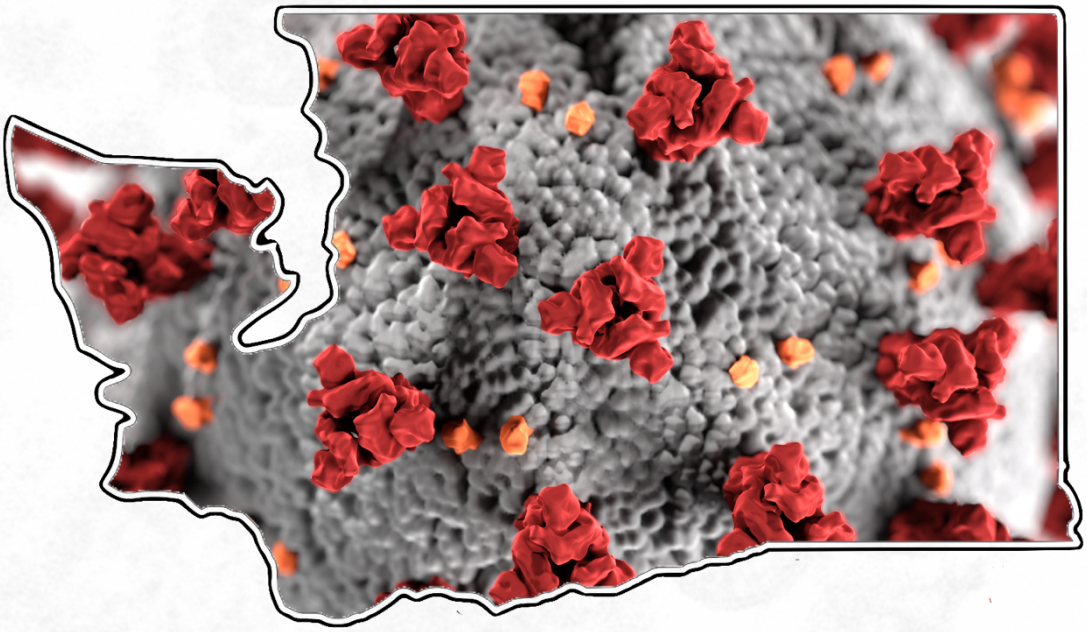


PARTICULATE MATTER 2.5 AS A CONTRIBUTING FACTOR TO COVID-19 EVENTS IN WASHINGTON STATE: A SUPERVISED LEARNING APPROACH

ABSTRACT

Pollution has been linked to increases in both Cardiovascular Disease and respiratory illness, but little investigation has been performed to date regarding the impact of daily fluctuations in Particulate Matter (PM) on COVID-19 case counts, hospitalizations, deaths and recovery times. Due to the smaller size of PM 2.5, it is able to efficiently bypass our defenses, enter our bloodstream and damage our cells causing increased susceptibility to viral infection contraction. This study aims to perform a short-term time series analysis of COVID-19 event rates in relation to PM 2.5 levels based on both state-level and county-level scales. Environmental Protection Agency (EPA) approved and Volunteered Geographic Information (VGI) air quality data was collected and compared in terms of reliability and consistency, from which correlation between COVID-19 incidents and PM 2.5 was deduced through various statistical and data visualization analyses. Ultimately, long-term PM 2.5 level averages were found to have a more direct impact on cumulative COVID-19 positive cases, but daily collated COVID-19 data played a role in training and testing Machine Learning (ML) algorithms to predict its occurrence based on PM 2.5 and other variables included within the study.

KEY WORDS: PM 2.5, COVID-19, Supervised Machine Learning Algorithms

STUDENT	MARIE ANN LENIHAN-CLARKE
PROGRAM	MSc GEOINFOTMATICS, THESIS
SUPERVISOR	JAMAL JOKAR ARSANJANI
DATE OF SUBMISSION	JUNE 4, 2021

Acknowledgements

I would first like to thank my thesis supervisor, Professor Jamal Jokar Arsanjani at Aalborg Universitet, who made it known that he was always available to help me and gave me confidence in my ability to complete this project and others. I would like to thank Professor Carsten Keßler, also at Aalborg Universitet, who enabled me to build a strong foundation of knowledge with respect to Geoinformatics and who together with Professor Jamal Jokar Arsanjani, pushed me to become a better student and citizen by encouraging my continuous learning.

I would also like to acknowledge Marina Georgati and all staff at Aalborg Universitet in the Geoinformatics (Surveying and Planning) department, who were available to answer any questions I had as a student who was new to both Denmark and the Danish education system.

I would also like to thank my peers, especially Ana Cristina Mosebo Fernandes, Rebeca Quintero Gonzalez and Ezra Francis Leslie Trotter, who showed me what exceptional teamwork can accomplish and who have remained supportive to one other throughout our academic ventures.

Finally, I would like to express my profound gratitude to several friends and family members who have provided me with advice and consistent support throughout the duration of my studies; thank you to my husband Seb, my parents Bill, Ann, Don and Edith, my brothers Tom and Luc, my friends Grace, Kate, Mak, Allie, Faisal, Kat and Sey, and our pup Roxanne, who offered endless companionship and love.

Marie Ann Lenihan-Clarke.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

API	Application Programming Interface
AQI	Air Quality Index
AQS	Air Quality System
ASCII	American Standard Code for Information Interchange
AUCROC	Area Under Curve – Receiver Operating Characteristic
BAM	Beta Attenuation Mass
BR	Bayesian Ridge
CBSA	Core Based Statistical Areas
CDC	Centers for Disease Control and Prevention
CHARS	Comprehensive Hospital Abstract Reporting System
CM	Confusion Matrix
COVID-19	Coronavirus
CSV	Comma-Separated Values
CV	Cross Validation
CVD	Cardiovascular Disease
DT	Decision Tree
ED	Emergency Department
EHD	Environmental Health Disparity
EIOS	Epidemic Intelligence from Open Sources
EPA	Environmental Protection Agency
EPSG	European Petroleum Survey Group
FEM	Federal Equivalent Method
FN	False Negative
FP	False Positive
GAM	Generalized Additive Model
GCS	Geographic Coordinate System
GIS	Geographic Information System
ICD	International Classification of Diseases
IDE	Integrated Development Environment
IQR	Interquartile Range
LM	Linear Regression Model
MAE	Mean Absolute Error
MERS	Middle East Respiratory Syndrome
ML	Machine Learning
MSE	Mean Square Error
OSM	Open Street Map
PCR	Polymerase Chain Reaction
PCS	Projected Coordinate System
PM	Particulate Matter
PRISM	Parameter-elevation Regressions on Independent Slopes Model Climate Group
RF	Random Forest
RHINO	Rapid Health Information Network
RSEI	Risk-Screening Environmental Indicator
RMSE	Root Mean Squared Error
SARS	Severe Acute Respiratory Syndrome
SHP	Shapefile
SVM	Support Vector Machine
TIFF	Tag Image File Format
TN	True Negative
TP	True Positive
TSDF	Treatment, Storage and Disposal Facility
VGI	Volunteered Geographic Information
WDRS	Washington Disease Reporting System
WHO	World Health Organization

INTRODUCTION

According to the international weekly journal of science, Nature, paper submissions focused on modeling the Coronavirus (COVID-19) pandemic and its outcomes (e.g., testing, diagnostics and hospitalizations) peaked in April 2020, plateauing shortly thereafter. At present, topics of interest concerning the pandemic that shook the world in early 2020, fall under mental and public health categories (Else 2020). With the continuing accumulation of data on a global scale, and with increasing numbers of variants redefining disease strains causing its persistent spread, it could be in the interest of science to revisit modeling the virus and further investigating the factors contributing to increased cases, hospitalizations, deaths and recovery rates.

Pollution has been linked to COVID-19 in a positive manner due to the reduction of emissions occurring worldwide as a result of government prompted lockdowns. However, as with respiratory illness and Cardiovascular Disease (CVD), pollution may be a cause of surges in positive COVID-19 rates and prolonged recovery periods. This study examines the relationship between the two in further detail.

COVID-19

On December 31, 2019, a Wuhan Municipal Health Commission media statement that reported on an episode of viral pneumonia, was translated and shared with the World Health Organization (WHO) (World Health Organization 2021). Simultaneously, WHO's Epidemic Intelligence from Open Sources (EIOS) department also received reports of the situation and basic level emergency response protocol was initiated. Within ten days, Chinese authorities had identified the outbreak as being caused by a disease in the coronavirus sub-family, a group of RNA viruses including Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS) and the common cold, that usually share alike symptoms such as respiratory tract infections (Cleveland Clinic 2020). According to the Centers for Disease Control and Prevention (CDC), COVID-19 case incidence rapidly occurred from the time of virus identification to its spread throughout North American states and territories. The first US case was detected on January 21, 2020, through the use of Polymerase Chain Reaction (PCR) testing performed on a Washington state resident, by which time multiple COVID-19 deaths had occurred in China. As of mid-March 2020, all US states had encountered cases. At the time of this paper, approximately 168.6 million cases have been confirmed and over 3.5 million deaths have resulted from the contraction and spread of this disease globally (World Health Organization 2021). The United States accounts for just below 20% of COVID-19 cases and just over 16% of associated deaths (The New York Times 2021).

Due to the highly contagious nature of COVID-19, national-scale lockdowns, preventative measures and mandates have been implemented to reduce case rates. Such measures include keeping surfaces clean, wearing face coverings, frequently using hand sanitizer and washing hands, keeping a social distance of six feet apart, avoiding crowds and poorly ventilated spaces and when readily available, getting vaccinated (Centers for Disease Control and Prevention 2021). Although these measures aim to diminish the threat of COVID-19 spread, certain environmental factors are being researched in terms of their impact on the spread and severity of the disease. As the surface of the virus enables it to travel the respiratory tract and latch on to and invade healthy cells, particularly those in our lungs (WebMD LLC. 2021), it is important to consider what could incapacitate the body's ability to fight it off and ultimately result in higher case counts, hospitalizations, deaths and elongate recovery times. Current research on respiratory diseases and linked environmental variables places strong emphasis on pollution, its spatial variance and how it could inevitably change the course and nature of disease in certain locations dependent on its levels.

Pollution, Air Quality and Impact to Respiratory Disease

According to the Environmental Protection Agency (EPA), five major air pollutants that are regulated by the Clean Air Act, are monitored for Air Quality Index (AQI) purposes. These pollutants consist of ground-level ozone, carbon monoxide, sulfur dioxide, nitrogen dioxide and particle pollution, also known as Particulate Matter (PM), each of which possess a national air quality standard aimed towards protecting public health (AirNow 2021).

Table 1 Major air pollutants and sources

Leading Pollutant	Chemical Abbreviation	Source
Ozone	O3	Vehicle exhaust, industry emissions
Carbon Monoxide	CO	Vehicle exhaust, various fuel combustion
Sulphur Dioxide	SO_2	Fossil fuels, power plants, industrial facilities
Nitrogen Dioxide	NO_2	Fossil fuels, vehicle exhaust, power plants
PM 10		Road dust, wildfires, vehicle exhaust
PM 2.5		Factories, vehicle exhaust, waste, wood burning

PM is a primary source contributing to air pollution levels, either liquid or solid in form, suspended throughout the air and invisible to the naked eye. Rather than being defined by chemical composition as other pollutants often are, PM is expressed by size (Natural Resources Defense Council, Inc. 2014) in microns, and consists of PM 2.5 and PM 10. The smaller of the two, PM 2.5, which is made up of combustion particles and dust, has the ability to disperse up to hundreds of miles in distance in comparison to its counterpart, PM 10, which is limited to approximately 30 miles in range (Pima County 2021). Additionally, the smaller the particle, the more efficiently it is able to bypass our defenses, enter our bloodstream, occupy internal areas of our body and consequently trigger inflammatory responses and cause damage to our cells (measuring a mere six microns comparatively). By causing damage to internal organs and systems, contraction of, and heightened symptoms experienced from, mainstream influenza are common among populations who are at risk of exposure to higher levels of PM 2.5, PM 10 and other pollutants such as diesel exhaust emissions (IQAir 2021). In an attempt to address this issue among others, the EPA established the National Ambient Air Quality Standards for PM 2.5 in 1997. This has since been reviewed twice, with the most recent occurring in 2012, leading to a short-term acceptable standard of 35 micrograms per cubic meter of air (μg/m³). PM 2.5 is significant in its contribution to AQI level, whereby a reading of the first 10 micrograms

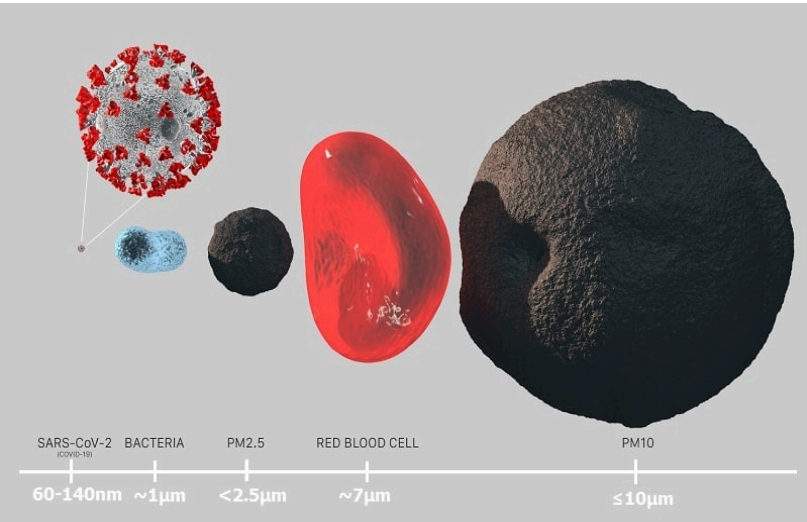


Figure 1 Size of PM, source: (Cralle 2020)

designates 42 index points (Talhelm 2020) out of a possible 500.

Due to the ongoing monitoring of air quality and pollutant levels, data has become more easily and openly accessible to health professionals. Studies underway are able to compile vast amounts of data that include a combination of granular information for age, gender, race, zip code, dates and diagnostic codes of patients in addition to timeseries analyses of air pollution. According to Francesca Dominici, a biostatistician and Harvard T.H. Chan School of Public Health professor, hospitalization rates and viral deaths increase where higher readings of PM 2.5 are collected (National Geographic 2021), specifically where exposure to such has occurred over the course of decades. According to Dominici's team, particle pollution accounted for 15% of COVID-19 related deaths on a global scale and peaked at 27% in East Asia. Research continues to show correlation between harmful levels of air pollution and human health and, furthermore, can often shed light on certain political and regulatory conduct.

