Identifying New Factors in COVID - 19 AI Case Predictions

Lynn Pickering, Javier Viaña, Xin Li, Anirudh Chhabra,
Dhruv Patel and Kelly Cohen

February 26, 2021

2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI)
Stockholm, Sweden, 2020, pp. 192-196. (Preprint and Prereview Version)

Identifying New Factors in COVID - 19 AI Case Predictions

Lynn Pickering*, Javier Viaña†, Xin Li‡, Anirudh Chhabra§, Dhruv Patel¶, and Kelly Cohen‖
University of Cincinnati
Cincinnati, Ohio 45219, U.S.A.
Email: *pickerln@mail.uc.edu, †vianajr@mail.uc.edu, ‡KratosOmega@iCloud.com, §chhabrad@mail.uc.edu,
¶patel4db@mail.uc.edu, ‖cohenky@ucmail.uc.edu

***Abstract*—Many machine learning methods are being developed to predict the spread of COVID - 19. This paper focuses on the expansion of inputs that may be considered in these models. A correlation matrix is used to identify those variables with the highest correlation to COVID - 19 cases. These variables are then used and compared in three methods that predict future cases: a Support Vector Machine Regression (SVR), Multidimensional Regression with Interactions, and the Stepwise Regression method. All three methods predict a rise in cases similar to the actual rise in cases, and importantly, are all able to predict to a certain degree the unexpected dip in cases on the 10th and 11th day of prediction**

***Index Terms*—COVID - 19, Infectious Diseases, Artificial Intelligence, Support Vector Machine, Correlation**

I. Introduction

A. NASA Space Apps

"A thriving example of crowd sourcing and citizen science, the NASA International Space Apps Challenge is a yearly event in which teams around the world use NASA's open data to develop innovative solutions to challenges in Earth and space science and technology" [1]. The NASA Space Apps COVID-19 Challenge brought together over 15,000 people from 150 countries [2]. The authors were inspired to accept the "Human Factors" challenge, and sought to identify variables and patterns that could help predict new disease hotspots. After the 48 hour hackathon was over, the team continued to work on solving the challenge, and focused on identifying those critical and measurable factors in disease spread.

B. Background

COVID - 19 is a pandemic that has affected the lives of all humans, regardless of where they live. The virus does not yet have a vaccine, and even once a vaccine is created, companies will not be able to immediately manufacture the amount required. Long term health effects from COVID - 19 are not yet known , nor do scientists know much about human immunity to the virus. [3] In other words, it is unclear if surviving the virus makes humans immune to future COVID-19 infections. Several countries praised for their efforts are experiencing an uptick in virus cases. To successfully navigate this novel coronavirus pandemic, governments and citizens must identify indicators of virus outbreaks so that communities may take action to stop the outbreak before it begins. Citizens need this data so they can trust their actions will have a meaningful impact on mitigating the virus, so that they do not lose hope and become risky to their community.

This paper examines three methods to predict the evolution of cases in a given city. In doing so, crucially, the authors look to identify those variables that may indicate the future outbreak of a COVID hotspot. The approaches compared are the Support Vector Machine Regression (SVR), Multidimensional Regression with Interactions, and the Stepwise Regression approach. A number of inputs are examined here in depth, but the real value of the method is the ability to remove those variables that have no clear effect on hotspot prediction, and replacing them with new variables to be tested. Many artificial intelligence methods look to predict the spread of COVID-19, including methods developed during the 2020 NASA Space Apps COVID - 19 challenge [4]. Our model will provide valuable data on what variables may be added to those AI methods to improve their prediction accuracy.

C. Objective

The current study focuses at the county level spread of the disease. The goal is to successfully predict the future COVID-19 cases in US counties with sufficient precision in the short term. Ultimately, this tool will be part of a data-driven decision-making software that can help the government and local authorities during the pandemic.

D. Hypothesis

The preliminary hypothesis assumed, is that there exists a correlation between the data gathered and the spread of disease at the county level.

Furthermore, it is not only expected that this data is correlated but also that it has a good predictive power in the forecasting of new COVID-19 cases.

The lockdowns caused by the disease and the fear of infection, tend to reduce the human activity. Thus there are less pollutants emitted to the air. At an atmospheric level, the ESA has proven a reduction in the concentrations of different

*† Corresponding authors. E-mail: pickerln@mail.uc.edu, vianajr@mail.uc.edu

pollutants [5]. This paper, among other things, evaluates the same effect with the ground air-quality.

