type'='Hickory',pollen_data.'level',null)) as 'level_Hickory',avg(if('pollen_data'.'tree_type'='Juniper',pollen_data.'level',null)) as 'level_Juniper',avg(if('pollen_data'.'tree_type'='Juniper/Cedar',pollen_data.'level',null)) as 'level_Juniper/Cedar',avg(if('pollen_data'.'tree_type'='Linden',pollen_data.'level',null)) as 'level_Linden',avg(if('pollen_data'.'tree_type'='Locust',pollen_data.'level',null)) as 'level_Locust',avg(if('pollen_data'.'tree_type'='Maple',pollen_data.'level',null)) as 'level_Maple',avg(if('pollen_data'.'tree_type'='Mesquite',pollen_data.'level',null)) as 'level_Mesquite',avg(if('pollen_data'.'tree_type'='Mulberry',pollen_data.'level',null)) as 'level_Mulberry',avg(if('pollen_data'.'tree_type'='Nettle',pollen_data.'level',null)) as 'level_Nettle',avg(if('pollen_data'.'tree_type'='Oak',pollen_data.'level',null)) as 'level_Oak',avg(if('pollen_data'.'tree_type'='Pecan',pollen_data.'level',null)) as 'level_Pecan',avg(if('pollen_data'.'tree_type'='Pine',pollen_data.'level',null)) as 'level_Pine',avg(if('pollen_data'.'tree_type'='Pollen Count Unavailable',pollen_data.'level',null)) as 'level_Pollen Count Unavailable',avg(if('pollen_data'.'tree_type'='Privet',pollen_data.'level',null)) as 'level_Privet',avg(if('pollen_data'.'tree_type'='Ragweed',pollen_data.'level',null)) as 'level_Ragweed',avg(if('pollen_data'.'tree_type'='Russian Olive',pollen_data.'level',null)) as 'level_Russian Olive',avg(if('pollen_data'.'tree_type'='Sagebrush',pollen_data.'level',null)) as 'level_Sagebrush',avg(if('pollen_data'.'tree_type'='Salt Cedar',pollen_data.'level',null)) as 'level_Salt Cedar',avg(if('pollen_data'.'tree_type'='Scorpion Weed',pollen_data.'level',null)) as 'level_Scorpion Weed',avg(if('pollen_data'.'tree_type'='Scrophularia (phacelia crenulata)',pollen_data.'level',null)) as 'level_Scrophularia (phacelia crenulata)',avg(if('pollen_data'.'tree_type'='Sedge',pollen_data.'level',null)) as 'level_Sedge',avg(if('pollen_data'.'tree_type'='Sycamore',pollen_data.'level',null)) as 'level_Sycamore',avg(if('pollen_data'.'tree_type'='Unidentified',pollen_data.'level',null)) as 'level_Unidentified',avg(if('pollen_data'.'tree_type'='Unidentified',pollen_data.'level',null)) as 'level_Unidentified',avg(if('pollen_data'.'tree_type'='Walnut',pollen_data.'level',null)) as 'level_Walnut',avg(if('pollen_data'.'tree_type'='Willow',pollen_data.'level',null)) as 'level_Willow' from pollen_data group by year(pollen_data.'submission_date'),month(pollen_data.'submission_date') order by month(pollen_data.'submission_date')

```

Figure 1. SQL statement used to produce figure 1: Each tree type had an average aggregate called on its level attribute. Null values were also accounted for and all the tree types were renamed for figure 1.

