Exploring the Link Between Gas Emissions and Air Quality in the United States

Sean Bray, Parker Fromm and Korede Ogundele

Department of Chemistry, University of California Berkeley

DATA 200S: Principles and Techniques of Data Science

Dr. Narges Nourouzi and Dr. Fernando Pérez

December 8, 2023

Video presentation:

https://youtu.be/vRIkJ_CuCuI

Abstract

Air quality is a critical factor influencing human health and environmental well being. This study investigates the relationship between gas emissions from facilities and air quality measures across counties and years in the United States. The distribution of emitting facilities, their emission types, and corresponding air quality measures were analyzed in order to understand the association between air quality and metrics and emission levels. The goal of this study is to create predictive models: one that estimates air quality based on emission types and one that builds upon that approach using distance from facilities to major cities and coastlines. We also added a model to predict how many emitting facilities a state would have in a given year.

Introduction

The purpose of this project is to explore questions about the relationship between gas emissions and air quality measurements in different counties and years in the United States. Air quality is an important determinant of human health; poor air quality has been linked to an increased number of asthma hospitalizations (Nadali et al, 2022) and increasing prevalence of respiratory diseases (Vassari-Pereira et al, 2022 and Yin, n.d.). Air quality measures also have environmental impacts. Pollutants such as CO2 and CH4 trap heat in the atmosphere, leading to global warming (Ward and Laing, 2015). Thirdly, air pollution has been shown to be connected with agricultural yields, creating a direct economic impact (Bishop, 2022).

First, we explore the distribution of emitting facilities throughout the United States and their average emission values per state. We then explore air quality data to see if there is a correlation between our emission data and the air quality values. We aim to make a model that allows us to predict the air quality of a region based on which gas and how much of it is released within a county. We also got some interesting results that prompted us to model the number of reporting facilities in each state as a function of the year, with Texas showing up as the outlier with the most reporting facilities. We frequently encountered issues with incomplete insufficient data.

Description of Data

We worked primarily with two datasets. The first was us_air_quality_measures.csv which contains over 200,000 air quality data measurements from many different states and counties. Each row represents air quality data from a single day. These records range from 1999 - 2011. Air quality measurements can be found in three columns/measurement types: Average, Counts, and Percent. Counts represent the number of days with a certain ozone concentration above a set standard, Percent represents the percentage of days with air quality levels above the National Ambient Air Quality Standard (NAAQS), and Average represents the average concentration of PM2.5 (particles smaller than 2.5 microns in diameter in micrograms per cubic meter. Some of these values are measured (known as ‘Monitored’ in the dataframe) and others predicted by a model. The second dataset is us_greenhouse_gas_emission_direct_emitter_gas_type.csv , which contains information on emitting facilities including their location (street address, city, county, and latitude and longitude) and types of gasses they emit. These records range from 2010 - 2019.

The variables being observed and used are average air quality measure (from the air quality dataset), gas type (from the gas emissions dataset). Distance to nearest major city and distance to coast are two variables that we calculated and added to the model.

The specific research questions investigated in this study are:

1. How does air quality relate to the number of emitting facilities and level of emissions in the area?
2. Is there a correlation between air quality and emission type and quantity?
3. Can we train a model that accurately predicts air quality based on emissions?
4. How has the number of reporting centers evolved over time, and how might those numbers look in the future?

Methodologies

We began our project by exploring and cleaning the data. Both data frames initially had column descriptors that were difficult to interpret, so we renamed all columns. Next, we made all state, county and address columns capital letters so that when performing pandas operations, we would not have any data left out due to case sensitivity. We next examined the values inside of our data frames by exploring by year, gas type, and emission values. We looked at distributions, and the number of unique values. From the emission data frame, we dropped entries with no location data, as we would not be able to use these in our study.