Project Goals and Objectives

PM 2.5 has been associated with increased respiratory illness and side effects. The purpose of this project is to examine various aspects of one such disease, COVID-19, including case totals, hospitalizations, deaths and potentially recovery times, in relation to pollution data, primarily PM 2.5 levels, both as a short-term timeseries analysis and as an averaged value. An analysis of whether PM 2.5 readings from EPA approved or Volunteered Geographic Information (VGI) air quality sensors, based on county or zip code respectively, will be performed to evaluate if one type is more capable than another in showing correlation between respiratory disease and pollution intensity. The inclusion of other meteorological, environmental and socio-economic variables should also be evaluated and compared to PM 2.5 as a leading cause for increased COVID-19 presence. Such variables may provide a complementary, or more comprehensive, explanation as to the rate of respiratory illness related events. Both classification and regression techniques should be considered in estimating current and future presence and absence of COVID-19, positive rates of the viral disease, and finally, duration of recovery time, all based on location, with certain investigations incorporating timeseries analysis.

Problem Statement

Connections between pollutants and respiratory disease have been established over time through investigative research. However, it is less clear whether short-term fluctuations in PM 2.5 can have as equal an impact on respiratory disease occurrence and intensity as cumulative exposure can.

Less research has been focused on whether standard-driven data collection provides more reliable results in comparison to VGI when analyzing PM 2.5 levels at higher temporal resolutions.

Although recent research has been conducted to analyze the correlation between COVID-19 case rate and PM 2.5 levels, due to restrictive data and methodologies, less attention has been placed on COVID-19 recovery times in relation to pollutant levels.

Research Questions

This project aims to answer the following research questions:

1. Does there appear to be significant difference in the results of pollutant-related, spatially dispersed data, whereby one set is the product of EPA approved air quality sensors and another is acquired through open-source VGI?

2. How accurately can COVID-19 occurrence/non-occurrence be estimated using classification-based Machine Learning (ML) algorithms commensurate with pollution level data and other alternate variables?
3. Can regression-based ML algorithms accurately approximate COVID-19 case rates, hospitalizations, deaths and recovery times using PM 2.5 and additional environmental variable data; can this lead to the assessment of susceptible areas at higher risk of extreme outcomes?

BACKGROUND

State of the Art

Analysis of Short-Term Effects of Air Pollution on Cardiovascular Disease Using Bayesian Spatio-Temporal Models, 2020

(Liu, et al. 2020)

This study provides insight into short-term exposure to outdoor emissions and air pollution as a contributing factor to the exacerbation of CVD.

Hospital admission data was collected from both inland and coastal cities in order to compare heavily polluted, densely populated regions with more rural areas within China. Bayesian spatio-temporal models were applied to estimate underlying pollution levels in each of the cities and probability was calculated to represent uncertainty prior to the application of a Generalized Additive Model (GAM). This nonlinear regression model assessed the effects of pollution on health, demonstrating that overall ambient air pollution in heavily polluted cities had an inverse impact on CVD hospital admissions, whereas short-term dramatic increases in air pollution levels in coastal regions led to excessive rates of CVD. This result was explained by discussing the underlying biochemical processes and bodily functions that enable adjustment to the exposure of high toxic pollutants levels in larger, more polluted cities.

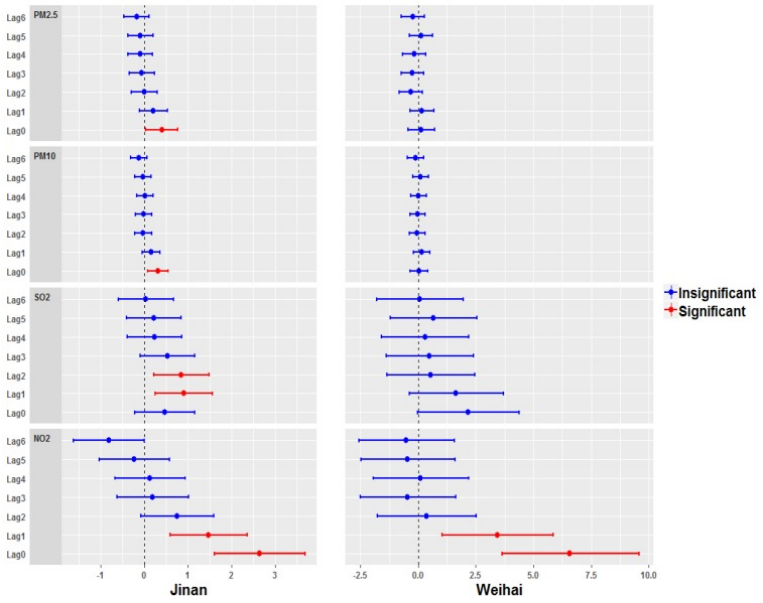


Figure 2 Percent change in CVD hospitalization risk per 10 μg/m³ pollution level increases, source: (Liu, et al. 2020)

This study was able to include the use of spatial variability by using daily air pollution concentration data that were read from specific areas rather than using a set of aggregated data for an entire administrative region. This ensured that measurements were not averaged over spatial locations. The air pollution and meteorological data source was governmental and consisted of 15 and nine monitoring stations, for each the inland and coastal city respectively, with hourly readings for pollutants PM 2.5, PM 10, SO_2 and NO_2 over the same time period. This data included daily mean

pollutant measures, highest and lowest values of monitored objects, average daily temperature and humidity, wind speed and pressure.

Hospital data spanned two years from each study area and included the name of the hospital, date of service, patient age, gender, date of birth and address and an International Classification of Diseases (ICD) code.

An additional consideration that was given to the study was time lag; a zero-to-six-day lag was used when comparing results of the pollutant levels versus CVD hospital admission rate.

This paper offered valuable introductory information in relation to GAMs as the most commonly used model in epidemiological studies as well as raising discussions on unexpected result outcomes. Physical geography of each area was called into question as a potential contributing factor of CVD hospital admission rates, in addition to pollutant and meteorological variables. As one study area was a basin, it was acknowledged that such topography could make it more difficult for emitted, ambient air pollution to drift away from the city, ultimately leading to inversion type conditions.

The Association between Respiratory Infection and Air Pollution in the Setting of Air Quality Policy and Economic Change, 2019

(Croft, et al. 2019)

Using case-crossover methods, this study attempts to estimate respiratory infection rate in adults aged 18-65 years in relation to acute increases in surrounding PM 2.5 concentrations. The time period for this study is divided into three segments: before (2005-2007), during (2008-2013) and after (2014-2016) the implementation of air quality policy and economic changes.

Data was collected for both the population being studied and air pollution and weather variables. Hospital admission and Emergency Department (ED) visit statistics were collected for all New York residents in the age range specified, who lived within a 15-mile radius of pre-identified air pollution monitoring sites that spanned a decade of operability. Six EPA-approved monitoring stations provided hourly PM 2.5 concentration readings in addition to temperature and relative humidity data. Each patient was assigned the values of the nearest corresponding station site prior to a statistical analysis being performed. Both hospital admissions and ED visits associated with each Interquartile Range (IQR) increase in PM 2.5 concentration, were estimated by fitting a logistic regression model and using a 'time-stratified, case-crossover design,' executed in R.

Results from this study show that over the decade long study period: the number of ED visits have increased in general, no changes pertaining to respiratory illness and seasonal patterns have been observed, increased rates for both hospitalizations and emergency room visits relating to culture-negative pneumonia increase along with augmented PM 2.5 concentrations after a two-to-seven-day time lag consideration, and no correlation was found between bacterial pneumonia and PM 2.5 levels. Overall, the hypothesis that increased rates of hospital admissions and visits were positively correlated with amplified PM 2.5 concentrations, was correct. Other geographic areas in which similar studies have been undertaken, such as the Wasatch Valley, Utah, validate this paper's findings.

One result piqued interest and led to further discussion within the study. When analyzing changes in rates over time, change in PM 2.5 composition was identified as a possible reason for increased respiratory illness-associated hospital admission and ED visits, rather than an increase in mass itself. Upon evaluation of the study, authors state that certain observed differences may be due to a misclassification in the degree of exposure and underestimation of PM 2.5 by sectional time periods studied. It is suggested in the paper that future expansion of the study include further investigation of this.

From this paper, knowledge was gained in regard to EPA and regulatory air monitoring in addition to one possible way in which values for PM 2.5 can be designated to patients. This paper also introduced the idea of using classification over regression by categorizing variables.

Effects of Air Pollution and Other Environmental Exposures on Estimates of Severe Influenza Illness, Washington, USA, 2020

(Somayaji, et al. 2020)

By incorporating pollution and various other environmental covariates into ecologic models that are normally excluded from national and global models of influenza disease burden, this study analyzes how such variables may account for associated hospitalizations.

Administrative hospitalization, respiratory virus surveillance and daily environmental exposure data were collected over the time period for 2001-2012 in three counties in Washington state: King, Pierce and Snohomish. Rates for severe respiratory and circulatory hospitalizations were estimated using the variables and a standard CDC negative binomial regression ecologic model in R. These age-specific models were then fitted to daily events in each of the counties with the following covariates: time (day expressed as fraction), daily respiratory and circulatory hospitalizations in a given county on a given day, the corresponding county's population size in the year of incidence, percentage of positive test results for influenza, environmental variables (e.g., PM 2.5, temperature, humidity and dew point) and variables accounting for seasonal trend.

For sensitivity analyses, three alternative models were used for a comparative examination. For the first, a one-day lag was applied when modeling the association between environmental factors and influenza hospitalization. For the next, a linear model was opted for in place of a cubic B-spline. Lastly, an evaluation was performed for the model running weekly aggregates rather than daily events. These three alternative models demonstrated no significant changes to any original results.

Ultimately, results demonstrated enhanced forecasting capacity when integrating environmental variables into the model to predict influenza occurrence but was near-negligible. The authors discuss the reasons behind the improvement of results which involve the increased ability of the virus to survive and spread when optimal temperature and humidity levels have been reached. Furthermore, escalating pollution levels per winter season also indicate higher rates of influenza due to increasing susceptibility of contracting respiratory disease as a consequence of weakened immune response systems.

Albeit that results from this study were not as advantageous as hypothesized, both county level study areas, and the shorter timeframe used, validated that varying scales and durations can be used to study the relationship between meteorological, environmental and health variables.

Air pollution and respiratory diseases: ecological time series, 2016

(Costa Nascimento, et al. 2016)

The objective of this study is to assess how exposure to PM relates to hospitalization rates associated with respiratory disease among residents in the industrious town famed for steel making, Volta Redonda, Brazil. PM 2.5 is central to this study's focus due to it being the most threatening pollutant to the human body's respiratory system out of O_3, CO, SO_2 NO_2 and other volatile organic compounds.

Admission data for all nine hospitals in the region were collated for one year for diagnoses including pneumonia, acute bronchitis and asthma. PM 2.5 concentrations were estimated using an atmospheric data tool that utilizes mathematical modeling and simultaneously accounts for additional variables such as wildfire and traffic to simulate weather and climate. The CCATT-BRAMS model was able to estimate additional values for other meteorological variables including temperature and humidity. The average PM 2.5 level output was compared to acceptable levels delineated by WHO standards.

GAM Poisson regression modeling was implemented in Statistica software where the analysis took place, taking the daily number of hospitalizations as the dependent variable, the PM 2.5 as independent, and adjusting for the remaining meteorological variables as well as seasonality and day of week so that weekend admission could also be accounted for. Cubic smoothing spline functions were applied for temperature and humidity to account for their non-linear nature. A lag of 0-7 days was also applied for inclusion when calculating Pearson's correlation to evaluate the possible relationship between pollution and admission rate.

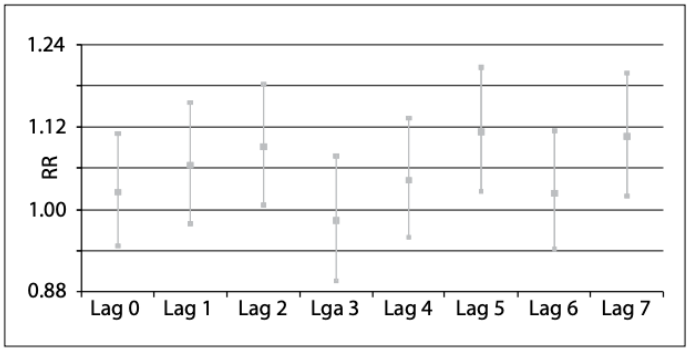


Figure 3 Increase in relative risk per 5 μg/m³ increment in PM, source: (Costa Nascimento, et al. 2016)

The exponent result along with averages, minimums, maximums and standard deviations for number of admissions, PM 2.5 concentration, temperature and relative humidity, demonstrate that a reduction in pollutant levels is connected to a potential decrease in annual hospitalization expenditure. Despite the support shown in favor of the hypothesis, the study concludes by questioning age as a factor that could alter results. All age groups were considered in this study as the hospital data collected did not specify patient date of birth, however, other research indicates that children younger than five years old are more likely to be admitted to hospital for reasons pertaining to pollution. This would be an additional consideration in future expansion of this study.

This study delved deeper into explanations of model quality using statistics that evaluate error, bringing awareness of the importance in data variability, covariance and prediction inaccuracies.

Methodology

Study Area Selection

Several factors led to the selection of Washington state as this project's study area. Firstly, Washington ranks 27th out of the 50 US states in terms of COVID-19 positive case incidents, and 31st in terms of death toll (USAFacts 2021). This places Washington at the approximate halfway mark, with median values in terms of COVID-19 events, suggesting that it would provide average, non-skewed COVID-19 positive case numbers, hospitalizations and death statistics.

Furthermore, based on prior research, the topography and physical geography of Washington is diverse, consisting of coastal regions, rural areas and larger metropolitan cities, with a state-wide elevation ranging from sea-level to 14,411 feet (NETSTATE 2016); other studies have implied that geographical variance can add a new dimension to explorative analysis.

