Air Quality and Mortality in the United States

Joseph McAndrews, Colleen Morse, Sunil Prasad, Valerie Schnapp, Jonathan Yang

1 INTRODUCTION

Breathing clean air is vital to life. From traditional medicinal practices to modern disease control, access to fresh air has long been integral to a healthy life. Each pollutant has a different chemical makeup and each is therefore likely to affect the human body differently. This project seeks to explore the relationship between specific air pollutants and the possible mortality effects they may have on the population that breathes the contaminated air.

1.1 Motivation

There are many organizations dedicated to tracking air pollution. These organizations' activities are mainly focused on environmental concerns like climate change. There are also a number of organizations dedicated to tracking the health trends of populations. These organizations include international bodies like the World Health Organization, national bodies like the Centers for Disease Control (CDC), and state agencies like the state Centers for Disease Control. However, during the course of our research, we have not found a tool that attempts to show correlation between air quality and health results. Although we are not public health professionals nor environmental scientists, we were motivated to create a tool to help decision makers in those fields to quickly and easily see how their data collection efforts coincide. It is our hope that these visual tools will foster collaboration and insight into how these diverse fields may overlap and compliment one another.

2 LITERATURE SURVEY

Historically, poor air quality has caused serious damage. The EPA was established in the United States after a town near Pittsburgh saw dozens of deaths following a thick smog that lasted multiple days [4]. It was agreed a combination of air pollutants, wind pattern changes, and a unique air entrapment were the reasons for the smog, however Brimblecombe adds it was difficult to find clear evidence of the smog's long term effects. Rapid urbanization and globalization in India was found to lead to severe degradation of air quality.[14] Venkatesan and Thirumal found the air severely impacted human and animal health. California saw a similar link between air quality and health in recent years. As air quality improved, lung-function development in children with and without asthma also improved.[8] Health effects are not just respiratory and studies have shown links to other negative health outcomes like obesity. [1]

Particulate matter (PM) was found to significantly contribute to health issues, however the authors admit to finding it difficult to separate specific pollutant causes and instead suggested a multi-faceted case involving socioeconomic status and weather. [6] In 2002, Pope et al found fine particulate matter and sulfur-oxide pollution were associated with higher rates of mortality for those with cardiopulmonary complications compared to the mortality rates for those without complications [11]. These health effects have been seen to be associated with financial costs, however the authors admit to a possible over-simplification of the association between air pollution and human mortality.[7]

Sometimes "pollution" is too general a term. To help solve the problem of air pollution, the exact cause helps to determine effective solution. Low cost sensors deployed in a network at London Heathrow airport helped to pinpoint the majority of Nitrogen Dioxide (NO2) pollution was actually coming from non-airport components. [12] Further research found that NO2 levels fluctuate throughout different times of day as well as different years in the early 2000s. The authors found that fluctuations were heavily dependent on topography, traffic, and wind patterns among other variables. [2] Researchers also need more tools to determine cause and effect. Ozone layer researchers know that certain chemicals cause damage, but there is less clarity on what exactly is emitting these chemicals. [13]

Smartphone geolocation has been validated to provide accurate enough data to correspond with air pollution models in a given city. This way, a person can know an accurate measure of their own air pollution exposure just by carrying around a powered-on smartphone. [9] Openair is an R package developed for rapidly and easily importing and displaying air quality data so that users can easily change how they visualize data. This allows decision makers more flexibility in how data is presented, improving insight into multi-faceted problems.[5]

Improving accessibility to air quality data is not a new feat. Traditional sensors are typically local based and do not provide ease of data connectivity. Proposals for sensor improvements include using cloud infrastructures to connect and stream detail air quality data near real time to allow for continuous monitoring. [16] Other monitoring proposals involve using Unmanned Aerial Vehicles (UAVs) to capture a wider scope of air quality data. [15] The United States currently uses an Air Quality Index value that is generated by the EPA. Other organizations around the world have developed different methods for calculating air quality that may improve the interpretation of the values for a specific purpose or population. [3] In these cases, improving data delivery and sharing will greatly help the public and those in leadership become informed in air quality.