Looking at the air quality data set we ran into some issues. The way the columns were set up had measure names in one column which consisted of 6 different possibilities, and then the measure type separated by three columns. The data years also did not align well with our emission data so we had to pick a method for combining the data. Comparing value counts of

years between the air quality dataset and the emission dataset, we decided to base our primary model on data from 2011, as this had the most data of all the overlapping years. Most of the air quality data was collected in the years before our emission data was collected, so 2011 was the best choice in order to get the most data from both sets.

We went ahead and started to answer some initial questions through graphing. We looked at the distribution of the facilities by mapping a density plot over the United States (**Figure 1**). We found that a majority of our facilities were on the northern part of the east coast and around Texas. We found a much smaller distribution on the northern section of the west coast. We also looked at the average CO2 emission values and found the states with higher emission values correspond to the states with more facilities, which would be as expected (**Figure 2**).

We explored the first two research questions using correlation indexes in our preliminary EDA. For the third question, we created a machine learning model which predicts average air qualities from emission quantities of a few different gasses. We combined our data frames of gas emissions and air quality for the year 2011. We attempted to model air quality based on individual gasses alone, but given the limited amount of data and other factors, we also made a model with locational data. Since we had latitude and longitude values for where the emission was coming from, we decided to add additional features such as the distance to coastline, as regions closer to the coast tend to have better air quality. We also added the distance to the nearest major city, as major cities have larger populations and also more air pollution leading to poorer air quality. The additional data sets used to do this can be found in our references. All work with latitude and longitude coordinates were done with the package geo pandas. To create a data frame for the modeling, we added these features and then performed a form of one hot encoding to our gas type and CO2 emission column. This data column was a bit misleading at

first, as it actually holds the emission value for the gas type name in the corresponding column. We separated the gas types into separate columns and filled in the rows of data not corresponding to the column's gas type with 0s, and the remaining with the "CO2" emission values as each column was actually representing one gas type.

Now that we had a data frame for the modeling it was time to actually build a model. We started with a linear model to predict air qualities based on gas types, which we will call Model 1. We used facilities distance to a major city in meters (feature 1) and facility distance to coast in meters (feature 2), and all the gas types that were available in the original emission data frame: BIOCO2, CH4, CO2, HFC, HFE, N2O, NF3, PFC, SFC6 and other. We split our data in 80% and 20% to save the 20% for testing. We then split the 80% again into 80% and 20% so we can have a training and a validation set. We ran our model with cross validation, then checked the loss using our test data we had put aside. We used the root square mean as our loss function because we wanted changes in our data to be minimized. We also normalized our distances by converting to kilometers so that our distance values were smaller and easier to interpret.

Summary of Results

Our first trial of Model 1 was bad, and we reported a loss of over 1000. Seeing this value, we realized we likely were including features that were unimportant. Our first thought was that some of the gasses and their emission values were too limited to be helpful features. We counted the number of non zero values in each gas type and decided to omit the ones with small amounts of data. This left us using only BIOCO2, CH4, CO2, and NO2. Using these values we were able to get a better loss, but still not great considering the range of our average air quality measures. We got losses ranging in the 20s. We then created a new data frame column called Log Average

with air quality measures converted to logarithmic values. Using this column, we were able to get much better results using the same features as above. We got losses around 0.3 - 0.5, which is much better than before.

Distance to coast and distance to nearest major city were selected because we suspected that the farther away from a major city or closer to coast, the lower the air quality measurement. A quick look at the correlation coefficient between air quality and distances showed that there was in fact a negative correlation. CH4, CO2 and N2O were selected as features because these were the gas types most commonly reported (and thus with the most data entries). We chose not to use columns with mostly zeros so that the model would not correlate zeros with air quality (when the zeroes in the data were more likely due to that gas type being omitted/unreported). Average air quality measures (y value) were transformed logarithmically. The loss metric used for this model was root mean squared error (RMSE). We chose this loss function because RMSE is easy to interpret, robust to outliers, and continuous (making it sensitive to small changes in prediction accuracy).