Another important factor in study area selection was data accessibility, transparency and cost. During initial phases of this study, a shortlist of states was compiled and researched in terms of data availability. Although dashboard data for COVID-19 was abundant across many cities, counties, states and on a national scale, raw data was restricted for several reasons. Firstly, the time period of March 2020-March 2021 being so recent, meant that data was not due to be shared online by certain entities until June 2021. Secondly, associated download costs came in at an excess of $1,700 for certain regions. Upon contacting various government agencies to request further information, it became evident that many data analysts working with COVID-19 data were currently non-responsive

to inquires due to the intensity of their workload. Additionally, due to a privatized US health care system, varying privacy regulations obscured certain data in order to prevent over-sharing of detailed patient information, consequently withholding valuable information for this study (e.g., date of hospital admission). Data obtainability could also have been affected by various unknown factors, such as a state's political affiliation, independent laws, socioeconomic vulnerability, etc.

Finally, the first case in the WHO region of the Americas, was reported in mid-January 2020, where a man in his 30s, who had returned from a trip to Wuhan, developed symptoms (Taylor 2021); the confirmed case came from Washington State. As Washington was the primary state to experience COVID-19 presence in North America, it is reasonable to infer that data for cases, hospitalizations and deaths span longer than some other states.

Prior Considerations

There are various technologies that enable measurements of PM 2.5 to be taken: infrared spectroscopy, Beta Attenuation Mass monitoring (BAM) and laser diffraction are all viable methods of pollutant data collection. Infrared can be useful for quick, on-the-spot assessments but lack in ability to determine what particles are causing the result when poor air quality levels occur (kaiterra 2017). On the other hand, BAM monitoring is a Federal Equivalent Method (FEM) for measuring ambient air pollutant concentration levels and is deemed reliable, accurate and in compliance with government regulations as a result. BAM monitors use filter paper to trap pollutant particles prior to exposing them to short bursts of beta radiation. As pollutants absorb the radiation on one side, the other side can then be measured, and calculations translate the difference into a mass measurement (kaiterra 2017). When lasers, commonly found in smaller, professional hand-held devices, are well-calibrated, measurement results have been reported to be just as accurate as the technology used inside a BAM monitor. Laser beams use sensors that analyze light scatter intensity and angle (Malvern Panalytical 2021) which, is then run through an internal algorithm that computes size and number of particles present in air or water (kaiterra 2017). In speaking with Matthew Harper, Air Monitoring Team Lead for Puget Sound Clean Air Agency, EPA air quality sensors are reliable not only in terms of equipment and technologies used, but also in regard to regular maintenance, service and quality control; this eliminates the issue of bias in readings.

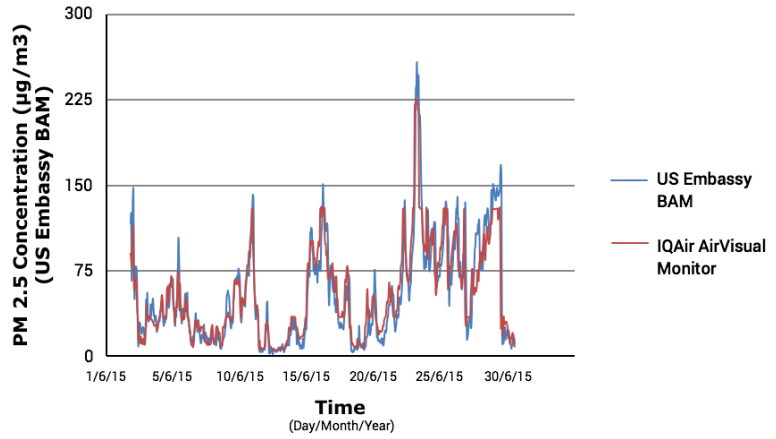


Figure 4 An example of BAM versus laser reading results, source: (IQAir 2021)

For this study and the investigation into potential connection between COVID-19 events and PM 2.5 on a state level (per county), air quality and pollutant level data was therefore extracted from the EPA Air Quality System (AQS) database. Approximately 70 EPA, state, local and tribal agency sensors provided daily PM 2.5 concentration level data (EPA 2020), all of which met FEM criteria.

However, when performing analysis at a higher spatial resolution and focusing on zip code level, there was a significantly lower availability of EPA approved air sensors to obtain information from. PurpleAir, a national scale air quality monitoring network, provided ample data but their sensors differ in two ways to that of regulatory PM sensors. Firstly, PM 2.5 mass is calculated from laser particle counters within the PurpleAir sensor system based off particle count and using PM

average density (PurpleAir 2021). Occasionally, this can result in higher-than-average readings. Secondly, rather than computing daily PM 2.5 as an average of hourly readings, PurpleAir sensors report PM 2.5 levels at every two-minute interval which becomes averaged out over the course of 24 hours. Due to large fluctuations in pollutant levels throughout the day, this can also contribute to higher-than-expected readings. Despite these differences, PurpleAir sensor data has been used government agencies and still adheres to the use of AQI scale and color legend.

As this was pre-emptively discussed with certain professionals prior to this study, the reliability of EPA versus VGI air quality data was incorporated into *Project Goals and Objectives*. However, had this project anticipated combining EPA and VGI readings within a singular study area, it would have been important to decipher whether the difference in sensors would have caused bias in results.

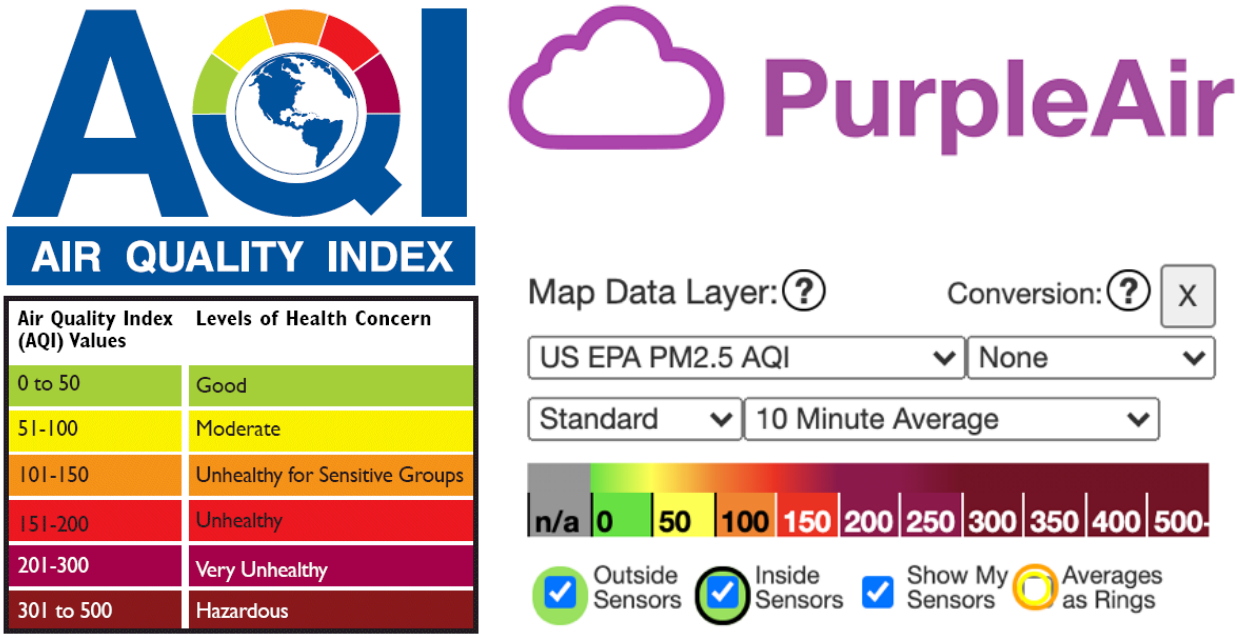


Figure 5 AQI and PurpleAir scales and legends, source: (Brotsky 2014), (PurpleAir 2021)

Workflow

A general workflow method was followed; a simplified version is shown in Figure 6.

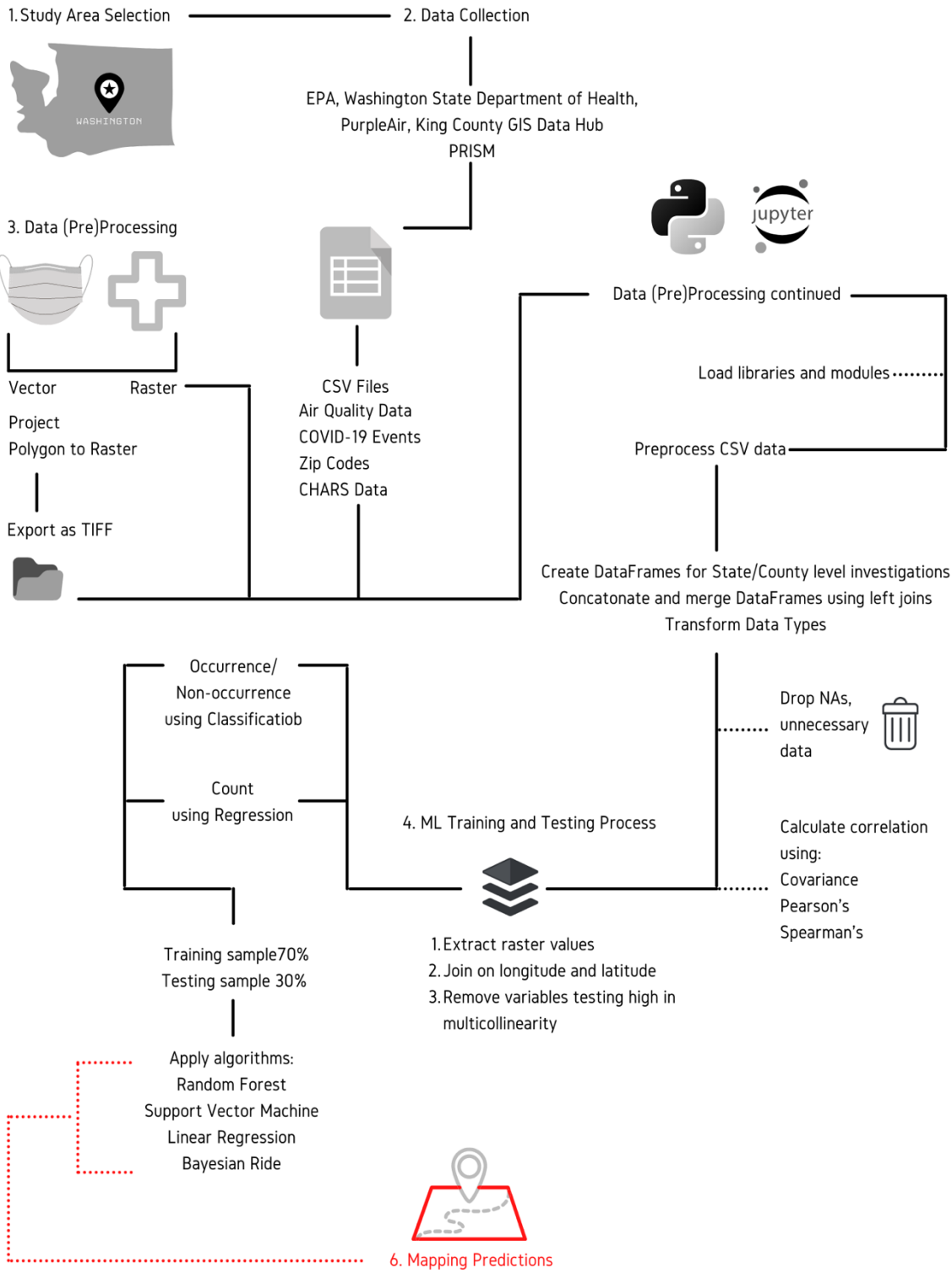


Figure 6 Workflow diagram

Software

The main software programs used for this project consisted of ArcGIS Desktop program ArcMap and Python Integrated Development Environment (IDE), Jupyter Notebook. These were both selected due to license coding availability and cost. ArcGIS Desktop software was downloaded directly from the ESRI user login portal and installed to a Microsoft Windows environment. Jupyter Notebook was launched from Anaconda Navigator which was downloaded as a 64-Bit Graphical Installer from its parent product webpage.

ArcGIS Desktop

ArcGIS Desktop is a Geographic Information System (GIS) software owned by ESRI and designed to manage data, enable spatial analysis and produce cartographic visualizations. The software is comprised of several programs including ArcMap, ArcPro and ArcScene. This desktop software permitted for the initial visual investigation of geographic data, the preprocessing of certain raster data, automation of workflows for repetitive tasks via the ModelBuilder tool and for the creation and export of any related maps.

Jupyter Notebook

Jupyter Notebook was used for almost all data visualization and manipulation and coding of ML algorithms along with result analysis in Python language. Jupyter Notebook's user-friendly interface, as well as its capacity to instantly visualize data and explore command documentation via shortcut use, was also reasoning for its selection over command line use. This open-source software was downloaded via the installation of Anaconda Navigator, free of charge.

Data

To explore the potential impacts to COVID-19 patients as a result of PM 2.5 levels (e.g., positive case count, hospitalizations, deaths, etc.) data was collected concerning COVID-19 incidents both on county and zip code levels. Air quality and pollutant level was also collected in various forms. PM 2.5 concentration was acquired from EPA monitoring sites for a state-level analysis, from VGI monitoring sites for zip code level assessment and as a shapefile for averaged readings per census tract. Additionally, for a more in-depth investigation and comparison of pollutants, information at state level was gathered for proximities to waste facilities and heavy traffic, and weighted concentrations of toxic release from facilities to air. Ethnicity, employment, English competency level and vulnerability to health disorder data was also compiled in order to analyze whether other factors played a significantly heavier role in the intensity of COVID-19 events. From data collected in shapefile (SHP) format, a raster files were generated based off of certain environmental rankings or percentages to represent population susceptibility to COVID-19. An overview of all data that was downloaded is listed in Table 2.