Collecting the data is only the beginning as the analysis for air quality data is relatively a new feat. New machine learning methods using neural networks have attempted to predict air pollution levels for varied pollutants. [10] These methods are still in development and future studies would benefit from exploring their uses further.

3 PROBLEM DEFINITION

Our study is two-fold. First, we aim to show Air Quality Index (AQI) data from the Environmental Protection Agency (EPA) at the state and county levels as far back as 1980 in an attempt to show trends across the US states. Second, we aim to show whether or not there is any correlation to specific air pollutants and cardiopulmonary-related mortality based on yearly aggregate data from the EPA and CDC.

It is hypothesized that Air Quality is a significant factor in predicting cardiopulmonary mortality rates.

4 METHODS

Our project sourced its data from the EPA's Air Quality Index as it was believed to be the most reliable data available. This was merged with mortality data from the CDC database. From these data, we identified and visualized the relationship between air quality and cardiopulmonary mortality rate and used Machine Learning algorithms to classify which pollutants were worst for mortality rates and predict negative mortality outcomes. Our concept of operations is summarized in Figure 1.

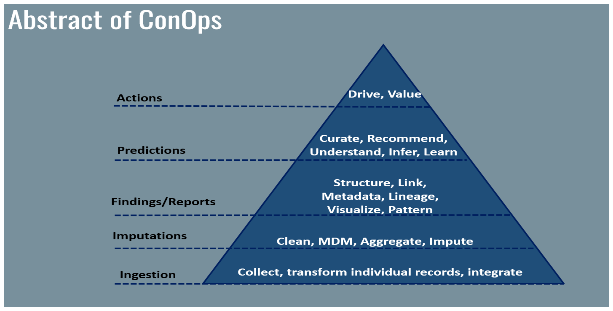

Abstract of ConOps
Actions
Drive, Value
Predictions
Curate, Recommend, Understand, Infer, Learn
Findings/Reports
Structure, Link, Metadata, Lineage, Visualize, Pattern
Imputations
Clean, MDM, Aggregate, Impute
Ingestion
Collect, transform individual records, integrate

Figure 1

4.1 Intuition and Innovations

Despite the clear link between air quality and health, there is no resource that combines the two related fields of study in a visual and interactive way. Our project will improve over the status quo in three distinct ways. First, this will be the only associated visualization that merges air quality (with specific pollutants) and mortality outcomes. We feel that this alone is a significant improvement over what is available now and will be useful to many different communities. Next, we will digitize old data going back to the 1980s. Although this data exists, there is no interactive database that shows visualizations of how mortality rates and air pollution have changed over the years. Finally, we will apply machine learning techniques to make connections that are not easily observable by humans. These insights could lead to better health outcomes for people in need.

4.2 Description of Approach

Our approach to this problem is outlined in Figure 2.

4.2.1 Sourcing Data. Because air quality and mortality information is very important to the government, there is readily available long-form data from reputable government agencies. For air quality data, we are using the EPA's Air Quality Index. For mortality data, we are using the CDC's National Center for Health Statistics Mortality Data. Both datasets are freely available to the public on each respective website. Since the data is near 700 MB, we hosted it in the Google Cloud Platform File Browser. For easy querying, we ingested the data into a