For the question of "which gasses contributed most to the PM2.5 particle density air quality measures" we once again used the data from us_air_quality_measures.csv, as well as us_greenhouse_gas_emission_direct_emitter_gas_type.csv. We faced some challenges as we analyzed our data: The "CO2E_EMISSION" column of the us_greenhouse_gas_emission_direct_emitter_gas_type dataset was only a direct measurement for rows where the GAS_NAME value was BIOGENIC CO2. For rows with entries referring to other gas types (e.g. Methane, Nitrous Oxide, etc), the CO2E_EMISSION column referred to some conversion from an emission quantity for that gas to "CO2 equivalents". We were unable to verify the method by which this conversion was performed, but our initial investigations

indicated that this method was performed so that all the different gas emission quantities could be compared on some kind of unified scale to measure "global warming potential". These two datasets only overlapped for 4 years: 2010, 2011, 2012, and 2013. This caused some difficulty in making predictions that associated these two datasets, and also caused some concern about data sparsity.

However, by analyzing data county-by-county level granularity, we hope we've generated enough data points to have train/test splits of sufficient size for model training and validation. It seemed a bit surprising that the average measurements of PM2.5 particle density (reported in units of micrograms per cubic meter) seemed to decline over the years data were provided, 1999 to 2013 (we initially expected particle densities to have increased over that timespan). (**Figure 4**). This sparked a set of questions & about the methods by which this data was gathered. We decided to add a model to predict the number of reporting facilities in each state, with the year as an input. We trained a model for each state individually (with scalar normalization on the years), and predicted the # of emitting facilities that would be reporting in the year 2050 (**Figure 3**). Texas, which has ~10x more facilities than the average state and has been increasing its number of facilities year by year, is a far outlier in the predicted number of reporting facilities.

Which gasses had their emissions measured varies highly between year/county/location. This has constrained many of the models we've attempted. In one dataframe we created which measured the total reported emissions for each gas, for each county, for each year, eight of the total twelve gasses had NaN in over 10000 of the ~12000 rows. For the relevant model, we decided to drop those 8 gasses, and then for the remaining 4, drop the rows with NaN rather than try to interpolate the values. Ultimately, predictions with this model yielded an RMSE of ~1.7 for both the train and test sets (**Figure 5**).

Discussion

We employed machine learning models to predict air quality based on emission types and quantities; we got good results with a logarithmic transformation of air quality measures. This is likely due to the fact that many cause-effect relationships are logarithmic rather than linear (Ball, 2008). The inclusion of distance to coast and distance to nearest major city as features proved beneficial, and highlights the importance of geographical characteristics when predicting air qualities. This suggests that geographical factors play a significant role in air quality, as air quality tends to be better in areas further away from major cities and coastlines (as we saw from our negative correlation value. Additionally, focusing on fewer gas types helped eliminate noise (from gas types with many zero values) and improve Model 1's performance.

We faced some limitations while working with these datasets. The data sparsity for 8 out of the 12 prevented us from incorporating all relevant emission data into our model. The limited overlap between years between the two datasets also limited the amount of data our model had access to, which could hinder model performance and make it more prone to overfitting. While we attempted to address these limitations by using sufficient train/test splits, further research with more complete data would likely lead to more robust and generalizable models.

A third limitation is the unknown method by which CO2 equivalents were calculated in the dataset. Although we tried looking this up, we were unable to find explanations that tied into these datasets specifically.

Conclusion

This study has demonstrated the feasibility of using machine learning models to predict air quality based on emission types and geographical factors. Despite challenges, we were able to address our research questions using correlation coefficients and linear regression models. Ways to improve this study would be to find data for more gas types, get emissions data for more overlapping years, find out what gas types are included in the “OTHER” category, or learning how CO2 equivalents were calculated. Also, more uniform measurement & reporting standards across facilities, counties, and/or states could lead to more accurate predictions and helpful conclusions.