Table 2 List of collected data, type and source

No.	Data (Official Name)	Type, Unit	Format	Source
1.	Outdoor Air Quality Data: PM 2.5	Int*, $\mu g/m^3$ LC	CSV*	EPA
2.	WA COVID-19 Cases, Hospitalizations, Deaths (*weekly*)	Int	CSV	Washington State Department of Health

3.	PurpleAir Sensor Data	Int, µg/m^3 LC	CSV	PurpleAir
4.	Daily Counts and Rates by Zip Code	Int	CSV	King County GIS Data Hub
5.	CHARS Public Use Data File 2020	Int	CSV	Washington State Department of Health
6.	All Zip Codes and PO Box as Centroids for King County	Int	SHP*	King County GIS Data Hub
7.	Proximity to Hazardous Waste TSDFs*	Int, Kilometers (km)	SHP	Washington Geospatial Open Data Portal
8.	Populations near Heavy Traffic Roadways	Int	SHP	Washington Geospatial Open Data Portal
9.	Toxic Releases from Facilities (RSEI* Model)	Int, weighted concentration	SHP	Washington Geospatial Open Data Portal
10.	Proximity to Wastewater Discharge	Meters (m)	SHP	Washington Geospatial Open Data Portal
11.	People of Color	Str*	SHP	Washington Geospatial Open Data Portal
12.	Unemployed Population	Str	SHP	Washington Geospatial Open Data Portal
13.	EHD* Sensitive Populations (Theme Ranking)	Int	SHP	Washington Geospatial Open Data Portal
14.	Limited English	Str	SHP	Washington Geospatial Open Data Portal
15.	Temperature (*4km*)	Int, Celsius (^{o}C)	Raster, ASCII*	PRISM*
16.	PM 2.5 Concentration	Int, µg/m^3 LC	SHP	Washington Geospatial Open Data Portal

*Int *integer*, CSV *Comma-Separated Values*, SHP *shapefile*, TSDFs *Treatment, Storage and Disposal Facilities*, RSEI *Risk-Screening Environmental Indicators*, Str *string*, EHD *Environmental Health Disparity*, ASCII *American Standard Code for Information Interchange*, PRISM *Parameter-elevation Regressions on Independent Slopes Model Climate Group*

Data Collection

The EPA Outdoor Air Quality Data portal enables users to download daily data based on various fields including pollutant type, year, geographic area (e.g., state, county or city) and monitor site. The tool queries all data summary statistics and allows for download of selected criteria in Comma-Separated Values (CSV) format. From this portal, PM 2.5 data was downloaded for all 70 EPA approved air monitoring sites located throughout Washington state. This averaged a PM 2.5 reading per an approximate 1,000 square miles. Despite the ability of PM 2.5 to travel distances ranging hundreds of miles, research indicates that it would only travel further in the lower troposphere when a cold surge hits (Wang, et al. 2017). As it is not entirely unusual to experience weather below 10^{o}C at low altitude in Washington state, and to avoid compromising data validity, data was left as is and was not added to with additional non-EPA approved readings.

Weekly aggregated COVID-19 positive case incidences, hospitalization counts, and death tolls were downloaded from the Washington State Department of Health COVID-19 Data Dashboard webpage. This dataset dates back to the earliest specimen collection date, is updated on a weekly basis, and, for each event type, counts are broken down by county and age group. Cases include both probable (positive antigen test result obtained) and confirmed (positive molecular test result obtained) instances. Hospitalizations are defined as a Washington state resident reported in the

Washington Disease Reporting System (WDRS) or Rapid Health Information Network (RHINO) as having been admitted to hospital with confirmed or probable COVID-19 (Washington State Department of Health 2021). Deaths come from the Washington Health and Life Events Systems official vital records database and have been reported by health care providers and departments, medical examiners and coroners (Washington State Department of Health 2021). Both rates and counts are provided for each event; for this study, counts were selected as the dependent variables. Counts took precedence due to rates accounting for time lags and population estimates which, would differ from other study areas and may impact results of the ML model if implemented in a different time-space in the future. The data was downloaded as a collection of three excel spreadsheets which were then each exported separately as their own CSV file.

King County was deemed an appropriate study area to further investigate the relationship between PM 2.5 and COVID-19 at a higher spatial resolution. This was to provide insight into result consistency and comparison between EPA and non-EPA air quality data. A zip code shapefile was downloaded from King County GIS Open Data hub. For each of the 85 zip codes, a manual search was performed on the PurpleAir online map. Each location was visually inspected with respect to the corresponding boundaries and centroid points, reflected in the shapefile, to ensure accuracy of the website's location mapping. Outside sensors with the earliest initialized PM 2.5 reading dates, that lay within the zip code boundary, were selected for download. Each PurpleAir sensor has two channels, A and B, for which each possesses a primary and secondary dataset. The primary data contains information on pollutant type levels measured in $\mu g/m^3$ (particle mass concentration), whereas the secondary reports on particle count (Public Lab 2021). As Channel B essentially acts as a backup of Channel A, only Channel A primary data was necessary to download per sensor for this study. Due to the migration away from direct data access via the Thingspeak Application Programming Interface (API) to a new one, the simplest way to collect this data was through the website. When doing so, the date search parameters were set to March 1st, 2020 to March 2nd, 2021 with a requested PM 2.5 value averaged over 1440 minutes (24 hours).

To match the higher level of PM 2.5 reporting, King County specific data for COVID-19 incidents were also downloaded. Daily counts and rates for positive cases, hospitalizations and deaths were organized by zip code. Factors to consider in regard to this dataset are addressed on the county website, with location being the most common of all errors. Precautions are taken to avoid this prior to data release by cross-referencing results with those from hospitals, geographic systems and other databases (King County 2021). According to King County's Daily COVID-19 outbreak summary webpage, these errors are usually fixed within a matter of days and if a location field remains blank, it is filled with the zip code from the location where testing occurred. This data followed the same procedure as the Washington State Department of Health data, whereby it was downloaded as a whole prior to being exported as individual CSV files.

In order to conduct a comparative analysis of PM 2.5 and other variables to see which carry more weight in terms of impact on COVID-19 events, several other datasets were downloaded.

Proximity to Hazardous Waste Treatment, Storage and Disposal Facilities (TSDFs), Populations near Heavy Traffic, Toxic Releases from Facilities and Proximity to Wastewater Discharge were all downloaded to inspect whether PM 2.5 ambient air pollution held more weight in impacting COVID-19 events in comparison to particular zones where high-volume pollution release occurs.

People of Color data represents the sum of ethnic groups and races with the exception of those identifying as white. The categories of this dataset are: Black, American Indian/Alaskan Native, Asian, Native Hawaiian-Other Pacific Islander, Two or more races and Spanish/Hispanic/Latino. Research reports that several ethnic minorities experience COVID-19 contraction rates between 4.7-5.3 times higher than that of non-Hispanic white populations (William F. Marshall 2020). Therefore, this data could be helpful in discerning what may be causing increased rates of COVID-19 events in

areas should PM 2.5 levels be low, or, prompt further investigation if levels read high, ensuring accurate analysis.

Unemployed Population data was included as a comparative variable due to studies showing an increased risk of COVID-19 contraction in the workplace. Should unemployment rates be low in certain regions, it could imply more movement and human-human interactions at work, in turn suggesting the reason for higher rates of COVID-19 positive case results, hospitalizations and deaths could be based on behavior rather than environmental factors.

The Limited English shapefile was also considered for inclusion for various reasons. Firstly, populations who speak English less than 'very well' (Washington State Department of Health 2021) usually represent those who have limited access to healthcare. Consequently, if an area of high PM 2.5 readings had considerably lower than expected COVID-19 event occurrences, this data could provide an explanation as to why. Conversely, should a spike of COVID-19 incidents occur in an area of high limited English proficiency ranking, it could be associated with the inaccessibility to pertinent information due to lack of understanding, rather than the hypothesized impact of pollution being the primary contributing factor.

All aforementioned shapefiles were downloaded from the Washington Geospatial Open Data Portal for the entirety of the State's area so that they could be examined at both state and county-levels.

Daily temperature (°F) and humidity (%) were both included within PurpleAir's Channel A primary data. This meant that temperature could be extracted as its own variable when analyzing correlation among others. As this was the case for the county-level study, monthly average temperature raster files in American Standard Code for Information Interchange (ASCII) format were downloaded from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) website, to be processed and used for state-level assessment.

The final dataset collected was provided by CHARS and downloaded from the Washington State Department of Health. This publicly accessible database provides deidentified hospital patient discharge information that has been collected and preprocessed by public and private hospital billing systems accordingly throughout the State. This dataset was downloaded as a CSV file and included information on zip code, age, gender, duration of hospital stay, diagnostic code and associated charges. To upkeep patient privacy standards, no specificities regarding date were provided, only a year was given. Although this data is difficult to use in conjunction with weekly aggregated data due to its lack of high temporal resolution, it was kept for potential regression analyses at a later date.

Data Processing

Much of the data pre-processing was performed in Python code and for which walk-through descriptions are included in section *Implementation* under *Python Code 3 Output*. Pre-processing in Python occurred for both Washington state and King County level data corresponding to COVID-19 events and air quality data.

For each Purple Air PM 2.5 CSV that was downloaded, an additional field labeled Zip was inserted and auto populated with a repeating value to identify the location of the area using excel. Upon manual inspection of the data as downloads took place, CSVs for zip codes 98034, 98056, 98057, 98178, 98195 and 98354 either contained virtually none, or no data at all and were dropped from inclusion in the study. These areas were compared to online maps which demonstrated their smaller, less residential extents. Therefore, they were considered unwarranted in use, especially as over 93% of the remaining files contained robust enough information to provide sufficient insight in terms of results.

For the shapefiles Proximity to Hazardous Waste TSDFs, Populations Near Heavy Traffic Roadways, Toxic Releases from Facilities Risk-Screening Environmental Indicators (RSEI Model), Proximity to Wastewater, People of Color, Unemployed Population, Environmental Health Disparity

(EHD) Sensitive Populations (Theme Ranking) and Limited English, ArcMap was used to convert the Washington state level data to a Tag Image File Format (TIFF) using the Polygon to Raster tool. Prior to this, for the shapefiles People of Color, Unemployed Population and Limited English, a field was added to the attribute table as type 'Float.' This was then populated with the corresponding values from columns representing the percentages of population. This ensured that raster files could be generated on the correct value, as percentages were originally data type 'String.' Each shapefile was then used as its own input and value fields were selected according to Table 3. The cell assignment value was selected to be `CELL_CENTER` and cell size remained as the default (0.014). Each raster was checked to ensure it possessed a Geographic Coordinate System (GCS) of `WGS_1984`, equivalent to European Petroleum Survey Group (EPSG) 4326; this was used for two reasons. Firstly, a GCS was selected over a Projected Coordinate System (PCS) as data occurrence location took priority over how to map the data while in the Python IDE. Secondly, due to the size of Washington state and depending on datum preferences and units of measure, up to 14 EPSG reference options are given. With a high spatial resolution and accuracy, and ability to be used for other study areas in case of project expansion, `WGS_1984` remained the GCS of choice.

Table 3 Polygon to raster value fields

Shapefile	Value Field
Proximity to Hazardous Waste TSDFs	EHD Rank*
Populations Near Heavy Traffic Roadways	EHD Rank*
Toxic Releases from Facilities (RSEI Model)	EHD Rank*
Proximity to Wastewater	EHD Rank*
People of Color	Percent People of Color
Unemployed Population	Percent Unemployed Population
EHD Sensitive Populations (Theme Ranking)	Environmental Sensitive Populations Rank
Limited English	Percent Limited English

* EHD ranks, based on distance and concentration, provided standardized readings for these variables

Two temperature raster files in TIFF format were additionally generated from downloaded PRISM data. This data included climate variable mean temperature per month for the period March 2020 through March 2021 and was collected in ASCII format. Both new files were created using the Mosaic to New Raster tool in ArcMap; the raster layers representative of singular months were each loaded as an input raster and assigned a spatial reference of `GCS_WGS_1984` and pixel type `32_BIT_FLOAT`. The number of bands was designated as one while the mean was prescribed as the mosaic operator. One newly produced raster spanned the duration March 2020 through December 2020, while the other covered March 2020 through March 2021. Each timeframe was selected to align with the temporal resolution of datasets being used for each the state and county level analyses and were used accordingly.

Implementation

Programming Language

Python, a powerful programming language, is consistently used across a broad spectrum of applications including web development and database access. Through understandable syntax, vast

amounts of data can be processed, extracted for analysis, merged on similar identification values and used to train ML algorithms to predict results for both scientific and mathematical investigation. A diversified selection of libraries and modules serve to meet the needs of different user groups, most of which are open source, as is the programming code itself. Python was opted for the programming language of choice for this study for various reasons: Python documentation is both extensive and detailed; the language allows for flexibility in execution of tasks; final code can be easily committed to, and expanded upon, through GitHub and community users; and, it has been steadily and dependably used by healthcare industry professionals indicating that this project could be effortlessly shared and interpreted by individuals who this study is relevant to.

Table 4 Python libraries and task capabilities

Library	Description	Project purpose
OS	Setting operating system path	Read in working directory
Pandas, pd	Data manipulation/analysis	Call on/manipulate dataframes
Geopandas, gpd	Geospatial operations	Read in files
Numpy, np	Mathematical function	Data location within dataframe, calculate averages
Datetime	Date/time manipulation	Convert date and time formats
Seaborn, sns	Statistical data visualization	Visualize data in form of heatmaps, scatterplots, histograms, etc.
Matplotlib, plt	Data plotting	Graph and map plots
Plotly	Interactive data visualization	Visualize data in form of interactive bubble plots, scatter plots, etc.
Cufflinks, cf	Connection gateway between Plotly and Pandas	Assist with Plotly tasks
Rasterio	Read/format GIS files	Opening raster files
Glob	Bulk file retrieval	Retrieve multiple same-format files from directory location
Folium	Interactive geospatial data mapping	Map data layers
SKlearn	Machine learning application, data analysis	ML algorithm model implementation, optimization and metric evaluation

Supervised Machine Learning and Algorithms

Regression modeling utilizes a mapping function to predict a continuous value based on independent input variables (Brownlee, Difference Between Classification and Regression in Machine Learning 2019). The output is numeric and often representative of a price, size or quantity. This statistical approach, used in conjunction with an IDE and coding language, allows for the quantitative analysis and visualization of correlation between variables. As part of this study investigates the relationship between positive case rates of COVID-19 and associated hospitalizations, deaths and potentially recovery times, it is a viable option to use for predicting values for any such group. A regression models' reliability is assessed as an error in its predictive capability, most commonly by calculating the Root Mean Squared Error (RMSE) (Brownlee, Difference Between Classification and Regression in Machine Learning 2019) which, is often helpfully denoted in the same unit as the value being predicted.

In some instances, classification can be used in place of regression if a group of values is converted to categorical variable through discretization (e.g., the number of observations is

converted to an occurrence/non-occurrence class). This type of modeling is often used to solve for discrete value classifications (Brownlee, Difference Between Classification and Regression in Machine Learning 2019) that belong to two (binary) or more (multi-class) classes. For this reason, it can also be used in this study for a simplified prediction of whether or not COVID-19 cases, hospitalizations or deaths will occur and at what intensity based on PM 2.5 values. The proficiency of a classification model is determined by accuracy which takes into account all true and false positive and negative predictions. Nonparametric classification results can be examined by generating a Confusion Matrix (CM) which not only allows the user to examine accuracy, but also additional performance capabilities of the algorithm, such as precision.

