CISC 451 – Topics in Data Analytics

Midterm Progress Report

Team Members: Will Murray (20054564), Emma Ritcey (20007918), Rylen Sampson (20051918)

Project Title: Analysis of Air Quality for Preventive Actions and Public Health

Problem Statement

Air quality, or a lack thereof, is emerging as a new and undoubtedly dangerous threat to public health. This threat is an even more prominent issue in developing countries where there is an access barrier to professional healthcare and in countries that are largely dependent on sectors with excessive industrial pollution. International health groups such as the World Health Organization have put tremendous work into performing broad-level analysis to estimate the effect a country's air quality may have on its citizens. WHO's analysis can provide insight into which countries require the help of data analytics the most in the fight against deteriorating air quality.

We plan to use air quality data made public by OpenAQ, a non-profit organization, and data science algorithms to provide actionable information for citizens living in countries with poor air quality on how to better protect themselves from this growing issue.
Air quality, with the given circumstances, is a pressing issue as COVID-19 heavily affects the respiratory system. We plan to use the most recent air quality data available as this could play a part in saving lives where COVID-19 is still out-of-control.

With the focus of data analytics heavily skewed towards developing complex and *fancy* models, there is a gap in developing real-world projects composed of explainable models with actionable results. This is where we plan to take our project. While traditional statistical work has been historically used for decision-making, it lacks a certain degree of precision when it comes to predictive modelling and assessing future scenarios. We propose using a culmination of data analysis methods to advise government or policy groups on the risk associated with certain decisions and how it may influence the concentration of harmful air pollutants in their region.

Data Description

As mentioned in the *Problem Statement* our main data source is OpenAQ. They've done the dirty work of collecting air quality data from numerous unformatted sources. Our challenge is how to effectively utilize this data in a data science manner for predictive or prescriptive analysis.

OpenAQ provides air quality data for a total of 12667 locations across the world. For the scope of this project we have decided to only focus on a subset of these locations.
WHO, similarly to OpenAQ, has air quality data but not on the same granularity. The data hosted on the Global Health Observatory (GHO), a data repository run by WHO, is of yearly averages and estimated joint effects of air pollution. We used this data to inform our decision on which countries to select, these countries represent a widespread group composed of those with low, medium and high air quality.

To further the problem we wanted to find data representing various country indicators such as weather data, commitment to the environment and general population health indicators. To access this type of data we used the *The World Bank's* database.

Challenges

Through the OpenAQ API, the furthest back data we could obtain was 90 days, so using the web portal we were able to collect data for countries with PM2.5 measurements dating back two years (2018-2020). The more data we have, the more accurately and confidently we can model the problem. The challenge of limited data is difficult to overcome with any modelling techniques and resulted in us spending much more time than expected on data collection and considering which data and sources we should use.

One evident challenge are biases. From the motivating factors behind this project, public health, and our initial exploration into the data it is clear there may exist demographic bias within how we choose which countries to look at. Developing countries and those with lower quality healthcare systems will likely have elevated death rates compared to those with a similar level of air quality, but who also have better quality of life and healthcare. Another source of bias is found within the geographic details of the locations as geological features like proximity to water bodies can affect air flow and in turn the air quality. Lastly, unseen factors such as local proceedings may affect the air quality. An example of this could be for a location where this data is being recorded, there are also numerous factories producing air pollutants.

As OpenAQ is a non-profit organization, their data is guaranteed to contain imperfections. The pressing imperfections apparent to us are drastically different sample sizes, missing or incorrectly recorded data, and the lack of features for newly added locations. We can employ outlier detection and imputation methods to efficiently handle any erroneous data. The differing sample sizes require a more thought-out solution, one that may become more apparent during further exploratory analysis.

Another problem we expect to encounter is the inability to find supplementary data on the same frequency as the obtained air data quality. When the data being dealt with contains irremovable time dependencies such as ours, it can be difficult to select or construct an appropriate model. This is because different algorithms will attempt to model this time dependency in different ways and without knowing an approximate function which fits this dependency, model selection starts as a sort of guessing game.

Detailed Timeline

Week	Deliverable	Additional Notes
3	Final Project Proposal	Review feedback on the initial proposal and incorporate it into our plan.
4	Finish Collecting Additional Data	Depending on the complexity of collecting additional data this may take longer than expected.
6	Exploratory Analysis	Complete things such as: - Validate model assumptions and features assumptions - Initial feature selection - Preprocessing data
7 & 8	Preliminary Analysis	- Begin to try different model types - Obtain an idea of the results
9 & 10	Finalize Experiments	- Finish proposed model - Obtain results for all experiments - Clean up loose-ended code - Produce any coded items needed for the report
11	Report Rough Draft	
12	Finish Report	- Any final editing / formatting - Double check plots and results - Get feedback from other groups

Methodology

The challenges brought up earlier in the report will be dealt with individually and then collectively as we refine our model further. They will also motivate the choice of different models.

Dealing with biases will likely be the most difficult problem, our plan is to gauge the effectiveness of different preprocessing strategies to mitigate the influence on our analysis. Erroneous data will be handled with outlier detection and efficient imputation methods, this was worked on for this progress report and we believe we have a good understanding of the data we now wish to use.

Sample sizes and the frequency of the collected data may cause significant issues and it is still a difficult problem to solve. A quick way to deal with it may be adjusting sampling frequency or applying bin transformations to ensure similar data and equivalent frequencies. Deciding which approach is the best will likely be a direct result of applying numerous models and moving forward in the direction of the model which performs best.

To validate the data that we currently have, we began with understanding the proportion of outliers within the data.

- Clearly, negative concentrations of PM2.5 can't exist and so these were identified.
- From research, we found that concentrations above 200 were not realistic and so these were included in the definition of outliers for this data.
- Finally, after excluding this data, we looked at outliers through the IQR method to find measurements that could be erroneous but were still recorded as possible values.

To understand how these countries compared in terms of air quality we took an initial metric that counted the number of measurements exceeding concentrations of PM2.5 above 35 for each country; the advised level of unhealthy air quality as stated by the *WHO*.

Evaluation

Given that we're focusing on a real-world problem and not one of creating the best model for pre-collected and cleaned data, we have been heavily focused on data selection, feature engineering and validating various types of data. We've yet to build any predictive models for this reason but we envision ourselves building models to predict trend direction of air quality. As of now we feel very confident in our preprocessing and exploratory analysis of the air quality data. Furthermore, after researching what could be key factors in air quality levels we have identified numerous possible features that will be very decisive in future model performance.

Future Work

Our future work requires the fleshed out guide written in our methodology. From these insights we can gain a much clearer idea on how we are going to move forward with our model and how we can best turn this into a useful estimator for the general public and not just in a data science sense. Unfortunately our steps beyond exploration are hindered without knowledge to be gained; however, the picture will become significantly clearer as we make headway and fulfill the goals we have set for ourselves.