Figures

Figure 1: PDE plot, density distribution of number of facilities across the United States

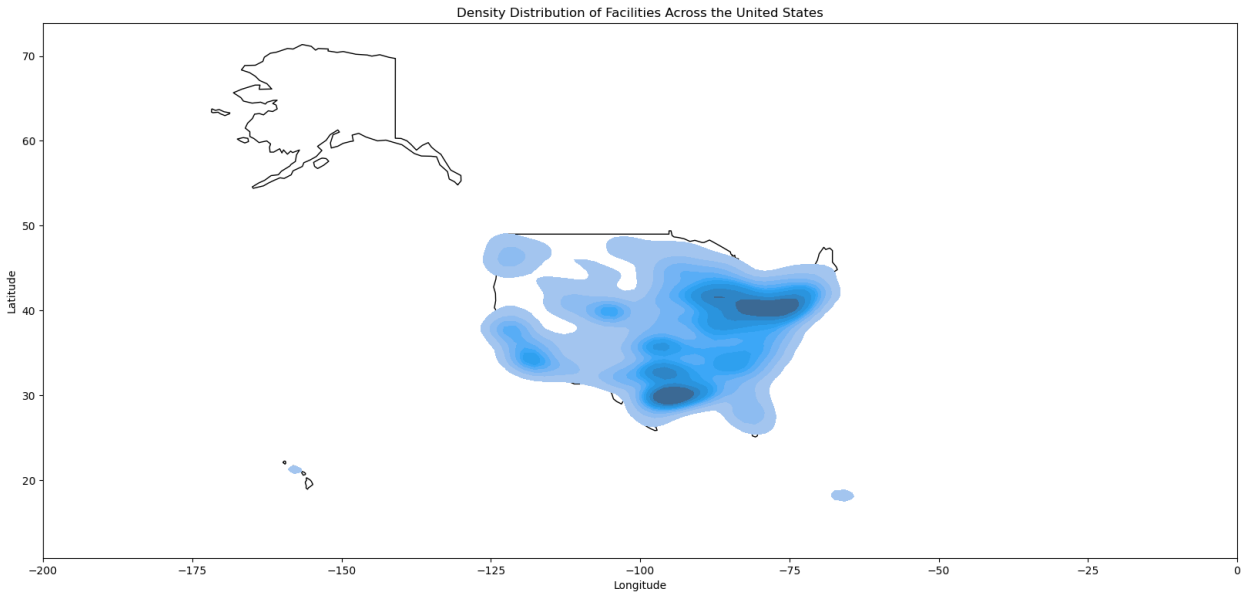


Figure 2: Average CO2 emission per States

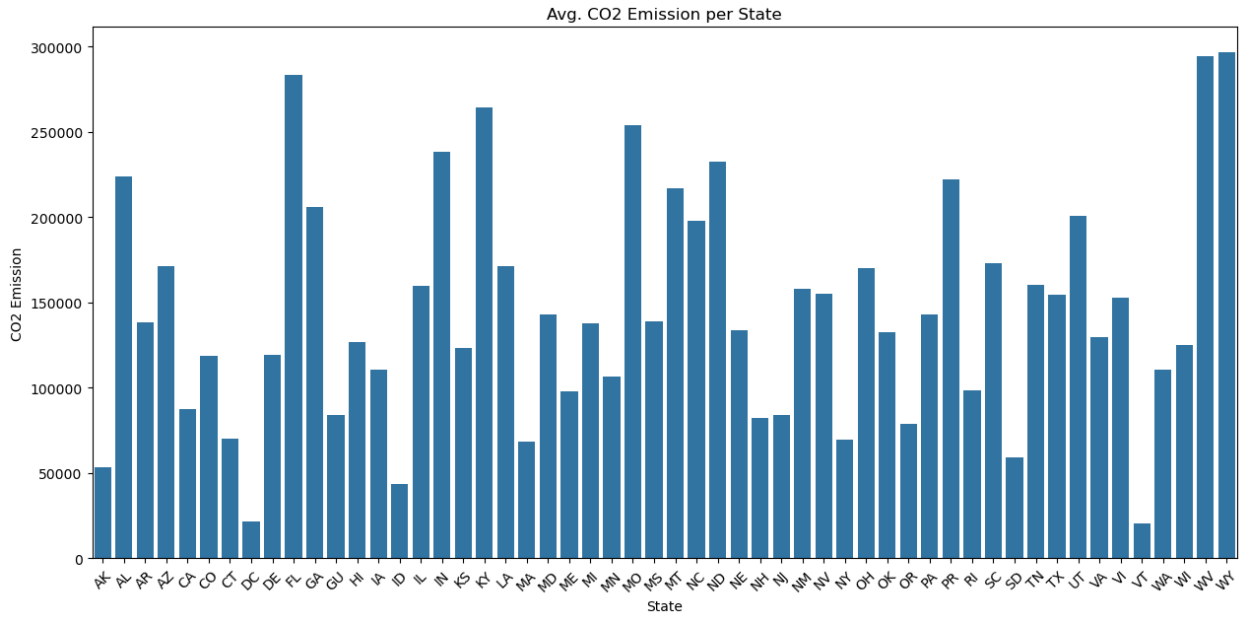