As data is provided prior to the execution of ML algorithms, this study will be taking a supervised learning approach. Inspection of the algorithm's parameters based on the training data, facilitates the search of statistical similarities which, ultimately assist in final prediction of values.

The ML algorithms used in this study are described in Table 5.

Table 5 ML algorithms

Type	Algorithm	Description
Classification, Regression	Random Forest (RF)	An ensemble of Decision Trees (DT) where top-down division enables label assignment (Donges 2019). A higher rate of randomness can be applied by adjusting threshold levels, increasing model reliability even though the parameters are similar to that of a DT algorithm.
Classification, Regression	Support Vector Machine (SVM)	A search for best fitting hyperplanes in *n*-dimensional space (where *n*=number of features) (Gandhi 2018). Low computational power still produces high levels of accuracy.
Regression	Linear Regression (LM)	Also referred to as Ordinary Least Squares Regression. Uses coefficients to fit a model, predicting targets by using linear approximation (scikit-learn 2021). Some drawbacks include poor performance in the presence of outliers and depend heavily on linear relationship of data.
Classification, Regression	Bayesian Ridge (BR)	Addresses some of the pitfalls of LM by imposing penalties based on the size of coefficients. This model adapts well when data may be insufficient or when a lack of linearity in data distribution occurs (Bora 2020).

Python 3 Code Output

Washington State Level Analysis

After loading in appropriate libraries and modules and setting the working directory, the Python environment was successfully prepared to run an investigation of correlation between PM 2.5 and COVID-19 events, on a state-level scale. imposing

PM 2.5 daily concentration readings per county were loaded into the Jupyter notebook as a dataframe from a CSV file where data cleaning then took place.

AQS parameters were checked for acceptable statuses to ensure data quality and integrity according to EPA standards. Codes for Core Based Statistical Areas (CBSA) were converted from unique values to a numeric where micropolitan was represented by '1,' metropolitan was expressed

as '2' and other areas (including rural) were identified as '0.' This enabled an area's designation to be assessed in terms of correlation to PM 2.5 and COVID-19 occurrences at a later point. After review of the dataframe, unnecessary columns were deleted leaving Date, Daily Mean PM 2.5 Concentrations, Daily AQI Value representative of general air quality levels, CBSA conversions, County name and Longitude and Latitude.

Data types were transformed appropriately (e.g., date transformation from object to datetime, etc.) to prevent errors from occurring when later aggregating data from daily to weekly averages; duplicate columns that resulted from this were dropped. Dates were set to match available COVID-19 case, hospitalization and death data for Washington state which, ran beginning of March 2020 through the end of December 2020. Additionally, PM 2.5 readings were aggregated into weekly counts to match the arrangement of COVID-19 incidents using the `.groupby` and `.resample` functions. This enabled the user to maintain the categorical variable 'County' during the process rather than losing it.

Using `lambda`, the county names were combined with the date to create a unique key on which other data possessing the same information could be joined on later. For each of the COVID-19 case, hospitalization and death CSVs read in as single dataframes, the county name strings were stripped of 'County,' unnecessary columns and start dates listed as 'Unknown' were removed. Remaining dates were converted to data type datetime and age group segmentations were deleted, only keeping total counts of COVID-19 incidents. Having performed these tasks, identical unique keys, `id_value`, were created and all three COVID-19 dataframes were merged with the PM 2.5 data using a left join. After the join, the `id_value` was removed and `NaN` values for event counts were assigned '0's.

A time lag of one week was applied for each COVID-19 event, resulting in nulls for first week counts within the dataframe due to data carry-over, therefore leading to its elimination. A search for any remaining `NaN` values was performed; two rows were dropped.

A quick check for correlation between all current variables using the seaborn library heatmap was employed and visually inspected. COVID-19 events inclusive of lag times appeared to hold slightly higher correlation with variables, suggesting that lag times approximating seven days were more significant than consideration of events without a lag time applied.

After initial assessment, a geometry was created for the dataframe based on latitude and longitude values, and a projection of EPSG:4326 was applied. Nine raster files were then read in for which values were extracted to new columns created in the dataframe based on location of COVID-19 events. Following this, another heatmap was plotted to further examine the relationship between variables so that those highly correlated to one another could be dropped. Limited English as a variable was both removed due to its strong connection to the data representing ethnicity. Unemployment appeared insignificant in terms of correlation to COVID-19 cases. Furthermore, this variable had already been scrutinized in previous research as having a possible bidirectional connection to COVID-19 and was removed in part to prevent confusion around which variable of the two causes an effect on the other. Severity of AQI level was also removed in favor of PM 2.5, as was Wastewater Discharge due to its strong correlation with Proximity to Hazardous Waste TSDFs. The final variable to be eliminated was Populations near Heavy Traffic; Toxic Releases from Facilities was kept. When refreshing the heatmap, it appeared as though a lack of correlation existed between PM 2.5 and COVID-19 cases with a one-week lag. This was confirmed by a scatterplot visual along with additional statistics including covariance, Pearson's correlation coefficient and Spearman's rank-order correlation (Table 6).

Covariance is the average of the difference between two variables, which can be calculated as

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\mathrm{N} - 1}$$

This result enables evaluation of whether two variables move in the same or opposing directions and can be ranked in a standardized or non-standardized way, where '0' shows no correlation between them. In ML, a covariance matrix can be produced for review but can be difficult to interpret without other supporting statistics (Brownlee, How to Calculate Correlation Between Variables in Python 2018). Covariance for this part of the study measured low, showing little correlation between the two variables of interest.

Pearson's correlation coefficient is another assessment of the strength of a linear correlation and is one of the most popular among mathematical studies and investigations. It is calculated as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})(y_i - \bar{y})}}$$

Due to the use of mean and standard deviation, as with covariance, this equation assumes a normal-like distribution (Brownlee, How to Calculate Correlation Between Variables in Python 2018). This correlation produces values between -1 and 1. Negative and positive correlation between the variables is represented best at or below -0.5 and at or above 0.5, respectively. The result between PM 2.5 and COVID-19 incidents is negligible for this part of the study as seen from the result -0.015.

Spearman's rank-order correlation considers non-linear relationships between variables and is calculated as

$$\rho = 1 - \frac{6\Sigma d_i^2}{\mathrm{n}(n^2 - 1)}$$

This statistic is also scored on a scale of -1 to 1, following the same rationale in terms of negative or positive correlation as Pearson's correlation coefficient. This is likely the most meaningly statistic for this part of the study due to the unknown nature of the relationship between the two variables in question. Although the resulting value is low, it does convey a small chance of a positive relationship between PM 2. 5 and COVID-19 positive case events with a one-week lag.

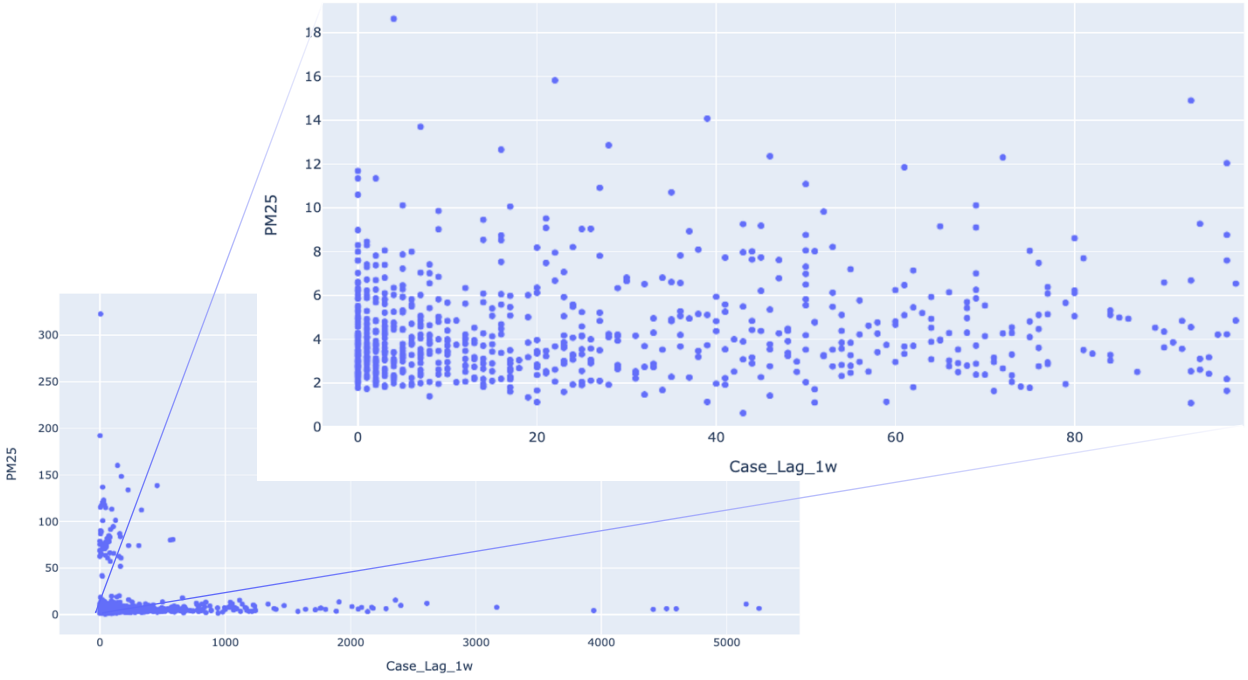


Figure 7 A scatterplot of correlation between PM 2.5 and COVID-19 case count with a one-week time lag, Washington state

Table 6 Statistical results for COVID-19 incidents with one week lag and PM 2.5 at state-level

Data	Statistic	Value
Case_Lag_1w (Positive cases with one week lag), PM 2.5	Covariance	[[2.1444817e+05 1.49552435e+02] [-1.49552435e+02 4.68984381e+02]]
Case_Lag_1w, PM 2.5	Pearson's correlation coefficient	-0.015
Case_Lag_1w, PM 2.5	Spearman's rank order correlation	0.287

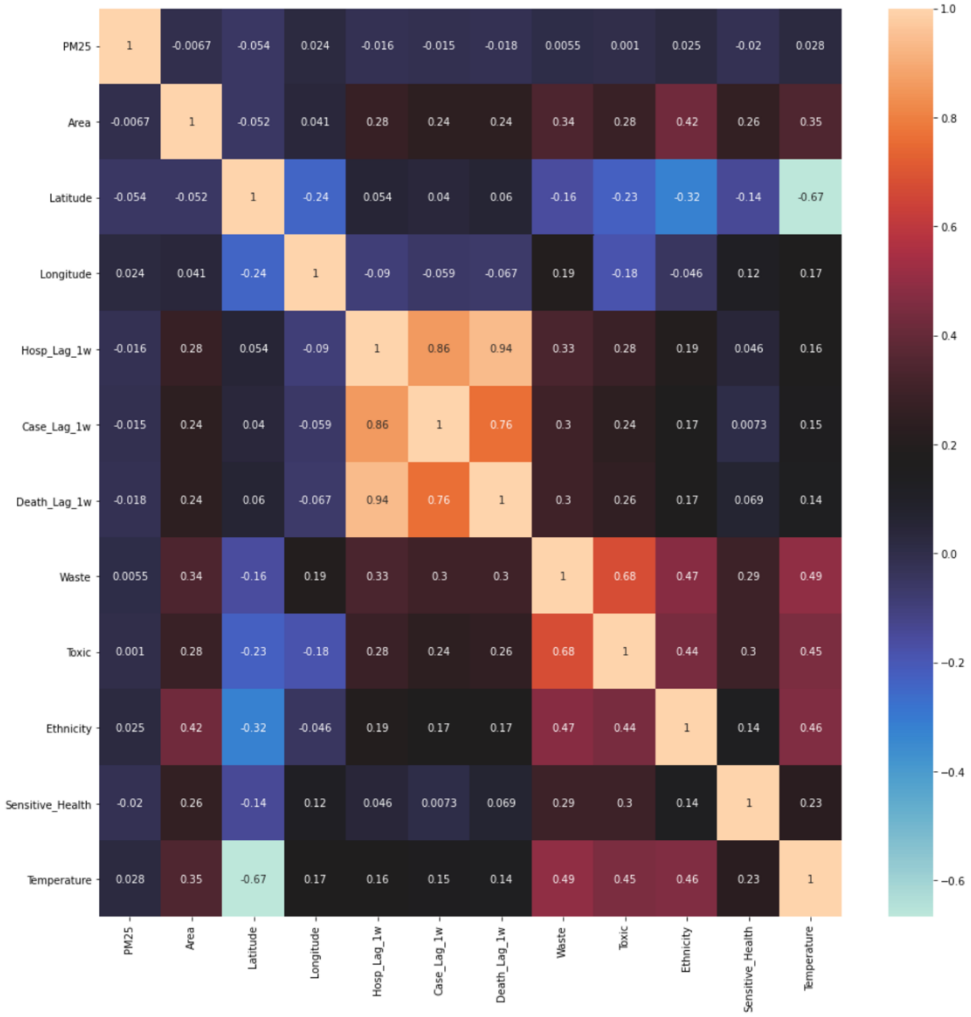


Figure 8 Heatmap showing correlation between selected variables at state-level

Ultimately, reviewing the results of the time-series analysis on a statewide level showed insignificant correlation between COVID-19 events and PM 2.5 levels as hypothesized. Due to this, a large-scale area was selected for further explorative analysis.

King County Level Analysis

The same approach for set up and reading in of files was taken for the correlative analysis of PM 2.5 and COVID-19 events on the King County zip code level.

All PurpleAir Channel A Primary data per zip code were loaded and concatenated into one dataframe. The date column values were converted to datetime data type after stripping 'UTC' from the original strings. Based on the new date format and the zip code, a unique key named `id_value`, identical to that of the county dataframe, was generated per row. All data was inspected for null values which, were deleted along with the removal of unnecessary columns. Date, PM 2.5 readings, temperature values and humidity percentages remained intact.

COVID-19 cases, hospitalizations and deaths, based on zip code, were individually read in to Jupyter notebook and underwent data type transformations as appropriate, in addition to being assigned a corresponding `id_value` key based on matching dataframe column values. Time lags of 24 hours, 48 hours and 7 days were created for later comparative analysis before merging via left joins on to the PM 2.5 data using alike keys.

