

Correlating Lockdowns, Mortality Rates and Air Pollution: A Deep Learning Imbued Study of COVID-19

by

Tahia Tabassum
17301183

Saiham Rahman
17101116

Moosfiqur Hassan Mahmood
17101105

Md. Fahim Siam
20141040

Sadia Anika Mumu
20141032

A thesis submitted to the School of Data and Sciences
in partial fulfillment of the requirements for the degree of
Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2021

© 2021. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tahia Tabassum

Tahia Tabassum
17301183

Saiham Rahman

Saiham Rahman
17101116



Moosfiqur Hassan

Moosfiqur Hassan Mahmood
17101105

Siam
Md. Fahim Siam
20141040

Sadia Anika Mumu

Sadia Anika Mumu
20141032

Approval

The thesis/project titled “Correlating Lockdowns, Mortality Rates and Air Pollution: A Deep Learning Imbued Study of COVID-19” submitted by

1. Tahia Tabassum (17301183)
2. Saiham Rahman (17101116)
3. Moosfiqur Hassan Mahmood (17101105)
4. Md. Fahim Siam (20141040)
5. Sadia Anika Mumu (20141032)

Of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 11, 2021.

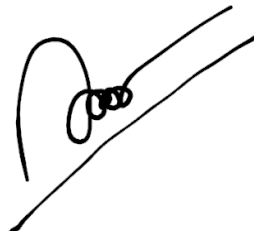
Examining Committee:

Supervisor:
(Member)



Dr. Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Dr. Muhammad Iqbal Hossain, PhD
Assistant Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Nationwide lockdowns implemented in consequence of the devastating COVID-19 pandemic, caused noticeable improvements in air quality throughout the world. This paper implements a multivariate long-short term memory network to forecast changes in the Air Quality Index and Particulate Matter 2.5 (PM2.5) concentration for 26 cities in India, and 50 cities in Europe, had their lockdown not occurred or been extended. A linear regression model was used to correlate confounder-adjusted PM2.5 values with COVID-19 mortality rate in the U.S.A. Heat maps were visualized with K-Means Clustering that signified the correlation between increased air pollution with higher COVID-19 cases and mortality rates. Our results indicate that 76% of the European cities in our dataset underwent at least a 40% improvement in air quality as a result of their lockdowns, whereas 17 out of the 26 Indian cities observed 20%. Adjusted PM2.5 was seen to be a statistically significant contributor to increasing mortality rate, with a single unit increase contributing to 3% more deaths due to COVID-19, at a 95% confidence level.

Keywords: COVID-19; LSTM; Air Pollution; K-Means Clustering; COVID-19 Mortality; Regression; COVID-19 Lockdowns

Acknowledgement

This thesis is an embodiment of a group's relentless hard work. We thank our honorable supervisor, Dr. Md. Golam Rabiul Alam, and our respected co-supervisor, Dr. Muhammad Iqbal Hossain for the invaluable guidance and support they have provided us throughout our journey.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vii
List of Tables	viii
Nomenclature	ix
1 Introduction	1
1.1 The Pollution Problem	1
1.2 COVID-19 and The Correlation with Air	2
1.3 Research Methodology	2
1.4 Research Objectives	3
2 Related Work	4
2.1 Studies Conducted Upon India	4
2.2 Studies Conducted Upon USA	6
2.3 Studies Conducted Upon Italy	7
2.4 Studies Conducted Upon Other Countries	7
3 Analysis of Datasets	10
3.1 Collection	10
3.2 Methods and Results	12
4 Predictive Analysis Using Multivariate RNN and LSTM	16
4.1 Model Summary	16
4.2 Method and Implementation	18
4.3 Results	19
4.4 Evaluation	25

5	Correlating COVID-19 and Air Quality with Linear Models	28
5.1	Methods	28
5.2	Results	30
6	Factor Visualization With K-Means Clustering	32
6.1	Model Summary	32
6.2	Method	32
6.3	Implementation and Analysis	35
7	Future Work	40
8	Conclusion	41
	Bibliography	42
	Appendix A. Tables of Lockdown Dates Around the World	46
	Appendix B. Percentage Tables from Forecasting Model	48

List of Figures

3.1	Seasonal variation of pollutant concentration for Indian cities	11
3.2	Percentage decrease in AQI and PM2.5 from 2018 and 2019 to 2020 .	12
3.3	Percentage decrease in PM10 and SO ₂ from 2018 and 2019 to 2020 .	13
3.4	Percentage decrease in CO and NO ₂ from 2018 to 2020 and 2019 to 2020	14
4.1	General architecture of a Recurrent Neural Network model	17
4.2	General architecture of a Long Short-Term Memory RNN	18
4.3	Flowchart depicting LSTM process	19
4.4	Air quality in Ahmedabad with and without the COVID-19 lockdown	20
4.5	National Air Quality Index of India	21
4.6	Air quality in Visakhapatnam with and without the COVID-19 lockdown	21
4.7	PM2.5 levels in Bengaluru with and without the COVID-19 lockdown	22
4.8	Percentage increase in PM2.5 values if no lockdown in India	22
4.9	Air quality in Lucknow if the COVID-19 lockdown been extended . .	23
4.10	Forecasted PM2.5 for Barcelona if lockdown had not occurred	24
4.11	Forecasted PM2.5 in Lisbon had the lockdown been extended	25
4.12	Forecasted PM2.5 in London had the lockdown been extended	25
4.13	Block diagram of a Linear Forecasting Model	26
5.1	Linear Model Summary for co-variates and Mortality Rate Analysis .	29
5.2	Correlation of PM2.5 concentration with COVID-19 mortality rates .	30
5.3	Scaled plot of adjusted PM2.5 and COVID-19 Mortality with Linear Models	31
6.1	Flowchart of the K-Means Clustering Model	33
6.2	Algorithm of the K-Means Clustering Process	34
6.3	Clusters of PM2.5, COVID-19 Cases, and Mortality in USA for 2020	35
6.4	County-wise PM2.5 and COVID-19 mortality cases in the U.S.A. . .	36
6.5	Percentage difference of averaged historical PM2.5 with PM2.5 in 2020	37
6.6	Map of Italy depicting NO ₂ concentration in 2019 and 2020	38
6.7	Map of Italy depicting COVID-19 mortality data	38
6.8	Map of Italy depicting COVID-19 cases	39

List of Tables

4.1	Mean Percentage Decrease in PM2.5 due to COVID-19 lockdown in Europe	24
-----	--	----

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AQI Air Quality Index

CO Carbon Monoxide

CO₂ Carbon Dioxide

COVID – 19 Coronavirus 2019

LSTM Long-Short Term Memory

NO Nitrogen Monoxide

NO₂ Nitrogen Dioxide

NO_x Mix of oxides of nitrogen

O₃ Ozone

PM10 Particulate Matter 10

PM2.5 Particulate Matter 2.5

RNN Recurrent Neural Network

SO₂ Sulfur Dioxide

USA United States of America

WHO World Health Organization

Chapter 1

Introduction

1.1 The Pollution Problem

One of the most concerning health and environmental issues in the world right now is air pollution. Every year seven million deaths are recorded worldwide due to air pollution [1]. According to the Global Burden of Disease Study, a significant contributor to cardiovascular disease mortality is air pollution; it was considered the main reason for nearly 5 million untimely deaths throughout the world in the year 2017 [2]. Judging from this, we can clearly understand the severity of air pollutants on our health. All types of pollution affect our health slowly but surely; among all the problems, air pollution is considered the most fatal.

All over the world, the severity of air pollution is measured through values of AQI. The AQI of a region is a measure of how healthy the overall air quality is; it is calculated by taking into account the concentrations of primary air pollutants such as Nitrogen Dioxide, Particulate Matter 2.5, Sulfur Dioxide and more. Countries around the world have different standards for the measurement of air quality. However, for all of the variations a higher value of AQI always indicates poorer air quality (for e.g., 125 units indicates worse air than 80 units of AQI).

The causes of air pollution can be parted into two sections: the first is indoor air pollution, which is mostly caused by burning brushwood, fuel, crop waste, coal for preparing food, and heating purposes. The second is outdoor air pollution caused primarily by engines of all automobile vehicles and industrial fuel burning. Other outside air pollution origins include windblown dust, smoke from forest fire, and biogenic discharge from vegetation.

According to the World Health Organization (WHO) 90% of people Breathe air that is worse than the WHO guideline limitations, with the most affected people being from low and middle-income countries [1]. The main pollutants in the air are Particulate Matter (2.5 and 10), Sulphur Dioxide, Carbon Monoxide, Ozone, and the different Oxides of Nitrogen, all of which we will be investigating in this paper. Air pollution affects our health in many ways that we do not directly notice. In a recent study, it was found that the percentage of deaths and disease from lung cancer is 29% due to air pollution. Similarly, from acute lower respiratory infection

the statistic is 17%, from stroke it is 24%, from ischemic heart disease 25%, and chronic obstructive pulmonary disease 43% [3].

1.2 COVID-19 and The Correlation with Air

The Coronavirus disease 2019 (COVID-19) is a highly contagious disease, originating from the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). The contagious disease was first detected in Wuhan, Hubei, China in December 2019 and has since evolved into a global pandemic. More than 31.4 million cases have been registered over 188 countries with more than 966,000 deaths till 22nd September 2020 [4]. Governments of many countries all over the world have declared lockdown periodically at different times in order to control the outbreak of COVID-19 as much as possible. Despite the repercussions of this, we have observed one constructive change due to the lockdown: drastic improvements in air quality.

In recent years, the whole world suffered immensely from acute air pollution problems which even led to life-threatening diseases. Just in China, some statistics suggest that 25 million healthy life years were lost because of air pollution (Kassebaum et al., 2014). However, due to the COVID-19 lockdown, air quality has improved in many countries. Recent researches also suggest that the correlation between air and COVID-19 is not limited to just the lockdown periods. Work conducted in the United States and in Northern Italy focused on the effect of historical air quality on deaths related to COVID-19.

India held one of the biggest nation-wide lockdowns in history by implementing stay-at-home orders and closing down almost all kinds of outdoor activities for almost 1.38 billion people, for over 60 days. This caused the lowest ever recorded air quality index in history for many different Indian cities. Similarly, the majority of cities throughout Europe implemented stay-at-home orders, in an effort to keep their citizens safe. When a lockdown was imposed on 23rd January 2020 in Wuhan, China till 8th April 2020 (total 76 days), along with helping China in reducing the number of infections to almost zero, it also drastically decreased the global carbon emission and level of nitrogen dioxide (NO_2) in the atmosphere.

1.3 Research Methodology

The negative effects of air pollution ranging from mild distress to death is a case that has been frequently discussed in all forms of media. However, the correlation between COVID-19 induced lockdowns and improvements in air quality is one that is just beginning to be explored.

The first segment of our paper focuses on performing statistical analysis on a dataset containing 5 years of daily air pollutant data (2015-2020) on 26 Indian cities and on another dataset containing the same for 50 European cities. Next, using a multivari-

ate LSTM RNN model, we decided to forecast what the Air Quality Index (AQI) and PM2.5 values in our chosen regions would have looked like if the COVID-19 lockdowns had not occurred and what they would have been like if the lockdowns had been extended. We added meteorological data like daily wind speed and temperature values to further improve the accuracy of our analysis.

Our next objective was to analyze possible correlations between COVID-19 mortality rates and air quality. Primarily, we wanted to understand whether increasing concentrations of historical air pollutant values contributed to higher mortality rates. In order to this, we collected PM2.5 concentration data for over 3000 counties in the U.S.A as well as COVID-19 county wise death data, population data, and data of 43 other confounding factors that could have affected mortality rates. We adjusted the PM2.5 values by regressing out the effect of the identified confounders and then implemented a linear regression upon the adjusted values (independent variable) with the COVID-19 mortality rate (dependent variable).

Finally, in an effort to understand the impact of air quality and other confounding factors we decided to spatially visualize their effects on a global scale. We added pollutant and mortality data from the country of Italy to our pool of datasets and created cluster diagrams correlating, for instance, adjusted PM2.5 values with COVID-19 cases using the K-Means clustering algorithm. The results of the performance of our analyses have been illustrated in the subsequent chapters of our paper.

1.4 Research Objectives

Our objectives for the paper can be summarized into 5 main sections. They are as follows:

1. Analyzing the air quality data of all of our chosen regions during their specific lockdown period, and the air quality data for the same duration in 2018 and 2019 to visualize and statistically compare how the air quality has changed due to the lockdown.
2. Forecasting AQI and PM2.5 values of our chosen cities with an RNN LSTM model for 100 days, considering the COVID-19 pandemic and the subsequent lockdown had not occurred and comparing them with actual values.
3. Forecasting AQI and PM2.5 values of our chosen cities with an RNN LSTM model for 30-60 days after the lockdown period, considering the lockdown was not lifted and comparing the values with the actual data to see whether it would have been better in an extended lockdown.
4. Correlating adjusted PM2.5 values with COVID-19 mortality rates and investigating the existence of a relationship between the two.
5. Clustering pollutant values, COVID-19 mortality indices, COVID-19 case and factor data to identify correlations among them.

Chapter 2

Related Work

2.1 Studies Conducted Upon India

Before starting our work, we looked at some similar papers that dealt with similar problems. In a research article, Mahato, Pal, Gosh (2020), worked on the air quality before and during lockdown phases by collecting the data of the main pollutants (PM10, PM2.5, SO₂, NO₂, CO, O₃ and NH₃) for 34 monitoring stations encompassing over the city of Delhi, India [5]. During the early lockdown (24th March - 14th April) in Delhi, they used the data of the seven pollutant parameters and used the National Air Quality Index (NAQI) to visualize spatial patterns of air quality.

Firstly, they worked on differences in the levels of the primary pollutants before and during lockdown phases by analyzing 24-hour median values of PM2.5, PM10, SO₂, NH₃, NO₂, and NAQI and 8-hour median daily maxima of CO and O₃ from 3rd March to 14th April in NCT Delhi; in doing so, they found variations between pre-lockdown and lockdown phase.

Secondly, they performed analysis on the spatial pattern of National Air Quality Index (NAQI) before and during India's COVID-19 lockdown and found a reduction in NAQI during the stay-at-home period. From the analysis of collected data of pollutants, six NAQI categories (CPCB, 2015) were used to evaluate the impact on our health by the seven pollutants according to their standards.

Thirdly, from the spatial concentration pattern they showed that major pollutants like PM2.5, PM10, NO₂ and CO changed excessively because of the lockdown and the quality of air improved much. Then they also separately analyzed PM10 and PM2.5 pollutant data from 2017 to 2019 and compared it with 2020 data for the same time period. Their results found that for the past three years it was higher compared to 2020. Furthermore, they also tried to find correlations between the ambient air pollutants, for example, whether the median value of PM2.5 is related with the median value of NO₂, CO and SO₂. Their results concluded that throughout the lockdown phase different sectors were shut down which resulted in improving the air quality in the city Delhi, India.

In a paper, Bera, Bhattacharjee, Shit, Sengupta, Saha (2020) analyzed the air quality of Kolkata throughout and before the COVID-19 lockdown of India by studying the concentration of six parameters, namely, SO_2 , CO , NO_2 , O_3 , PM_{10} and $\text{PM}_{2.5}$ [6]. They collected hourly emission levels of these pollutants from the State Pollution Control Board under the Govt. of West Bengal during the lockdown period from March 25th to May 15th, 2020 and similarly the same time period for the years 2017, 2018 and 2019 from different data stations. After the collection they analyzed the data, and calculated the monthly average and spatio-temporal variation of the pollutants to understand the changes in air quality.

Next, they performed statistical analysis like hierarchical cluster (HCA) and principal component analysis (PCA) to examine the similarity or dissimilarity between the pollutants and also identified the land surface temperature variation. From their analysis they found that during the lockdown phase (March 25th to May 15th) the average levels of CO , NO_2 and SO_2 remarkably reduced and the consistent dropping of PM_{10} and $\text{PM}_{2.5}$ was recorded in comparison for the same period from 2017 to 2019. On the other hand, the strength of the O_3 layer increased in Kolkata throughout the lockdown; the researchers also found a significant reduction of surface temperature during the lockdown compared with the previous years (2017 to 2019). Moreover, they showed using HCA and PCA that before the lockdown, period the pollutants O_3 , SO_3 and NO_3 had more or less similar values although PM_{10} , $\text{PM}_{2.5}$ and CO were correlated. On the contrary, during the lockdown period SO_2 had a remote correlation with other pollutants, while the other parameters behaved similarly. Lastly, on the basis of their analysis they suggested some environmental management plans which should be long-term sustainable.