Figure 3 : model predictions for # of emitting facilities in year 2050

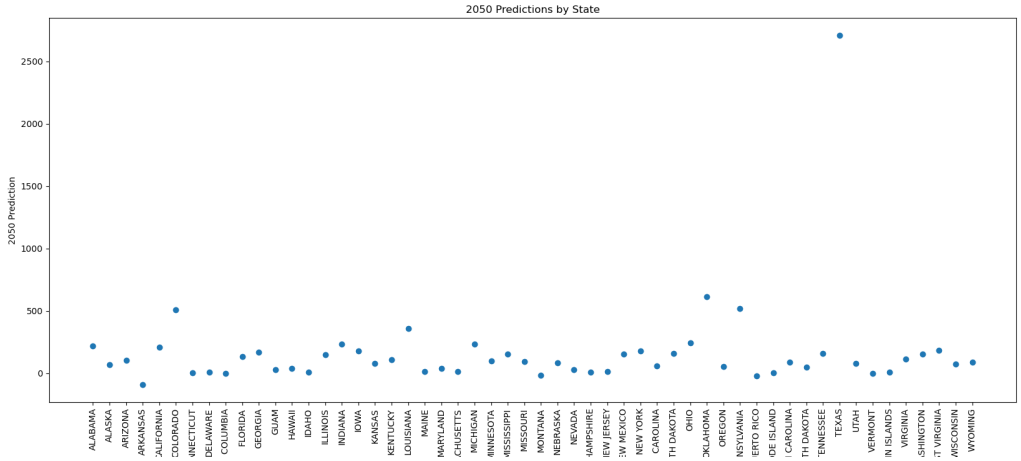


Figure 4: Average monitored values of PM2.5 particles in micrograms/cubic meter