Data spanning a slightly longer time period for both PM 2.5 levels and COVID-19 events permitted for a more extensive project date range. The timeframe ran from March 2nd, 2020 through March 1st, 2021, generating an entire year's worth of data. As time lags were made previous to date selection, no null values occurred as a result.

Finally, the `id_value` and city name columns were dropped.

A geometry was added to the dataframe, again based on latitude and longitude, in order to extract raster values as before from all of the same files but temperature, which, was already accounted for in PurpleAir's PM 2.5 datasets. Upon plotting and inspection of the preliminary heatmap, it could be deduced that immediate count reports and those where a time lag had been applied, were fairly comparative to one another. With this being the case, COVID-19 events with a one-week time lag were selected as final incident data. This allowed for a more reasonable comparison of results between state and county level analysis, as Washington state results used weekly aggregated data with a one-week time lag also. As a result, COVID-19 count data that had been generated with a 24- or 48-hour time lag were removed from the dataframe, but original, immediate, non-lag counts were kept for further review at a later stage. Along with Populations near Heavy Traffic, Proximity to Wastewater Discharge, Unemployed Population and Limited English data, PM 10 and PM 1.0 columns were also dropped. The same conclusions made from the heatmap in regard to highly correlated variables, continued to show consistency between datasets and existing correlation results.

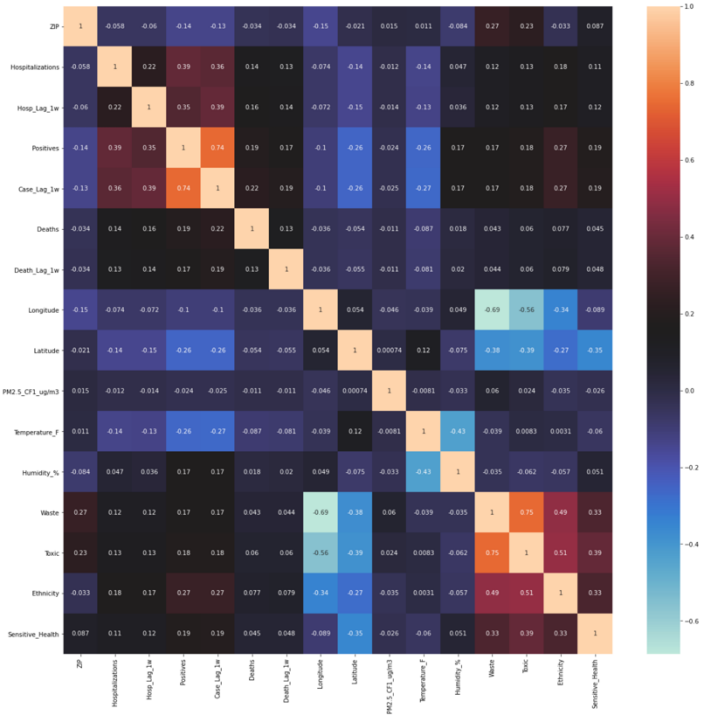
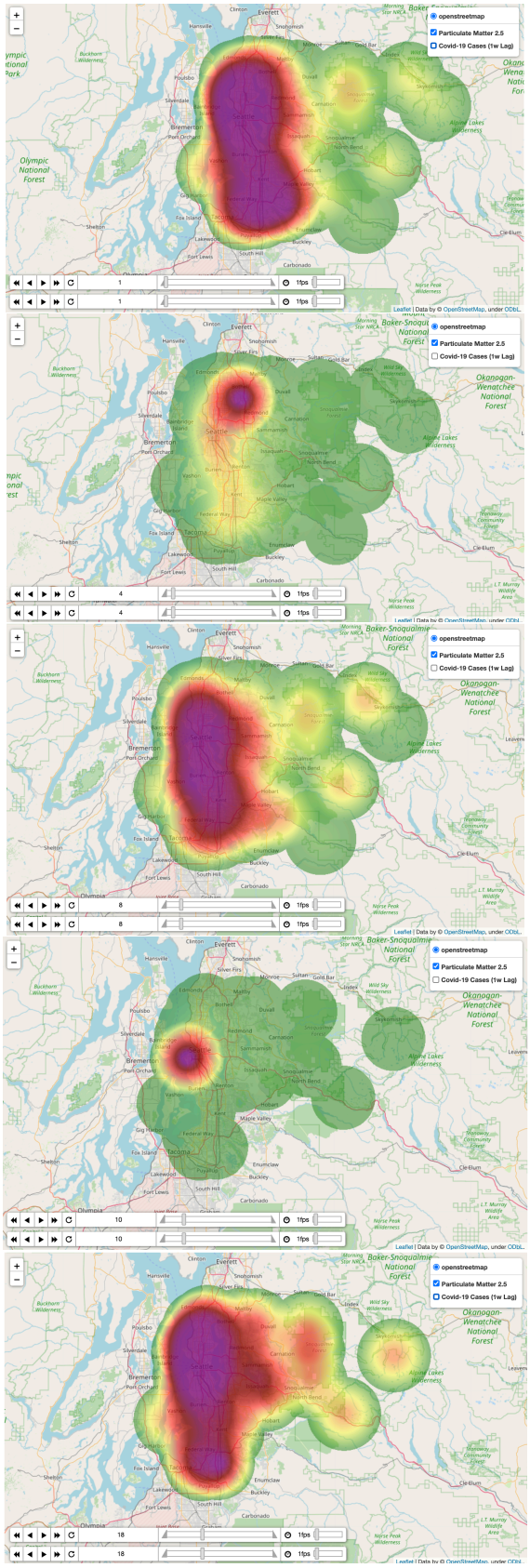


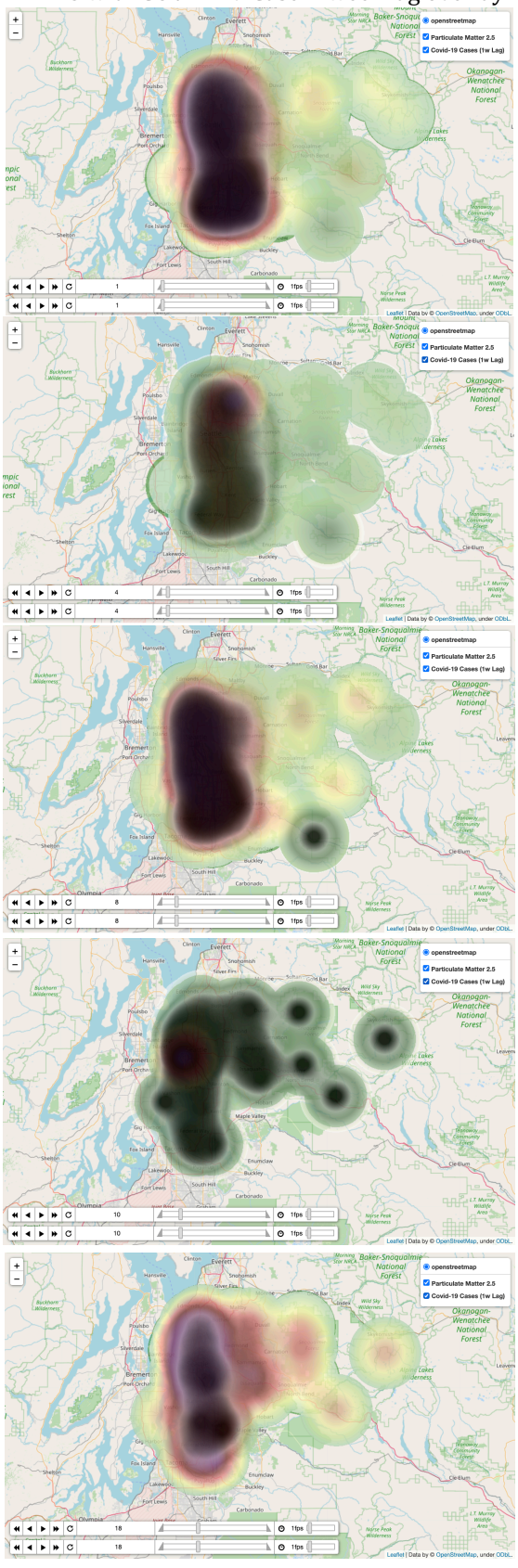
Figure 9 Heatmap showing correlation between selected variables for King County by zip code

After an updated heatmap was generated, further investigation between COVID-19 case count with a one-week time lag and PM 2.5 intensity ensued. Each variable's minimum and maximum values were printed to determine how best to represent legend breaks when using `folium` to visualize time-series correlation between the two. Data was aggregated into weekly averages using `lambda` within the `folium` mapping function. Each variable could be switched on and off on top of the Open Street Map (OSM) base map to inspect the data's spatial patterns. A varied selection of results, from ten out of the 52-week study timeframe, can be seen in Figure 10.

PM 2.5



PM 2.5 with COVID-19 Case 1-week lag overlay



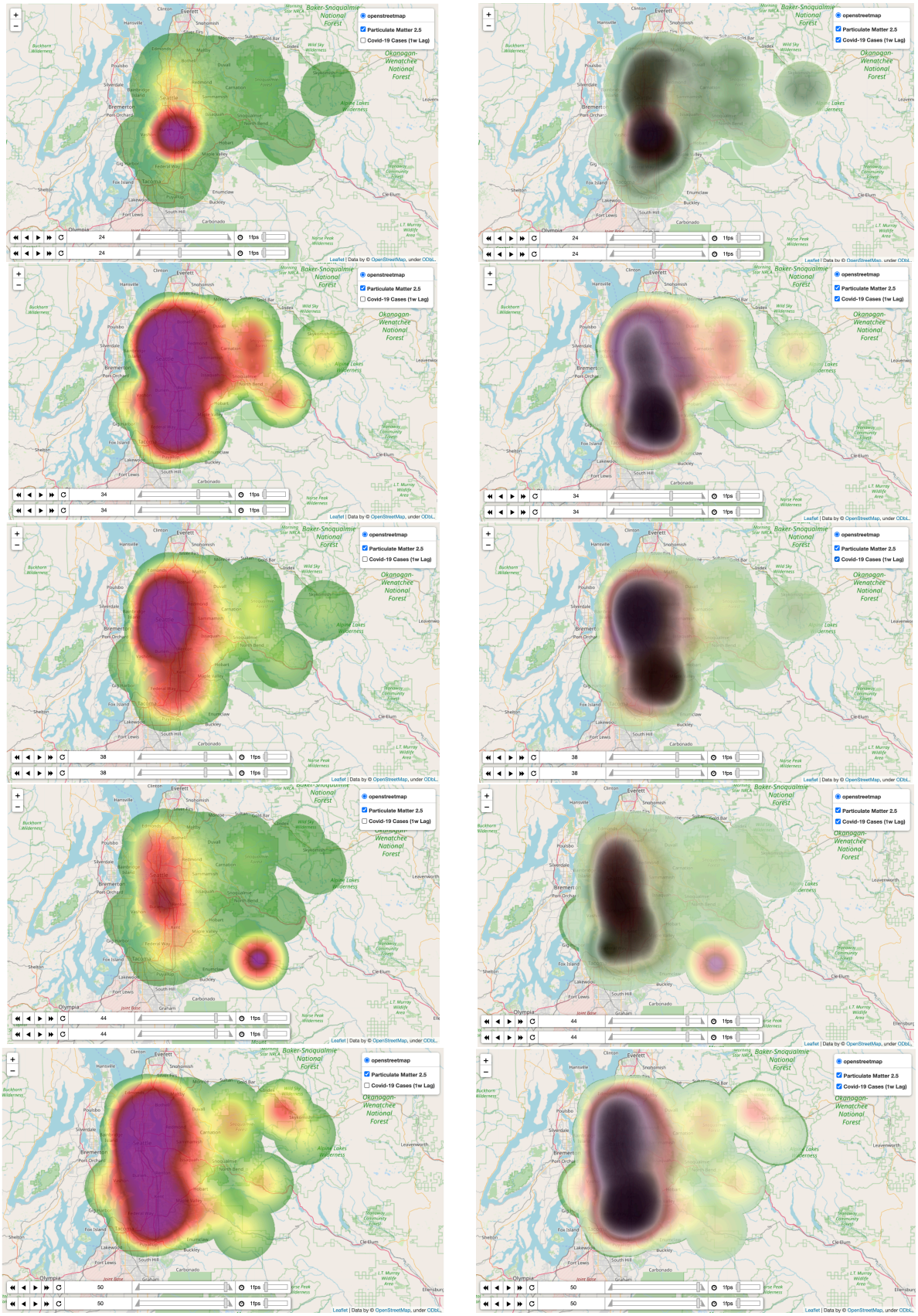


Figure 10 Mapped PM 2.5 levels with COVID-19 incidence overlay using folium, King County

Between heatmapping, visualizing spatial patterns using `folium` and a final review of statistical outputs, EPA air quality data on a state-level, and VGI PurpleAir data on a county-level showed similar results. Closely related values imply that both datasets at both levels of investigation communicate consistent findings.

Table 7 Statistical results for COVID-19 incidents with one week lag and PM 2.5 at county-level by zip code

Data	Statistic	Value
Case_Lag_1w (Positive cases with one week lag), PM2.5_CF1_ug/m3	Covariance	[[2.42862914e+01 -3.24047511e+01] [-3.24047511e+01 6.89323951e+04]]
Case_Lag_1w, PM2.5_CF1_ug/m3	Pearson's correlation coefficient	-0.025
Case_Lag_1w, PM PM2.5_CF1_ug/m3	Spearman's rank order correlation	0.083

The dataframe was then duplicated for ML algorithm test purposes. Departing from time-series analysis, classification was aimed to predict sole occurrence or non-occurrence of COVID-19 events. From the new dataframe, positive cases, hospitalizations, deaths, hospitalizations and deaths with a one-week time lag, zip and date were all dropped. Using `numpy`, positive case counts with the one-week lag were converted to values '1' and '0,' where '1' represented COVID-19 case occurrence and '0' represented no associated incidents. Case occurrence/non-occurrence values were extracted to their own dataframe, y, and the column was dropped from the previously duplicated dataframe, X. The test-train data split was set to a standard 30:70 ratio respectively, with a random state parameter also used for reproducibility reasons. ML classification algorithms RF and SVM were selected for partly due to the fact that they can be used for both classification and regression problems. The algorithms were applied whereby default and optimal parameters were each run on both models. Ideal parameters were obtained by using random search with Cross Validation (CV) for RF and using grid search with CV for SVM. Through this process, the range for each hyperparameter was narrowed down, an assortment of hyperparameter combinations were tested, and prevention of overfit was attempted. For all results, CMs were printed and compared; these are evaluated in Table 10.