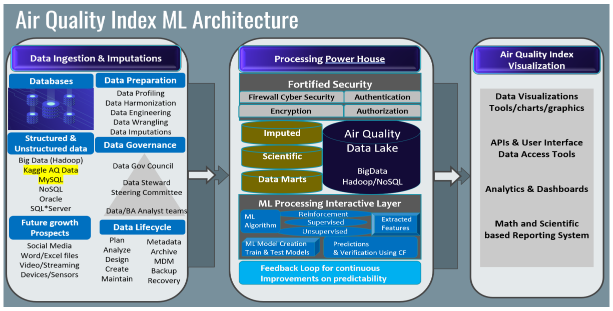


Figure 2

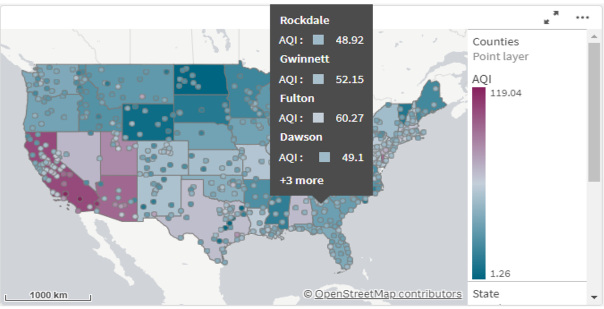


Figure 3

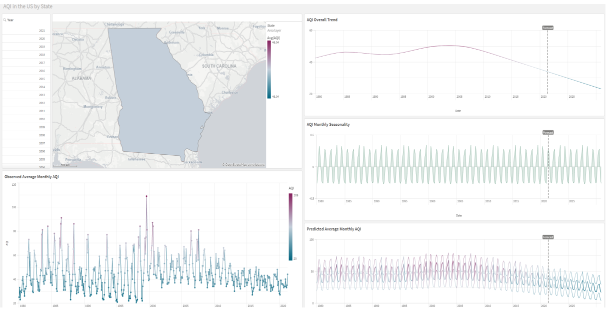


Figure 4

MySQL database, which has a direct connection to our Qlik instance.

4.2.2 Data Preparation. The AQI data, while extensive, contained air quality information from 1,048 counties, omitting over 2,700 counties nation-wide. Information in the data included the state name, county name, exact date of data collection, AQI number, AQI category, the pollutant measured, and the latitude and longitude of the measurement. Dates of collection range from January 1, 1980 through May 6, 2021.

The CDC's mortality data includes yearly death counts down to a county level. For AQI-mortality association analysis purposes, we decided to aggregate the data to the state level due to the AQI data missing so many counties. The CDC Wonder querying app was used to pull both cardiopulmonary-associated death counts as well as general death counts across each state in the continental US for each year available back to 1980. However, changes in coding (ICD-9 versus ICD-10) made exact comparison difficult prior to 1999. While the AQI data includes observations as recently as May of this year, the CDC was only able to provide mortality data as recently as 2016.

After our data ingestion was complete and data cleaned and prepared, we then hosted the Datalake on Qlik.

4.2.3 Data Visualization. Qlik gives our group a clean dashboard to do our project and also allows our group to quickly and easily collaborate on different visualization methods. We used this interactive space to allow users to generate maps showing pollutant type and death rate in a given year. Figure 3 shows this on Qlik, with the associated tooltip showing AQI value for some of Georgia's counties, including Fulton County. The county-level analysis was difficult to integrate with the rest of the project because there so many counties were missing data in the AQI data set. As you can see, not all counties are represented in the visual. This section leaves room for future analysis.

It is also possible to query a certain state and view that state's specific data. Figure 4 shows the AQI trend, AQI monthly seasonality and predicted average monthly AQI for Georgia.

That last piece of visual analysis we did was to visualize the relationship between air pollution and mortality. This is one of our major innovations because to date there is no web tool that provides this service. Figure 5 shows an example of an AQI vs Mortality scatter plot. Although the user can generate any relationship they want, this one happens to show the relationship between SO2 concentration and cardiopulmonary death rates in the state of Virginia. This graph is typical of what we saw across the US and the relationship is not as strong as we would have thought. It does show a slight increasing relationship between mortality and AQI, but there are many data points that do not conform to this. An AQI score less than 50 is currently considered to be "good" air quality, however all of these data points are associated with AQI levels below that threshold. Future work may benefit from exploring whether or not the