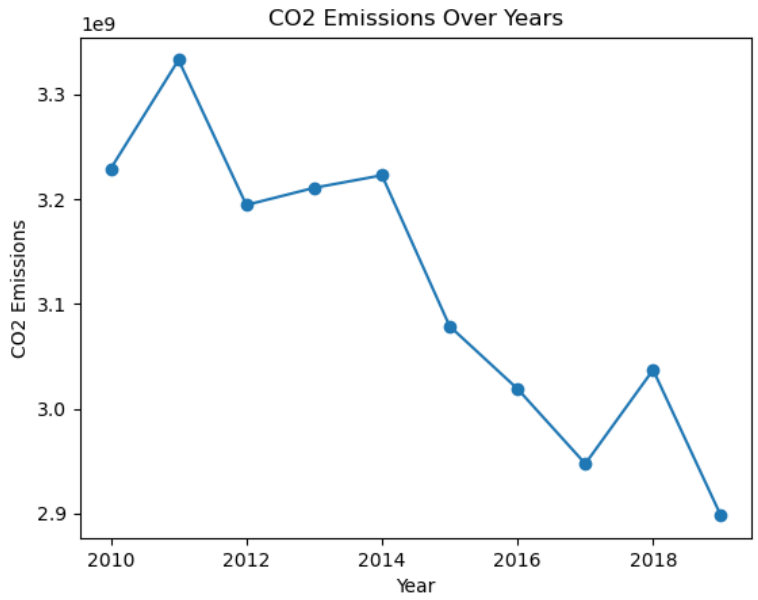


Figure 5: RMSE for air quality vs gas types for model that does not include location data

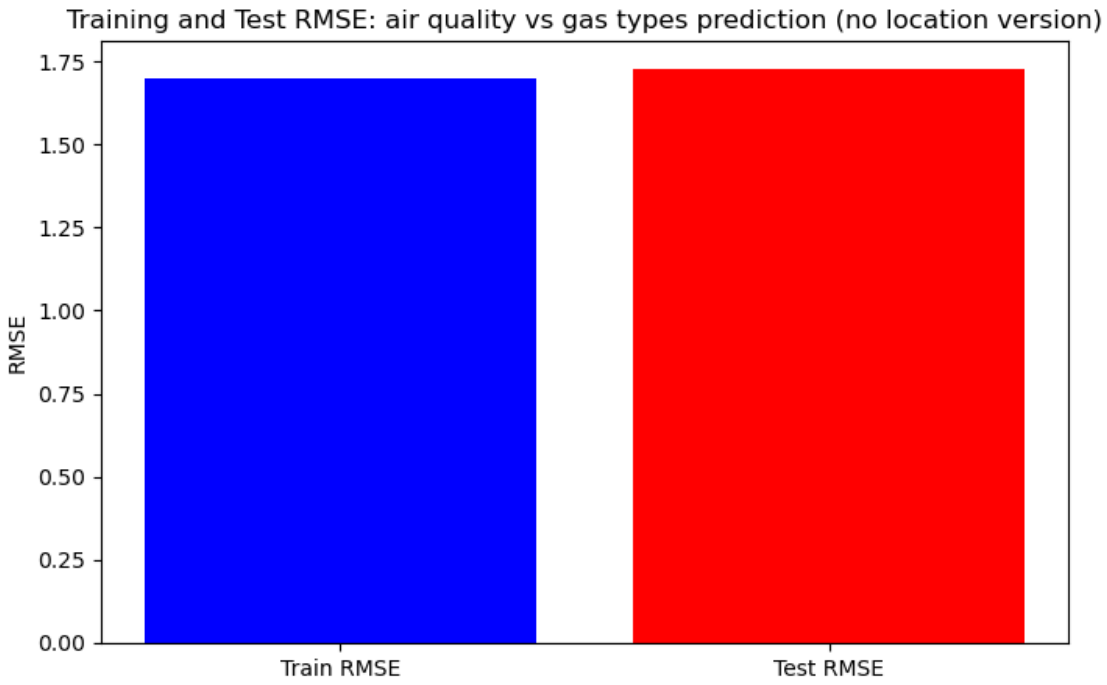
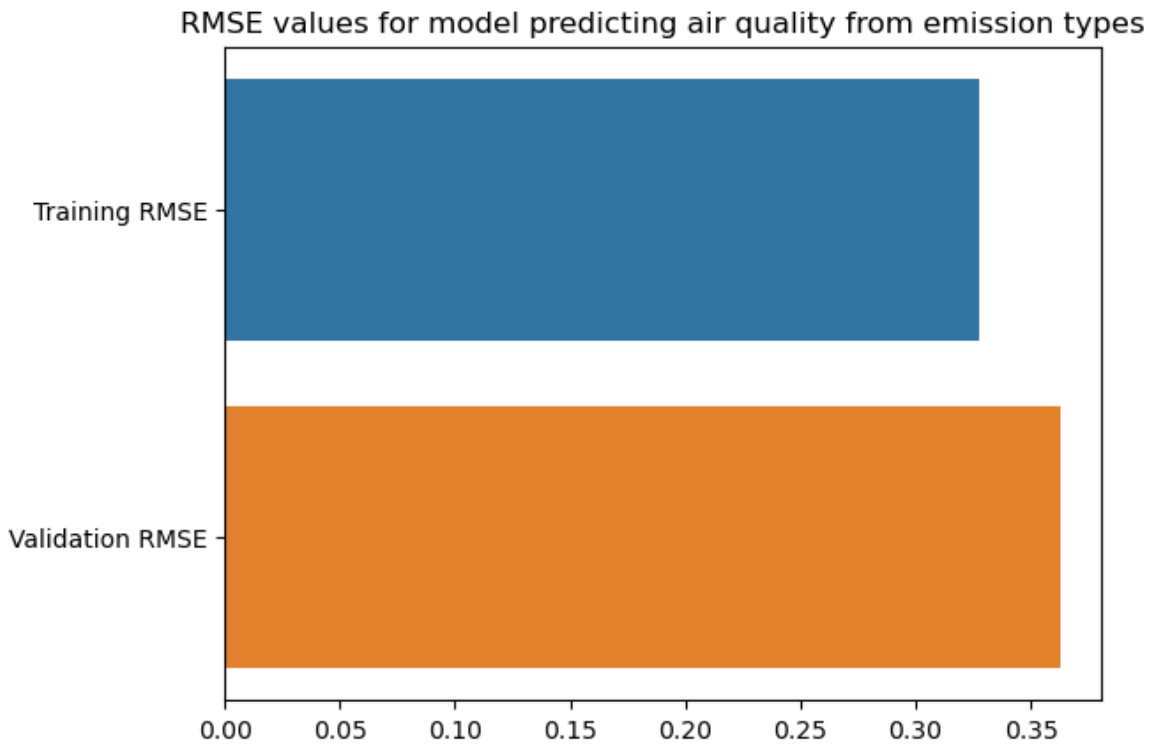


