Patterns in the Air: A Study on Air Pollution in South Korea

Brendan Prednis
University of Colorado, Boulder

Abstract

0.1 Problem Statement & Motivation

Air pollution is a problem that affects the entire world, whether we are aware of it or not. South Korea is one such country that has had a history of air pollution problems, and has struggled keeping it in check. By studying major pollution particles in the air in South Korea, we can determine when various particles in the air increase during different times of the year, in order to reflect on events and determine actions to control those harmful particles. This allows governments to make a clear plan, and inform the public of health impacts.

Keywords: Air Pollution, Data Science, Data Mining

1 Introduction

South Korea deals with various problems relating to air pollution, which has made an impact on the lives of the people living there. I want to study the relation between various particulate matter in the air during various times of the year and the safety of the air. My motivation for studying the pollution is partially inspired by my awareness of air pollution and it's widespread and long lasting impact on everything in the world. My girlfriend at the time of writing this paper is in South Korea, giving me the basis of inspiration.

1.1 Paper overview

In this paper I will explore air pollution in South Korea, how it changes throughout the year, and potential various prediction methods to predict impact as the year moves. By utilizing common data mining techniques such as clustering, correlation analysis, and linear regression, we will infer whether different kind of pollution can be predicted and offset during different times of the year.

[1]

2 Methods

2.1 Literature Survey

Various studies have been done on air pollution trends in South Korea. Some research the chemical and particle density in the air during different seasons, while some look at the economic and political states that impact air quality control. One study, done by Harvard University, Yonsei University, University of Maryland, and NASA's Climate and Radiation Laboratory, looks at the trends of air quality throughout the year. It investigates how air quality and air pollution change unevenly across the country, with various times of the year bringing in different pollutants. South Korea's GDP has grown extensively over the last 30 years, resulting in high emissions of CO and SO_2. The study states that roughly 30,000 premature deaths per year are attributed to air polution in South Korea [2]. Fine Particulate Matter ($PM_{2.5}$) is one of the main items studied within air quality and air pollution research. The joint research article shows that $PM_{2.5}$ has been decreasing, but Nitrogen Dioxide and Nitric Oxide (NO_x) density has not. In regards to O_3, it is highest during the summer, but due to summer monsoons, clean marine air is brought to the land, resulting in lower O_3 levels in July through August, compared to May through June [2]. The article concludes that CO and SO_2 levels have stayed below air quality standards since the late 1990s, while NO_x is now below the air quality standard at almost all AirKorea study sites [2].

Another article discusses a study led by POSTECH Professor Hyung Joo Lee, stating that NO_2 exposure levels are consistently higher in areas associated with higher socioeconomic status [1]. The article asserts that NO_2 is a key air pollutant emitted from combustion sources, such as vehicles and power plants, and is regulated by South Korea's Clean Air Conservation Act, stemming from NO_2's adverse impact on respiratory health. The team at POSTECH used special satellite data for sensing NO_2, which allowed them to produce a high resolution map of NO_2 exposure. With the goal to assess whether the nations current ground monitoring network accurately recorded the population's exposure to NO_2, their study revealed that national ground monitors underestimated NO_2 exposure by up to 11% in Gangwon-do, and overestimated exposure by as much as 61% in Jeju-do[1]. The study concluded that ground monitoring did not imporve accuracy just by adding more monitors, and that improving accuracy requires more monitoring methods and efforts to be put in place.

In relation to the political and economic side, an article in 2014 by Jongsik Ha of the Korea Environment Institute stated that South Korea's air quality standards were insufficient in terms of establishing procedure for managing air pollution effectively. In order to bring ideas to the table, the NAAQS of the US was examined in order to suggest ways, which consider health effects, to establish air quality standards in South Korea. The author concludes the study stating that "Realistically speaking, it is impossible to establish standards that protect an entire population from air pollution. Instead, it is necessary to find a balance between what should be done and what can be done." [3]. The study further says that few are aware of the dangers of environmental risk factors, despite the fact that exposure to them contributes considerably to disease prevalence and mortality rates in local communities [3].

The last major article researched looked at using Machine Learning to study and predict the health impact of $PM_{2.5}$.

Since there are limited sources for $PM_{2.5}$ exposure and health data, research on $PM_{2.5}$ at a national level was limited[4]. By using randomized sampling from a large scale data pool of participants 50 years of age or older from the National Health Insurance Service-National Sample Cohort (NHIS-NSC)[4], the researchers built a prediction model to identify patterns relating $PM_{2.5}$ to all-cause mortality and cause-specific mortality. The result of the study suggests a 1μg/m3 increase in $PM_{2.5}$ was associated with all-cause mortality hazard ratio (HR) 1.002 [95% confidence interval (CI): 0.998-1.007][4].