Looking at relatable papers we came across an interesting paper where Srivastava, Kumar, Bauddh, Gautam, Kumar (2020) scientifically analyzed elements connected to air quality of two of the biggest cities in India [7]. New Delhi and Lucknow were their primary focus as the impact of lockdown on these busy cities would be the most significant. They collected the $\text{PM}_{2.5}$, NO_2 , SO_2 and CO data from four locations in Lucknow and ten locations in the mega city Delhi. They wanted to be precise about their research so they collected data of 21 days before (01/02/2020 to 21/02/2020) the first phase of the India's lockdown and 21 days after the first phase of lockdown was lifted (25/03/2020 to 14/04/2020). They computed the average and mean values of all 14 points of their observation and also calculated the AQI mathematically by assuming $\text{PM}_{2.5}$, NO_2 and SO_2 as the primary pollutants. To further recognise the impact of the long-range movements of atmospheric pollutants on both of the cities they also analyzed the air mass back trajectory. Their initial hypothesis was found correct according to their findings as in almost all the sites the AQI significantly declined after the lockdown, compared to what it was 21 days before the lockdown. The major impacts were noticed in the differences of $\text{PM}_{2.5}$, NO_2 and CO levels. SO_2 showed fewer compelling results. In order to counter the drastic air pollution upsurge, they concluded by suggesting the adoption of short periodical lockdowns.

In one of the papers, Madaan, Dua, Mukherjee, Lall (2019), introduced an online real-time air pollution prediction system at five locations in Delhi designed by using the past historic air quality data and meteorological data [8]. They used a real-time air quality monitoring dataset from the Central Pollution Control Board which con-

tains various parameters such as air pollutant variables (SO_2 , NO_2 , $\text{PM}_{2.5}$, PM_{10} , CO and O_3) and meteorological parameters (temperature, humidity, wind speed and barometric pressure). All these variables were collected on an hourly basis (every 4 hours) from the past 24 hours and these collected data sets were used to predict the concentration of $\text{PM}_{2.5}$, PM_{10} and NO_2 . In this paper, they proposed a BiLSTM model consisting of a BiLSTM layer, with the attention mechanism layer having four modules (Input Feature Module, BiLSTM module, Attention module and Output module). They trained this model to be able to forecast the pollutant values for the next 24 hours on the test set which was then used for the evaluation of different models. They developed an algorithm that proposed an adaptive method which minimizes the errors of the algorithm by using hourly data for each location.

Furthermore, real-time air quality data was collected from external sources by scraping the web pages of the Central Pollution Control Board each hour. Google Cloud storage was used to store this massive collection of data, on which there is also Google virtual machine and Google Machine Learning Engine support. Moreover, after this updated data was fetched from Google cloud storage, it was used as training data for the machine learning model on Google Machine Learning Engine every week for all the stations. The online trained model minimized the mistakes made by the predictive model by replacing the model that was stored in the Google cloud storage. This predictive model was used to predict air quality and classify the threat levels for the next 24 hours.

2.2 Studies Conducted Upon USA

One of the most prominent works we found was done by Wu, Nethery, Sabath, Braun, and Dominici (2020). Five Harvard students collected data of more than 3000 counties of The United States and analyzed their collected data using multiple methods to bolster their claims [9]. Their initial hypothesis was that the long-term slight increase in $\text{PM}_{2.5}$ is a primary reason for increased mortality rate due to COVID-19. The results they got from their research showed that a surge of just $1 \mu\text{g}/\text{m}^3$ in $\text{PM}_{2.5}$ can result in a rise of 8% COVID-19 mortality rate in those counties with 95% confidence. In their research, they did make some assumptions for the lack of proper data availability. The corona deaths for each county separately were not available so they divided the deaths according to the population proportions of each county. They calculated the $\text{PM}_{2.5}$ data of each county by using regression on the satellite data of the entire continental United States and averaging the values of each zip code of the counties. Apart from their initial analysis they made some secondary analysis to further support their claims. Since New York was the most heavily affected state, they did their analysis once without considering New York. They analyzed the data without considering counties with less than 10 confirmed COVID-19 deaths; they obtained consistent results.

2.3 Studies Conducted Upon Italy

While researching we came across a study that was done by Coker, Cavalli, Fabrizio, Guastella, Lippo, Parisi, Pontarollo, Rizzati, Varacca, Vergalli (2020) upon the northern regions of Italy [10]. They primarily used street level long term PM2.5 for measuring the correlation between PM2.5 and COVID-19 mortality rate. They calculated the increasing deaths in the first quarter of 2020 using the death data of the previous 5 years (2015 to 2019) at the municipality level of the same time frame and used it to scale the official COVID-19 deaths registries to get a better granular representation. They believed that 6 years was the perfect period of time to consider in this case and considered some confounding factors in their study like the urbanization and population density of different regions. They used negative binomial regression to find that for every unit increase of PM2.5 the COVID-19 mortality rate increases at 9% at a 95% confidence interval.

2.4 Studies Conducted Upon Other Countries

In a research paper, He, Pan, Tanaka (2020), used timely and extensive data collected throughout all the prefectural cities in China from 1,600 monitoring stations to investigate how much the air quality has refined due to the government steps taken for COVID-19 [11]. They merged the station level data to calculate the city level and then added the temperature, weather variables etc. from the Ministry of Ecology and Environment and National Oceanic and Atmospheric Administration. They analyzed weekly city-wise data of 324 cities between January 1st and March 1st and noticed that after the lockdown (January 23rd to February 11th) AQI reduced roughly 34%. Firstly, they used two types of DID models and with the help of baseline regression they determined the relative change in air pollution metrics between the treated and control cities. Moreover, in all the regressions they clustered the standard errors at the city level. After that, they performed a comparison on the air pollution data of the same period between 2019 and 2020 within the control group to see if there were any dissimilarities between the trends. Then they showed their results using figures where different panels showed before and after the Chinese Spring Festival, where the differences in AQI decreased more in cities that were locked down. In another panel they revealed that the air pollution levels were little low in 2020 after the festival compared to the 2019 post festival period. After the combined analysis they found that the AQI reduced by 19.4 points (18%) and PM2.5 by 13.9 $\mu\text{g}/\text{m}^3$ (17%). They also worked on the heterogeneous impacts of city lockdown for example, the impacts were more remarkable in colder cities and AQI was reduced around 20 to 30 points on other hand 0 to 10 points for warmer and southern cities. They also worked on the consequences of air pollution on death rate and used seasonal agricultural straw burnings as the main variable for PM2.5, and estimated how PM2.5 influences mortality and showed that the total averted untimely deaths would be around 24,000 to 36,000 which were higher than the casualty caused by COVID-19 in China.

In another research paper, Wang, Su (2020), worked on data from China by analyz-

ing the dynamic impact of COVID-19 on the environment [12]. Firstly, they focused on a part where the government took all necessary steps so that the outbreak could be suppressed for example, several travel and movement restrictions and tried to compare railway and traffic lots condition by the Satellite images from Planet Labs of NASA captured scenes of traffic and parking lots near Wuhan Railway Station pre-lockdown and post-lockdown (January 12 and January 28, 2020). Secondly, they analyzed that not only China's coal, crude oil, energy consumption decreased drastically but also GDP decreased by 5.3% over the same period last year due to a lesser number of vehicles on road and industrial activities during the quarantine period. Additionally, it was shown that the level of NO_2 in the air had decreased drastically through the data received from NASA and ESA satellites. Furthermore, through the TROPOMI sensor on the Sentinel-5 Precursor platform that keeps track of the global atmosphere daily compared 2005-2019 NO_2 concentration of China with 2020 data which revealed that in eastern and central China the NO_2 emissions were lower than the normal level during the same period in previous years. Thirdly, they evaluated that six types of air pollutants (NO_2 , CO, PM2.5, PM10, SO_2 , O_3) declined during this epidemic through the data of the Ministry of Ecology and Environment of China. Moreover, they explored that the above impacts were short-term and it was because of the factors related with the quarantine due to COVID-19.

We also came across a paper, where Cole, Elliott and Liu (2020), briefly researched the effects of lockdown on the air pollution in Wuhan city, The origin of the Coronavirus [13]. They did the research in mainly two steps. First, they used machine learning to clean their dataset and clear out redundant information from their data to get a more accurate result. Secondly, they used an Augmented Synthetic Control approach to estimate the impact of the lockdown on the 12-day post lockdown period since February 3rd. They analyzed the 4 main concentrations of air quality (PM10, NO_2 , CO, SO_2). Most cities in China did not lockdown for 2 weeks like Wuhan and other cities have different air quality natural behaviors than Wuhan due to other factors impacting air quality. So, the first step they took was they used machine learning to do a forest-based weather normalization technique of 30 cities in China to get their desired hourly weather normalized air quality data. Then they aggregated those hourly data into daily values. After that they used a (ridge) augmented synthetic control method on this dataset to estimate how the concentrations in Wuhan have changed relative to the synthetic control. By doing this they created a synthetic Wuhan city air quality dataset that did not get impacted by the lockdown since they used data from 18th January 2013 to 29th February 2020. So, after using a SCM approach to creating a synthetic weather normalized Wuhan city dataset, they compared what the air quality values would be from January 21st 2020 by contrasting values for the Wuhan actual weather normalized dataset and the synthetic Wuhan dataset. They found that NO_2 levels and PM10 levels have dropped significantly to a reduction of almost 63% of NO_2 and 35% of PM10 levels at the end of the 12-day period they did their investigation on. The SO_2 and CO differences they found were quite insignificant though. Lastly, they did two Placebo tests to justify their findings and show that their estimations of the synthetic Wuhan were accurate. To do that they used a different time in December assuming a fake lockdown and they made different cities into synthetic cities and ran their ASCM model those cities assuming a fake lockdown happened there too. In both of their

placebo tests they showed similar outcomes.

In a research paper by Korunoski, Stojkoska, Trivodaliev (2019), presented a pollution model using spatial interpolation which identifies the pollution field evolution and position of potential sources by adding pollution measurements and meteorological parameters [14]. The system architecture of this model is made up of four subsystems and each subsystem is responsible for different system operations. Firstly, the central subsystem, Spatial interpolation is used to calculate the pollutant field based on the used dataset. Secondly, the Sources Determination subsystem that depends on short-term variations in the air pollution and considers the weather by which this system identifies the pollutant sources. Thirdly, to solve the forecasting problem another subsystem used two stacked layers of Recurrent Neural Network (RNNs) and Long Short-Term Memory (LSTM) cells. Lastly, another subsystem Time-to-Event Prediction is used to predict the future pollution level and hours until it exceeds the thresholds. By integrating the four subsystems the pollution model measures the concentration of different pollutants such as carbon monoxide (CO), nitrogen dioxide (NO₂), or ozone (O₃), particulate matter (PM 10 and PM2.5) and sulfur dioxide (SO₂) for the city Skopje. This whole system architecture can be observed by a user through a web service.

Chapter 3

Analysis of Datasets

3.1 Collection

We collected data from an extensive list of sources in order to accurately carry out our analysis. In order to make the methods followed in our paper replicable, we included links to all of the websites that we obtained data from in the reference section.

To begin, an open-source dataset containing hourly and daily values (in micrograms per meter cube for all and milligrams per meter cube for carbon monoxide) for 10 different air pollutants including PM2.5, PM10, NO, NO₂, CO, SO₂, and O₃ was collected for a duration of 5 years (from the 1st of January 2015 till the 31st of June 2020) for 26 cities in India.

The city data was calculated by aggregating values from various air quality monitoring stations situated throughout India. The data was obtained from the Central Pollution Control Board under the Government of India [15] and used in our multivariate RNN LSTM model to forecast air pollution. Temperature data obtained from Reliable Prognosis Weather Archive [16] was also added to the extracted subsets of city data, to increase the number of input factors in our analysis.

Next, individual pollutant data was collected from the United States Environmental Protection Agency website [17], which we used to perform univariate LSTM forecasting and visualize what the air quality would have been like had the COVID-19 lockdown not been implemented in the United States or if it had been extended. Moreover, in order to conduct our analysis between historical air pollution and COVID-19 mortality rates, we obtained a dataset of county-wise PM2.5 concentration values from the year 2001 to 2016 published by the Centers of Disease Control and Prevention [18] in the United States. For the calculation of COVID-19 county wise mortality rates, we gathered county-wise COVID-19 cases and death data (up until the writing of this paper on the 22nd of December) for over 3000 American counties [19-21] as well as estimated mask usage data (a confounder in our analysis) which was divided into five categories from ‘never’ to ‘always’.

We also gathered data for a number of other confounding factors including county-wise population demographics (percentage of working-age people from 15 to 64 years old, percentage of children who are less than 14 years old, percentage of seniors who are over 65 years old, percentage of Black, White, Hispanic, and Asian people as well as percentage of males and females in each of these groups) in each county. All of this was collected from the official United States Census Bureau website [22-23].

We performed forecasting using RNN LSTM upon five years of air pollutant data (PM2.5, PM10, O₃, NO₂ and SO₂) in micrograms per meter cube for European cities as well. A dataset containing these values for 50 European cities was gathered from the Copernicus Atmosphere Monitoring Service (CAMS) [24] and used in our models. Finally, we collected COVID-19 cases and death data as well as population and air quality [25-27] for the country of Italy to use in the K-Means clustering model implemented in Chapter 6 of this paper. Again, all of the links to the sources of this data have been included in our paper for fellow researchers to continue work in this domain through our methods.

Before implementing our chosen deep learning models, we performed a wide range of statistical analyses to visualize the pollution data for the regions of our choice. To begin, in order to account for major changes in government policy (such as the quadrupling of parking fees in India during the year 2017 to discourage people from taking out their cars [28] and hence reduce air pollution), we decided to consider pollution figures from the year 2018 to 2020.