Figure 6: RMSE values for air quality vs gas type prediction



References

Ball, P. (2008, May 29). Why we should love logarithms. Retrieved December 6, 2023, from https://www.nature.com/articles/news.2008.866

Bishop, S. (2022). The Two Way Relationship Between Agriculture and Air Pollution.

Nadali, A., Leili, M., Karami, M., Bahrami, A., & Afkhami, A. (2022). The short-term association between air pollution and asthma hospitalization: a time-series analysis. Air Quality, Atmosphere and Health, 15(7), 1153–1167. https://doi.org/10.1007/s11869-021-01111-w

Natural Earth "Blog Archive" Coastline - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales. (n.d.). Retrieved from https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/

United States Cities Database. (n.d.). Retrieved from https://simplemaps.com/data/us-cities

Vassari-Pereira, D., Valverde, M. C., & Asmus, G. F. (2022). Impact of climate change and air quality on hospitalizations for respiratory diseases in municipalities in the Metropolitan Region of São Paulo (MRSP), Brazil. *Ciência & saude coletiva*, *27*(5), 2023–2034. https://doi.org/10.1590/1413-81232022275.08632021

Ward, P. L., & Laing, D. B. (2015). What Really Causes Global Warming?: Greenhouse Gases or Ozone Depletion? (1st ed.). Morgan James Publishing.

Yi-Rong Lin, & 林宜蓉. (n.d.). A study on the relationship between air quality and respiratory diseases in Tainan area.