Regression modeling was then used to test the predictive power of ML algorithms in calculating *how many* cases would occur, based on the preloaded environmental and socio-economic variables. A new dataframe was constructed using the original, individual COVID-19 event CSV files for positive cases, hospitalizations and deaths. For each, a cumulative amount per incident type per zip code was calculated. The dataframes were then merged together with the previously curated zip code dataframe in order to retain locational information. For this analysis, two additional raster files were loaded. One TIFF file represented average temperature, similar to that used in the previously conducted state-level analysis, but with three additional months included for a more accurate value considering the timeframe over which the COVID-19 data was collected. The other represented average PM 2.5 levels. Having established little difference in correlative results between EPA and VGI data for PM 2.5, and having eliminated time-related components, a shapefile for average readings updated bi-annually was downloaded from the Washington Geospatial Open Data Portal. This was then converted to a raster file from which values could be extracted based on zip code latitude and longitude values. This was done to assess the impact of *average* PM 2.5 readings on total numbers of COVID-19 incidents, and to determine if long-term PM 2.5 levels hold more weight than smaller daily fluctuations.

Prior to model execution, a heatmap and statistical analyses were run, demonstrating a much stronger positive correlation between the two main variables when evaluating them cumulatively.

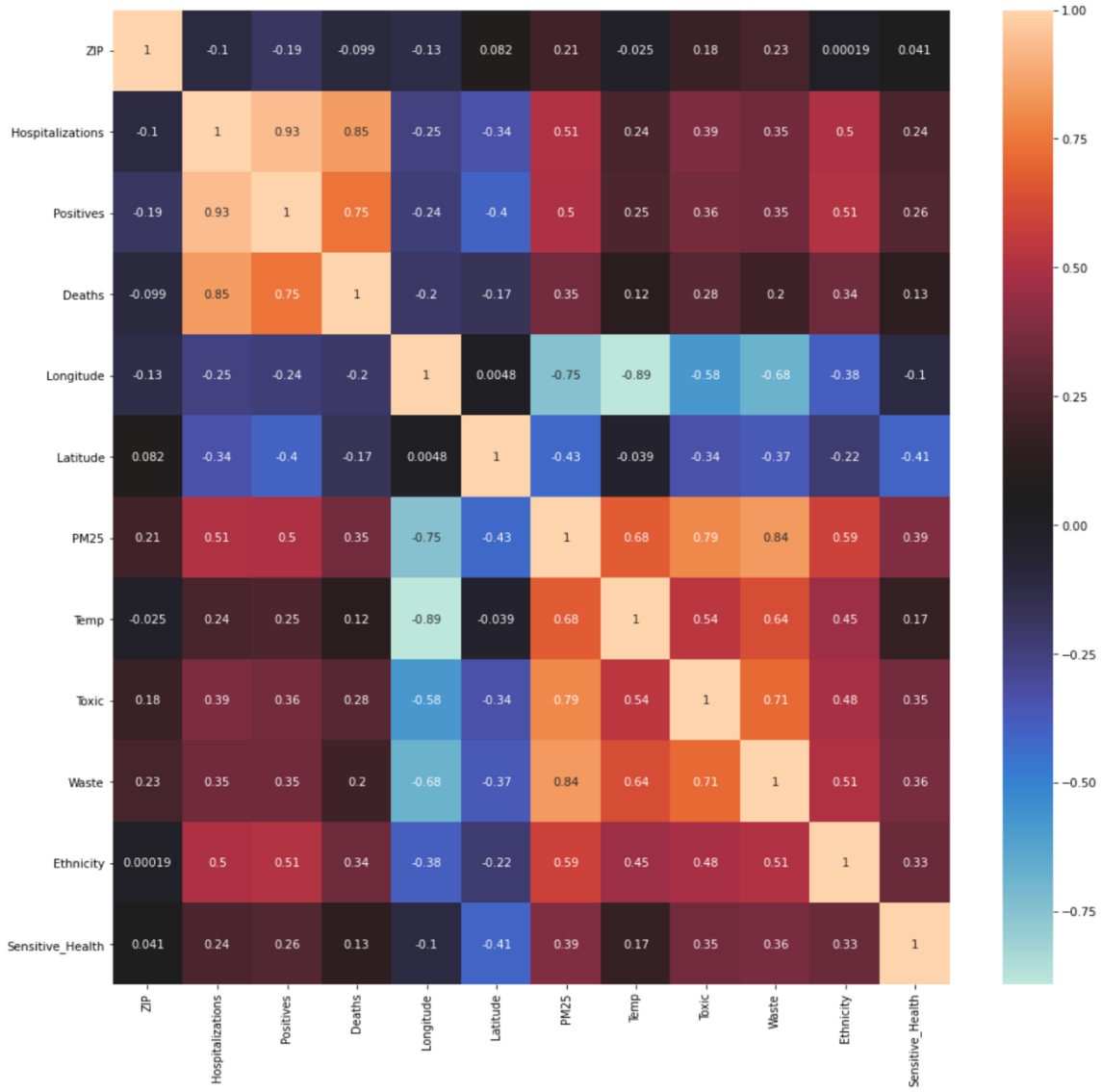


Figure 11 Heatmap showing correlation between cumulative COVID-19 case count and average PM 2.5 levels, King County

Table 8 Statistical results for cumulative COVID-19 incidents and average PM 2.5

Data	Statistic	Value
Case_Lag_1w (Positive cases with one week lag), PM2.5_CF1_ug/m3	Covariance	[[4.71048519e+05 3.62860000e+02] [3.62860000e+02 1.11233349e+00]]
Case_Lag_1w, PM2.5_CF1_ug/m3	Pearson's correlation coefficient	0.501
Case_Lag_1w, PM PM2.5_CF1_ug/m3	Spearman's rank order correlation	0.477

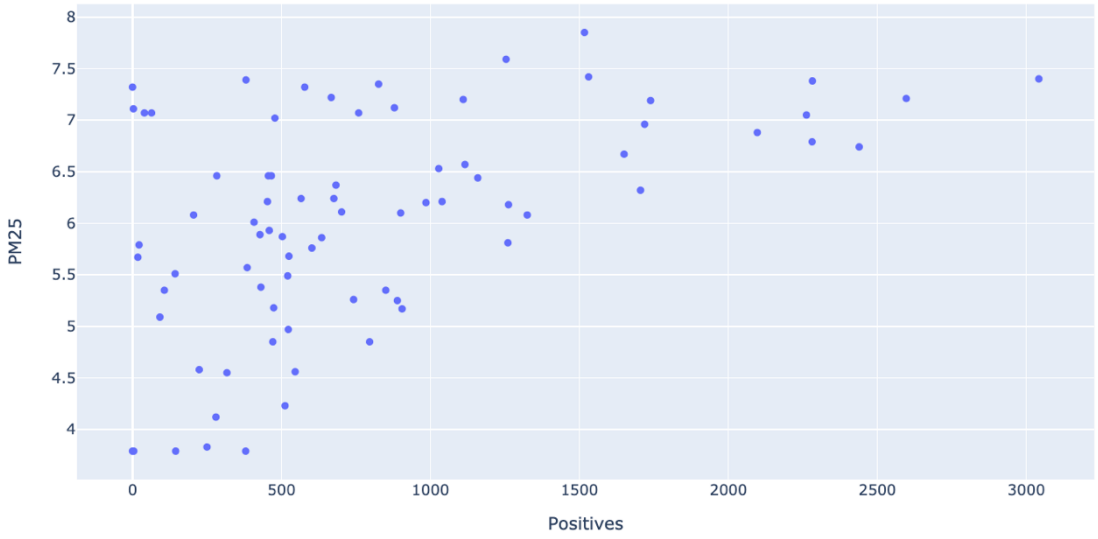


Figure 12 Correlation between cumulative COVID-19 case count and average PM 2.5 levels, King County

In preparation to run the four regression algorithms, positive case count, hospitalizations, deaths and latitude and longitude were removed from dataframe X, and all but case count totals were removed for dataframe y. The same 30:70 test-train split was applied to the data and default parameters were used to make initial predicted values for case counts dependent on other variables included. Statistical averages best representative of model performances were collected and are evaluated in Table 11.

Results and Discussion

The success of the study's results to date consists of both classification and regression model performance metrics listed in Table 9.

Table 9 Model performance metrics

Model Type	Performance	Definition
Classification	TP	True Positive (TP). The model's ability to correctly classify the positive class (i.e., correct identification of COVID-19 occurrence in an area).
	FP	False Positive (FP). The model's inability to correctly classify the positive class (i.e., incorrect identification of COVID-19 occurrence in an area).
	TN	True Negative (TN). The model's ability to correctly classify the negative class (i.e., correct identification of COVID-19 non-occurrence in an area).
	FN	False Negative (FN). The model's inability to correctly classify the negative class (i.e., incorrect identification of COVID-19 non-occurrence in an area).
	Avg. Precision	$TP/(TP + FP)$

		Total predicted positive class. This metric evaluates the cost of FPs (Shung 2018) and how precise the model is in predicting positive values.
	Avg. Recall	$TP/(TP+FN)$ Total actual positive class. This metric evaluates the cost of FNs (Shung 2018) by calculating the percentage of positives.
	Avg. F-1 Score	$2(\frac{Precision \times Recall}{Precision + Recall})$ This metric combines Precision and Recall performance values of a classifier and is helpful in comparing two or more.
	Overall Accuracy	$(TP+TN)/(TP+TN+FP+FN)$ A measurement of how frequently a correct classification/ prediction is made by the model.
Regression	R-Squared	$R^2 = 1 - \frac{\Sigma_i(y_i - \hat{y_i})^2}{\Sigma_i(y_i - \bar{y})^2}$ A measurement of 0-1 where higher values express fit between predicted and actual values (Wu 2020).
	MSE*	$MSE = \frac{1}{N}\Sigma_{i=1}^{N}\Sigma_i(y_i - \hat{y_i})^2$ An absolute number indicating deviation of predicted values from actual values (Wu 2020). This metric penalizes larger error over smaller error.
	RMSE	Root of the MSE, allowing for easier interpretation of deviation between predicted and actual values.
	MAE*	$MAE = \frac{1}{N}\Sigma_{i=1}^{N}\lvert y_i - \hat{y_i}\rvert^2$ A sum of errors (Wu 2020). This metric penalizes all error at the same rate.

*MSE *Mean Square Error,* MAE *Mean Absolute Error*

By way of supervised ML classification, COVID-19 occurrences were predicted with up to an accuracy of 78% based on environmental and socio-economic variables. Between 70-80% for overall accuracy is deemed 'good' whereas 80-90% is categorized as 'excellent.' Above this value implies the likelihood of overfit, therefore, there is room in this study for improvement and more experimentation with alternate variables and algorithms is advised. Disappointingly, the accuracy in prediction rate is unlikely due to any correlation between COVID-19 case count and PM 2.5. Based on statistical results and visualization of correlation, PM 2.5 had a negligible effect on variables across the board. In answering one of the three posed research questions, it would appear that despite reasonable performance metrics scores, COVID-19 occurrence/non-occurrence prediction accuracy as a result of classification modeling is *not* related to pollution level data.

Should a similar study occur, it is suggested, based off of results, that the RF algorithm be used to estimate predictive values. This is due to its consistent precision, recall, F-1 Score and overall accuracy. Additionally, it appears that default hyperparameters give a comparable initial result and an Area Under Curve – Receiver Operating Characteristic (AUCROC) validation score of 86% also demonstrates RF's capacity to exceed SVM's predictive power in all areas.

Table 10 Classification model testing phase performance metric results

Phase	Performance Metric	Prediction Model			
		RF, default	RF, optimal hyperparameters	SVM, default	SVM, optimal hyperparameters
Testing	TP	1662	1639	370	1359
	FP	763	786	2055	1066
	TN	3362	3351	3798	3356
	FN	610	582	135	577
	Avg. Precision	0.77	0.78	0.69	0.73
	Avg. Recall	0.77	0.77	0.56	0.70
	Avg. F-1 Score	0.77	0.77	0.52	0.71
	Overall Accuracy	0.78	0.78	0.66	0.74

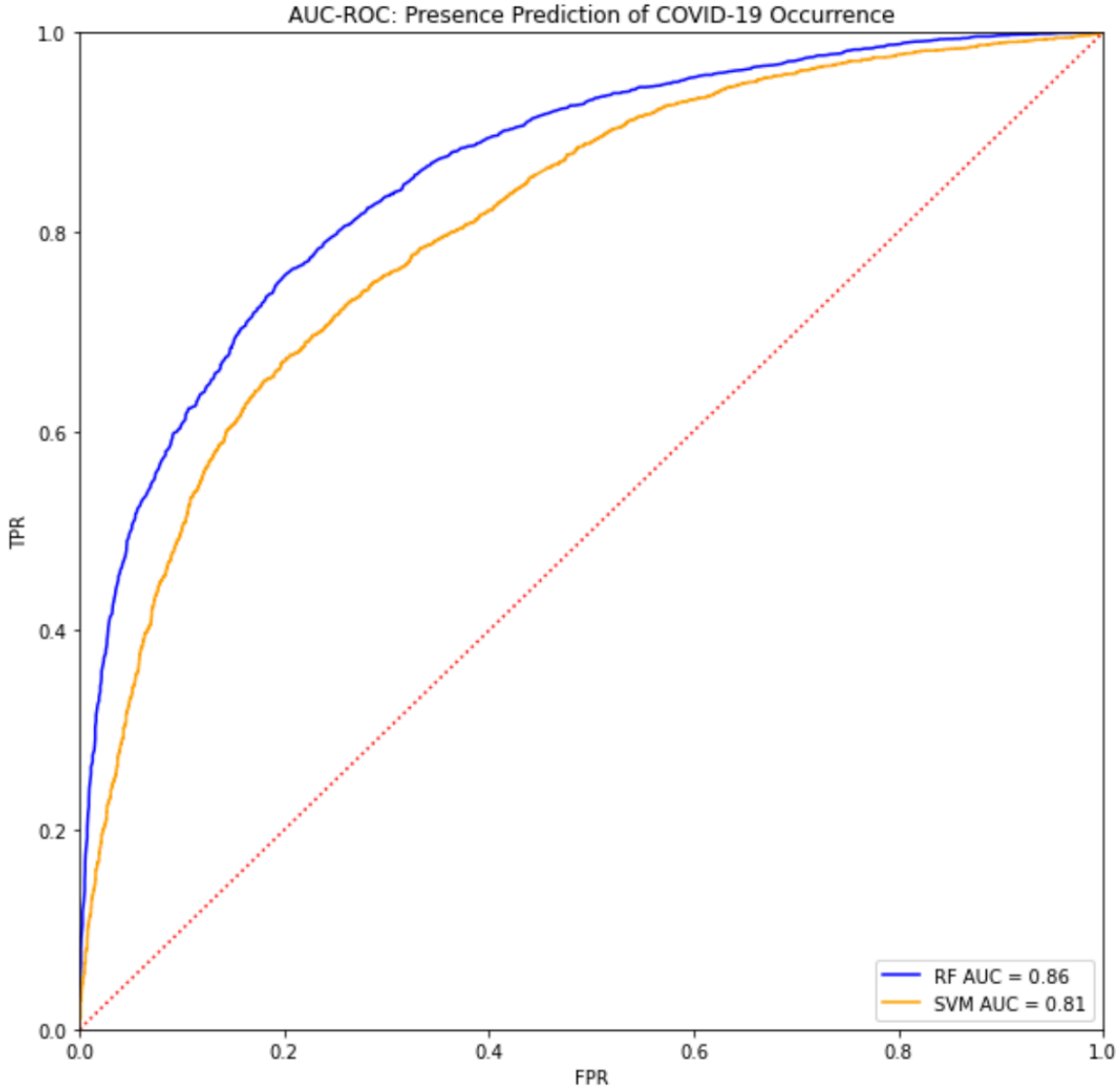


Figure 13 AUCROC results of RF and SVM classification models for predicting occurrence/non-occurrence of COVID-19 cases