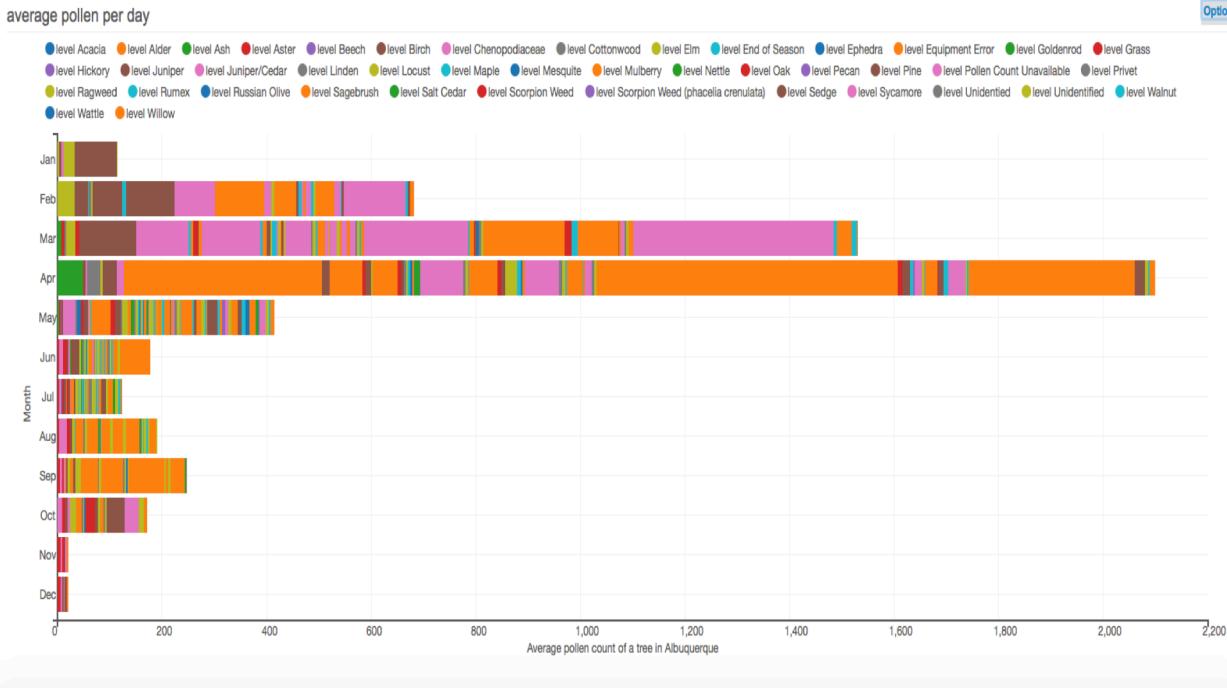


Figure 2. Average pollen count per tree in Albuquerque vs. the month of the year: Each column is made up of its respective tree pollen quantities shown by the different color make up of each column. April is the month with the average highest pollen count. November and December are the months with the lowest average amount of Pollen production.

drawn later on when the weather data is joined into the pollen data. Figure 3, graph (b) shows the number of pollen readings taken per day from 2004 to the end of 2015 for the Eastside. Notice that in 2010 less pollen data was taken implying that there was a low level of pollen based on the trend line for the number of records over all years. Figure 3, graph (c) shows the pollen levels from 2004 to the end of 2015 for the Westside of Albuquerque. Notice that the pollen levels are lower overall for the Westside of Albuquerque and also that the low point for pollen levels was in 2010. Figure 3, graph (d) shows the Number of records taken per day from 2004 to the end of 2015 of the Westside of Albuquerque. Notice that in 2010 there was a significant drop in readings per day.

A closer look at the data is required based on the results of figure 3. Specifically a look at why there is a much lower pollen count in the period of 2009-2014 when this is compared with the rest of the data in graphs (a) and (c). Before we factor in weather data to investigate a closer look at what trees are contributing the most will be conducted.