To simplify the analysis of the causal impact of the disease spread in the concentrations of the area, the following assumption is considered: The evolution of the concentrations in the current year compared to the average of recent years (the previous 2 years are selected) provides useful insights.

The concentrations of the pollutants are not the only features considered. For the remaining features (mentioned below) a similar assumption is made to measure the causal impact.

II. Related Work

The world has turned its attention to fighting the coronavirus, each discipline contributing expertise. Nath, Gary and Shepard-Smith [6] list some of the ways that Artificial Intelligence has been used in the fight against COVID - 19 such as mapping and predicting hot spots, contract tracing, research databases and search engines. Yang et al. modified a Susceptible - Exposed - Infectious - Removed (SEIR) model to derive the curve of the epidemic, and then used a Long-Short-Term-Memory (LSTM) neural network to predict future cases [7]. Yang et al. shows how consequential the effects of placing quarantine orders five days earlier or five days later would have been, for several provinces in China. SVR has been used in the prediction of COVID - 19 cases with the inputs to the model all taken from the Center for Systems Science and Engineering at Johns Hopkins University daily [8]. In another study, Linear Regression and Multiple Linear Regression models are compared using global data available daily from the World Health Organization(WHO) [9]. To build on the work that is being done, this paper focuses on identifying new sources of data and new variables that can make AI methods for predicting new hotspots of COVID - 19 more effective. Furthermore, the paper compares new cases of disease predictions across several models, given the same inputs.

III. Methodology

A. Data Collection

1) Flight Data: Traveling has had a major impact on the spread of disease across large and smaller distances, and is included as an input to the models. The time frame for the flight data collected is February and March of 2018, 2019 and 2020. The average of number of flights in 2018 and 2019 are used as a reference to reduce fluctuations these 2 years in the data. This average is compared to the 2020 flight data to get a better insight of how the traveling rate changes, as the changes in traveling rate to a specific city is more relevant when comparing cities that have vastly differing normal traveling rates.

The flight data is is obtained from Bureau of Transportation Statistics [10].

2) Hospitalizations Data: Hospitalizations also play an important role in disease forecasting. The emergency visits of five age ranges along with the average hospital visits of those age ranges, allow us to pick up some latent trends of potential outbreaks. Furthermore, by investigating the infected cases and number of hospitalizations with the death cases, it is possible to estimate the potential future outbreaks, as well as the dangerous potential of an overwhelmed medical system.

Since NYC had been impacted the most from COVID-19, NYC Health [11] provides valuable data that fits the disease forecasting purpose.

3) Traffic Data: The new driver application count dataset is used to assist the forecasting of disease. Unlike the flight data, which is well documented and can be easily tracked, ground traffic is very difficult to estimate. However, by studying the new driver applications, a latent trend of how people's will to drive is affected by the disease over a period of time can be discovered to help disease forecasting.

Such data is obtained from NYC Open Data [12] by selecting several criteria to narrow down the focused time period and region.

4) Climate Data: Climate data plays an important role in this research. The virus has greatly affected many areas of people's life, such as how they commute, work and eat. All these activities relate to the need for transportation. Furthermore, factories and other heavy users of energy have had to shut down or limit operations to allow for the safety of their workers in these times. As a result, pollutants, and therefore the climate, is affected during virus outbreaks. Although reduced pollution does not cause disease, climate data can be used to forecast future outbreaks as it reflects the level of transportation and commercial energy use. This is an important factor in disease spread.

The air quality data is obtained from the Aura Satellite (OMI instrument) [13] and from the Environmental Protection Agency [14].

B. Factor Correlation

A correlation matrix is obtained that quantifies how important the identified time variables are. Subsequently, a Principal Component Analysis algorithm is applied that filters the information, leaving only the variables with an identified high contribution to the evolution of the disease. Finally, three methods are used to obtain a prediction of the number of people currently infected by COVID-19; SVR, multidimensional regression with interactions and stepwise regression.

C. Support Vector Machine Regression

In this study, SVM (Support Vector Machine) is used for regression, hence its called SVR. Traditionally, SVM has been utilized for binary or multi-class classification problems. Despite the nature of the problem, there are just a few differences between these two techniques (SVM and SVR). Perhaps the most distinct factor is the use of the concept hyperplane [15]. In SVM this is the separation line that divides the different classes of data. But, in SVR the hyperplane is defined as the line that helps in the prediction of the target variable. Using the hyperplane as a reference, the boundary lines are defined, ultimately aiming to enclose as many points as possible within the margin of tolerance.