Although regression algorithms were used to approximate COVID-19 case rates, and technically could be used to determine hospitalizations and deaths in the same manner, results close to 0 for R-Squared indicate that variables do not depend on one another; thus, prediction of case numbers are most likely random. Once again, however, RF (regressor) had the most informative score of 0.1, which can be marginally noteworthy when dealing with vast amounts of real-world data. In the instance of regression modeling using PM 2.5 levels averaged over time, it appeared that longer term pollution data was more closely linked to cumulative counts of COVID-19 incidents. However, due to both information and time restraints, the assessment of susceptibility to prolonged recovery times as a consequence of poor air quality contributing factors, was not answered. Although this gives way to the possibility of future project expansion, the results do not rank high in terms of acceptability and a reassessment of how this might occur, along with what data may be more meaningful, should be considered.

Ultimately, the foundation of regression modeling for COVID-19 case count prediction is in place and although results were not as insightful as hypothesized, this general framework can be used to explore additional datasets.

Table 11 Regression model testing phase performance metric results

Performance Metric	LM	BR	RF (Regression)	SVM (Regression)
R-Squared	0.037	-0.025	0.100	-0.041
MSE	462454.681	492411.841	432150.829	500129.282
RMSE	680.040	701.721	657.382	707.198
MAE	475.487	508.933	486.487	539.429

Future Potential Expansion

Various ideas have been noted throughout this project regarding its potential future expansion. Firstly, the original format of certain COVID-19 incident data included cases, hospitalizations and deaths by age group. Focusing on location, age was eliminated as a variable from this part of the study but could easily be reincorporated at a later time. This may uncover certain patterns and relationships between variables that have gone unnoticed to date, providing further insight into the effect of PM 2.5 on certain populations.

Alternative ML algorithms should be used in the extension of regression modeling for PM 2.5 and COVID-19 events. GAMs are renowned for their utility within epidemiological studies but were not however, able to be included in this project due to time constraints and the need for data restructuring.

As time passes, more data concerning the novel Coronavirus will be released. This could enable the investigation between PM 2.5 levels and COVID-19 recovery times as well as the mapping of areas at risk of both COVID-19 event intensity and prolonged, projected recuperation periods as a result of environmental and socio-economic variables.

Conclusion

It appears from the results of this study that short-term fluctuations in PM 2.5 concentration do not have as equal an impact on COVID-19 related cases, hospitalizations or deaths in comparison to the cumulative effect of pollution on respiratory health. Although daily mean concentrations of PM

2.5 can be used in classification modeling to predict whether or not there will be occurrences of incidents in a certain area, it seems that other environmental and socio-economic variables hold more weight in overall effects to case counts.

It had been noted at the beginning of this study that less research has been focused on whether standard-driven data collection provides a higher reliability of results in comparison to VGI when analyzing PM 2.5 levels. One takeaway from this project are the similarities that unfolded between the two datasets. When using heatmapping to assess correlation, both EPA and VGI PM 2.5 data demonstrated significantly similar correlative values with the same variables. In conclusion, there does not appear to be significant difference in the results of pollutant-related, spatially dispersed datasets, whereby one is the product of volunteered information and another meets FEM criteria.

Finally, in response to the third research question, it is apparent that although ML algorithms are able to approximate COVID-19 case rates, hospitalizations and deaths using PM 2.5 and other variable data, they cannot do so as accurately as had been hoped. This is in part due to the non-linear relationships between variables and possibly the algorithms that were selected for use. Alternate algorithms should be explored for this part of the study, in addition to the incorporation of recovery times, as suggested in *Future Potential Expansion*. An analysis of whether this variable responds to the effects of PM 2.5 concentration levels in the same manner as other COVID-19 incidents could lead to an interesting comparison.

Bibliography

AirNow. 2021. *Air Quality Index (AQI) Basics .* Accessed May 2021. https://www.airnow.gov/aqi/aqi-basics/.

Bora, Alakesh. 2020. *Implementation of Bayesian Regression .* September 20. Accessed May 2021. https://www.geeksforgeeks.org/implementation-of-bayesian-regression/.

Brotsky, Edward. 2014. *Air quality index – an indispensable tool .* January 28. Accessed May 2021. https://healthymesacounty.org/blogs/air-quality-index-an-indispensable-tool/.

Brownlee, Jason. 2019. *Difference Between Classification and Regression in Machine Learning .* May 22. Accessed May 2021. https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/#:~:text=Fundamentally%2C%20classification%20is%20about%20predicting,is%20about%20predicting%20a%20quantity.&text=That%20classification%20is%20the%20problem,quantity%2.

Brownlee, Jason. 2018. *How to Calculate Correlation Between Variables in Python.* April 27. Accessed May 2020. https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/.

Centers for Disease Control and Prevention. 2021. *How to Protect Yourself & Others.* March 8. Accessed May 2021. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html.

Cleveland Clinic. 2020. *Coronavirus, COVID-19.* December 11. Accessed May 2021. https://my.clevelandclinic.org/health/diseases/21214-coronavirus-covid-19#:~:text=Coronaviruses%20are%20a%20family%20of,cause%20illness%20in%20humans.

Costa Nascimento, Luiz Fernando , Luciana Cristina Pompeo Ferreira Vieira, Kátia Cristina Cota Mantovani, and Demerval Soares Moreira. 2016. "Air pollution and respiratory diseases: ecological time series." *Sao Paulo Medical Journal* 134 (4): 351-321.

Cralle, Terry. 2020. *www.terrycralle.com.* June 26. Accessed May 2021. https://www.terrycralle.com/pm-2-5-filter-masks/.

Croft, Daniel P., Wangjian Zhang, Shao Lin, Sally W. Thurston, Philip K. Hopke, Mauro Masiol, Stefania Squizzato, Edwin van Wijngaarden, Mark J. Utell, and David Q. Rich. 2019. "The Association between Respiratory Infection and Air Pollution in the Setting of Air Quality Policy and Economic Change." *Annals of the American Thoracic Society* 16 (3): 321-330.

Donges, Niklas. 2019. *A Complete Guide to the Random Forest Algorithm.* 06 16. Accessed May 2021. https://builtin.com/data-science/random-forest-algorithm.

Else, Holly. 2020. *How a torrent of COVID science changed research publishing — in seven charts .* December 16. Accessed May 2021. https://www.nature.com/articles/d41586-020-03564-y.

EPA. 2020. *Air Quality System (AQS).* October 28. Accessed May 2021. https://www.epa.gov/aqs.

Gandhi, Rohith. 2018. *Support Vector Machine — Introduction to Machine Learning Algorithms.* Medium. 06 07. Accessed May 2021. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

IQAir. 2021. *Air pollution affects likelihood, severity of flu.* Accessed 2021. https://www.iqair.com/blog/health-wellness/air-pollution-affects-likelihood-severity-flu.

IQAir. 2021. *Comparison of PM2.5 measurements using the AirVisual Monitor and Beta Attenuation Monitor.* Accessed May 2021. https://www.iqair.com/blog/air-quality/comparison-of-pm25-measurements-using-the-airvisual-monitor-and-BAM.

kaiterra. 2017. *The 3 Types of Particle Detectors: How We See the Invisible.* September 18. Accessed May 2021. https://learn.kaiterra.com/en/air-academy/the-3-types-of-particle-detectors-how-we-see-the-invisible.

King County. 2021. *Daily COVID-19 outbreak summary .* Accessed May 2021. https://kingcounty.gov/depts/health/covid-19/data/daily-summary.aspx.

Liu, Yi, Jingjie Sun, Yannong Gou, Xiubin Sun, Dandan Zhang, and Fuzhong Xue. 2020. "Analysis of Short-Term Effects of Air Pollution on Cardiovascular Disease Using Bayesian Spatio-Temporal Models." *International Journal of Environmental Research and Public Health* 17 (3): 879.

Malvern Panalytical. 2021. *Laser Diffraction (LD).* Accessed May 2021. https://www.malvernpanalytical.com/en/products/technology/light-scattering/laser-diffraction.

National Geographic. 2021. "The Fight for Clean Air." *National Geographic.* April.

Natural Resources Defense Council, Inc. 2014. *The Particulars of PM 2.5.* November 14. Accessed May 2021. https://www.nrdc.org/onearth/particulars-pm-25#:~:text=Road%20dust%20and%20tiny%20bits,all%20major%20PM%202.5%20sources.

NETSTATE. 2016. *Washington.* April 20. Accessed May 2021. https://www.netstate.com/states/geography/mapcom/wa_mapscom.htm#:~:text=Washington's%20elevation%20runs%20from%20sea,coast%20in%20Washington%20and%20Oregon.

Pima County. 2021. "What is Particulate Matter?" *Pima County.* Accessed May 2021. https://www.webcms.pima.gov/UserFiles/Servers/Server_6/File/Government/Environmental%20Quality/Air/Air%20Monitoring/AAWhat%20is%20Particulate%20Matter.pdf.

Public Lab. 2021. *PurpleAir.* Accessed May 2021. https://publiclab.org/wiki/purpleair.

PurpleAir. 2021. May. Accessed May 2021. https://www.purpleair.com/map?opt=1/mAQI/a10/cC0#11/40.7311/-111.8565.

PurpleAir. 2021. *FAQ.* Accessed May 2021. https://www2.purpleair.com/community/faq#hc-how-do-purpleair-sensors-compare-to-regulatory-particulate-matter-sensors.

scikit-learn. 2021. *Linear Models.* Accessed May 2021. https://scikit-learn.org/stable/modules/linear_model.html#bayesian-regression.

Shung, Koo Ping. 2018. *Accuracy, Precision, Recall or F1?* Medium. 03 15. Accessed 01 2021. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

Somayaji, Ranjani, Moni B. Neradilek, Adam A. Szpiro, Kathryn H. Lofy, Michael L. Jackson, Christopher H. Goss, Jeffrey S. Duchin, Kathleen M. Neuzil, and Justin R. Oritz. 2020. "Effects of Air Pollution and Other Environmental Exposures on Estimates of Severe Influenza Illness, Washington, USA." *Emerging Infectious Diseases Journal* 26 (5): 920-929.

Talhelm, Thomas. 2020. *What Is the Difference Between the PM2.5 and AQI Measurements? .* Accessed May 2021. https://smartairfilters.com/en/blog/difference-pm2-5-aqi-measurements/.

Taylor, Derrick Bryson. 2021. *A Timeline of the Coronavirus Pandemic.* March 17. Accessed May 2021. https://www.nytimes.com/article/coronavirus-timeline.html.

The New York Times. 2021. *Coronavirus in the U.S.: Latest Map and Case Count.* May 30. Accessed May 2021. https://www.nytimes.com/interactive/2021/us/covid-cases.html.

USAFacts. 2021. *USA FACTS.* May 29. Accessed May 2021. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map.

Wang, Jianjun, Meigen Zhang, Xiaolin Bai, Hongjian Tan, Sabrina Li, Jiping Liu, Rui Zhang, et al. 2017. "Large-scale transport of PM2.5 in the lower troposphere during winter cold surges in China." *Scientific Reports* 7.

Washington State Department of Health. 2021. *COVID-19 Data Dashboard.* 6 2. Accessed 2021. https://www.doh.wa.gov/Emergencies/COVID19/DataDashboard#downloads.

Washington State Department of Health. 2021. *Population Age 5+ Speaking English Less than Very Well .* Accessed May 2021. https://fortress.wa.gov/doh/wtn/WTNPortal#!q0=620.

WebMD LLC. 2021. *Coronavirus: What Happens When You Get Infected? .* February 3. Accessed May 2021. https://www.webmd.com/lung/coronavirus-covid-19-affects-body#1.

William F. Marshall, III. 2020. *Coronavirus infection by race: What's behind the health disparities?* August 13. Accessed May 2021. https://www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-infection-by-race/faq-20488802.

World Health Organization. 2021. *Timeline: WHO's COVID-19 response.* WHO. Accessed May 2021. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline?gclid=CjwKCAjwzMeFBhBwEiwAzwS8zKHIP25zAk8qqllVMxR-34dilrXwAJg-0FJ2UJpM5CTNwoKyzjO7FRoCJ0gQAvD_BwE#event-9.

World Health Organization. 2021. *WHO Coronavirus (COVID-19) Dashboard.* Accessed May 2021. https://covid19.who.int/.

Wu, Songhao. 2020. *3 Best metrics to evaluate Regression Model?* May 23. Accessed May 2021. https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b.

Yse, Diego Lopez. 2019. *https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14.* Medium. 04 17. Accessed 01 2021. The Complete Guide to Decision Trees.

Appendix

GitHub Repository link: https://github.com/mlenih19/PM25_COVID19

Cover image: Cover and header images were taken from the CDC newsroom image library, which provides high-resolution, public domain imagery, allowed for free use and inclusion in publications and prints.