In Figure 3 we further explore what the major contributors are, in terms of tree type, to the pollen levels in the Albuquerque area. Firstly figure 3 shows that mulberry is the leading contributor to pollen levels, more so than any other plant type that was included in this figure. Juniper and cedar is the next big pollen producer in the Albuquerque area (keep in mind that this data reflects both data collected from the Westside and Eastside of the Albuquerque area). Surprisingly this graph also further supported the need to investigate why pollen levels dropped off so heavily from the

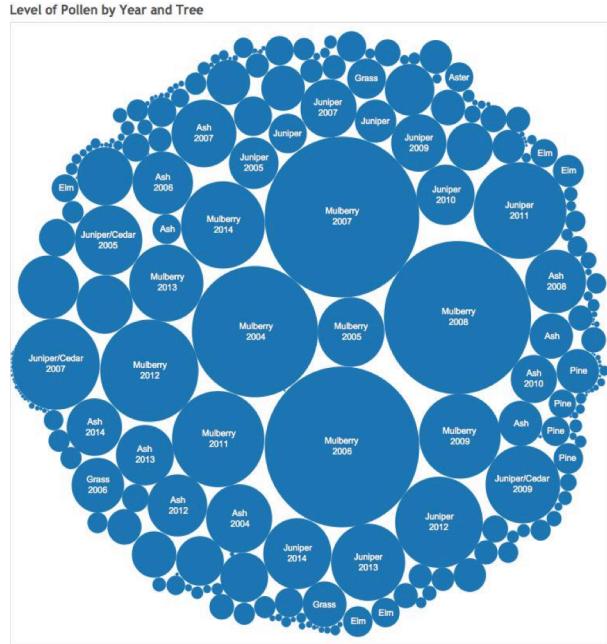


Figure 4. Diagram showing the highest pollen producing plants by level of pollen produced per year. The larger the circle the higher the amount of pollen, circle size is directly related to the amount of pollen produced that year for the specified plant.

Location Comparison Level of Pollen vs. Number of Records

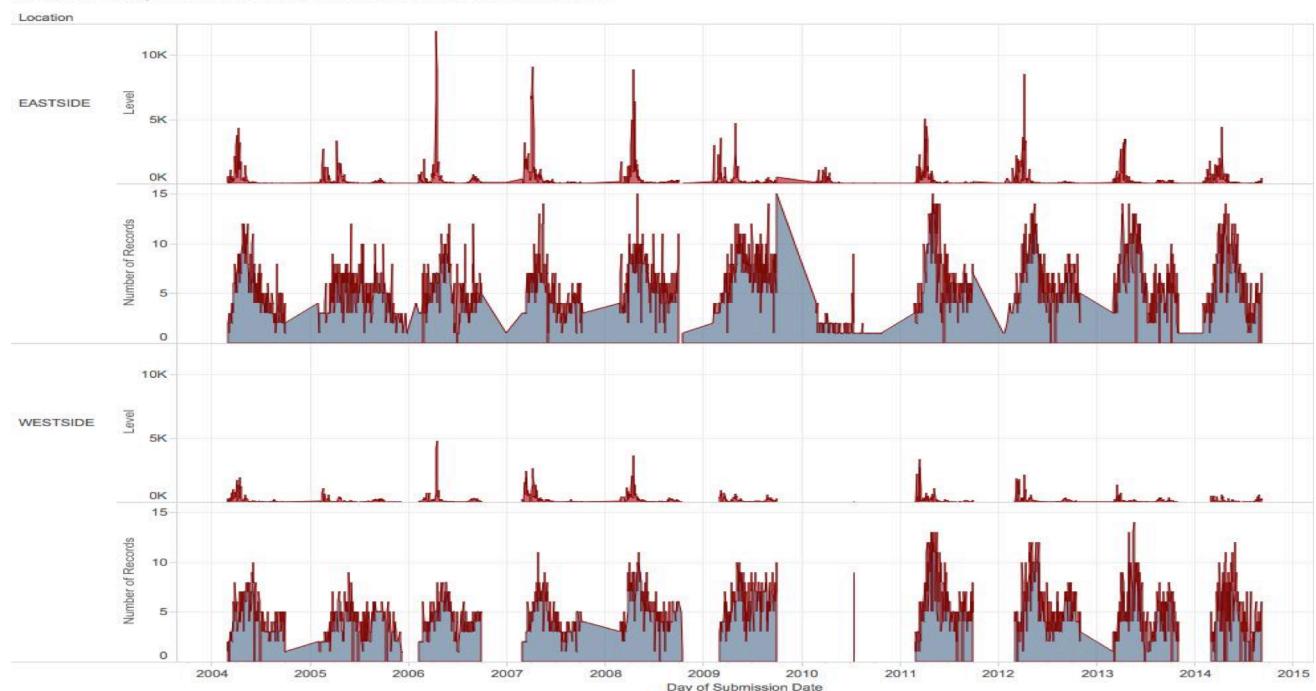


Figure 3. Location Comparison of the Level of Pollen and the number of records taken by day over the span of data acquisition and location. (a) Level of pollen at any point in time for the Eastside of Albuquerque. (b) Number of records taken per day for the Eastside of Albuquerque. (c) Level of pollen at any point in time for the Westside of Albuquerque. (d) Number of records taken per day for the Westside of Albuquerque.

