GAN, Prinz Romeo O.
HEBRON, Yenzy Urson S.
NICOLAS, Jack Vincent
OSCARES, Jim

CS 180
Proposal

Exploring the relationship between Air Quality and COVID-19 Infection Rates

Project Summary

This project aims to explore the relationship between Air Quality and COVID-19 infection rates on a per country basis. Two separate datasets — the *Ambient Air Quality* dataset and the *COVID-19 Cases and Deaths* dataset, both sourced from the World Health Organization (WHO) — will be merged using Country and Year indicators. Great consideration will be placed on determining the appropriate time steps between new cases. Some features, such as *Year* may have to be generated or extracted from existing samples. Additional datasets may also be sourced to augment the chosen ones if the need arises. Linear Regression will then be used to model the combined dataset with respect to the raw measures of air pollutants, type of settlement, and the cumulative cases, with the end goal of potentially predicting the number of new cases. Possible beneficiaries include government officials who would have more data to craft new policies in mitigating both air pollution and COVID-19, and also healthcare professionals who might be able to better track the progression of a COVID-19 outbreak, helping them institute more timely and appropriate measures.

A preliminary assessment of existing studies linking air pollution with COVID-19 were performed to check this project's novelty and soundness. The assessed studies did not use introductory linear regression as this project intends to, instead using more advanced methods such as ecological regression and multivariate logistic regression (Harvard University, 2020; Ali & Islam, 2020). Similar to labeling during preprocessing, the merger of datasets and the choice to use selective linear regression to possibly show which air quality features affect COVID-19 infection rates the most is also an unconventional approach in the project's context, enhancing its novelty.

The hypothesis is that because air pollutants can weaken the epithelial linings and interfere with the immunological functions of the respiratory tract, among other deleterious

effects, higher levels of air pollution would thus correlate with a greater chance that airborne and droplet-borne pathogens can bypass the body's defenses, increasing one's susceptibility to respiratory illnesses like SARS-CoV-2. This project is eagerly undertaken with hope that its results will contribute to safer air and healthier communities.

Background

According to WHO in 2018, roughly 7 million premature deaths can be attributed to air pollution each year, with ninety percent of the world's population continuing to breathe in dangerous levels of air pollutants every day. Recent findings have demonstrated an alarming connection between air pollution and wide-ranging health conditions, including respiratory problems, cardiovascular disease, and mental stress (Jiang et al., 2016). In response to these developments, WHO in 2021 has further tightened air pollution safety guidelines, lowering the tolerable exposure levels for most air pollutants.

Given air pollution's adverse impact to public health, the Air Quality Index (AQI) was developed as a practical means of informing the public about their local outdoor or ambient air quality. The AQI tool provides a color-coded scale with descriptors (ranging from "Good" to "Hazardous") and corresponding health advice (Cheng et al., 2020). There are multiple renditions of AQI, and the most commonly used is the US AQI which is measured from a scale of 0 to 500, wherein higher values represent greater levels of air pollution. It is used as a guide in forecasting what areas will be affected by bad air quality. The monitored air pollutants include PM2.5 and PM10 (fine particles with aerodynamic diameter of less than or equal to 2.5 and 10 micrometers, respectively), ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), and sulfur dioxide (SO2). This information helps civilians plan their activities around possible spikes of harmful air pollutant concentrations (ACT, 2023).

SARS-CoV-2 is the causative agent of COVID-19 whose spread has reached pandemic proportions. Although the disease no longer constitutes a public health emergency of international concern, periodic surges and new variants continue to pose a threat to everyday life (Taylor & Diamond, 2023). An infected person can produce virus-laden aerosols by talking, sneezing, or exhaling. These lightweight aerosols are capable of remaining suspended in the air by adhering to fine airborne particles (PM2.5 and PM10) for a prolonged period of time, establishing the disease's airborne route of

transmission (Thomson, 2019). Consequently, high amounts of air pollutants — which could come with more particulate vectors and could double as lung irritants — constitute a risk factor that can increase the infectiousness of the disease and thus its cumulative disease burden. Moreover, comorbidities such as hypertension and pulmonary disease are shown to have a strong correlation with the severity of COVID-19 (Liu et al., 2020). Many of these conditions are common outcomes from poor air quality, further indicating a potential link between air quality and COVID-19 infection.

As discussed, air pollution and COVID-19 appears to be related, and verifying the existence or non-existence of this relationship would be aided by imposing an air pollution dataset over a COVID-19 dataset, and then checking the accuracy of the model's predictions. In addition, a growing body of observational research and various synthesis suggests that air quality may play a role in the transmission and severity of COVID-19, further laying the groundwork for this exploration. As COVID-19 enters its endemic phase, with continuous circulation akin to Influenza, the importance of understanding its interplay with another chronic issue such as air pollution is made even more urgent.

Materials and methods

Datasets

This project will utilize the following datasets, restricted within the same time frame of 2020 to 2022:

(1) Ambient Air Quality dataset (WHO, 2023) (link)
(2) COVID-19 Cases and Deaths dataset (WHO, 2023) (link).

These datasets have been chosen for their appropriate feature sets, completeness (temporal and country-wise), reliability, and number of records. In particular, dataset (1) was selected over other options because it uses raw measures of air pollutants instead of discretized AQI scores, a characteristic that would hamper the planned linear regression. However, the project still considers the AQI scores to be good additions to a completed version. Meanwhile, dataset (2) was selected for its good formatting and for containing both new cases and cumulative cases as feature columns, improving compatibility with dataset

the first dataset. As a failsafe, we have also chosen reserve datasets in case our modeling for the previous pair of datasets does not work out. They are as follows:

(3) World Air Quality Index by City and Coordinates (Ramachandran, 2023) (link)
(4) total_cases.csv and total_cases_per_million.csv (Our World in Data, 2023) (link).

