Air Quality Fine Particulate Data Exploration for U.S. 2017

Deanna L. Schneider

University of Wisconsin – Eau Claire

Air Quality Fine Particulate Data Exploration for U.S. 2017

In the summer of 2018, my home became a host site for an air quality monitor sponsored by the local neighborhood association. The monitor collects a sample every 20 seconds and reports data online every 20 minutes at www.purpleair.com. My home sits next to a manufacturing plant and has been subject to numerous rounds of soil and soil vapor testing since 2003. However, this is the first time it's been a site for air quality monitoring. I know little about air quality monitoring but realized that this is an excellent opportunity to begin to do some data exploration. The monitor measures atmospheric particulate matter (PM) in micrograms/cubic meter.

Data Used

This exploration focuses on PM 2.5 data that is available from the EPA Air Quality System. Of the EPA Data, only 2017 data will be used (the most recent complete year). In total, the data is 7.26 GB in size. The majority of the data is available on the EPA website.

The following files are used:

- hourly_RH_DP_2017.csv
- hourly_TEMP_2017.csv
- hourly_PRESS_2017.csv
- hourly_WIND_2017.csv
- hourly_880101_2017.csv
- aqs_sites.csv

One U.S. Census Bureau file is used, which can be downloaded and saved as division.csv. Finally, Purple Air monitor data from 10/4/2018 to 10/17/18 for the SASY 1 monitor is used. The data can be downloaded from the Purple Air site and saved as Sasy1.csv for this project. A zip file of all data, appropriately named, can be downloaded from Google Drive.

Directions

To run this code, follow these steps:

1. Download and unzip that data files. If you are running this on Hortonworks, the data files should be stored in user/zeppelin. If you are running this on AWS, create an S3 bucket for the files.
2. Import the Schneider-Problem3-Scala.json file into Zeppelin. This will create a notebook called "Schneider-Problem 3."
3. Open the Schneider-Problem 3 notebook. If you are running this on AWS, update the path variable in the second paragraph to the URL for your S3 bucket.
4. Run the paragraphs in order.

Interactive Visualizations

Online, interactive visualizations of the questions in this report are available on Tableau Public.

Question 1

The Purple Air website reports PM2.5 data in different categories, which correspond to various levels of health concern.

PM2.5 Reading	Health Effects
0-12	Air Quality is Considered **Satisfactory** and air pollution poses little or no risk
12-35	Air Quality is **Acceptable**, however for some pollutants, there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution
35-55	Members of **sensitive** groups may experience health **effects**. The general public is not likely to be affected.

PM2.5 Reading	Health Effects
55-150	**Everyone** may begin to experience health **effects**; members of more sensitive groups may experience more serious health effects.
150-250	**Health Warning**s of emergency conditions. The entire population is likely to be affected.
250 +	**Health Alert**: everyone may experience more serious effects.

I am curious about how often readings fall into each of the categories, and whether that varied by geographic area. Initial data exploration indicated that between 96.3% and 98.8% of all readings by region fall into the Satisfactory category. With such a monumental skew towards the low end of the spectrum, it's hard to get a sense of where problem areas might be. Filtering out all the Satisfactory readings, we can see how the other categories are represented in each region.

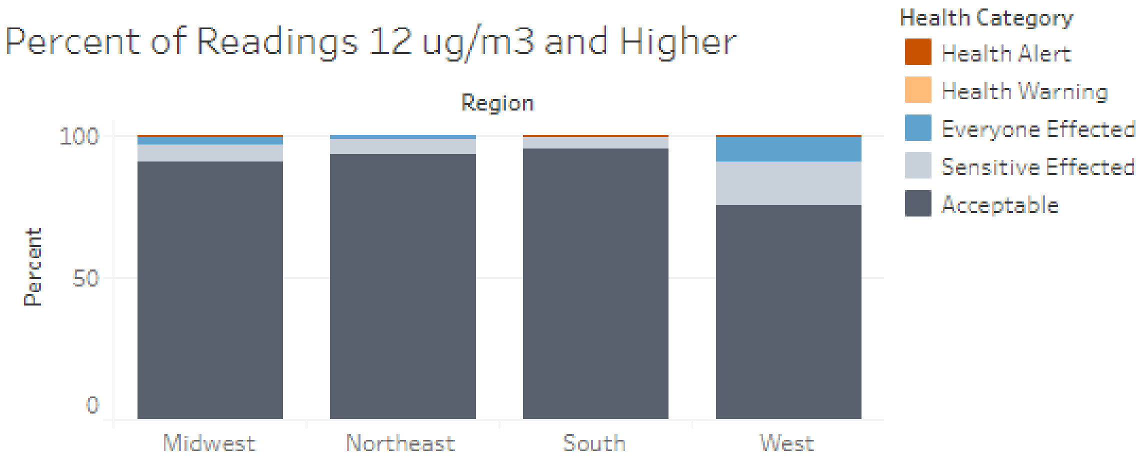


Figure 1 Percent of Readings 12 μg/m^3 and higher by U.S. Census Bureau Region.