TABLE I
INPUTS OF MODEL

Input	Metrics For a Given County
Air Quality Data	*Unusual trends in the daily concentrations of ground level CO, NO2, Ozone and SO2*
Hospital Data	*Daily number of emergencies attended for several age-ranges and number of hospitalizations*
Flight Data	*Total number of departing and arriving flights per dayC*
Traffic Data	*Number of new driving licenses issued each day*

D. Multidimensional Regression with Interactions

Using a simple regression approach might perform better than other methodologies for extrapolation purposes. In the case and period studied, the number of COVID-19 cases continue to reach limits that were not seen before. Thus, the problem requires a model with good extrapolation capabilities.

The training of the algorithm is carried out with the available data up to the day considered, March 31st, 2020. Nonlinear combinations of the input features were also evaluated to obtain a better prediction, generating the so called interactions. These are artificial features, that root in the original data.

E. Stepwise Regression

Unlike the multidimensional regression with interactions, this model automates the selection of features, both the original and the artificially created ones. It uses an iterative validation procedure to decide whether to subtract or add a given variable. The criteria to stop the process and come up with the best set of predictive features, is based on the results for the F-test [16].

F. Inputs to Models

The inputs to the model are summarized in Table I.

The air quality data was entered in the model as the difference of the smoothed evolution of the concentration of each air quality value. Raw data on these inputs for the city of New York was taken from the Aura satellite [13]. The processing of one of the air quality values, CO, is shown in Figures 1, 2, and 3. The data is smoothed from the raw data to obtain a useful seasonality variation (sudden peaks and valleys are removed), and the difference is taken from an average of the values over the past two years, because the deviation of these values from previous years is what provides the most meaningful information.

IV. Results

A. Factor Correlation and Disease Spread Forecasting

To complete the factor correlation, this method focuses on the city of New York (NYC). NYC is one of the cities with the highest number of coronavirus cases in the United States as well as the world(in the time period studied), and extensive city data is easily available. The methods have been tested given the information between February 1st and March 31st of 2020. The results of the correlation matrix are shown in Fig. 4 and Fig. 5.