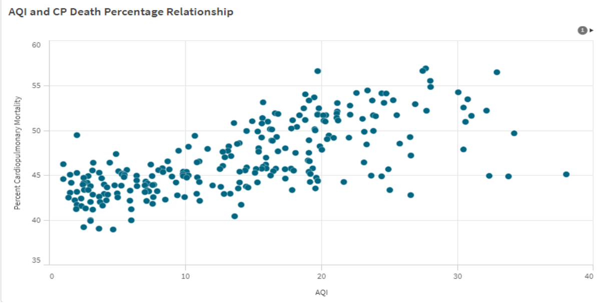


Figure 5

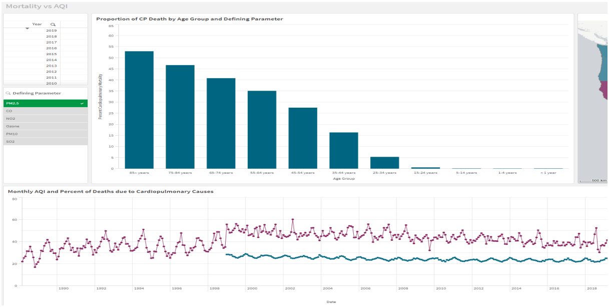


Figure 6

AQI threshold should be adjusted for different types of air pollutants.

Our Qlik dashboard can also display this same data in a line chart. This shows no clear relational effect between pollution and subsequent death, however both appear to be decreasing slightly over time. An example is provided in Figure 6.

4.2.4 Analytics. Fundamentally our project aims to discover if air pollution is correlated to mortality rates. If a relationship can be established, we will then analyze whether a certain pollutant is more harmful than others. However, as discussed above, there was no strong correlation between AQI and cardio pulmonary mortality. Indeed the most important factors as determined by Random Forest Regression were clearly age. As age increases, so does likelihood of cardio pulmonary mortality. Figure 7 below shows the most important factors.

We built a dashboard that allows the user to access the results from our Random Forest Regression model. On the left hand side of the screen, we show our train and test fitting. This is shown in Figure 8.

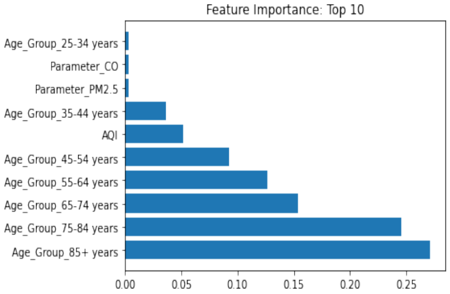


Figure 7

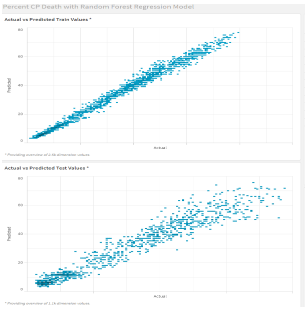
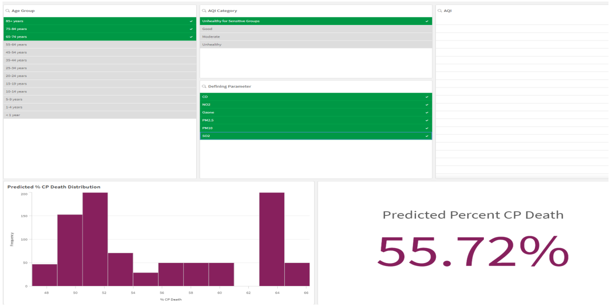

Percent CP Death with Random Forest Regression Model
Actual vs Predicted Train Values *
Predicted
80
60
40
20
0
Actual
* Providing overview of 2.5k dimension values.
Actual vs Predicted Test Values *
Predicted
80
60
40
20
0
Actual
* Providing overview of 1.1k dimension values.

Figure 8


Age Group
85+ years ✓
75-84 years ✓
65-74 years ✓
55-64 years
45-54 years
35-44 years
25-34 years
20-24 years
15-19 years
10-14 years
5-9 years
1-4 years
< 1 year

AQI Category
Unhealthy for Sensitive Groups ✓
Good
Moderate
Unhealthy

Defining Parameter
CO ✓
NO2 ✓
Ozone ✓
PM2.5 ✓
PM10 ✓
SO2 ✓

AQI

Predicted Percent CP Death
55.72%

Figure 9

On the right hand side of the screen, the user can choose the parameters of the model and access the predictive rate of cardiopulmonary mortality. In Figure 9, we show only age groups 65 and older with AQI in the "Unhealthy for Sensitive Groups" category, using all pollutants. This shows that there is a 55.72 percent chance of cardio pulmonary mortality, which may seem like a strong correlation. However, since the most important factor is age, the predicted value does not change very much for any concentration of air pollution. The user is free to draw as many graphs with as many different combinations as they would like. It is possible that an expert in the field could use these visualizations to draw a more nuanced conclusion than we were able to.