3 Proposed Work

3.1 Data Collection

The data collection for this research project was straight forward, as it was just downloaded from Kaggle. In order to clean the data, I concluded that using a mean fill would be appropriate for filling in NaN or empty numeric values. The data will be integrated in Polars, an alternative DataFrame library for Python that is fast and type-safe, providing the researcher (me) some semblance of sanity while working with Python. The data does need to be transformed into a more appropriate structure, as each row corresponds to a single particle, rather than utilizing each particle as a feature. This ends up making the dataset 9x taller than it needs to be.

3.2 Workflow

The workflow for processing the data is not complex, but multi-step and involved. It starts with reading through the dataset, cleaning any data that may cause errors in our model. Next, we analyze the dataset for any interesting patterns. We will look at finding any single-level associations between features, starting with each major particle/compound. After looking at single-level associations, we will look at deeper multi-level associations. This kind of data lends itself to cluster analysis, grouping similar data points together to further understand the relations between particles and health impacts.

3.3 Dataset

The dataset is a structured document collected from various public data from the Seoul Metropolitan Government, put onto Kaggle. The dataset in total has 647,512 rows, making it a slightly smaller dataset for this project. However, I believe the density and consistency of data provided is sufficient.

The dataset provides useful attributes to improve our study.

3.4 Preprocessing

In Table 1, we can see all the attributes collected in the dataset. While each attribute is important, the main focus will be on the measurement date, latitude, longitutde, and the particles/compounds that have been recorded. We don't require the sensor address for now, and will be filtering out any sensors that have abnormal codes. We will also be using mean averaging for filling in missing or NaN numeric values, and filtering out full rows with NaN values that aren't numeric. This ensures no data is empty, causing a skew in results or errors while iterating through the code. During preprocessing, it is also important to figure out common functions for processing data, and document them appropriately.

Name	Description
Measurement Date	• The date the measurement was taken, in hour segments.
Station Code	• The code designation given to the sensor station.
Address	• The South Korean address of the sensor station.
Latitude	• The latitude of the sensors location.
Longitude	• The longitude of the sensors location.
O_3	• Ozone concentration. • Measured in **ppm**.
CO	• Carbon Monoxide concentration. • Measured in **ppm**.
SO_2	• Sulfur Dioxide concentration. • Measured in **ppm**.
NO_2	• Nitrogen Dioxide concentration. • Measured in **ppm**.
PM_{10}	• particulate matter that consists of tiny particles or droplets in the air with a diameter of 10 micrometers or smaller. • This can be dust, smoke, and various emissions from vehicles and industrial machines. • Measured in **Mircrogram/m3**
$PM_{2.5}$	• Fine particulate matter that is 2.5 micrometers or smaller in diameter. • Major cause for respritory and cardiovascular health issues. • Significant portion of air pollution. • Measured in **Mircrogram/m3**

Table 1: The relevant attributes of the dataset

3.5 Evaluation Methods

Evaluation methods for this study have evolved over the course of investigating the data and researching more about the domain. Evaluating the data for this project originally was planned to use a standard supervised learning / classification model to determine which times of the year produce certain combinations of air pollution particles. This reduces the the need for external datasets, such as economic information or political changes. While these datasets could be useful, to reduce the scope of the project, we will isolate to

SO2	NO2	O3	CO	PM10	PM2.5
1.0	0.866	0.692	0.258	0.043	0.051
0.866	1.0	0.665	0.087	0.039	0.045
0.692	0.665	1.0	−0.064	0.003	−0.002
0.258	0.087	−0.064	1.0	0.143	0.178
0.043	0.039	0.003	0.143	1.0	0.225
0.051	0.045	−0.002	0.178	0.225	1.0

Table 2: Chemical Compound Correlations

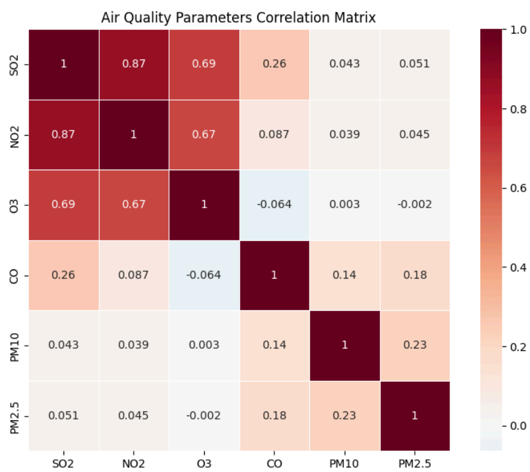

Figure 1: Compound Correlation Heatmap

pure sensor readings for now. While this method of evaluation is still the ultimate goal, I organized a few fundamental evaluation methods to find patterns.

3.5.1 Chemical Compound Correlation

The first method used is a simple correlation method between the chemical compounds emitted.

In Table 2, we see that SO_2 and NO_2 share a strong correlation coefficient, with O_3 following somewhat closely. What was surprising was that SO_2 and NO_2 are the only chemical compounds that do not have any negative correlations, showing that with an increase of those chemical compounds, all other chemical compounds raise a certain amount. This shows us that the contributors to the three chemical compounds NO_2, SO_2, and O_3, do not contribute to particulate material pollution. Continuing on our observation with the other compounds, none have a drastic negative correlation coefficient. Our largest observed negative correlation coefficient is −0.064, which, looking at a graph of the data points, does not seem to indicate a strong correlation either way.