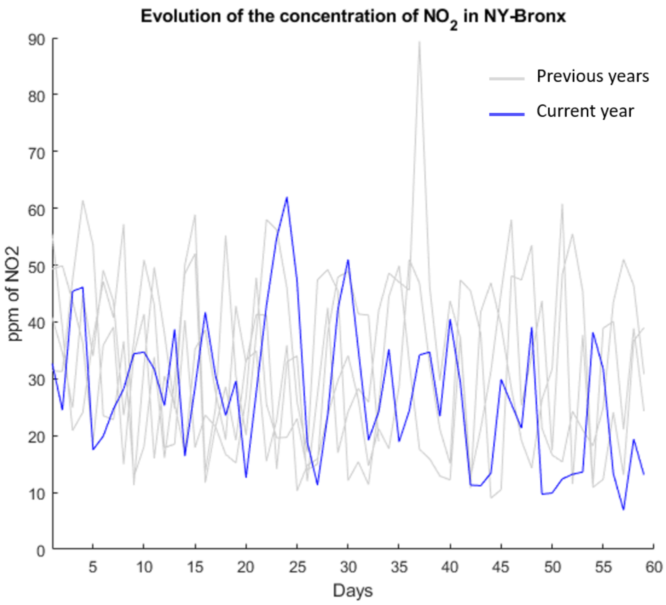

Fig. 1. Evolution of the concentration of NO2 in NY-Bronx

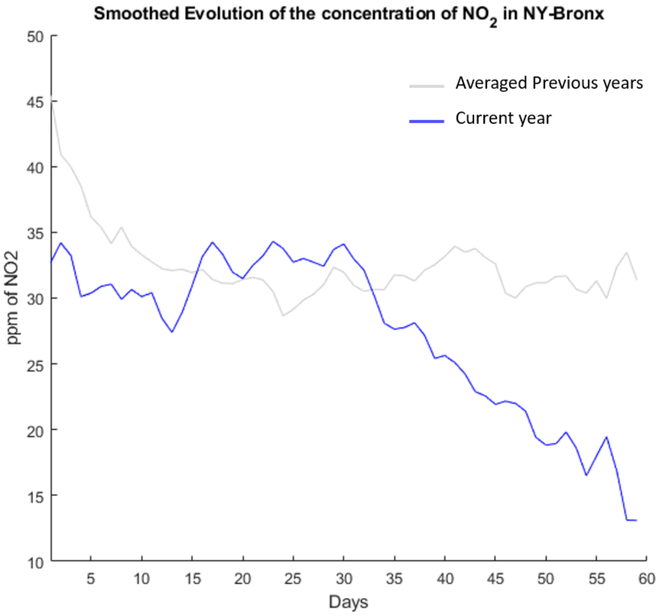

Fig. 2. Smoothed Evolution of the concentration of NO2 in NY-Bronx

First, the decrease of the human activity can be inferred from the CO and NO2 concentrations. As expected, these two features are highly correlated to the COVID-19 cases.

On the other hand, the SO2 and the Ozone concentrations do not seem to be affected. One study hypothesizes that ozone

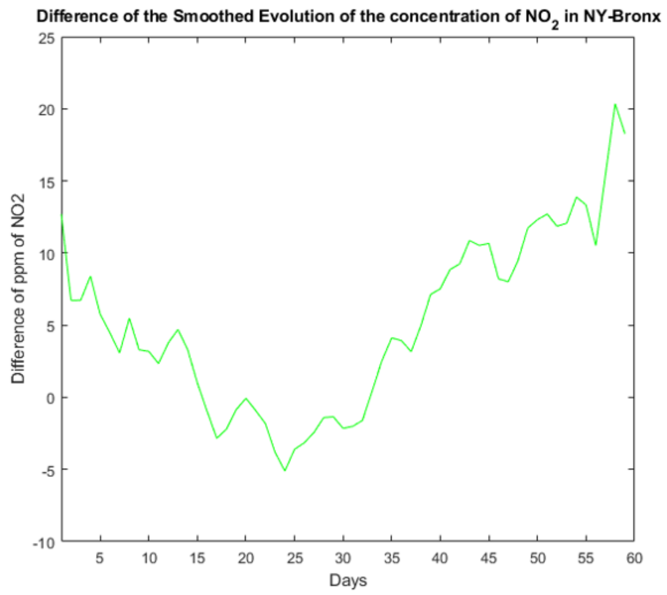

Fig. 3. Difference of the Smoothed Evolution of the concentration of NO2 in NY-Bronx

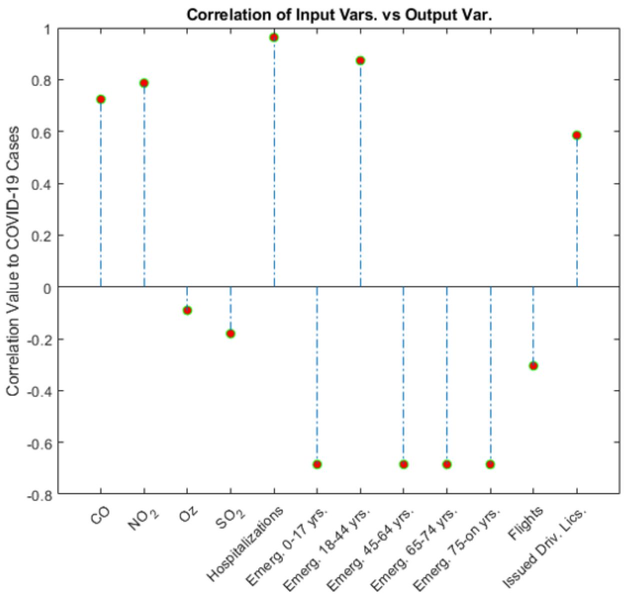

Fig. 4. Correlation of Input Vars. vs Output Var.

levels have not dropped during the shutdowns due to the fact that ozone is a secondary pollutant, and is dependant on many more sources than just traffic [17].

Both the total hospitalizations and the emergency hospitalizations for those 18-44 years old have a high correlation to the spread of the disease. Nevertheless, the rest of the features of hospitalizations (age ranges 0-17, 45-64, 65-74, 75-on) are not as useful as these two. This might be due to the presence of young adults in the city (average age of the NYC population in this time frame was 35.8 years).

In the period selected, the outbreak was still at its earliest stage, thus, not much effect was visible yet in the commercial flights of NYC.