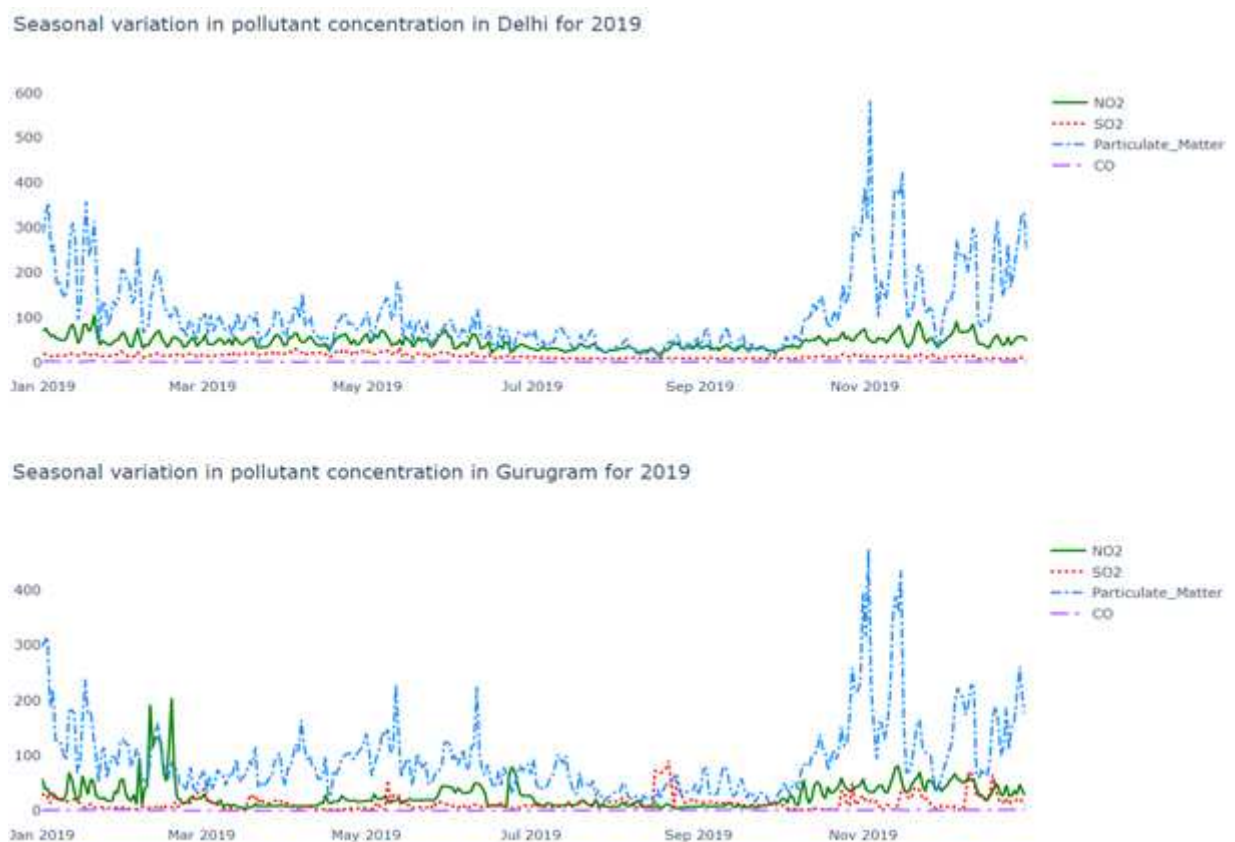


Figure 3.1: Seasonal variation of pollutant concentration for Indian cities

Furthermore, since the quantity of pollutants was seen to vary seasonally (as il-

lustrated in the graphs above), statistical comparisons were made only during the respective lockdown periods of each region which we outlined in Appendix A of this paper.

3.2 Methods and Results

In order to compare pollutant concentrations during the lockdown period in all three years, at first, we calculated the mean of each pollutant in the respective year. Next, the values obtained in 2020 were subtracted from the values in 2018, resulting in a difference table; the process was repeated for 2019 and the data visualized in the manner presented below.

It was overwhelmingly clear that the quality of the air in India during the lockdown period was far healthier than the past two years. The AQI values of 18 cities improved in 2020 compared to 2019, with Ahmedabad witnessing the biggest drop of 33.5% (426 ppm) within the years.

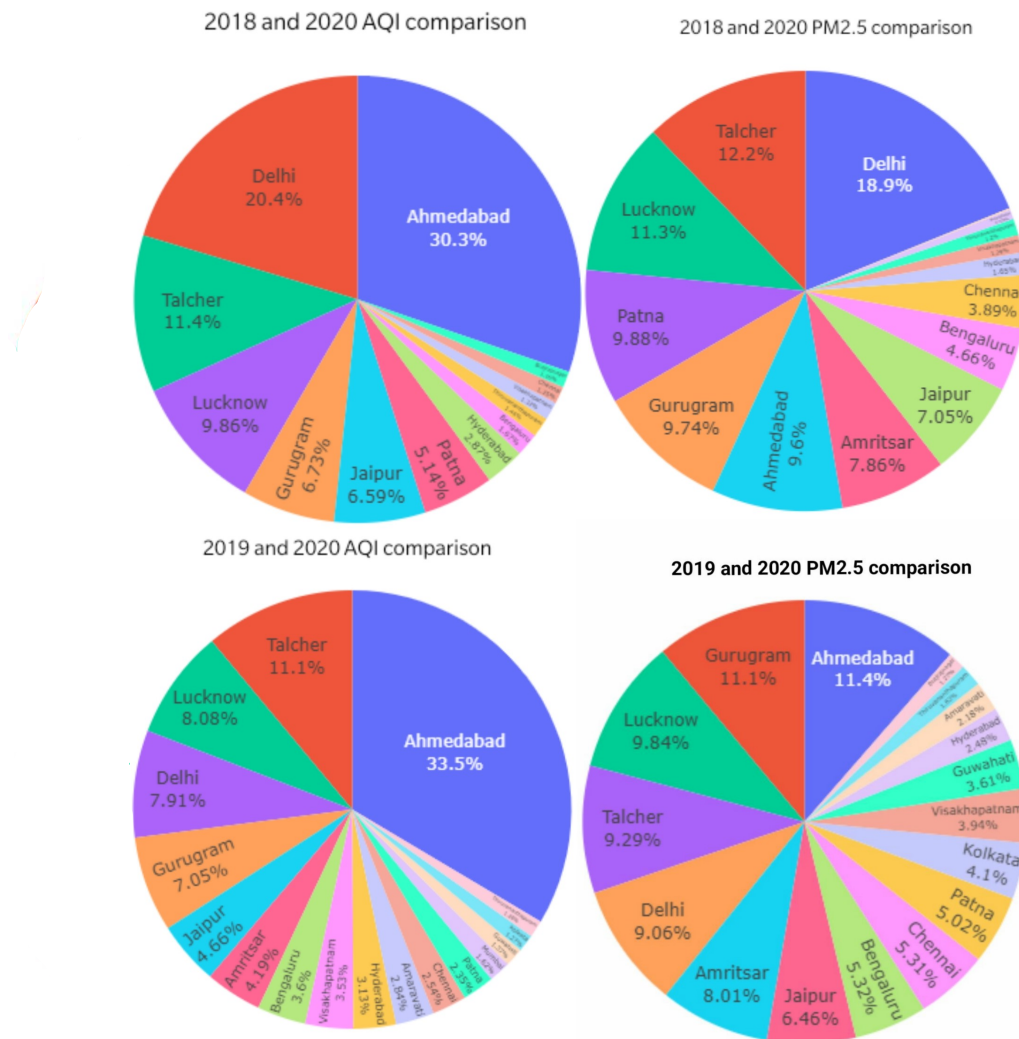


Figure 3.2: Percentage decrease in AQI and PM2.5 from 2018 and 2019 to 2020

Additionally, for the same city the AQI also shows improvement dropping from 486 units to 128 units (30.3%) from the year 2018 to 2020. For the city Talcher we see an improvement of AQI by 11.1% from 2019 and an improvement by 20.4% for the city Delhi from 2018 in the year 2020. From our results it is evidently clear that the AQI values have improved considerably during the lockdown for almost all the cities.

The PM2.5 charts of India also showed an improvement for most of the cities including Ahmedabad, Gurugram, Lucknow, Talcher, Delhi whose pollutant valued dropped by 11.4%, 11.1%, 9.84%, 9.29%, 9.06% respectively from 2019 to 2020. Ahmedabad again had the highest improvement in air quality with its PM2.5 pollutant concentration dropping from 76.48 ppm in 2019 to 28.8 ppm in 2020. The changes are even bigger from 2018 to 2020 since there is a huge drop of almost 20% in PM2.5 concentration in the city in that duration. Therefore, in correspondence with the AQI values, out of 25 cities, 18 witnessed a drop in the concentration of PM2.5, which is almost 80% of the entire dataset.

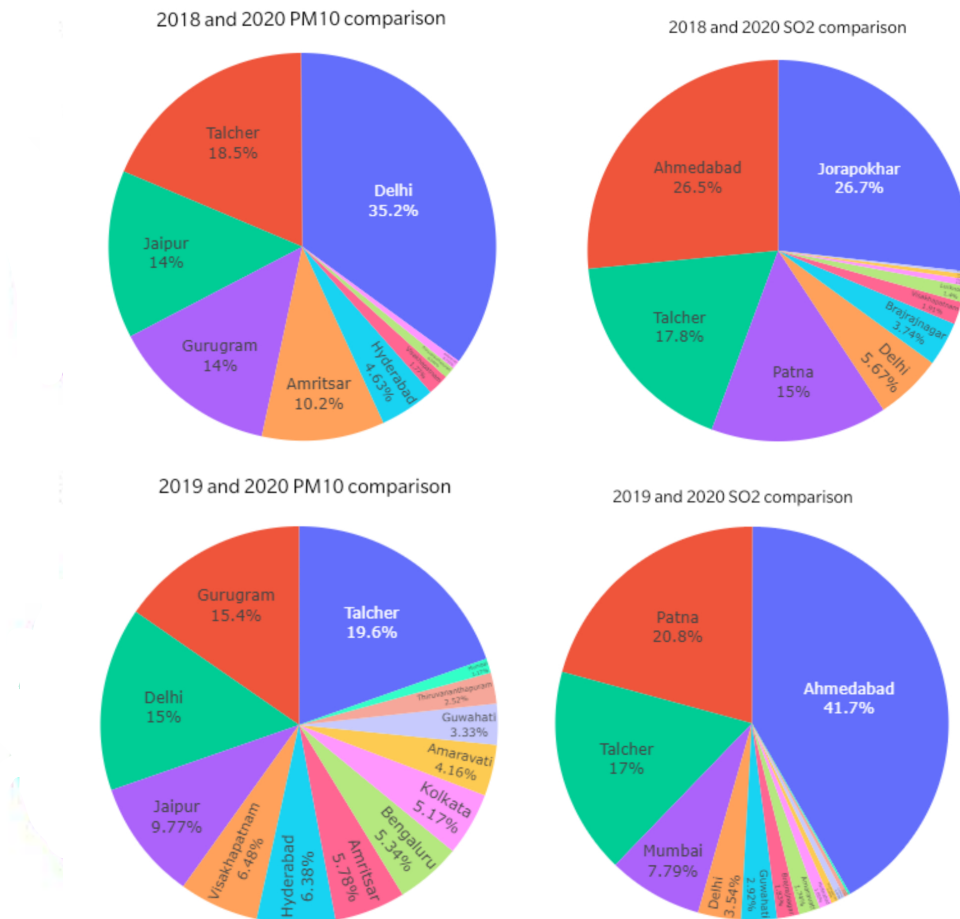


Figure 3.3: Percentage decrease in PM10 and SO₂ from 2018 and 2019 to 2020

For the country of India we constructed charts for the changes represented by all of major air pollutants in our dataset. From the above charts of PM10 for the cities such as Talcher, Gurugram, and Delhi we discovered a decrease in concentration by 19.6%, 15.4%, 15% compared to 2019 in 2020. Correspondingly, we see a decrease in concentration of PM10 for the city Delhi, Talcher, Jaipur, Gurugram respectively by

35.2%, 18.5%, 14%, 14% from 2018 in 2020. Furthermore, if we look at the difference of the mean values compared to 2018, in 2020 we notice an improvement for cities like Delhi, Talcher by 248.5 units and 131 units, and for 2019 by 120.3 units and 157.5 units. This result shows a remarkable improvement in PM10 values during the period of lockdown 2020.

The SO₂ charts for India also showed an improvement, the highest of which was for the city of Ahmedabad by 41.7% from 2019. Additionally, we see a decrease in concentrations for other cities such as Patna, Talcher accordingly by 20.8% and 17% from 2019 and for the cities Jorapokhar, Ahmedabad, Talcher, Patna respectively by 26.7%, 26.5%, 17.8% from 2018 which represents a considerable reduction of SO₂ concentrations in 2020.

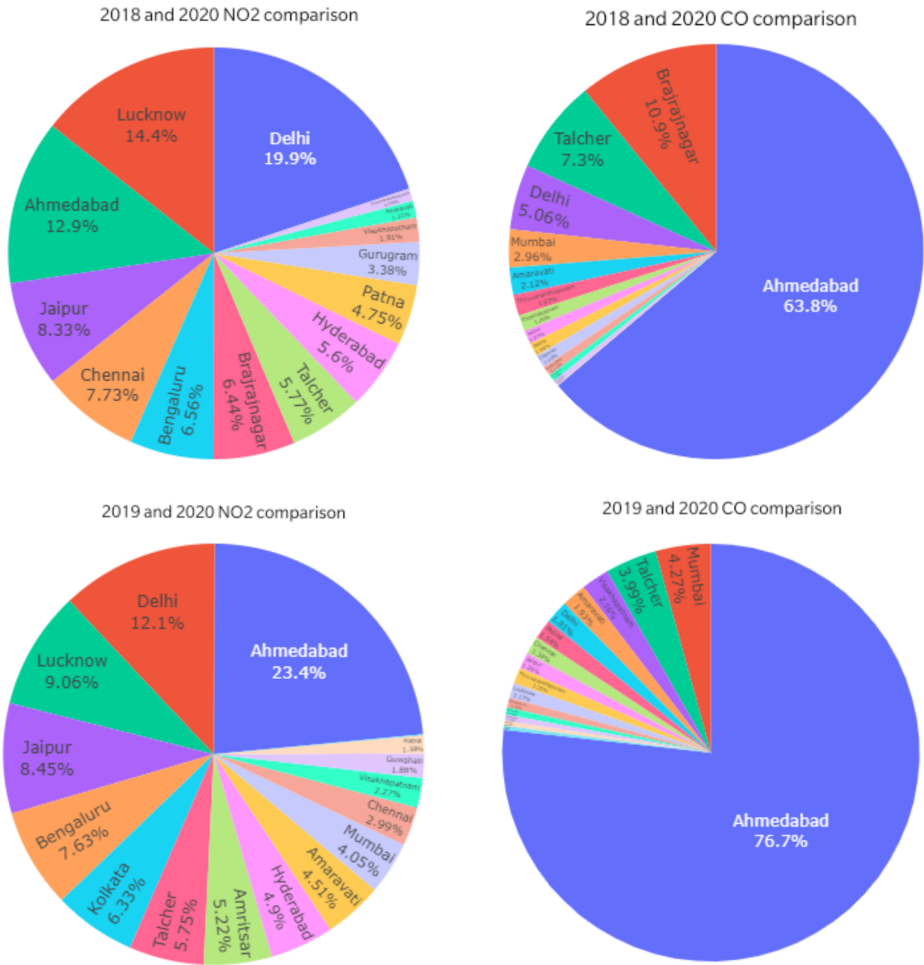


Figure 3.4: Percentage decrease in CO and NO₂ from 2018 to 2020 and 2019 to 2020

Finally, we present the charts for carbon monoxide and nitrogen dioxide emissions in the Indian cities and observe a reduction in the mean value of 22.88 units (76.7%) from 2019 and by 18.19 units (63.8%) from 2018 compared to 2020 for the city of Ahmedabad. In a similar way, we see improvements in the form of reduced NO₂ concentrations for cities like Ahmedabad, Delhi, Lucknow by 53.49 units (23.4%), 27.68 units (12.1%) and 20.71 units (9.06%) from 2019 and by 32.05 units (12.9%), 49.29 units (19.9%) and 35.64 units (14.4%) from 2018 accordingly.

Using the same methods as implemented for the Indian pollutant datasets, we observed the changes in pollutant concentrations for several counties in the United States and cities in Europe. We present our results in Chapter 6 in the form of heat maps using K-Means Clustering.

Statistical analysis of our accumulated datasets proves that the implementation of a nationwide lockdown did in fact improve air quality throughout many regions in the world. In the following chapter of our paper, we will analyze how extension of the lockdown could have further improved the air quality in these regions and forecast what the air quality would have been like had the lockdown not been implemented at all.

Chapter 4

Predictive Analysis Using Multivariate RNN and LSTM

A multivariate Recurrent Neural Network with Long Short Distance Memory cells was selected as the primary deep learning model to forecast our time series data. In this chapter we provide background about the model, detailed insight into our reasoning for its selection, and finally a thorough analysis of the results obtained.

4.1 Model Summary

Recurrent neural networks (commonly known as RNNs) are a popular class of neural network that allow the previous outputs of the model to be used as inputs using a set of hidden states. The feedback of information into the inner-layers enables RNNs to keep track of the information it has processed in the past and thus use it to influence the decisions it needs to make in the future. That is to say, RNNs have a memory which remembers all information about what has been calculated. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output; this in turn reduces the complexity of parameters, unlike other neural networks.