2009 period onward. As shown in figure 3 and figure 4, pollen levels drop off dramatically from their high point in 2008. Next strictly the leading pollen producers will be plotted in figure 5 much like how they are plotted in figure 4 after being run through more data filters.

Figure 5 clearly shows mulberry as the overall largest pollen producing plant in the Albuquerque area. With Juniper cedar and Ash being the other major pollen producing plants. Mulberry nearly accounts for half of all the pollen production.

The results of a more refined search on Eastside vs. Westside pollen data where the two fields are stacked on each other, is examined by figure 6. The data was plotted in this manner to highlight the large drop in readings and pollen levels in the year 2010. In the section on joining weather data and pollen data this issue will be addressed. One possibility that can't be overlooked is the there is a gap in the Albuquerque pollen data set. This was never entirely ruled out but later on we will see a strong correlation with the pollen levels and the weather data. Suggesting there is no gap in the pollen data but rather that the data showed an extreme low in the pollen levels due to a weather phenomenon.

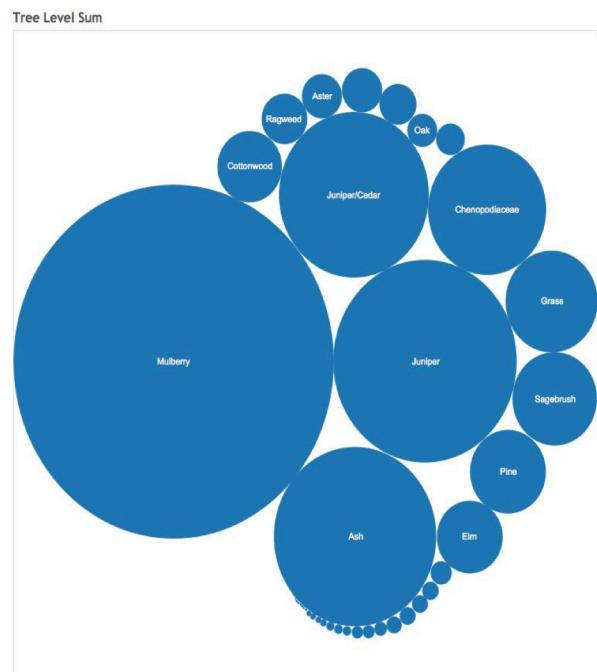