The Western region has the highest percentage of readings that have the capacity to cause health concerns (over 20% of the readings over 12 μg/m$^{3)}$ and the highest percentage in each of the health categories above Acceptable except the Health Warning category, where the Southern region has the highest percentage with .037%.

Question 2

Each of the monitor sites is coded as being in one of three settings: Rural, Suburban, or Urban and Center City. Given that most readings fall into the Satisfactory or Acceptable levels, for this question, both of those levels were eliminated and only readings that were at or above 35 µg/m^3 were considered. For those readings, I asked how many readings were found at each location type, by day of week. I was curious to know if there was a pattern – for instance weekends showing fewer high readings than other days of the week due to lower traffic volume or industrial businesses that do not operate on the weekends. However, no such pattern was noted.

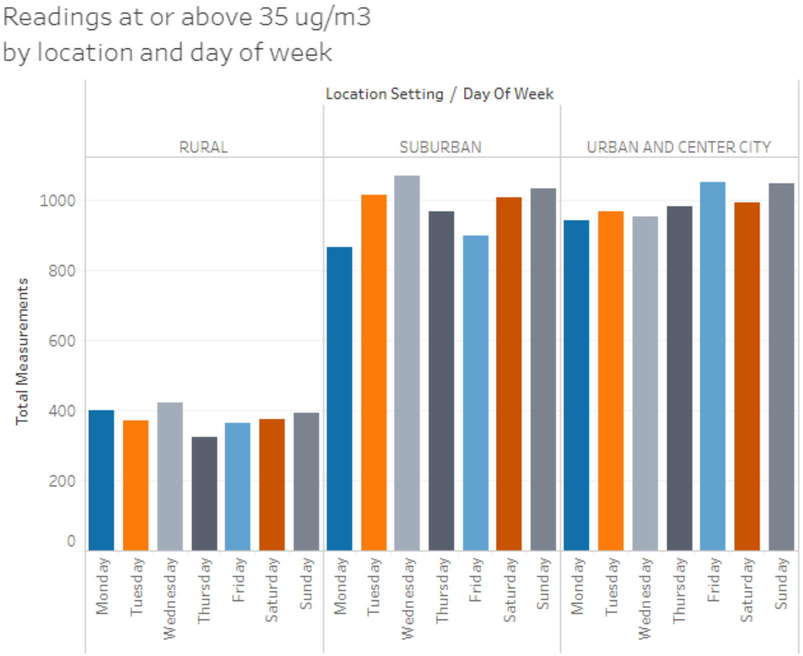


Figure 2 Count of readings 35 µg/m^3 and higher by monitor setting and day of week.

Question 3

Since acquiring a monitor of my own, I am particularly interested in monitor sites near me. For this question I asked the data for the 10 monitor sites nearest to my latitude and longitude that had readings at or above 55 $\mu g/m^3$, how many readings there were at or above that level, and what the location setting was for each of those monitors. The data showed that the closest monitor was approximately 1 km from my house, in an urban setting. It had 2 readings. Not surprisingly, there were several monitors in the nearby urban centers of Milwaukee and Chicago. Perhaps most surprising is a lone rural monitor in Grant County, Wisconsin.