Some advantages of using RNNs include the possibility of processing inputs of any length where the model size is not increasing with the size of the input, computation taking into account historical information, weights being shared across time etc. However, there are considerable problems to this model too. First of all, general RNNs suffer greatly when it comes to processing data from far back into the past. A phenomenon known as ‘the vanishing gradient problem’ occurs in which the values used to update the neural network’s weights slowly declines with time. As a result, when the gradient becomes too small, the network stops learning; this usually occurs during the earlier layers of training, this making RNNs unsuitable for training historically lengthy data. Moreover, RNNs are extremely tasking to train and generally take up considerable amounts of time.

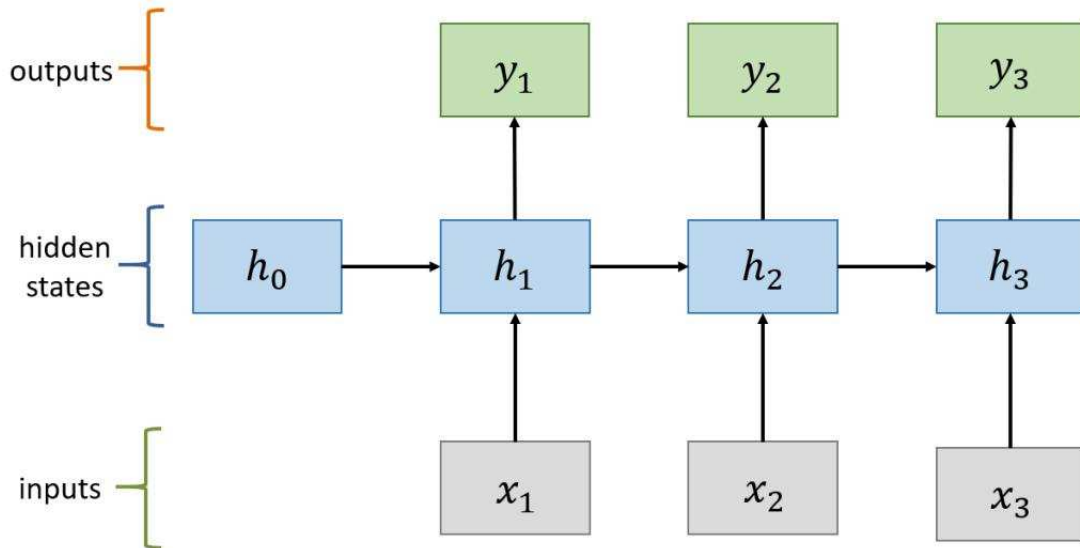


Figure 4.1: General architecture of a Recurrent Neural Network model

We decided to incorporate a specific kind of RNN called Long Short-Term Memory networks, (commonly known as “LSTMs”) into our work due to the advantage of the network learning long-term dependencies. LSTMs are explicitly designed to avoid the short-term memory problem which is generally faced by ordinary RNNs. They help preserve the error that can be backpropagated through time and layers. By maintaining a more constant error, they allow recurrent nets to continue to learn over many time steps (over 1000), thereby opening a channel to link causes and effects remotely.

LSTMs contain information outside the normal flow of the recurrent network in a gated cell. Information can be stored in, written to, or read from a cell, much like data in a computer’s memory. The cell makes decisions about what to store, and when to allow reads, writes and erasures, via gates that open and close. Unlike the digital storage on computers, however, these gates are analog, implemented with element-wise multiplication by sigmoid. Those gates act on the signals they receive, and similar to the neural network’s nodes, they block or pass on information based on its strength and import, which they filter with their own sets of weights. Those weights, like the weights that modulate input and hidden states, are adjusted via the recurrent networks learning process.

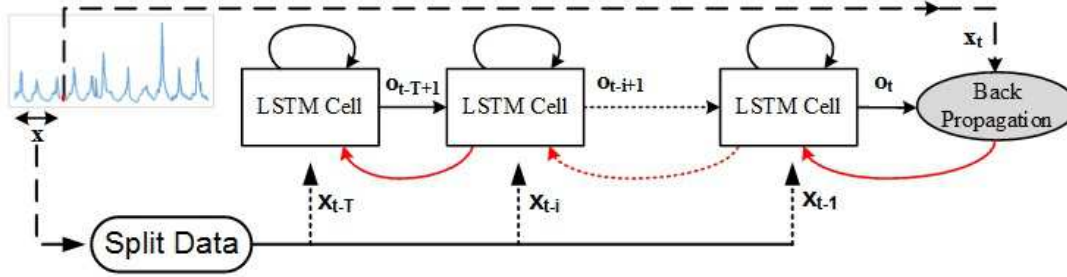


Figure 4.2: General architecture of a Long Short-Term Memory RNN

That is, the cells learn when to allow data to enter, leave or be deleted through the iterative process of making guesses, backpropagating error, and adjusting weights via gradient descent.

4.2 Method and Implementation

The aim of our model involving RNNs and LSTMs was twofold: first, to be able to forecast what the air quality in the regions of our choice would have looked like had the lockdown not occurred and second to forecast data in the event that the lockdown had been extended. We separated our analysis among two regions, that is, India and Europe. For the country of India, we already possessed data of 11 air pollutants as well as overall AQI over a 5-year period (2015 to 2020) and their subsequent meteorological data (temperature and wind). We researched the impact of these pollutants in the air, and chose to keep the five primary pollutants (PM_{2.5}, PM₁₀, NO₂, CO, SO₂) as training data in our analysis. From them, we performed multivariate RNN LSTM in order to forecast AQI values as well as the concentration of the primary pollutant in air, PM_{2.5}.

Our model consisted of 2 stacked LSTM layers with the first being an input layer. In order to forecast what the air quality in the cities would have been like if there was no lockdown, we chose to forecast 100 days (just over 3 months) into the future with a lookback value of 30 days into the past. That is, the AQI of the 31st day was forecasted using the values of the past 30 days as an input. The reason for choosing a duration of 30 days was to account for seasonal variations, which looking back further than a point would undoubtedly factor into consideration. It should also be noted that the values in our dataset were all assembled in a daily basis to do this.

Our first LSTM layer had 64 neurons whereas the second had 10; this one used the tanh activation function since it can overcome the vanishing gradient problem, is suitable for both positive and negative values, and can converge quicker. A dropout of 0.25 was added to account for overfitting and then finally, there was the output layer of the sequential model. It had only one neuron (since we have a single output AQI or PM_{2.5}) and had been given a linear activation function so that the outputs resemble the inputs as is suitable for a prediction problem. For compilation, the Adam optimizer was seen to produce the best fit with a learning rate of 0.01 and a loss calculation of mean squared error.

In this manner, we conducted training for each selected city/county, being careful to train each model according to the specific lockdown date of a region. For example, in the United States, the state of Colorado maintained stay-at-home orders from the 26th of March 2020 until the 26th of April of the same year; in contrast, New Hampshire observed stay-at-home orders from the 27th of March until the 11th of June. Therefore, when conducting analysis of an American county situated in Colorado, the model will train up to and begin forecasting from the 26th of March, whereas a county in New Hampshire will train up to and begin forecasting from the 11th of June.

Finally, once the forecasted values were obtained, for every city we calculated the percentage difference in real and forecasted value for each day. We then calculated the mean of the percentage differences, and tabulated the results for every region in our analysis.

Below is a flowchart depicting the entire process:

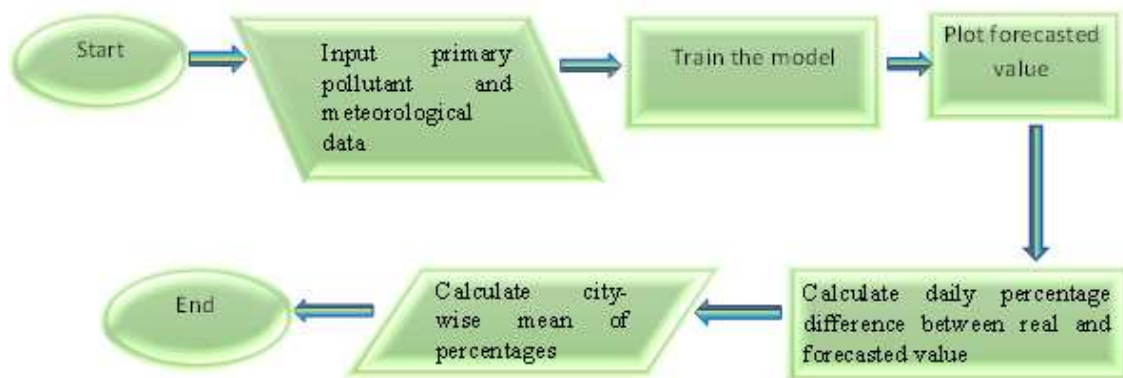


Figure 4.3: Flowchart depicting LSTM process

4.3 Results

We began by visualizing the Air Quality Index (AQI) values of the 26 cities in India had their COVID-19 lockdowns not been implemented. The graph below illustrates our results for the city of Ahmedabad:

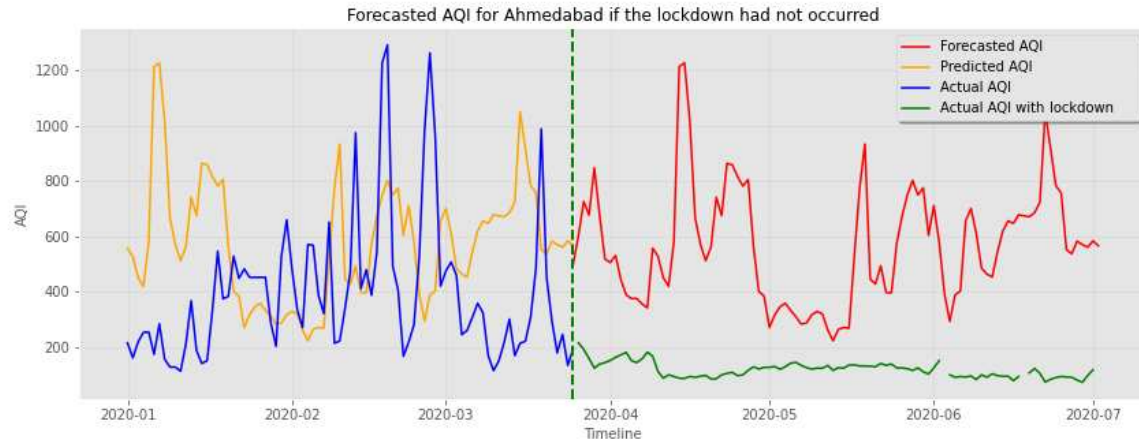


Figure 4.4: Air quality in Ahmedabad with and without the COVID-19 lockdown

Four lines have been plotted and clearly marked on the graph. As listed in the key, the blue line represents the AQI values in Ahmedabad from the beginning of 2020 until the onset of the lockdown. Our model trained on these values and the yellow line represents the values it predicted from the training set. The dotted green line indicates that the 25th of March represents the beginning of the lockdown period; the green line after is the true AQI of Ahmedabad that was recorded during the lockdown. Finally, we have the red line: this represents the forecasted values printed by our model assuming that all other factors remained constant (and accounting for seasonal variation) if the COVID-19 lockdown had not been implemented by India. Gaps in the graph indicate a region where there were gaps in the test data.

It can be deduced from our results that the air quality in Ahmedabad was much better during the lockdown period than it should have been in that frame of time. Even the smallest difference between the forecasted and actual values is approximately 100 ppm and occurs on the 12th of May. On the India AQI chart, this is a drastic improvement from the ‘Poor’ air quality zone (forecasted) to the ‘Moderate’ air quality zone (real).

AQI Category	AQI	Concentration range*							
		PM ₁₀	PM _{2.5}	NO ₂	O ₃	CO	SO ₂	NH ₃	Pb
Good	0 - 50	0 - 50	0 - 30	0 - 40	0 - 50	0 - 1.0	0 - 40	0 - 200	0 - 0.5
Satisfactory	51 - 100	51 - 100	31 - 60	41 - 80	51 - 100	1.1 - 2.0	41 - 80	201 - 400	0.5 - 1.0
Moderately polluted	101 - 200	101 - 250	61 - 90	81 - 180	101 - 168	2.1 - 10	81 - 380	401 - 800	1.1 - 2.0
Poor	201 - 300	251 - 350	91 - 120	181 - 280	169 - 208	10 - 17	381 - 800	801 - 1200	2.1 - 3.0
Very poor	301 - 400	351 - 430	121 - 250	281 - 400	209 - 748*	17 - 34	801 - 1600	1200 - 1800	3.1 - 3.5
Severe	401 - 500	430+	250+	400+	748+*	34+	1600+	1800+	3.5+

* CO in mg/m³ and other pollutants in µg/m³; 2h-hourly average values for PM₁₀, PM_{2.5}, NO₂, SO₂, NH₃, and Pb, and 8-hourly values for CO and O₃.

Figure 4.5: National Air Quality Index of India

We repeated this process for every city in India that was present in our dataset, and obtained similar results. Below is another graph depicting the changes forecasted for the city of Visakhapatnam:

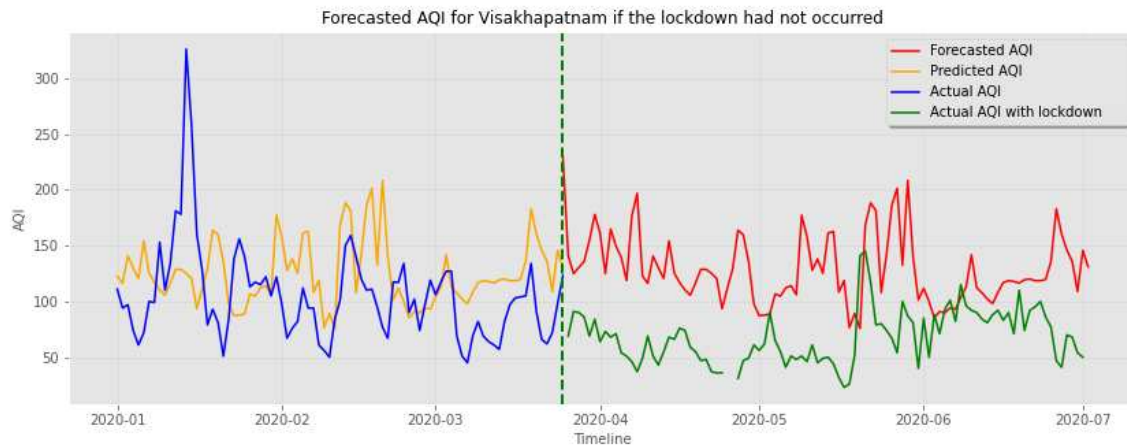


Figure 4.6: Air quality in Visakhapatnam with and without the COVID-19 lockdown

Similar to the graph obtained for Ahmedabad, we again see that the air quality would have been much worse during the India COVID-19 lockdown period in ordinary circumstances. To get a numerical measure for this, we calculated the mean of the percentage difference between daily forecasted AQI and daily real AQI during the lockdown period. For Visakhapatnam this value was 48.3%; for Ahmedabad, it was a shocking 75% improvement in air quality as a result of the lockdown. The mean percentage change in AQI value for every Indian city has been tabulated and included in Appendix B.

Next, we forecasted PM2.5 concentrations in the event that the COVID-19 lockdown in India had not occurred; the reasoning behind choosing this pollutant is that it's the biggest contributor to overall air quality in a region. Excess concentrations of PM2.5 can cause a myriad of health issues including by corroding the alveolar wall of lungs and impairing lung function [31].