Figure 5. Diagram showing the highest pollen producing plants by level of pollen produced over all the data. The larger the circle the higher the amount of pollen, circle size is directly related to the amount of pollen produced for

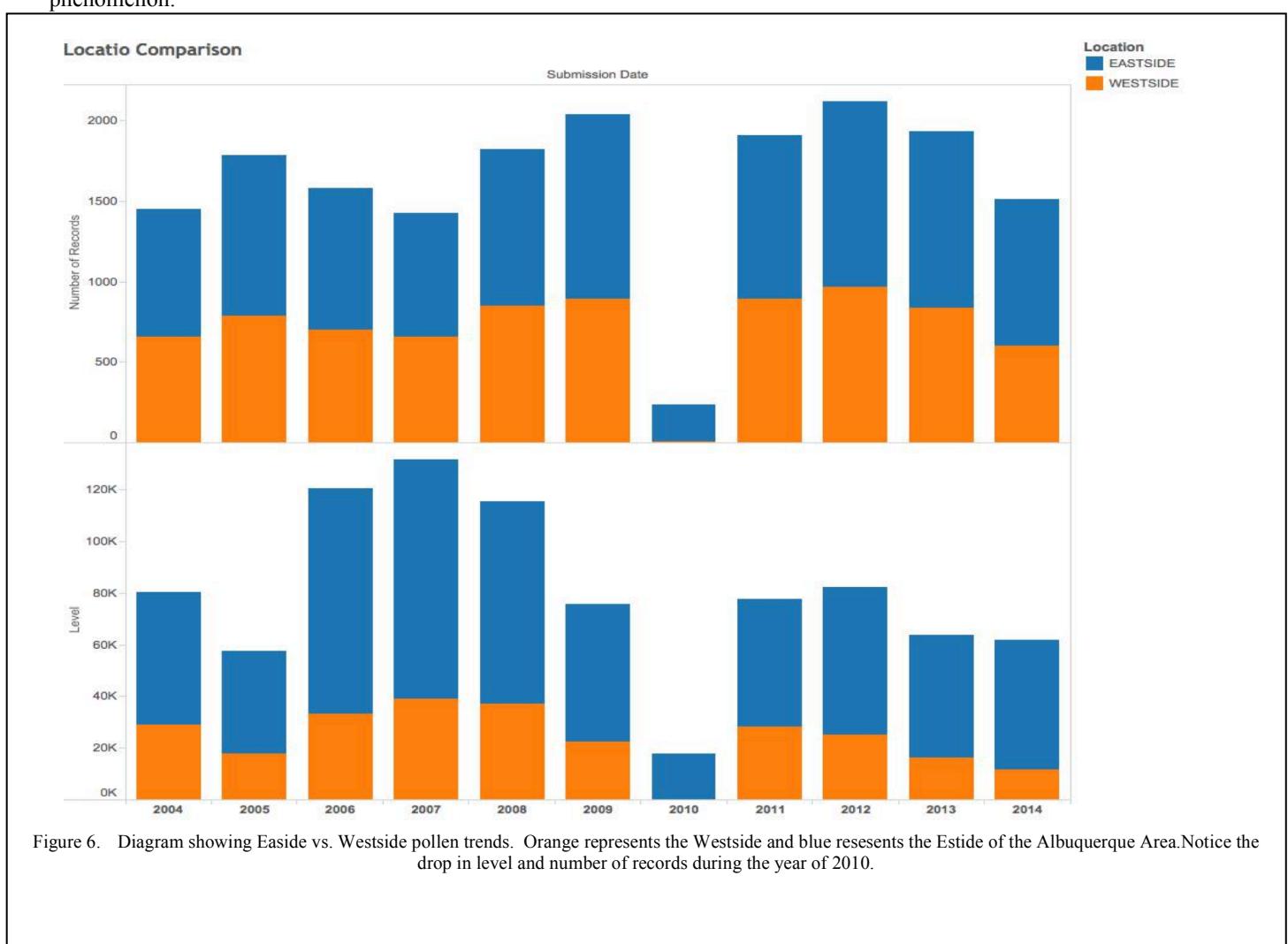


Figure 6. Diagram showing Easide vs. Westside pollen trends. Orange represents the Westside and blue represents the Estide of the Albuquerque Area. Notice the drop in level and number of records during the year of 2010.

Combining the Data Sets

In this section, of the results, correlation will be drawn through the joining of the pollen data set with the joining of the weather data set. Aspects of the previous results from the Pollen data set will be further addressed. The major topics covered are the large drop in pollen levels in the 2010 to 2014 time frame, what weather factors contribute to pollen levels, prediction of what years could potentially have elevated or depressed pollen levels and when should you purchase allergy medication based of these results. First we will strictly look at the weather data set and plot out the averages of a few key weather attributes.

Figure 7 shows three weather attributes averaged for each month and plotted as two-trend lines layered over a bar graph. The two trend lines represent the average high and low for each day and the bar represents average precipitation for the month.