Preprocessing

The tentative preprocessing details are listed here. Main steps are data sanitization, potential imputation, and dataset merging.

1. Dataset (1) and (2) are downloaded and converted as a workable csv file.
2. To accomodate the temporal restriction, majority of rows or samples from dataset (1), from 2000-2022 will be dropped. Meanwhile, dataset (2) will follow with minor truncations.
3. Duplicates in both Datasets will be dropped.
4. Irrelevant features will be determined and dropped from both datasets.
    a. Dataset (1) will drop features *who_region*, *iso3*, and *version*.
    b. Dataset (2) will drop features *Country_code* and *WHO_region.*
5. To avoid wasting samples with NA values, imputation of NA-containing rows will be performed for both datasets instead of reckless dropping. Imputation would generally try to estimate the most likely value of a particular NA depending on its neighbors or some other activation function. If imputation is non-viable for the dataset, wholesale dropping of NA-containing rows will be performed.
6. The *Year_reported* feature will be extracted from the *Date_reported* feature of dataset (2), and augmented back to the database.
7. Dataset (1) and dataset (2) will be merged together, using *Country* and *Year* equality as conditions for the merge.
8. The combination of the two datasets would then be analyzed and pre-processed using Python in Jupyter Notebook.
9. Due to the nature of a merged table from different datasets, dimensionality reduction may have to be used to tackle lengthy feature sets.

Modeling Method

Linear Regression will be used to describe data and to explain the relationship between the Air Quality and COVID-19. From the combined dataset, Air Quality will be plotted against the amount of COVID-19 Cases. Further, specific pollutants of Air Quality can also be tested to show which may have the strongest effect towards COVID-19 Cases. Any statistically significant relationships can be found using linear regression, which can also serve as a basis for prediction.

If there exists a significant relationship between the target feature (new cases) and some key features (air pollutants raw measures and cumulative cases), then it is likely that the spread of COVID-19 is indeed exacerbated by air pollution.

Results and Prediction

The gathered results will reflect the relationship between air quality and COVID-19. A more specific pollutant might be shown to have a much stronger effect towards the spread and severity of the virus. Overall, finding a significant relationship would support the hypothesis statement; Air Quality heavily exacerbates the effects of COVID-19. This would be beneficial for policy makers to strategize and to plan methods in mitigating the decreasing air quality. Otherwise, not finding a significant relationship would still help future researchers dealing with air quality and COVID-19 by indicating a need to revise the approach used in this study.

References

ACT (2023) Air pollutants and sources. https://health.act.gov.au/about-our-health -system/population-health/environmental-mon itoring/air-quality/air-pollutants-and

Ali, N., & Islam, F. (2020). The Effects of Air Pollution on COVID-19 Infection and Mortality—A Review on Recent Evidence. *Frontiers in Public Health*, *8*. https://www.frontiersin.org/articles/10.3389/fpubh.2020.580057

Cheng, W. L., Chen, Y. S., Zhang, J., Lyons, T. J., Pai, J. L., & Chang, S. H. (2007). Comparison of the revised air quality index with the PSI and AQI indices. Science of the Total Environment, 382(2-3), 191-198.

Harvard University. (2020, May 19). Coronavirus and Air Pollution. *C-CHANGE* | *Harvard T.H. Chan School of Public Health*. https://www.hsph.harvard.edu/c-change/subtopics/coronavirus-and-pollution/

Jiang, X.-Q., Mei, X.-D., & Feng, D. (2016). Air pollution and chronic airway diseases: What should people know and do? *Journal of Thoracic Disease*, *8*(1), E31–E40. https://doi.org/10.3978/j.issn.2072-1439.2015.11.50

Liu, H., Chen, S., Liu, M., Nie, H., & Lu, H. (2020). Comorbid chronic diseases are strongly correlated with disease severity among COVID-19 patients: a systematic review and meta-analysis. Aging & Disease, 11(3).

Rodríguez-Urrego, D., & Rodríguez-Urrego, L. (2020). Air quality during the COVID-19: PM2. 5 analysis in the 50 most polluted capital cities in the world. Environmental Pollution, 266, 115042.

Taylor, A., & Diamond, D. (2023, May 5). WHO ends covid global health emergency—The Washington Post. *Washington Post*. https://www.washingtonpost.com/ world/2023/05/05/who-covid-global-health-emergency/

Thomson, E. M. (2019). Air Pollution, Stress, and Allostatic Load: Linking Systemic and Central Nervous System Impacts. *Journal of Alzheimer's Disease*, *69*(3), 597–614. https://doi.org/10.3233/JAD-190015

World Health Organization. (2018). *9 out of 10 people worldwide breathe polluted air, but more countries are taking action*. https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

World Health Organization. (2020). *Coronavirus disease (COVID-19)* https://www.who.int/health-topics/coronavirus

World Health Organization. (2021). *What are the WHO Air quality guidelines?* https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines

Datasets

World Health Organization (2023) *Ambient Air Quality Database*. https://www.who.int/publications/m/item/who-ambient-air-quality-database-(update-2023)

World Health Organization (2023) *COVID-19 Global Table Data.* https://covid19.who.int/WHO-COVID-19-global-table-data.csv

Ramachandran (2023) *World Air Quality Index by City and Coordinates.* https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates

Our World in Data (2023) *Total Cases* https://github.com/owid/covid-19-data/tree/master/public/data/cases_deaths