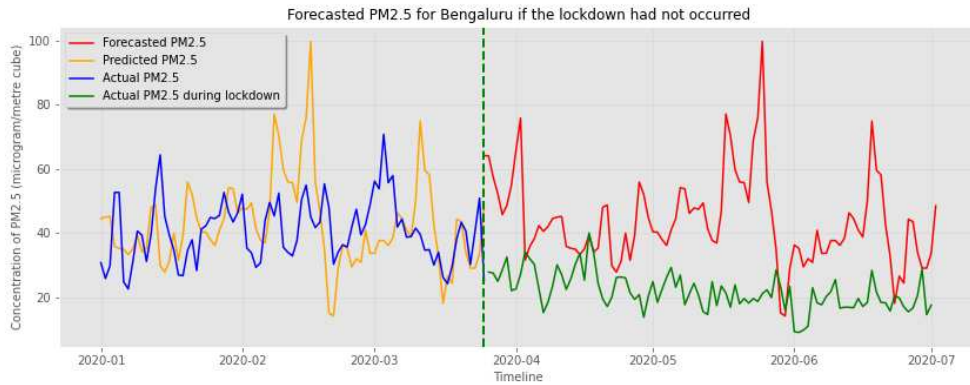


Figure 4.7: PM2.5 levels in Bengaluru with and without the COVID-19 lockdown

The line graph above depicts our results for the city of Bengaluru; all of the color codes have been kept the same. Similar to our graphs forecasting AQI, we observed that PM2.5 values would have also been higher had the COVID-19 lockdown in India not been implemented. The mean forecasted value of PM2.5 with no lockdown was calculated to be 49% lower than the actual value. Below is a map depicting the percentage increase in PM2.5 values if the COVID-19 lockdown had not occurred for every Indian city (aggregated state-wise):

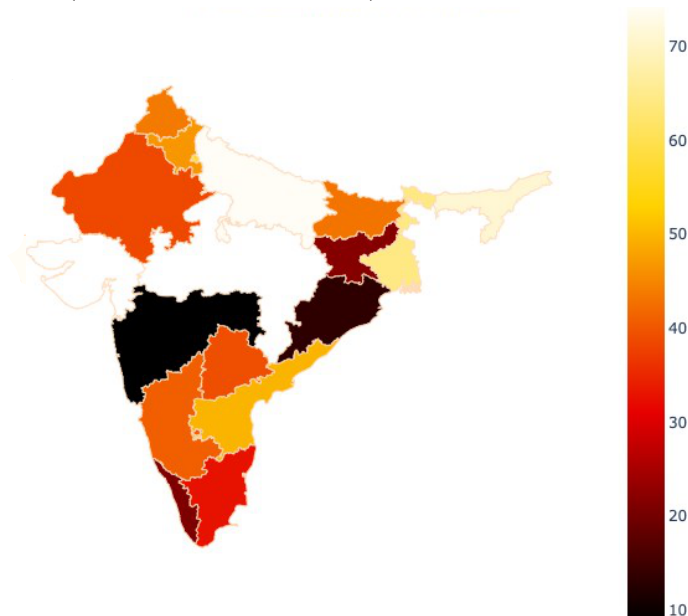


Figure 4.8: Percentage increase in PM2.5 values if no lockdown in India

The second part of our analysis focused on another pressing question in the study between air quality and lockdowns: what if the COVID-19 lockdown had been extended? In that case would the air quality have been even better? Below is our result in forecasting AQI for the city of Lucknow:

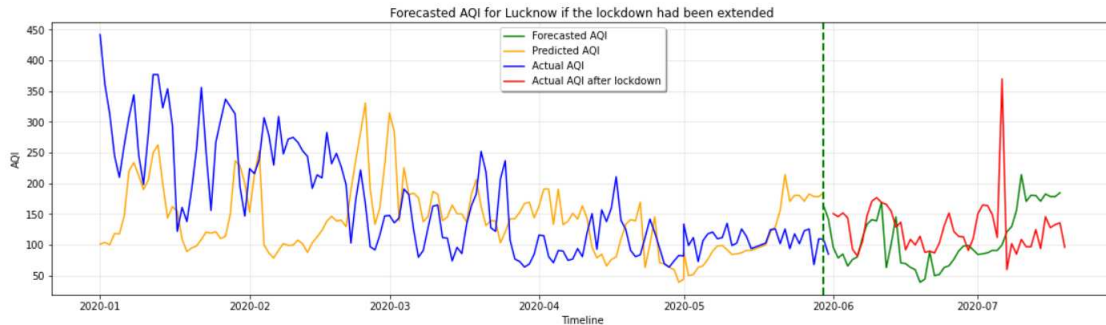


Figure 4.9: Air quality in Lucknow if the COVID-19 lockdown been extended

We switched the color codes in this graph, so that the red line depicts the actual AQI in Lucknow after their lockdown period ended on the 31st of May, and the green line depicts the forecasted values of our model. In terms of the National Indian Air Quality chart, if the lockdown had not occurred, Lucknow’s overall air quality would have been in the ‘Moderated Polluted’ zone; the reality was far worse since instead the air was in the ‘Poor’ zone. The cause of the unexpected fluctuation can again be hypothesized with the end of the 2-month lockdown; as stores and jobs open, the number of vehicles on the roads increase and the citizens again venture outside, the quality of the air worsens.

The results of forecasting air quality if the lockdown had persisted beyond the month of May, indicate that the air quality would have continued in a similar trend and would have been healthier for the citizens of India. However, due to the restrictions being lifted, this was no longer the case; instead, the real air quality was approximately 30-50 units higher than it would have been. By the beginning of July, we can see a drastic spike in the readings where the actual AQI value is 375 units which is 273 units higher than the forecasted pollution value at the same date.

Our results for the country of India clearly depicted an improvement in the city’s air quality if the lockdown had not occurred; out of the 26 cities we tested, 20 of them showed improvement air quality (both AQI and PM2.5) due to the lockdown. However, the results were not as extensive when we forecasted values assuming that the lockdown was extended. Therefore, before drawing conclusions from our analysis, we repeated our process on a dataset containing air quality data of 50 European cities.

City	Estimated increase in forecasted PM2.5 with no lockdown (%)
Paris	45.295328
Milan	55.551869
Naples	34.436787
Rome	41.556863
Turin	62.464479
Nicosia	29.045986

Table 4.1: Mean Percentage Decrease in PM2.5 due to COVID-19 lockdown in Europe

Above, we display a portion of the table summarizing the mean percentage increase in forecasted PM2.5 concentration assuming that the COVID-19 lockdown had not occurred, and have included the entire table in Appendix B.

Among the 50 cities we analyzed, 38 displayed characteristics in concurrence with the trend we obtained for the Indian cities. The lockdowns implemented in these cities reduced their PM2.5 values by an estimated 40% than what they usually would have been. For example, in the graph of Barcelona we see that the forecasted lockdown PM2.5 concentration is higher than the real one on several occasions, especially after the last week of May when the forecasted values lie almost continuously above the predicted ones. For both regions, we noted a unique characteristic in the analysis, in the fact that the results coincided more with the hypothesis of the air pollution being worse if the COVID-19 lockdowns had not been implemented.

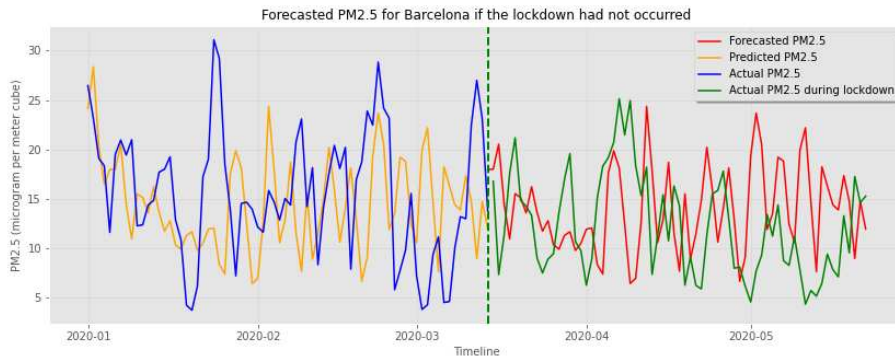


Figure 4.10: Forecasted PM2.5 for Barcelona if lockdown had not occurred

Our second investigation regarding whether the air quality would have remained healthier if the lockdown had been extended, yielded fewer results. For example, the graph we obtained for an extended lockdown in the city of Lisbon in Portugal shows that the forecasted PM2.5 (in green) would have been less than the real, no-lockdown values from the 16th May to the 3rd of June, with few exceptions. On the rest of the points however, the real lockdown values are lower than the forecasted ones.

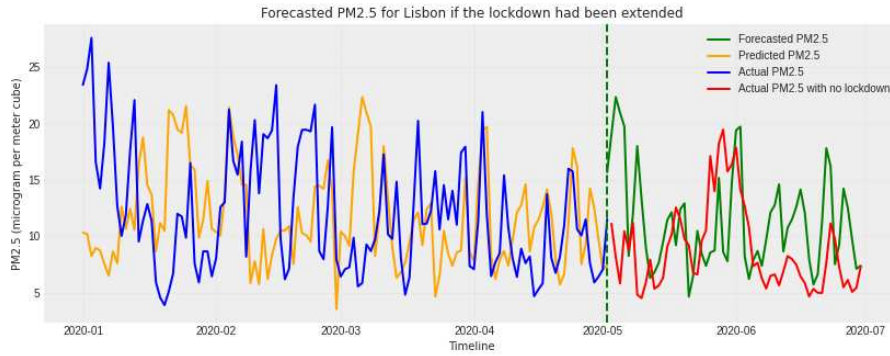


Figure 4.11: Forecasted PM2.5 in Lisbon had the lockdown been extended

The same can be said for the air quality graph of an extended lockdown in London that's shown below. Even though there are places where an extended lockdown appears to have lowered forecasted PM2.5 values, in the majority of places, the real values appear to still host healthier air. After analyzing our findings, we correlate this trend to the possibility that even though the lockdowns were officially announced to be ended at specific dates such as the 31st of May, even after that period, residents in cities continued to obey stay-at-home and quarantining measures, which continued to contribute to the further improved air quality depicted in our results.

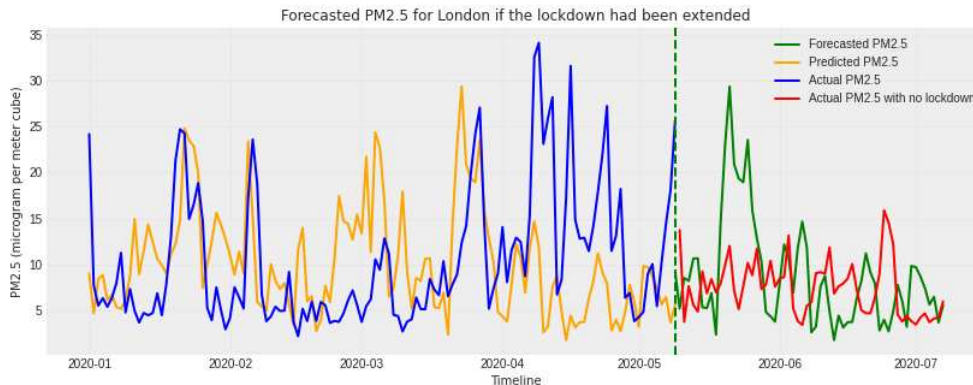


Figure 4.12: Forecasted PM2.5 in London had the lockdown been extended

4.4 Evaluation

According to a research paper published by Armstrong (2001) [32], there are several ways to evaluate the quality of a forecasting model. First of all, there's replication: repeating the procedure on different datasets with the exact same conditions and then drawing up a comparison of the results obtained. This is why we chose to conduct our LSTM forecasting on so many cities (total of 76) located throughout the country of India and continental Europe. The results obtained from both forecasts are uniform, with air quality forecasted to be better if the lockdown had not occurred for the majority of cities, and the air quality forecasted to be better if the lockdown

were extended for some cities. The latter forecast was not observed as widely, which we correlated to the possibility of people still maintaining stay-at-home orders even after their period of lockdown ended (thus real air quality values continued to improve and the model’s forecasted values tended to be higher in some cases instead of lower).

Next, we evaluate potential biases in our study. Before conducting our research, we had initially thought that the forecasting would yield worse pollutant values without a lockdown and better ones if the lockdown were extended; for the majority of the cities, this turned out to be true. However, in order to eliminate potential researcher bias, we decided to showcase the entirety of our results in Appendix B., even the cities which did not follow the overwhelming trend.

Another method of evaluation is to compare the performance of our chosen model with other popular forecasting algorithms to check whether the performance is better. We did this with the simple linear forecasting model in which the output of each time step is independent of the other ones; although this produced some accurate values, it was not feasible for the 5-year long time-series data that we worked with. Single step models cannot learn the shape and characteristics of the data, which was an essential requirement in our study.

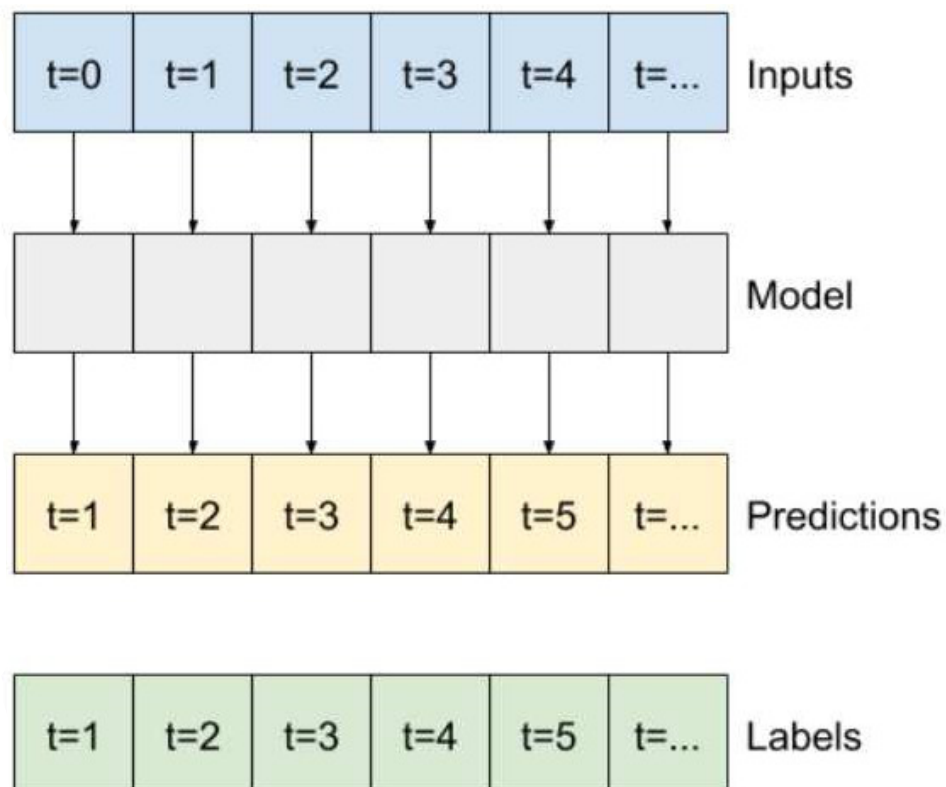


Figure 4.13: Block diagram of a Linear Forecasting Model

One of the main assumptions in our study is that the model, upon training of so many years of data, will learn to accommodate seasonal changes in air quality and forecast accordingly. LSTM RNN’s were the right way to do this due to their

ability to store essential information regarding the dataset. Using the multivariate procedure, we could train the model for several contributing factors like pollutant concentrations and temperature in order to get our final output.

There were several conditions in our problem statement which we maintained rigidly. For example, individual lockdown dates were gathered for every city in our dataset and each time, the model was trained up to the onset of each specific city's lockdown. We excluded cities in our model that had over 10% missing values, and replaced missing values with the arithmetic mean for the remaining dataset. We included weather data such as temperature and wind speed in our analysis as well to compensate for meteorological factors.

Finally, we have included all of the sources of our data in the bibliography section of this paper; this is to simplify the replication of our procedures for future researchers.

Chapter 5

Correlating COVID-19 and Air Quality with Linear Models

Studies have been conducted in various regions around the world, dedicated to analyzing the correlations (if any) between air quality and COVID-19 mortality rates. After gathering data upon over 40 potential confounding factors as well as cumulative COVID-19 related death counts for 3000 counties in the United States, we also decided to implement correlating the variables. Our methods and results are outlined in the sections below.