Oddly a direct correlation with the average precipitation cannot be drawn from figure 7 that aligns with the data presented in figure 2. In figure 2 we see that the highest pollen months are February, March, and April. In figure 7 we see that the months with the most precipitation are July and August. No remarkable connection was made with precipitation and pollen level from this. One thing that does stick out is that when temperatures begin to rise, pollen levels begin to rise along with them. This suggests that there is a correlation between rising temperatures and the start of allergy season. In order to further understand factors that control pollen levels the two data set were join on their

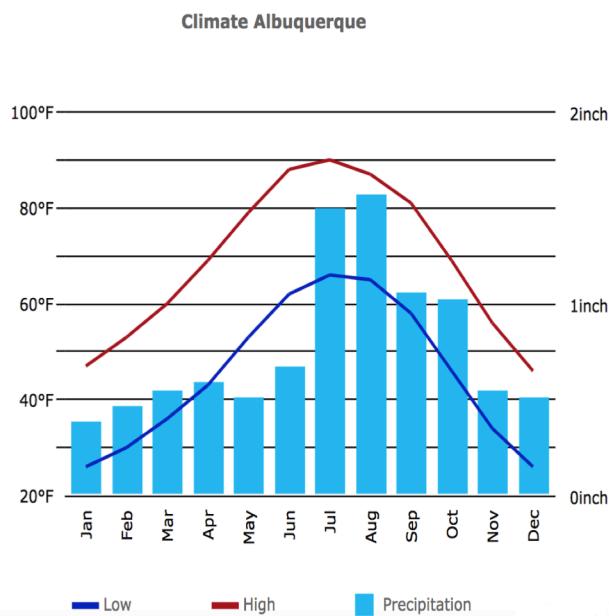


Figure 7. Plotting of weather data averages for the Albuquerque area, with roughly 100 years worth of data. The red line represents the average highs at any given time of the year. The blue line represents the average low for any given time of the year. The blue bars represent the average precipitation for any given month.

respective date attributes.

The resulting set from joining the data on the date attribute was used to make figures 8 and 9. In figure 8 and 9 we can see a direct correlation between the amount of pollen produced for a year and the amount of rain the Albuquerque area got for that year. Another trend that is a little less obvious than the previous one has to do with location. This trend is more easily seen on figure 9. The west side of Albuquerque gets less rain than the east side of Albuquerque, which is reflected in the overall pollen level averages by the respective regions.

The overarching factors to pollen count can now be inferred from figures 8 and 9. The major contributor to pollen levels appears to be whether or not the Albuquerque area is in a drought. The lower pollen levels in the years from 2009 to 2014 can be explained by this postulation. Officially the national weather service published that the Albuquerque area started to fall into a drought pattern in 2009 [7]. Both the pollen data and weather precipitation data corroborate the publish statistic from the National Weather Service. A few postulations can be made now that we have all the data in a form that is understandable.

First we can make a prediction about when to purchase medication. Based off table 1 and figure 1 we can predict when to buy your medication, with certain caveats for drought years. The best time of year to purchase antihistamines is January. Furthermore, on drought years it is best to wait and see whether or not you should purchase medication as the pollen level is a factor of 10 smaller then the average year.

TABLE I. SHELF LIFE OF THE TWO MOST POPULAR ORAL ANTIHISTAMINES ZYRTEC AND CLARITIN. NOTICE MONTHS MOD 12 IS EQUAL TO 0.

Antihistamine	Months
Claritin	36
Zyrtec	60

V. DISCUSSION

Obtaining and Cleaning the Data

Data was obtained from the City of Albuquerque public data set website, ABQ data [5]. The data was presented in the form of an XML file with 80,000 data points. This xml file was parsed and clean for insertion into a relation SQL database implemented with the MySQL engine. A second set of weather data for Albuquerque was obtained. This data set had millions of data points. This data set was obtained as a CSV file from Wunderground and was parsed, cleaned, shrunk to a 10-year span, and inserted into the MySQL database using java [6]. At this point the data was ready for manipulation, interpretation, and prediction. The main reason to upload these two data sets into a MySQL database was for that ability to run inner joins on the data off of the date attribute.

Data Manipulation Tools

Four different data tools were used for manipulation, prediction, and presentation of the data. The first tool used was Ubiq, which is an Analytics tool that integrates with MySQL. This tool was chosen because of its powerful filters and functions and its ability to represent joins on a bar chart. The second analytic tool used was tableau. Tableau is an analytics prediction tool. Tableau's prediction tool was used

for its ability to handle multi thousand line query results, as Ubiq can only handle up to 1000 rows of data. The third manipulation tool was java. Java was used for parsing and cleansing of the data. The final data manipulation tool was MySQL. MySQL is a key player for analyzing big data because it can be used to leverage operations such as joins and the SQL language itself for key insights into Big Data sets.