The NO2 concentration and the issued drivers licenses decrease significantly as the disease spreads. It can be inferred that the traffic diminished, as more people in the city was opting for telematic working and there was a smaller necessity to use ground transportation for commuting purposes. In some places, the offices issuing the drivers licenses were considered non essential businesses, and shut down, causing a decrease in new drivers licenses as well.

B. Prediction

For the prediction phase, the future 12 days are considered. All the three methods obtain a successful forecasting of the daily cases.

The result of the multidimensional regression with interactions approach, support vector machine regression approach, and stepwise regression approach for the purposes of predicting the number of COVID-19 cases are shown in Fig. 6.

The metric used to compare the efficiency is the RMSE (Root Mean Squared Error). The result of this figure for each method is shown in Table II.

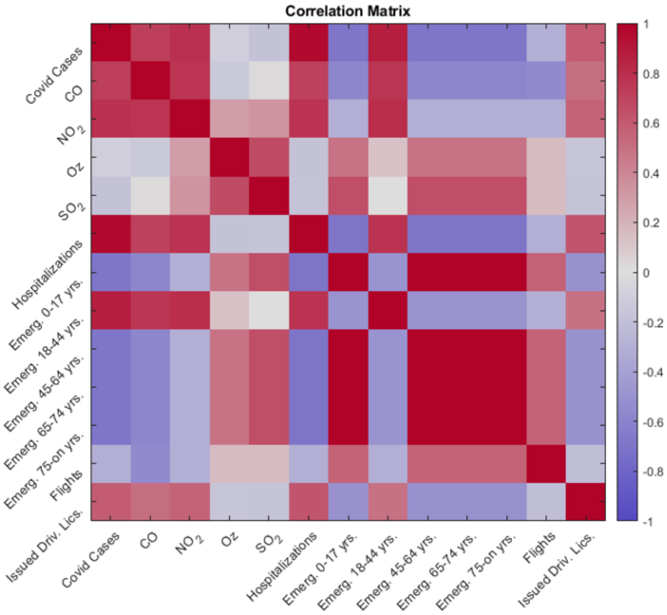

Fig. 5. Correlation Matrix

V. Conclusion and Future Work

The correlation matrix shows interesting results. Particularly interesting is that emergencies that led to hospitalizations, categorized by age group, only show positive correlation with the number of COVID - 19 cases in the 18-44 age group. The correlation matrix was used to test time dependant data. To better evaluate the validity of the matrix, it must be tested on a wider range of cities, and on a greater number of inputs. The models used to predict future COVID - 19 cases include the SVR, multidimensional regression with interactions, and

TABLE II
RESULTS OF MODELS

Method	**RMSE**
Support Vector Machine Regression	796.42
Multidimensional Regression with Interactions	911.81
Stepwise Regression	1129.43

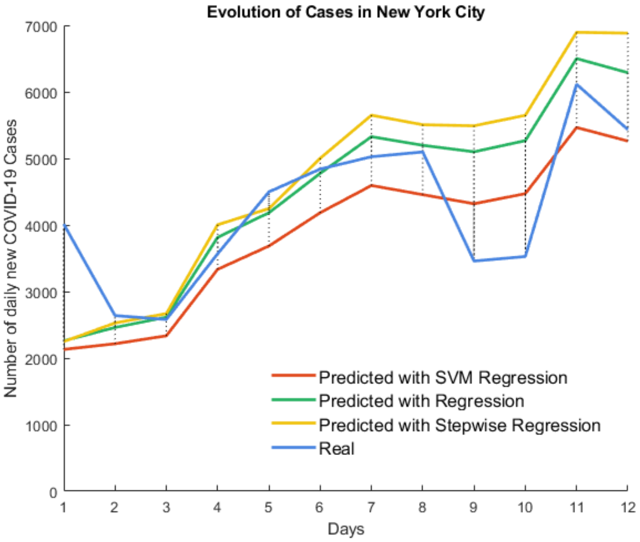

Fig. 6. Results

stepwise regression approaches. Using the RMSE as a measure of accuracy, the SVR method performs best at predicting the next 12 days of case numbers. All three methods are able to predict to a certain degree the unexpected dip in cases on the 10th and 11th day of prediction. This experience leaves room for future exciting work. The factor correlation and disease spread forecasting component will be expanded to incorporate more variables, such as stock data, prices of different products, number of events carried out in the city, etc, and the model will be tested on an increasing number of cities. Furthermore, the model will be expanded to include categorical data, not just numerical data. Furthermore, a separate study to develop an efficient data collection system can be conducted in order to make such data available seamlessly for future studies. The results of this study show the positive correlation that data such as certain pollution data and issued driver licences have with COVID - 19 cases, and highlights the importance of including more data in disease prediction models, expanding on data found just in the Johns Hopkins database, and other such sources.