5 EXPERIMENTS AND EVALUATION

Our testbed is made up of data taken from the CDC and EPA hosted on Qlik. We used Qlik for rapid prototyping and testing of various visualization techniques. Our best techniques and visualizations were included in our final Qlik dashboard and were used to illustrate the relationships between air pollutant and cardiopulmonary mortality rates on a state level. Although we were not able to achieve a county-level analysis in our final project, we have laid the ground work for more detailed analysis in future studies when additional data might exist. In the following sections we will detail the experiments we did and how we arrived at our final conclusions. The visualizations on Qlik form our descriptive analytics portion. For predictive analytics, we utilized Python with Pandas, Numpy, and statsmodels packages. We determined which factors led to worse mortality outcomes and made a prediction of mortality based on AQI.

5.1 Questions To Answer

Our project seeks to answer the following questions:

- How does air pollution correlate to cardiopulmonary-associated mortality rates?
- Are certain air pollutants more associated with mortality rates than others?
- Can mortality predictions be made from air quality data?

5.2 Details of Experiments

5.2.1 Experiments. Many potential relationships are available to explore in the fields of air quality and mortality. Air pollution can cause health effects much more diverse than just cardiopulmonary-related mortality. However, our team has decided to set aside those potentially interesting relationships while we explore the relationship of pollutant vs. cardiopulmonary-related mortality. Exploring additional health conditions and additional factors of those conditions are outside the scope of our current research but could be valuable topics for future research.

At first we wanted to do a county-level analysis. This would have allowed a very localized approach to this problem. We thought that the more local we could make our project the better because air pollution, although it can travel many miles, would be most harmful to the population living closest to the source. If we analyzed at the state level, the majority of the population might not be affected by pollution that was absolutely devastating to the small group of people living nearby.

Ultimately, we were not able to do a county level analysis because there was far too much missing or inconsistent data at the county level. We found that the federal government and state governments had very good data for the state level. This is likely because most states have similar agencies responsible for this data collection, while counties vary drastically in their resources and agencies. The only thing we were able to do at the county level was allow a user to view the AQI for a given county in a given year. If data was unavailable, it was not displayed. This can be seen in our United States Air Quality Dashboard.

Using state-by-state analysis, we were able to complete all of the tasks we set out to accomplish. For visual analytics, we created a dashboard that allows a user to view any state's AQI data for a given year. We also used the FBprophet package which takes a probabilistic approach to predict the data using Bayesian techniques to predict the trend and seasonality of AQI data. Markov chain Monte Carlo methods were used to sample the probability distribution to estimate the mean prediction. This will allow decision makers to quickly understand the air quality in their states.

We also created the Mortality vs. AQI dashboard to show a user how pollution correlates to cardio pulmonary mortality. The user is able to draw line and scatter plots showing the relationship between air pollution and cardio pulmonary mortality. This is our best achievement in this project because no other tool available does this. Most of the experimentation for this section came in the form of data wrangling and merging. It was difficult to aggregate the large datasets from the CDC and EPA in order to show the relationship. We ended up solving this issue by hosting the large datasets on Google Cloud and using SQL queries to merge and manage the data. Qlik's interfaces allowed for the customizable drawings made by the user.

The final task of this project used machine learning to see if there were any strong predictors of mortality using AQI data. For this, we experimented with many techniques but ultimately decided to use random forest regression. This is because random forest regression is a robust algorithm for large data sets and is able to handle missing data. As we saw using just our visual approach, link between AQI and mortality was weak and inconsistent across age groups and pollutants, which in itself is a big discovery. This is contrary to what we thought would happen. We also used a basic linear regression model which did show a strong link between air quality and mortality from cardio pulmonary causes when all factors were aggregated. Although not hosted on our dashboards, this model can be seen in our GitHub repository. This same relationship was not clearly seen in our random forest regression model, which accounted for individual states, pollutants, and age groups.