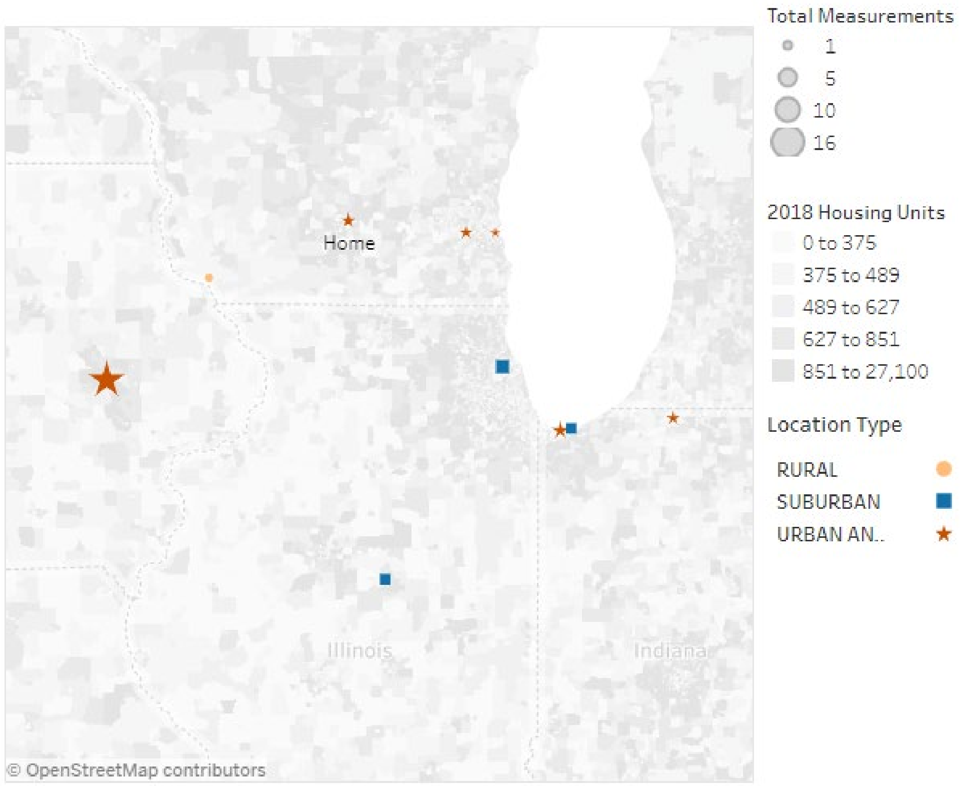
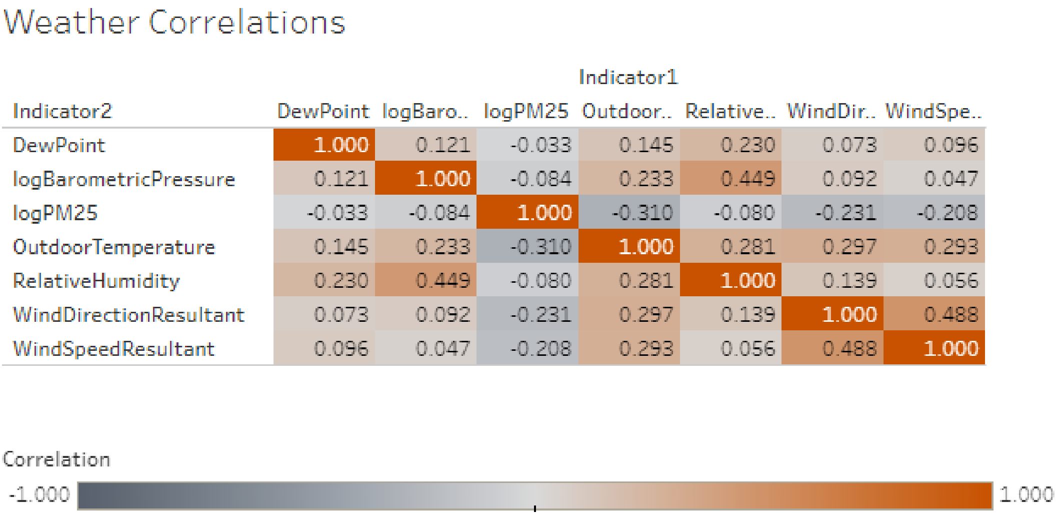


Figure 3 Ten monitors nearest my home with readings over 55 $\mu g/m^3$ during 2017. Map base indicates housing density.

Question 4

Most monitors in the EPA Air Quality program also collect basic weather information. I was curious as to whether there were any correlations between weather variables and particulate matter readings. In exploring the data, it was clear that both the particulate matter readings and the barometric pressure readings had extreme skew. Those two variables were log transformed and Pearson's correlation coefficients were calculated. There were no strong correlations between logged particulate matter readings and weather variables.



Weather Correlations

				Indicator1			
Indicator2	DewPoint	logBaro..	logPM25	Outdoor..	Relative..	WindDir..	WindSpe..
DewPoint	1.000	0.121	-0.033	0.145	0.230	0.073	0.096
logBarometricPressure	0.121	1.000	-0.084	0.233	0.449	0.092	0.047
logPM25	-0.033	-0.084	1.000	-0.310	-0.080	-0.231	-0.208
OutdoorTemperature	0.145	0.233	-0.310	1.000	0.281	0.297	0.293
RelativeHumidity	0.230	0.449	-0.080	0.281	1.000	0.139	0.056
WindDirectionResultant	0.073	0.092	-0.231	0.297	0.139	1.000	0.488
WindSpeedResultant	0.096	0.047	-0.208	0.293	0.056	0.488	1.000

Figure 4 Correlations between logged PM2.5 readings and weather readings.