5.1 Methods

To begin, we gathered historical county-wise PM2.5 data extending from the year 2001 to the year 2016 for 3000 counties in the United States. This was then averaged city-wise to find a mean historical value of the air quality in the specific region. Current populations were also obtained for these regions, as well as cumulative death counts. Our mortality rate calculation is:

$$M = \frac{d * 1000}{p}$$

Where,

M = County-wise Mortality Rate (per thousands of people)

d = Cumulative COVID-19 deaths

p = Population

Directly plotting the PM2.5 concentrations against the Mortality Rates was not a feasible option since there are many confounding factors which could be skewing our analysis. For example, research has show that seniors are more likely to be hurt by

the COVID-19 pandemic [33]; therefore, if a county has an increased percentage of the senior demographic, a graph plotting PM2.5 against mortality rate may show a skewed value for the county's data and report higher mortality rates as a result of the age confounder.

In order to account for this, we began by at first conducting sensitivity analyses to identify potential confounders in our dataset; we already had values for approximately 43 confounders, whose data we provide in the bibliography section of this paper. In line with ideal statistical convention, our procedure to identify a potential confounder in the dataset involved performing linear regression of each potential confounder and then observing the change in the regression coefficient of PM2.5 brought about by the confounder. If the change was greater than or equal to 10% in either direction, we identified the factor as a confounder in our model. In this way, we identified and adjusted for significant confounding factors in our model, including, the percentage of young people in a county (aged 14 or younger), the percentage of seniors in a county (65 years old or older), the ratio of cumulative COVID-19 cases (cumulative counts divided by population), the percentage of working aged male people (between 14 to 65 years old) who were racially black, the percentage of male seniors who were racially black, the percentage of people in each county that always wore masks, the percentage of people that never wore masks, and more. The tabulated figure below represents the 95% confidence interval values as well as the p-values for all of the variables involved in our analysis:

Names	Estimate	SE	95% Confidence Interval		β	df	t	p
			Lower	Upper				
(Intercept)	0.9575	0.01005	0.93775	0.9772	0.00000	2272	95.2898	< .001
mean_pop_25	0.0165	0.00641	0.00389	0.0290	0.05438	2272	2.5685	0.010
case	0.0116	5.26e-4	0.01056	0.0126	0.48412	2272	22.0247	< .001
Y_TOT_POP	-5.4163	1.27338	-7.91335	-2.9191	-0.25443	2272	-4.2535	< .001
O_TOT_POP	15.3191	1.80668	11.77620	18.8620	1.18089	2272	8.4792	< .001
W_BA_MALE	-1.0744	0.14661	-1.36188	-0.7869	-0.16993	2272	-7.3283	< .001
O_BA_MALE	0.4354	0.25693	-0.06841	0.9393	0.03841	2272	1.6948	0.090
O_TOT_MALE	-21.5600	1.90780	-25.30124	-17.8188	-1.69674	2272	-11.3010	< .001
ALWAYS	0.1193	0.17001	-0.21408	0.4527	0.03096	2272	0.7018	0.483
NEVER	-0.6816	0.25942	-1.19033	-0.1729	-0.06645	2272	-2.6274	0.009
RARELY	-0.2366	0.28555	-0.79660	0.3233	-0.02233	2272	-0.8287	0.407
SOMETIMES	-0.3843	0.26106	-0.89621	0.1277	-0.03869	2272	-1.4720	0.141
W_BA_FEMALE	0.0271	0.12919	-0.22622	0.2805	0.00401	2272	0.2099	0.834
W_WA_MALE	-7.5701	1.09215	-9.71182	-5.4284	-0.51140	2272	-6.9314	< .001
O_H_MALE	2.7592	0.50878	1.76147	3.7569	0.14358	2272	5.4232	< .001
O_H_FEMALE	0.0349	0.40130	-0.75204	0.8219	0.00224	2272	0.0870	0.931
W_H_MALE	0.1486	0.22438	-0.29143	0.5886	0.01559	2272	0.6622	0.508
W_H_FEMALE	-0.7225	0.24820	-1.20920	-0.2357	-0.06636	2272	-2.9108	0.004

Figure 5.1: Linear Model Summary for co-variates and Mortality Rate Analysis

Our model deemed concentration of PM2.5 as a statistically significant factor in our analysis with an RMSE of 0.33.

5.2 Results

Below is a graph depicting the correlation with unadjusted historical PM2.5 values with mortality rate for pollutant concentrations less than 7.5 micrograms in value:

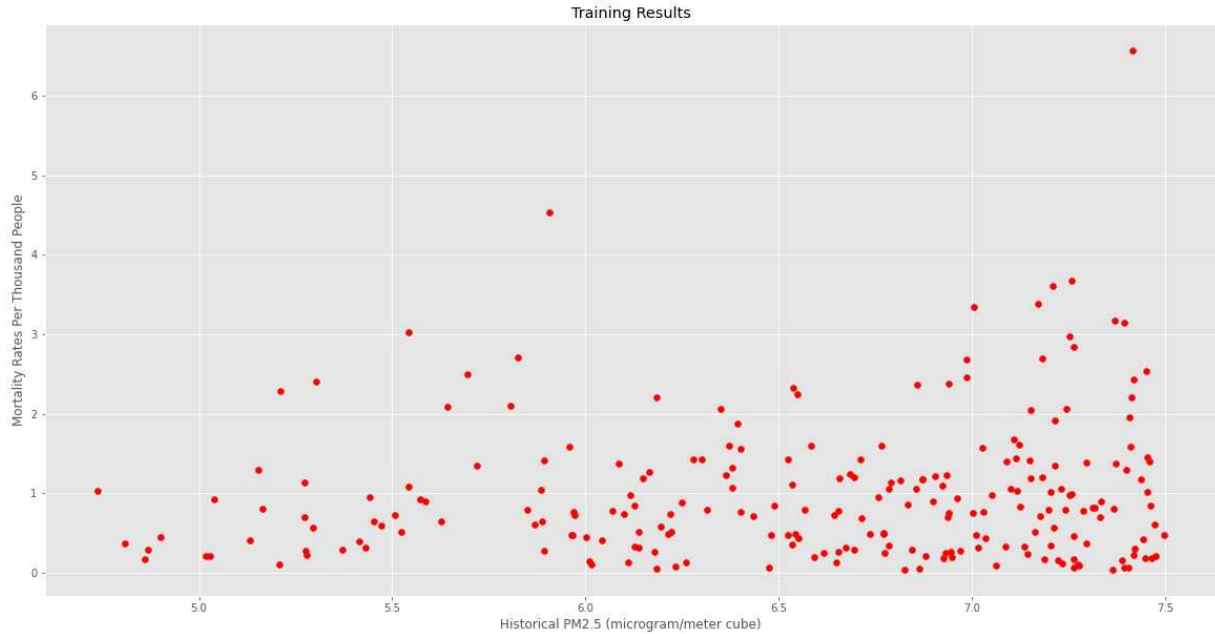


Figure 5.2: Correlation of PM2.5 concentration with COVID-19 mortality rates

We adjusted the raw values by regressing them out; we multiplied a matrix of confounding variable values with a corresponding matrix of regression coefficients.

$$X(\text{adjusted}) = X - C\beta$$

Where,

C = Matrix of confounders

β = Matrix of regression coefficients

Keeping mortality rate as the dependent variable, we then used our remaining variables as inputs into a linear model, to observe its correlation with increasing mortality rates. Our plot indicates a uniquely positive linear relationship between the two factors. From the model summary information, we deduced that a single unit increase in PM2.5 concentration can cause a 3% increase in the number of people likely to perish from COVID-19 at a 95% upper bound confidence interval.

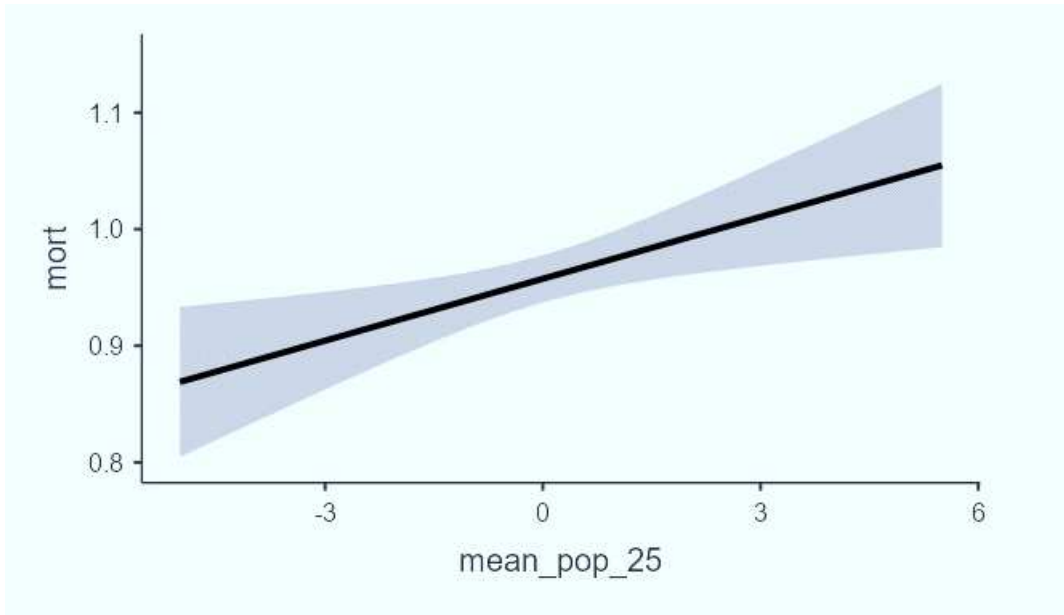


Figure 5.3: Scaled plot of adjusted PM2.5 and COVID-19 Mortality with Linear Models

The limitations of a model should always be addressed and in this case, it is no different. Our results displayed a direct correlation between adjusted PM2.5 values and mortality rates for the 43 confounding variables we tested for and took into account. However, since data limitations regarding, for example, the number of ventilators at the hospitals of each county existed, we envision that the model could be further improved when confounder-testing is repeated at a time where more data is available. Another point to note is that although we used historical county wise pollution data, the model could be even further refined if city-wise data in the USA was collected and the model repeated with city-wise mortality rated instead. We believe that soon, when the statistics regarding COVID-19 have solidified even further, our model can be directly replicated to both confirm and build upon our results.

Chapter 6

Factor Visualization With K-Means Clustering

6.1 Model Summary

Our main objective in forming clusters was to find the similar trends of air pollution in correlation with COVID-19 cases and mortality based on pollutants concentration. Since the pollutants were not dependent on each other and had raw concentration levels, we used an unsupervised clustering algorithm to form clusters of data. Out of all the unsupervised algorithms we used the K-means Clustering algorithm [34] to determine the clusters based on the parameters we selected.

Clustering is defined as the process of classifying an assortment of objects into different groups. Put simply, it involves dividing a dataset into smaller subsets known as clusters, in such a way that data in a single subset shares some common traits. The division is usually done based on a pre-defined distance measure [35].

6.2 Method

We formed clusters based on some key factors. Using these factors we prepared the dataset of key pollutants for some major areas in order to identify which cluster they belonged to. Using mathematical methods, we identified the key points of the intended result for the algorithm.

The whole pipeline for our cluster analysis is given below:

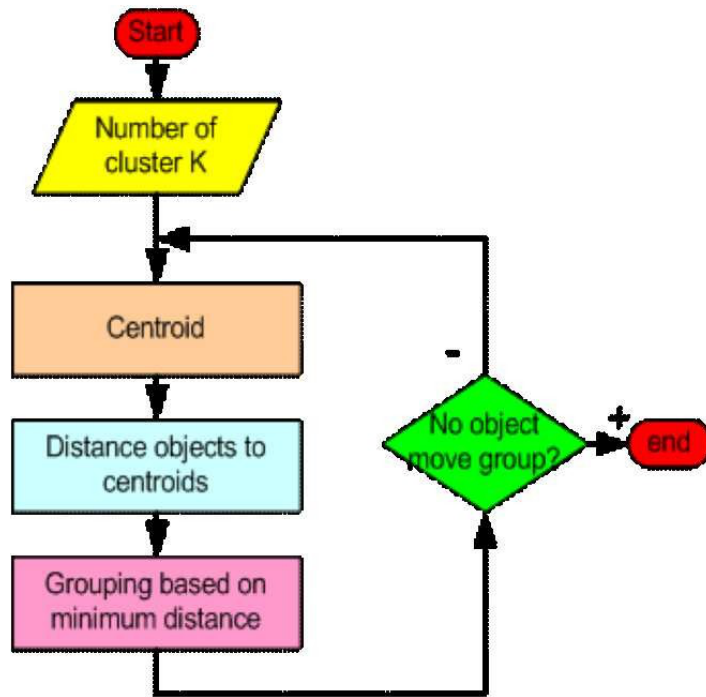


Figure 6.1: Flowchart of the K-Means Clustering Model

Our first step for identification was the number of clusters to be used. The identification was done with respect to the concentrations of a pollutant for a given number of areas. Then by applying K-means clustering we identified which data point belonged to which cluster and then grouped them accordingly. For a more significant cluster, we applied the algorithm after taking confounding variables into consideration (identified through sensitivity analyses) and adjusting for them.

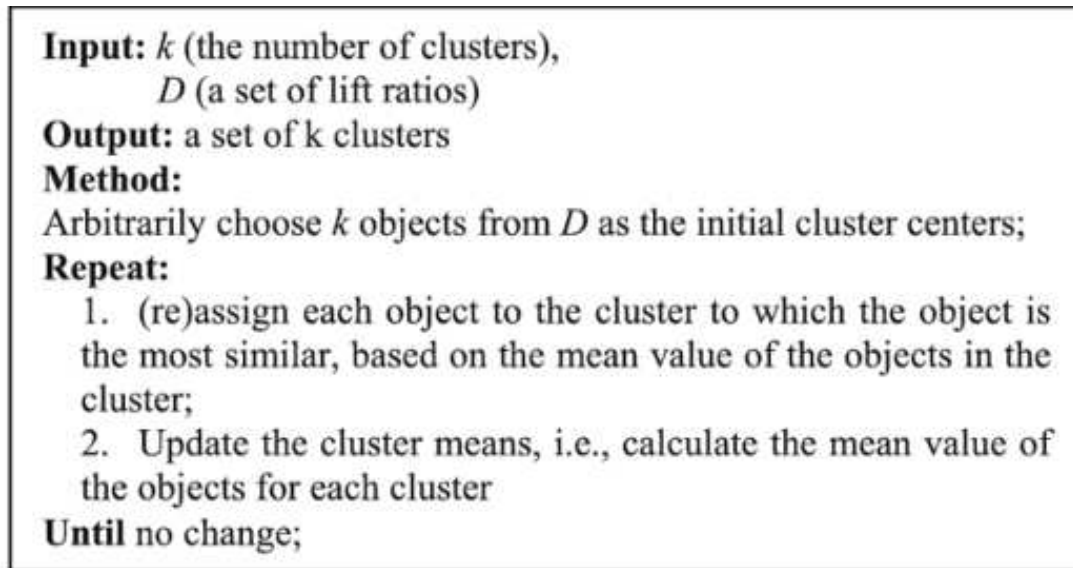


Figure 6.2: Algorithm of the K-Means Clustering Process

After preparing the dataset we then applied the K-means algorithm. Using the K-means function in scikit-learn [36] library we clustered out the data points. The two main steps of the algorithm were:

1. **Initialization:** We began with the decision on the number of clusters k . By the help of the Elbow Method, we identified the required number of clusters, the algorithm selects k centroids at random for identifying the next clusters. These centroids were then used for the calculation of the distance measure and each datapoint was assigned to a cluster.

Next, the centroids were used to calculate the distance from each data point as discussed below.

2. **Quantization:** For the distance calculation we use the Euclidean distance between the centroids and the subsets

$$d(i, j) = \sqrt{\sum_{i=0}^m (x_i - y_i)^2}$$

To minimize the Sum of Squares Error (SSE) the objective function becomes,

$$J(\alpha) = \sum_{i=0}^k \sum_{n \in S_j} |x_n - \mu_j|^2$$

Where,

k = number of disjoint subsets

x_n = vector representing the n^{th} data point

μ_j = centroid of the data points in subsets S_j

After determining closeness to a centroid by Euclidean distance (SSE), the data points were then grouped based on the minimum distances from the centroids and then again after updating the centroids. Thus, we repeated until convergence is achieved. The result depicted a list of data points identified by the clusters through which we understood similar trends within the data points and correlated different attributes to it.