5.2.2 Evaluations. The biggest question we are left with after this project is why we did not see the strong correlation between pollution and mortality that we were expecting. We think this is because the limitations of our data - specifically the geographic size of each state compared to the relatively localized effects of air pollution. For example, a large state like California was found, on average, to have poor air quality but low rates of cardiopulmonary mortality. However, these numbers come from people living in a large state. People in northern California may not be as affected by air pollution in southern California, compared to those living in Los Angeles. Furthermore, cities on the boarders of states can be more effected by pollution on the opposite side, depending on where pollution is concentrated.

Strangely, these shortcomings were also seen in small states like Delaware and the District of Columbia, where the affect of "dilution" would be less pronounced. This leads us to believe that future analyses could benefit from a much more localized approach.

Finally, we only analyzed cardiopulmonary mortality rates because we thought it would be the most correlated with air pollution. Future studies would likely benefit from exploring additional health-related conditions and causes of death. Our project could then be used as a model for faster development. It is also possible that pollution may show stronger correlations to quality of life measures and non-fatal illnesses. Those data could also be substituted in future research.

Our own work can also be reviewed in the future because the CDC did not provide mortality data past 2016. We used an ARIMAX model to predict cardio pulmonary mortality for those years, and although those years have already happened we do not have mortality data for these years. When using aggregated AQI values as the exogenous factor, the time series forecast for cardio pulmonary mortality follows the AQI trend. This model can be seen in our GitHub repository.

5.3 Observations

The data demonstrates the following:

- AQI data currently does not have data for 2,771 counties, making direct county-to-county comparisons with mortality data not possible and unreliable.
- Counties names, exact parameters, and data collection conventions have changed over the years studied.
- We are only considering the US mainland due the drastic change of air conditions in Alaska and Hawaii. Further, the CDC does not currently have mortality data for Puerto Rico and the Virgin Islands to compare with AQI scores.
- There is no significant correlation between air pollution and cardiopulmonary mortality rates when controlling for the various parameters in the data.
- Machine learning is a powerful tool for seeing relationships that are not obvious, but our conclusions could have been just as easily described just using visual analytics.
- Our innovative public health project provided new graphics and data hosting capabilities. However, since we could not show a strong correlation between pollution and mortality, future research will be needed to benefit policy makers.

6 CONCLUSION

Our team has created multiple Qlik dashboards to host our visualizations of air pollution and assess air quality and mortality associations. We a also performed machine learning to classify which pollutants are most linked to mortality and predict which states may need additional assistance to mitigate their air quality situations. A conclusive relationship could not be established. Our project improves on the visualizations currently available to the public by showing air quality data at the state level and, where available, at the county level. Combining air quality data and mortality rates proved to be a complex and innovative task. All team members have contributed a similar amount of effort.

Our project can be found at the following web address: https://github.com/schnappv/CSE_6242_Project

REFERENCES

[1] Ruopeng An, Mengmeng Ji, Hai Yan, and Chenghua Guan. 2018. Impact of ambient air pollution on obesity: A systematic review. *International Journal of Obesity* 42, 6 (2018), 1112–1126.

[2] Margaret Carol Bell, Fabio Galatioto, Ayan Chakravartty, and Anil Namdeo. 2013. A novel approach for investigating the trends in nitrogen dioxide levels in UK cities. *Environmental Pollution* 183 (2013), 184–194. https://doi.org/10.1016/j.envpol.2013.03.039 Selected Papers from Urban Environmental Pollution 2012.