Question 5

I am curious about how my monitor's readings compare to the monitor closest to me. EPA data is only released every 6 months, so there are no overlapping date readings yet. However, given the lack of correlation with weather variables, I compared the readings from the same dates in October, using 2017 data for the EPA monitor and 2018 data for the monitor at my home. My monitor collects a sample every 20 seconds and reports samples every 20 minutes.

The EPA data is using 8-hour averages. To compare equal time frames, both datasets were averaged by day and compared. For this brief snapshot in time, my monitor has reported higher daily averages on 11 of the 14 days, with the largest difference being on October 9th, when there was a 11.21 µg/m^3 difference. On that date, my monitor would have had a daily average in the Acceptable category, while the nearby monitor would have been in the Satisfactory category. A difference large enough to span categories happened on 2 of the 14 days compared.

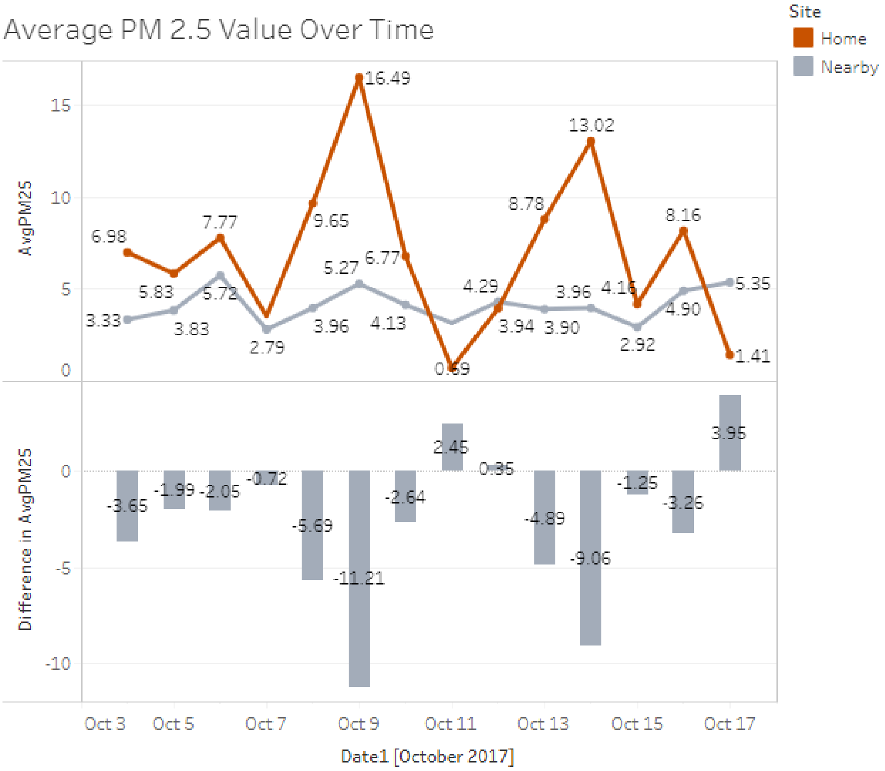


Figure 5 Difference in nearby monitor and home monitor readings - 2017 and 2018 comparison.

Conclusion

This analysis was exploratory only. No conclusions should be drawn about causality or predictability from any of the question answers. Further analysis would be required to understand the “why” associated with the “what” of these questions.

References

Air Quality Monitoring. (n.d.). Retrieved from https://www.purpleair.com/

AirData website File Download page. (n.d.). Retrieved October 1, 2018, from https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw

Barrett, P. (2017). Skewness and Pearson Correlations: Attenuation of coefficient size as a function of skewed data. Retrieved from http://www.pbarrett.net/techpapers/skewness.pdf

Halpert, C. (2014, June 23). Census Regions. Retrieved October 2, 2018, from https://github.com/cphalpert/census-regions

Rahman, M. (n.d.). How can I deal with negative and zero concentrations of PM2.5, PM10 and gas data? Retrieved from https://www.researchgate.net/post/How_can_I_deal_with_negative_and_zero_concentrations_of_PM25_PM10_and_gas_data