6.3 Implementation and Analysis

In the case of USA, we aimed to gain a fresh perspective on the COVID-19 issue by correlating historical pollutant data (PM2.5) with that of Covid-19 cases and deaths through clustering. After preparing the dataset by doing sensitivity analyses, we applied the algorithm for k=4 number of clusters. At first, we tried to find the clusters of mortality rate and their relation with historical long-term exposure to PM2.5.

2020 US County Wise Mortality Rate
Clustered by PM2.5 Exposure and Covid-19 Cases

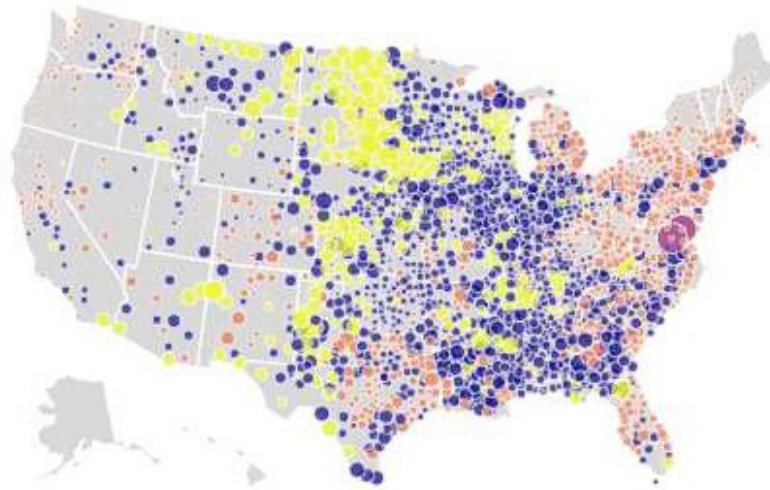


Figure 6.3: Clusters of PM2.5, COVID-19 Cases, and Mortality in USA for 2020

In the figure, the four groups of clusters are identified by their separate color. The same color indicates similar trends in COVID-19 cases. The clusters are defined as follows: those in purple represent that these counties had the highest long-term exposure to PM2.5 which resulted in a considerably large mortality rate; the most frequent ones in blue represent the second highest values, also resulting in high mortality rate. The clusters in yellow are third in the order, whereas the ones in orange represent that these areas had comparatively less exposure to PM2.5 with lesser amount of mortality rate.

2020 US County Wise Cases correlated to long-term PM2.5 exposure



Figure 6.4: County-wise PM2.5 and COVID-19 mortality cases in the U.S.A.

In this figure, we displayed the spread of COVID-19 cases in correlation to long-term PM2.5 exposure over 10 years. Compared with our clustered points of county-wise mortality we see that the counties with the most exposure to PM2.5 had more Covid-19 cases and also this exposure had more impact in the mortality of people in those counties.

In another cluster diagram, we calculated the percentage decrease in PM2.5 in 2020,

compared to 10 years of historical PM2.5 exposure. First, we calculated the median value of 10 years of PM2.5 exposure, and then found the percentage change of concentrations in 2020.

$$P_{ij} = \frac{(C_i - C_j)}{C_i} * 100$$

Where,

P_{ij} = Percentage change of a pollutant for a given area

C_i = 10 years averaged concentration of pollutant

C_j = Concentration of pollutant in 2020.

Thus, for each city the dataset has columns consisting of the percentage change values for the given pollutant. Now by plotting the difference in the map we see that some states have higher decrease in pollutant concentration than others which indicate that people in those states were conscious of the outbreak and maintained a sort of self-isolation, avoided crowded places and reduced transportation pollution. In accordance to mortality clusters found we can see that states with the least amount of change had more mortality rates than the rest.

**Percentage difference of averaged historical PM2.5
With PM2.5 in 2020**

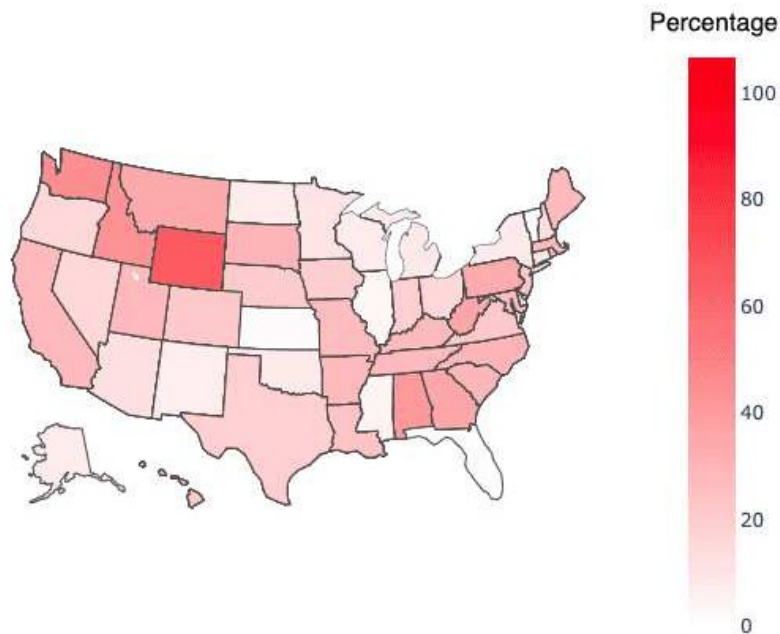


Figure 6.5: Percentage difference of averaged historical PM2.5 with PM2.5 in 2020

The next region we focused on was the country of Italy. In the figure below [38] we see that in the year 2019, the concentration of the pollutant Nitrogen Dioxide was far beyond the amount in 2020; it is yet another indicator of how the COVID-19 lockdown improved the air quality of another region. Moreover, it can be deduced from the map that in the years before the lockdown, Italy had a substantially high amount of pollution, concentrated near the city of Milan.

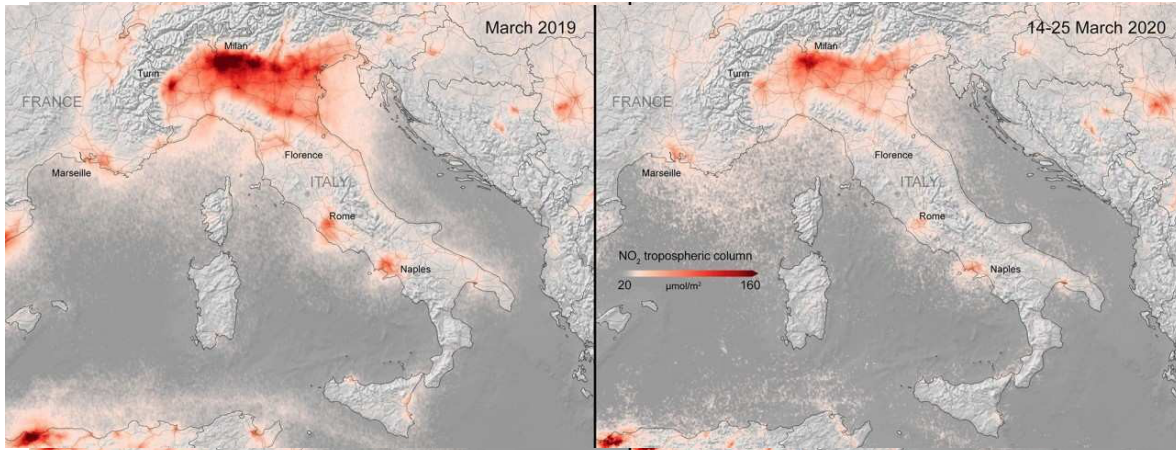


Figure 6.6: Map of Italy depicting NO₂ concentration in 2019 and 2020

In order to compare the spread of pollution with the spread of COVID-19, we clustered region-wise mortality rates in Italy as well as COVID-19 case data. From the figure given below, it is apparent that Italy had suffered a huge loss of life mostly in the city of Milan in the region of Lombardia.

Italy Mortality Rate Per Thousand
Due To Covid-19 in 2020

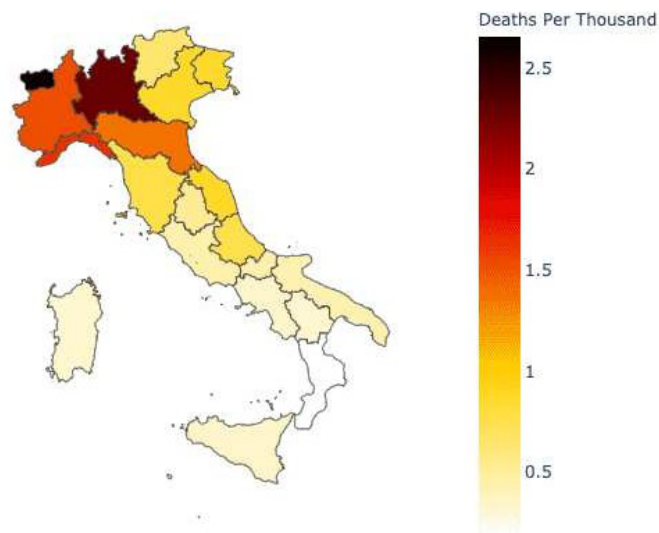


Figure 6.7: Map of Italy depicting COVID-19 mortality data

Therefore, in comparison with our Italy pollutant map, we can say that the long-term exposure of pollutants caused a larger amount of deaths in that region, since it coincides with the exposed area of NO₂ pollutant. Following the same trend with the rate of active cases in Italy as shown below, we conclude that the first COVID-19 clusters were found in the highest pollutant exposed regions of Italy.

Italy Total Positive Cases Due To Covid-19 in 2020

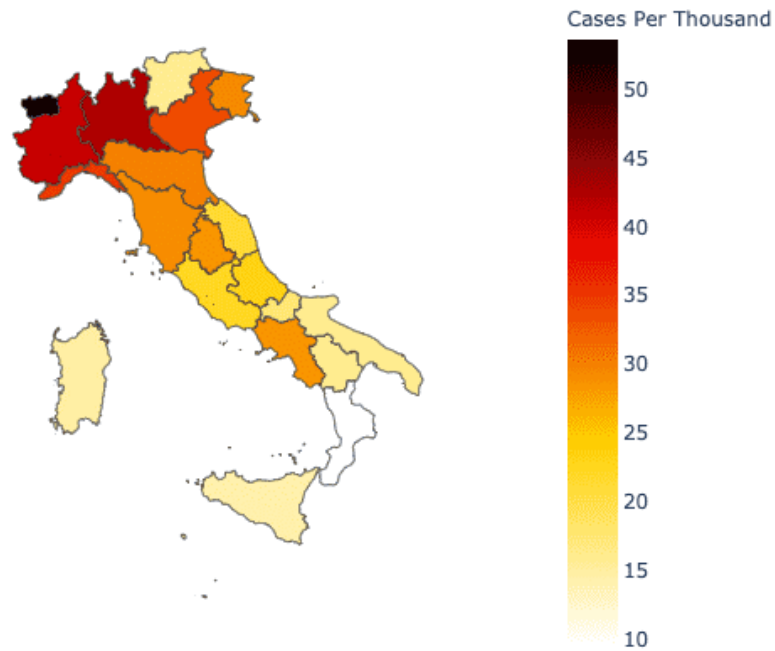


Figure 6.8: Map of Italy depicting COVID-19 cases

Our results help us to identify the trend of air pollution as clusters and we conclude our remarks on identifying which counties and areas had more cases and mortality rates by long-term pollution exposure.

We evaluated each of the results using null hypothesis testing. By taking normally distributed samples from the population and performing both z-tests and t-tests at 95% confidence level we fail to accept the null hypothesis of the pollutant mean concentration not being changed and accept the alternate hypothesis of pollutant mean concentration being less than the previous. Similarly, we did a hypothesis testing on correlation of COVID-19 cases and deaths with pollution and found that the alternate not to be true.

Chapter 7

Future Work

At the time of our analysis, the COVID-19 pandemic is still internationally at large. New research warns of a mutated strain of the virus and an ominous second wave to wash over the world. Understandably, data regarding the virus's spread as well as the multitude of factors that correlate with it is still unavailable for many of the world's regions. This leaves adequate room for potential future work in our thesis.

For example, our paper focused on county-wide correlation of adjusted PM2.5 values with COVID-19 mortality rates; if instead, a dataset of city-wise air quality values was curated and a corresponding city-wise death count data was created for U.S.A., an analysis of even larger proportions can be carried out. To continue the point regarding data availability, it can also be feasible to carry out our analysis on even more regions, thus covering more ground in terms of the virus's spread, if their air quality datasets were maintained and kept up to date.

Another potential expansion of our work is to conduct it for different air pollutants. Our motivation behind choosing Particulate Matter 2.5 was its pronounced effect in debilitating lung function when present in a region in hazardous concentrations. However, the other primary components of polluted air such as the oxides of nitrogen and sulfur dioxide can have equally devastating consequences if they rise too far. Studies analyzing the impact of COVID-19 lockdowns on the other pollutant concentrations, as well as observing the spread of COVID-19 (cases and deaths) with pollutant spread therefore, carries potential for further research.

Finally, there is the matter of confounding factors. Though we completed our analysis by taking into account sensitivity analyses of 43 such potential confounders, conducting research by taking even more into account is also a reasonable avenue.

Chapter 8

Conclusion

The aim of our research was threefold: first, to analyze the effect of lockdowns upon air quality by quantifying the amount that a region's air may have improved (or deteriorated), second, to identify correlations between a widespread air pollutant and COVID-19 mortality, and third, to spatially visualize pollution and the spread of COVID-19 in order to identify trends. Our results concluded that the majority of cities underwent at least a 40% improvement in air quality due to the lockdown. We also, found a linear correlation between adjusted PM2.5 and mortality rate values, where 1 unit increase in the former increased the chance of the latter by 3%. Finally, the results of our clustering model clearly visualized the relationship between regions of high air pollution and increased deaths and cases of COVID-19 mortality rates. Our results concur with those obtained by other researchers [9-10] in the fact that we found positive linear correlations between air pollution and COVID-19 deaths and improvement in air quality due to lockdowns.

Air pollution has been considered an international hazard for years. It has resulted in the death of countless people. However, despite increasing awareness measures, it still took a worldwide pandemic's lockdown for air quality to drastically improve in regions and reach safer levels. We hope that our results augment the awareness campaign regarding short and long-term dangers of poor air quality, now that findings have been made connecting it to increased death due to a viral disease. We also illustrated the effectiveness of lockdowns in mitigating poor air quality, and expect that regions will take this evidence into account when they are planning policies to improve the health of the environment. Our research is a timely one since much about the characteristics of the COVID-19 viral strain are still yet to be discovered. We conclude therefore by achieving our goal to bolster the fight against the virus by lending meaningful findings to the cause.

Bibliography

- [1] Air pollution- World Health Organization,
<https://www.who.int/health-topics/air-pollution> accessed September 23, 2020
- [2] Health Effects Institute. State of Global Air 2019,
www.stateofglobalair.org accessed April 23, 2020
- [3] Ambient air pollution: Health impacts,
<https://www.who.int/airpollution/ambient/health-impacts/en/>
accessed September 23, 2020
- [4] Corona disease 2019 - Wikipedia,
https://en.wikipedia.org/wiki/Coronavirus_disease_2019 accessed September 23, 2020
- [5] Mahato, S., Pal, S., & Ghosh, K. G. (2020). Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. *Science of The Total Environment*, 730, 139086. doi: 10.1016/j.scitotenv.2020.139086
- [6] Bera, B., Bhattacharjee, S., Shit, P. K., Sengupta, N., & Saha, S. (2020). Significant impacts of COVID-19 lockdown on urban air pollution in Kolkata (India) and amelioration of environmental health. *Environment, Development and Sustainability*. doi: 10.1007/s10668-020-00898-5
- [7] Srivastava, S., Kumar, A., Bauddh, K., Gautam, A. S., & Kumar, S. (2020). 21-Day Lockdown in India Dramatically Reduced Air Pollution Indices in Lucknow and New Delhi, India. *Bulletin of Environmental Contamination and Toxicology*, 105(1), 9-17. doi: 10.1007/s00128-020-02895-w
- [8] Dua, R. D., Madaan, D. M., Mukherjee, P. M., & Lall, B. L. (2019). Real-Time Attention Based Bidirectional Long Short-Term Memory Networks for Air Pollution Forecasting. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). doi: 10.1109/big-datSERVICE.2019.00027
- [9] Wu, X., Nethery, R. C., Sabath, B. M., Braun, D., & Dominici, F. (2020). Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study. doi:10.1101/2020.04.05.20054502
- [10] Coker, E., Cavalli, L., Fabrizi, E., Guastella, G., Ippolito, E. N. R. I. C. O., Parisi, M. L., Pontarollo, N., Rizzati, M., Varacca, A., & Vergalli, S. (2020).