[3] Biswanath Bishoi, Amit Prakash, and V.K. Jain. 2009. A Comparative Study of Air Quality Index Based on Factor Analysis and US-EPA Methods for an Urban Environment. 9, 1 (2009), 1–17. https://doi.org/10.4209/aaqr.2008.02.0007

[4] P. Brimblecombe. 2017. *Air Pollution Episodes*. World Scientific Publishing Company. https://books.google.com/books?id=jpA4DwAAQBAJ

[5] David C Carslaw and Karl Ropkins. 2012. openair — An R package for air quality data analysis. *Environmental modelling software : with environment data news* 27 (2012), 52–61.

[6] Cliff I. Davidson, Robert F. Phalen, and Paul A. Solomon. 2005. Airborne Particulate Matter and Human Health: A Review. *Aerosol Science and Technology* 39, 8 (2005), 737–749. https://doi.org/10.1080/02786820500191348 arXiv:https://doi.org/10.1080/02786820500191348

[7] World Health Organization. Regional Office for Europe and World Health Organization. 2006. *Air Quality Guidelines: Global Update 2005 : Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide*. World Health Organization. https://books.google.com/books?id=7VbxUdlJE8wC

[8] W. James Gauderman, Robert Urman, Edward Avol, Kiros Berhane, Rob McConnell, Edward Rappaport, Roger Chang, Fred Lurmann, and Frank Gilliland. 2015. Association of Improved Air Quality with Lung Development in Children. *New England Journal of Medicine* 372, 10 (2015), 905–913. https://doi.org/10.1056/NEJMoa1414123 arXiv:https://doi.org/10.1056/NEJMoa1414123 PMID: 25738666.

[9] Mark L Glasgow, Carole B Rudra, Eun-Hye Yoo, Murat Demirbas, Joel Merriman, Pramod Nayak, Christina Crabtree-Ide, Adam A Szpiro, Atri Rudra, Jean Wactawski-Wende, and Lina Mu. 2016. Using smartphones to collect time-activity data for long-term personal-level air pollution exposure assessment. *Journal of exposure science environmental epidemiology* 26, 4 (2016), 356–364.

[10] Malgorzata Pawul. 2019. Application of neural networks to the prediction of gas pollution of air. *New Trends in Production Engineering* 2, 1 (2019), 515–523. https://doi.org/doi:10.2478/ntpe-2019-0055

[11] C. Arden Pope III, Richard T. Burnett, Michael J. Thun, Eugenia E. Calle, Daniel Krewski, Kazuhiko Ito, and George D. Thurston. 2002. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA* 287, 9 (03 2002), 1132–1141. https://doi.org/10.1001/jama.287.9.1132 arXiv:https://jamanetwork.com/journals/jama/articlepdf/194704/joc11435.pdf

[12] Olalekan A.M. Popoola, David Carruthers, Chetan Lad, Vivien B. Bright, Mohammed I. Mead, Marc E.J. Stettler, John R. Saffell, and Roderic L. Jones. 2018. Use of networks of low cost air quality sensors to quantify air quality in urban settings. *Atmospheric Environment* 194 (2018), 58–70. https://doi.org/10.1016/j.atmosenv.2018.09.030

[13] J. Staehelin, N. R. P. Harris, C. Appenzeller, and J. Eberhard. 2001. Ozone trends: A review. *Reviews of Geophysics* 39, 2 (2001), 231–290. https://doi.org/10.1029/1999RG000059 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1999RG000059

[14] G. Venkatesan and J. Thirumal. 2019. *Global Perspectives on Air Pollution Prevention and Control System Design*. IGI Global. https://books.google.com/books?id=GQyjDwAAQBAJ

[15] Tommaso Francesco Villa, Felipe Gonzalez, Branka Miljievic, Zoran D. Ristovski, and Lidia Morawska. 2016. An Overview of Small Unmanned Aerial Vehicles for Air Quality Measurements: Present Applications and Future Prospectives. *Sensors* 16, 7 (2016). https://doi.org/10.3390/s16071072

[16] Kan Zheng, Shaohang Zhao, Zhe Yang, Xiong Xiong, and Wei Xiang. 2016. Design and Implementation of LPWA-Based Air Quality Monitoring System. *IEEE Access* 4 (2016), 3238–3245. https://doi.org/10.1109/ACCESS.2016.2582153