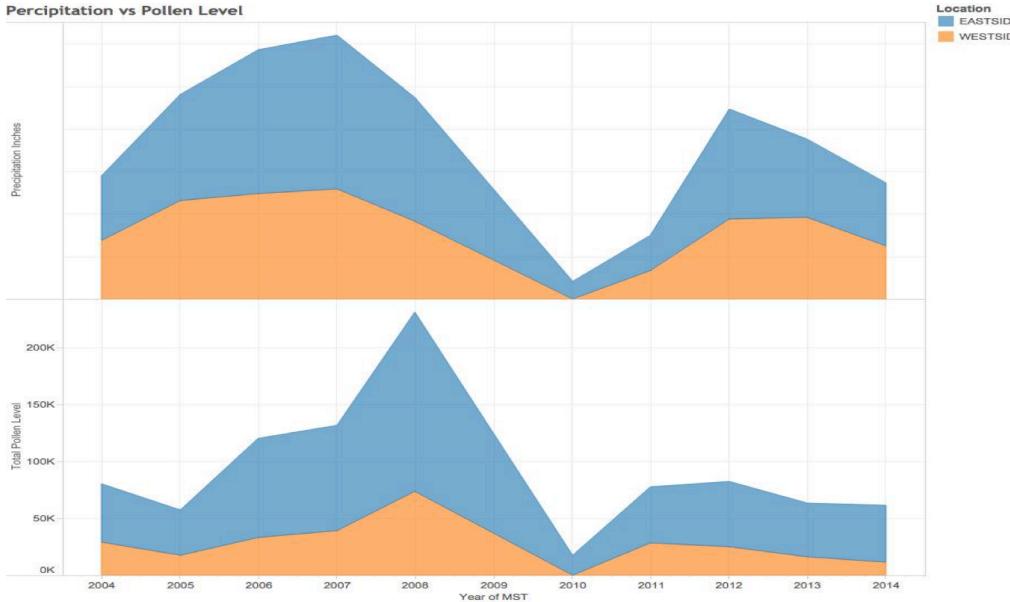


Figure 8. Plotting of precipitation weather data attribute values with the pollen data level attribute joined on dates over the last 10 years. (a) The top graph represents the precipitation for the Eastside and Westside areas of Albuquerque stacked on top of each other. (b) The lower graph represents the pollen levels over the specified range of years. Note the similarity of the two graphs.

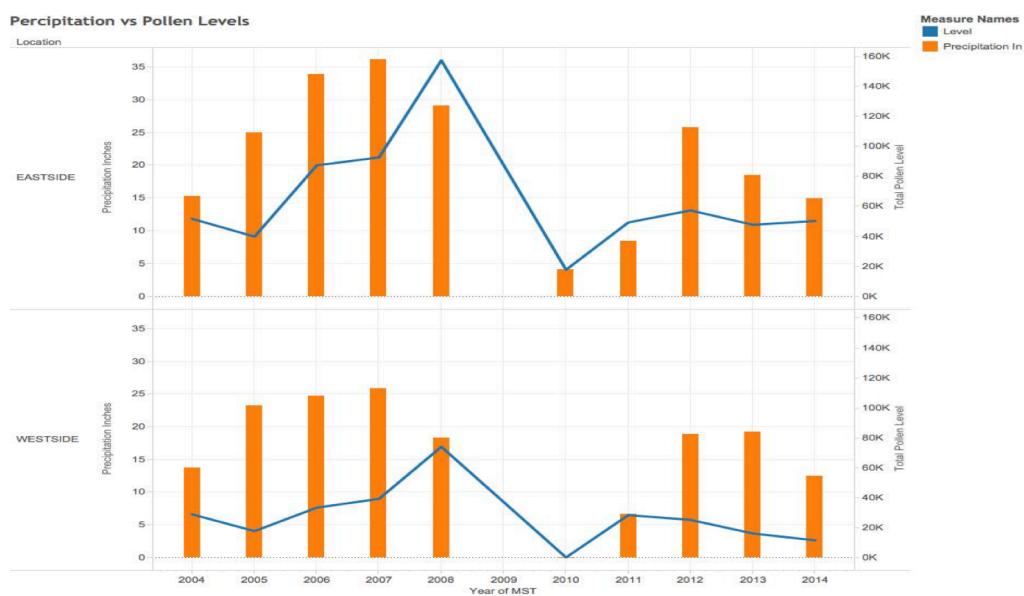


Figure 9. Plotting of precipitation weather data attribute values with the pollen data level attribute joined on dates over the last 10 years. The blue line represents the level of pollen. The orange bars represent the level of precipitation recorded for the specified year.