The Effects of Air Pollution on COVID-19 Related Mortality in Northern Italy. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3700797>

- [11] He, G., Pan, Y., & Tanaka, T. (2020). COVID-19, City Lockdowns, and Air Pollution: Evidence from China. doi: 10.1101/2020.03.29.20046649
- [12] Wang, Q., & Su, M. (2020). A preliminary assessment of the impact of COVID-19 on the environment – A case study of China., *Science of The Total Environment* 1-10.
- [13] Cole, M. A., Elliott, R. J., & Liu, B. (2020). The Impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach. *Environmental and Resource Economics*, 76(4), 553-580. doi:10.1007/s10640-020-00483-4
- [14] Korunoski, M., Stojkoska, B. R., & Trivodaliev, K. (2019). Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization. IEEE EUROCON 2019 -18th International Conference on Smart Technologies. doi: 10.1109/eurocon.2019.8861609
- [15] R. Rohan. (2020). Air Quality Data in India (2015 - 2020), Version 12. Retrieved 2020-08-01 from <https://www.kaggle.com/rohanrao/air-quality-data-in-india>
- [16] Weather for 243 countries of the world (2020, October 06). Retrieved October 06, 2020 from https://rp5.ru/Weather_in_the_world?fbclid=IwAR0siW56p9JSOYr7Tn1yaEA8Dm4byiGWFqwcO96jdXvUZDRzfGML9_nTigo
- [17] U.S. Environmental Protection Agency Report an environmental violation. (n.d.). Retrieved January 07, 2021 from <https://www.epa.gov/?fbclid=IwAR1rV1vXQWhmdiG7bdLr6sZX39hRUDorSC4kX866TC9ARgLOE70wJQaJ0h0/>
- [18] Daily PM2.5 Concentrations All County, 2001-2016. (2020, May 22). Retrieved January 07, 2021 from https://catalog.data.gov/dataset/daily-pm2-5-concentrations-all-county-2001-2016?fbclid=IwAR0NTpulzZAbgYhZsmOMG5VVuKqFUA934YQPdB0lqcsjP4bp-mir_XdeGI
- [19] Nytimes. (n.d.). Nytimes/covid-19-data. Retrieved January 07, 2021, from https://github.com/nytimes/covid-19-data?fbclid=IwAR0_tzp9_tza7NeY2J-fl8SwO-Ix76XJ_znoPeQaEbrvL1G8tNNzOdEA2Wk
- [20] Owid. (n.d.). Owid/covid-19-data. Retrieved January 07, 2021, from <https://github.com/owid/covid-19-data/tree/master/public/data>
- [21] US Coronavirus Cases and Deaths. (2021, January 07). Retrieved January 07, 2021, from https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/?fbclid=IwAR3kbR1wCNsLvsofDwQqDShc0Ohsp2coxtgArJKmCRL_Yjrle0SEkgGuspY

- [22] Bureau, U. (2020, June 22). County Population by Characteristics: 2010-2019. Retrieved January 07, 2021, from <https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html?fbclid=IwAR0j09XsKsBSiwdJm0U0t9xr-O29YQ3m8UeoRaDdrNT3CH9ToDIQs4Xw8i8>
- [23] COVID-19 Hospital Capacity Metrics. (2021, January 07). Retrieved January 07, 2021, from <https://healthdata.gov/dataset/covid-19-hospital-capacity-metrics?fbclid=IwAR08Y6avJOpUuGHj7kpQ32gxK3aZoupLo1sBnH1jndpT-aRG1LBp5AGjfs4>
- [24] CopernicusAtmosphere. (2021, 01 7). air-quality-covid19-response. Retrieved from Github: <https://github.com/CopernicusAtmosphere/air-quality-covid19-response/pulse>
- [25] Srk. (2020, December 07). COVID-19 in Italy. Retrieved January 07, 2021, from https://www.kaggle.com/sudalairajkumar/covid19-in-italy?select=covid19_italy_region.csv
- [26] Italy: High Resolution Population Density Maps + Demographic Estimates. (n.d.). Retrieved January 07, 2021, from <https://data.humdata.org/dataset/italy-high-resolution-population-density-maps-demographic-estimates>
- [27] Pcm-Dpc. (n.d.). Pcm-dpc/COVID-19. Retrieved January 07, 2021, from https://github.com/pcm-dpc/COVID-19?fbclid=IwAR0NTpulzZAbgYhZsmOMG5VVuKqFUA934YQPdB0lqcsjP4bp-mir__XdeGI
- [28] PTI. (2017, November 7). Delhi pollution: Quadruple parking fees, cut metro fares, sa .. Retrieved from The Times of India: <https://timesofindia.indiatimes.com/city/delhi/delhi-pollution-quadruple-parking-fees-cut-metro-fares-says-green-body/articleshow/61546194.cms>
- [29] U.S. state and local government responses to the COVID-19 pandemic. (2021, January 04). Retrieved January 07, 2021, from https://en.wikipedia.org/wiki/U.S._state_and_local_government_responses_to_the_COVID-19_pandemic?fbclid=IwAR0j09XsKsBSiwdJm0U0t9xr-O29YQ3m8UeoRaDdrNT3CH9ToDIQs4Xw8i8
- [30] Coronavirus lockdowns and stay-at-home orders across the U.S. (2020, April 06). Retrieved January 07, 2021, from <https://www.nbcnews.com/health/health-news/here-are-stay-home-orders-across-country-n1168736>
- [31] Fine Particles (PM 2.5) Questions and Answers. (2018, February). Retrieved from New York State: https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm#~:text=Exposure%20to%20fine%20particles%20can,as%20asthma%20and%20heart%20disease.
- [32] Armstrong, J. S. (2001). Evaluating Forecasting Methods. Retrieved from http://repository.upenn.edu/marketing_papers/146

- [33] National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. (2020, December 13). Older Adults at Greater Risk of Requiring Hospitalization or Dying If Diagnosed With COVID-19. Retrieved from Center for Disease Control and Prevention: <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html>
- [34] J. A. Hartigan & M. A. Wong (1979) "A K-Means Clustering Algorithm", *Applied Statistics*, Vol. 28, No. 1, p100-108.
- [35] J. Gu, J. Zhou & X. Chen, "An Enhancement of K-means Clustering Algorithm," 2009 International Conference on Business Intelligence and Financial Engineering, Beijing, 2009, pp. 237-240, doi: 10.1109/BIFE.2009.204.
- [36] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830
- [37] Austin, E., Coull, B. A., Zanobetti, A., & Koutrakis, P. (2013). A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. *Environment International*, 59, 244–254. <https://doi.org/10.1016/j.envint.2013.06.003>
- [38] Reuters. (2020, March 30). Air pollution decreases across Europe as coronavirus locks down major cities. Retrieved January 08, 2021, from <https://nypost.com/2020/03/30/air-pollution-decreases-across-europe-as-coronavirus-locks-down-major-cities/?fbclid=IwAR1QwHliKx6clC79gl2Yob7IYaEpRGcxibdRTanntkzmbFWurQllkHDQ2A>

Appendix A

Tables of Lockdown Dates Around the World

We manually collected official lockdown dates of the regions involved in our analysis. The government of India declared the nation under lockdown from the 25th of March 2020 until the 31st of May 2020. The lockdown was split into four distinct phases as outlined below:

Phase Number	Start Date	End Date	Duration
1	25 th March 2020	14 th April 2020	21 days
2	15 th April 2020	3 rd May 2020	19 days
3	4 th May 2020	17 th May 2020	14 days
4	18 th May 2020	31 st May 2020	14 days

In Europe the lockdown periods were distinct in different countries. Countries like Turkey and Macedonia followed stay at home only for four to five days.

Table of lockdown dates for countries involved in our European city LSTM Analysis:

Country	Lockdown Start	Lockdown End
Turkey	23 April, 2020	27 April, 2020
Greece	23 March, 2020	4 May, 2020
Greece	7 November, 2020	7 January, 2021
Switzerland	17 March, 2020	27 April, 2020
Slovakia	16 March, 2020	14 June, 2020
Denmark	13 March, 2020	14 April, 2020
Finland	16 March, 2020	13 May, 2020
Serbia	21 April, 2020	4 May, 2020
Albania	13 March, 2020	1 June, 2020
Bulgaria	13 March, 2020	15 June, 2020
Monaco	17 March, 2020	15 April, 2020
Bosnia	17 March, 2020	26 April, 2020
Herzegovina	17 March, 2020	26 April, 2020
Macedonia	17 April, 2020	21 April, 2020

Below are the lockdown start and end dates of all 50 states in the U.S.A:

States	Lockdown Start	Lockdown End
Alabama	4 April, 2020	30 April, 2020
Alaska	28 March, 2020	20 May, 2020
Arizona	31 March, 2020	30 April, 2020
Arkansas	None	None
California	19 March, 2020	5 May, 2020
Colorado	26 March, 2020	26 April, 2020
Connecticut	23 March, 2020	20 May, 2020
Delaware	24 March, 2020	15 May, 2020
Florida	3 April, 2020	4 May, 2020
Georgia	3 April, 2020	30 April, 2020
Hawaii	25 March, 2020	31 May, 2020
Idaho	25 March, 2020	30 April, 2020
Illinois	21 March, 2020	30 May, 2020
Indiana	25 March, 2020	1 May, 2020
Iowa	None	None
Kansas	30 March, 2020	4 May, 2020
Kentucky	16 March, 2020	10 April, 2020
Louisiana	30 March, 2020	15 May, 2020
Maine	2 April, 2020	30 April, 2020
Mayland	30 March, 2020	315 May, 2020
Massachusetts	24 March, 2020	18 May, 2020
Michigan	24 March, 2020	2 June, 2020
Minnesota	27 March, 2020	3 May, 2020
Mississippi	3 April, 2020	27 April, 2020
Missouri	6 April, 2020	3 May, 2020
Montana	28 March, 2020	26 April, 2020
New Hampshire	27 March, 2020	11 June, 2020
Nevada	17 March, 2020	7 May, 2020
New Jersey	21 March, 2020	9 June, 2020
New Mexico	24 March, 2020	15 May, 2020
New York	22 March, 2020	15 May, 2020
North Carolina	30 March, 2020	8 May, 2020
North Dakota	None	None
Ohio	23 March, 2020	1 May, 2020
Oklahoma	28 March, 2020	16 April, 2020
Oregon	23 March, 2020	15 May, 2020
Pennsylvania	1 April, 2020	8 May, 2020
Rhode Island	28 March, 2020	8 May, 2020
South Dakota	None	None
South Carolina	7 April, 2020	4 May, 2020
Tennessee	31 March, 2020	30 April, 2020
Texas	2 April, 2020	30 April, 2020
Vermont	25 March, 2020	15 May, 2020
Virginia	30 March, 2020	10 June, 2020
Washington	23 March, 2020	4 May, 2020
West Virginia	23 March, 2020	4 May, 2020
Wisconsin	25 March, 2020	13 March, 2020
Wyoming	25 March, 2020	26 May, 2020

Appendix B

Tables of Mean Percentage Improvement from Forecasting Model

This section has the tables of mean percentage difference we calculated during the RNN LSTM forecasting implementation of our paper. Below is the mean percentage improvement in AQI values for the cities of India due to the COVID-19 lockdown:

State	Mean Percentage Improvement in AQI because of COVID-19 Lockdown
Ahmedabad	75.028532
Amaravati	49.884083
Amritsar	43.78713
Bengaluru	40.950826
Brajrajnagar	36.741188
Chennai	32.820339
Delhi	57.757272
Gurugram	46.824925
Guwahati	71.310643
Hyderabad	39.210099
Jaipur	38.566265
Jorapokhar	21.731452
Kolkata	64.462707
Lucknow	74.207516
Mumbai	8.882186
Patna	43.140466
Talcher	13.279866
Thiruvananthapuram	21.022388
Visakhapatnam	48.335516

Mean percentage improvement in PM2.5 values for the cities of India due to the COVID-19 lockdown:

State	Percentage Improvement In PM2.5 due to Lockdown
Ahmedabad	62.924785
Amaravati	57.431952
Amritsar	69.274141
Bengaluru	48.991553
Brajrajnagar	20.881379
Chennai	55.401948
Delhi	63.600911
Gurugram	58.114694
Guwahati	77.076724
Hyderabad	31.027116
Jaipur	67.88777
Jorapokhar	37.244206
Kolkata	72.070542
Lucknow	61.759285
Mumbai	4.833274
Patna	61.354718
Talcher	2.832942
Thiruvananthapuram	27.01976
Visakhapatnam	63.098269

Since we obtained overall improvements in air quality for few Indian cities for an extended lockdown, the table of percentages for an extended Indian lockdown was not created; all of the reasoning behind the process (as well as the result) has been included in Chapter 4.

Mean percentage improvement in PM2.5 values for the European cities due to the COVID-19 lockdown:

City	Improvement in PM2.5 due to lockdown (%)
Amsterdam	45.838172
Athens	35.673675
Barcelona	9.520128
Madrid	46.950791
Berlin	53.039617
Cologne	46.369243
Hamburg	55.437038
Munich	47.608461
Birmingham	43.290077
London	41.836291
Brussels	51.187231
Bucharest	39.468686
Budapest	47.773294
Lisbon	39.081402
Ljubljana	65.091463
Luxembourg	44.615846
Lyon	37.523217
Marseille	49.233453
Paris	45.295328
Milan	55.551869
Naples	34.436787
Rome	41.556863
Turin	62.464479
Nicosia	29.045986
Riga	45.825108
Tallinn	49.190903
Vienna	50.932159
Vilnius	61.027943
Warsaw	56.448789
Zagreb	62.564697
Bern	43.523
Bratislava	58.071843
Copenhagen	48.440104
Helsinki	49.033187
Monaco	56.881842
Sarajevo	44.730476
Sofia	44.849794
Tirana	34.417691

Mean percentage improvement in PM2.5 values for the European cities due to an extended lockdown:

City	PM2.5 No Lockdown
Amsterdam	-72.452596
Athens	17.863904
Barcelona	-48.320401
Madrid	43.489455
Valencia	-22.30111312
Berlin	-69.992623
Cologne	28.826007
Hamburg	-437.156817
Munich	19.267423
Birmingham	-38.896023
London	-125.66302
Brussels	35.088441
Bucharest	53.608198
Budapest	31.819212
Dublin	11.515408
Lisbon	-45.934126
Ljubljana	39.870731
Luxembourg	41.260799
Lyon	40.074675
Marseille	39.7319
Paris	36.259917
Milan	35.072519
Naples	42.05974
Rome	30.77186
Turin	38.367731
Nicosia	25.371726
Riga	56.584688
Tallinn	30.705175
Vienna	40.011368
Vilnius	34.609372
Warsaw	38.765959
Zagreb	49.865327
Bern	26.409144
Bratislava	-32.855947
Copenhagen	37.928175
Helsinki	24.302639
Monaco	28.483403
Sarajevo	35.21004
Sofia	48.029244
Tirana	-42.918734

The negative values in this table represent locations where the forecasted air quality was better than the real one; as with India, for most of the European cities we also noted that the real AQI values were in fact healthier than our forecasted ones,

possibly due to the continued limitations in public transport and venturing outside after the lockdown period was over.