ACKNOWLEDGMENT

The authors would like to thank the NASA International Space Apps Challenge for a call to action in these challenging times, as well as Dr. Gautam Pillay and Dr. Scott Petersen for their interest and encouragement throughout this research.

REFERENCES

[1] Hemmings, S. N., et al. “The NASA Space Apps Challenge: Leveraging the World’s Largest Hackathon for Crowdsourcing and Citizen Science.” NASA/ADS, 2020, ui.adsabs.harvard.edu/abs/2019AGUFMIN51E0683H/abstract.

[2] Talbert, Tricia. “Space Apps COVID-19 Hackathon Brings 15,000 Together Worldwide.” NASA, NASA, 9 June 2020, www.nasa.gov/feature/space-apps-covid-19-hackathon-brings-15000-together-worldwide.

[3] “The Long-Term Health Effects of COVID-19.” Gavi, the Vaccine Alliance, 2020, www.gavi.org/vaccineswork/long-term-health-effects-covid-19.

[4] “G.I.D.E.O.N. : An Integrated Assessment.” covid19.Spaceappschallenge.org, 2020, covid19.spaceappschallenge.org/challenges/covid-challenges/integrated-assessment/teams/gideon/project.

[5] “Coronavirus Lockdown Leading to Drop in Pollution across Europe.” ESA, 2020, www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-5P/Coronavirus_lockdown_leading_to_drop_in_pollution_across_Europe.

[6] Nath, Hemanta and Gary, Todd and Shepard-Smith, Andrew, Artificial Intelligence Tools and Models Used by the Scientific Community to Address the COVID-19 Pandemic (July 21, 2020). Available at SSRN: https://ssrn.com/abstract=3657855 or http://dx.doi.org/10.2139/ssrn.3657855

[7] Yang Z, Zeng Z, Wang K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020;12(3):165-174. doi:10.21037/jtd.2020.02.64

[8] Yaohao Peng, Mateus Hiro Nagata, An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data, Chaos, Solitons and Fractals, Volume 139, 2020, 110055, ISSN 0960-0779, https://doi.org/10.1016/j.chaos.2020.110055.

[9] Smita Rath, Alakananda Tripathy, Alok Ranjan Tripathy, Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, Volume 14, Issue 5, 2020, Pages 1467-1474, ISSN 1871-4021, https://doi.org/10.1016/j.dsx.2020.07.045.

[10] Bureau of Transportation Statistics [Dataset]. Available: https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236. [Accessed: May 30, 2020].

[11] NYC Health [Dataset]. Available: https://www1.nyc.gov/site/doh/covid/covid-19-data.page. [Accessed: May 30, 2020].

[12] NYC Open Data [Dataset]. Available: https://opendata.cityofnewyork.us. [Accessed: May 30, 2020].

[13] Aura Atmospheric Chemistry, Ozone Monitoring Instrument. [Dataset]. Available: https://aura.gsfc.nasa.gov/omi.html. [Accessed: May 30, 2020].

[14] EPA, Outdoor Air Quality Data [Dataset]. Available: https://www.epa.gov/outdoor-air-quality-data/download-daily-data. [Accessed: May 30, 2020].

[15] Burges, C. J. C. ”A Tutorial on Support Vector Machines for Pattern Recognition.” Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998, pp. 121-167. SCOPUS, www.scopus.com, doi:10.1023/A:1009715923555.

[16] Harper, Jeffrey F. “Peritz’ F Test: Basic Program of a Robust Multiple Comparison Test for Statistical Analysis of All Differences among Group Means.” Computers in Biology and Medicine, Pergamon, 16 Mar. 2004, www.sciencedirect.com/science/article/pii/0010482584900441.

[17] PM2.5 and Ozone Air Pollution Levels Have Not Dropped Consistently Across the US Following Societal Covid Response [Online]. Available: https://chemrxiv.org/s/3299fafedd485e00c885. [Accessed: Aug 8, 2020].