A determination was made on whether to use MongoDB or MySQL. MongoDB has the advantage of map reduce functionality, which is ideal for reducing the size of the data. MySQL was chosen over MongoDB for the ability to run joins on data sets, which puts the data sets in the context of each other. Next was determining a language to write the parsers in.

As noted in the Design and Approach section, java was chosen as the main programming language for the project because of the java SAX XML parser. Java was further used to parse the CSV weather data and import it into the MySQL database. For all the graphs and data manipulation, the SQL language was used. This language was chosen not just because the data was moved to a MySQL database but for the ability to join data sets. This allowed for a way to combine the data sets.

Map reduce was considered for clustering the pollen data on the weather data. It was determined through research on the National Weather Service website, joining the data was determined to be an adequate way of determining trends and predicting future events. This is why joins were chosen for the pollen prediction. Once the data was in the MySQL database the ability to do joins on the datasets made interaction with the data very flexible justifying the effort for integration of the data into the MySQL database.

Conclusion

Through inferences and correlations, predictions can be made on the data sets. The major correlation that was shown, with figures 8 and 9, is that pollen levels are directly correlated with drought years. A prediction can be made using this result; on drought years a person should wait to buy medication. This is because of the drastic reduction in pollen levels, on the order of a factor of 10 (see figure 8). For non-drought years a more general prediction can be made.

On non-drought years, medication should be purchased in January. This is based on figure 2, where January is the month where the pollen levels begin to rise, on average. When combining this information with table 1, we see that antihistamines have a shelf life in increments of 12 months. Therefore, if a person buys their antihistamine medication in the month of January, they will be able to make it through the entire allergy season without having the medication expire.

Lastly it can be said that a small fraction of plants produce the majority of the pollen. Most notably mulberry making up 40% of all pollen produced in the Albuquerque area. Due to this fact, it is recommended that if a person is allergic to mulberry they should either buy double the amount of allergy medication in January or live in the

Westside of Albuquerque as this portion of town has much lower pollen levels.

ACKNOWLEDGMENT

The purpose of this research paper was as a final project for the CS 591 Big Data class at the University of New Mexico. Special Thanks to Trilce Estrada; she was a very knowledgeable teacher for Big Data and made the content interesting.

REFERENCES

- [1] D'Amato, G., et al. (2007), Allergenic pollen and pollen allergy in Europe. *Allergy*, 62: 976-990. Doi: 10.1111/j. 1368-9995.2007.01393.x
- [2] Mensing, A. M., et al. (2001), A Holocene pollen record of persistent droughts from Pyramid Lake, Nevada, USA. Doi:10.1016/j.yqres.2004.04.002
- [3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
- [4] Kaszuba SM, Baroody FM, deTineo M, et al. Superiority of an intranasal corticosteroid compared with an oral antihistamine in the as-needed treatment of seasonal allergic rhinitis. *Arch Intern Med*. 2001;161(21):2581-7. [PubMed]
- [5] "ABQ Data." City of Albuquerque. City of Albuquerque, n.d. Web. 06 Oct. 2014. <<http://www.cabq.gov/abq-data/>>.
- [6] "Albuquerque, NM." Weather Forecast & Reports. N.p., n.d. Web. 29 Nov. 2014. <<http://www.wunderground.com/>>.
- [7] "Weather History for Albuquerque." *National Weather Service*. N.p., n.d. Web. 30 Nov. 2014. <<http://forecast.weather.gov/>>.