Figure 1 shows the correlation distribution across each chemical compound, and its positive correlation coefficient. The code used to calculate this was straight forward:

```
particle_corr = df.drop([DATE, STATION_CODE, LATITUDE,
LONGITUDE]).corr().select(pl.col(pl.Float64).round(3))
```

Name	Freq.	Mean Sev.	Std Sev.
SO2	13,799	2.942	25.207
CO	63,720	1.066	4.709
PM10	30,185	2.7	9.048
O3	12,936	1.651	15.63
PM2.5	34,411	2.639	8.419
NO2	8,811	1.145	26.595

Table 3: Compound Frequency and Outliers

This makes use of Polars excellent expression library and fully typed methods. We can run an expression across all fields that match the type Float64 and rounding those values to three decimal places for easier reading.

3.5.2 Statistical Outliers

Next, I looked at outliers within the data. I took the frequency of each compounds outlier, with it's mean and standard deviation value. The total number of outliers was **163,862**, or about **25.636%** of the total rows in the dataset.

Table 3 shows the distribution of outliers for each chemical compound, with the modal class being CO.

The code to analyze and display the outliers is as follows:

```
outlier_df = detect_compound_outliers(df, compound_list)
compound_outliers_df = aggregate_outliers(outlier_df, compound_list)
outlier_clusters =
cluster_outlier_frequency_severity(compound_outliers_df)
total_outliers = outlier_clusters.select(pl.col("frequency").sum())
```

We have the functions **detect_compound_outliers** and **aggregate_outliers** filtering and organizing outlier data, and **cluster_outlier_frequency_severity** groups each outlier by chemical compound. By printing **outlier_clusters**, a DataFrame object, we are able to see Table 3.

3.6 Tools and Technologies

The tools used in this project are mostly tried-and-true tools used in the Data Science community.

The language of choice for extracting, transforming, and loading the data is Python, due to its extensive libraries and ease of use. The trade off with Python is that it is dynamically typed, which slows the iteration process as the project grows. However, for this project, the scope is maintainable and relatively small compared to a larger, multi-person project.

Handling DataFrame processing, I decided to use the Polars library. An alternative to Pandas, it is arguably faster than Pandas in certain areas, is type safe, and has a unique way of forming transformation expressions. I have used it in the past, and feel more comfortable with it and it's clear, detailed documentation. Polars allows the user to create complex aggregation logic using their built-in DSL-like expressions. For example, take this clustering method:

```
def cluster_outlier_frequency_severity(df: pl.DataFrame) ->
pl.DataFrame:
```

```python
return df.group_by("compound_name").agg([
    pl.len().alias("frequency"),
    pl.col("outlier_severity").mean().alias("mean_severity").round(3),
    pl.col("outlier_severity").std().alias("std_severity").round(3),
])
```

pl.col is part of Polars expression library, which is run through their expression engine behind the scenes. This allows for performance improvements over Pandas, as well as a cleaner,typed, and more consistent syntax.

For displaying data, I have chosen Matplotlib and Seaborn. Matplotlib is a classic library, providing easy access to displaying clean graphs of all kinds. Seaborn is a new library I found about, and it is useful for creating heat maps. Heat maps have been helpful in visualizing correlation between chemical compounds, as shown above.

For writing reports and this project proposal, I am using Typst in place of LaTeX, for easier iteration and faster compile times. Typst is a fantastic progression in the world of typing systems, with useful features, a straight forward syntax, and excellent documentation. It's compiler is much faster than LaTeX, and its tooling is all encapsulated in one binary. This allows it's ecosystem to be consistent across platforms.

References

[1] Kim N. R. and Hyung J. L., "Is Air Pollution Exposure Equal Across South Korea?," Feb. 13, 2025. [Online]. Available: https://www.postech.ac.kr/eng/research/research_results.do?mode=view&articleNo=18646&title=Is+Air+Pollution+Exposure+Equal+Across+South+Korea%3F+

[2] Oak Y. J. *et al.*, "Air quality trends and regimes in South Korea inferred from 2015–2023 surface and satellite observations," Mar. 17, 2025. [Online]. Available: https://acp.copernicus.org/articles/25/3233/2025/

[3] Jongsik Ha, "Applying policy and health effects of air pollution in South Korea: focus on ambient air quality standards," Oct. 01, 2014. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4199289/

[4] Jeongmin Moon, Ejin Kim, Aims Create Team, Whanhee Lee, and and Ho Kim, "Effects of long-term exposure to air pollution on all-cause and cause-specific mortality in South Korea," Sep. 17, 2023. [Online]. Available: https://ehp.niehs.nih.gov/doi/10.1289/isee.2023.OP-019