Temperature and other external factors affect the spread of COVID-19

Tanvi Kulkarni
University of Maryland, Baltimore County
Baltimore, USA
Email: t66@umbc.edu

***Abstract*—Some of the studies pointed out that the pandemic is experiencing seasonal patterns in its spread, incidence and nature of the distribution. In this report, we try to study data for temperature and external factors. We performed data shredding and pre-processing, and used Spark MLlib to extract the relationship between temperature, humidity and population with the spreading rate of COVID-19. The deep learning algorithm based on ridge regression model measures the correlation between the environmental factors and the novel COVID-19 spread. Our algorithm predicts the approximate number of covid cases in certain regions of United states.**

Keywords—Ridge regression; Linear regression; Deep Learning; Prediction ; COVID-19 data

I. Introduction

COVID-19 is a SARS-COV-2 mediated respiratory pandemic. We are well aware of the effects and symptoms of COVID. Initially, it first came to attention in a series of patients with pneumonia of unknown etiology in Wuhan City, China, and spread to all other countries of the world due to geographical proximity and major travel ties[3]. More recently, this has been declared a pandemic by the World Health Organization and it will possibly remain for a long time and people have to adapt how to treat and prevent it before a validated vaccine is available. The main issue now is we have to increase our understanding of situations and also try to figure out key factors to minimize the problem[4].

If we evaluate the graph based on regular covid status, we can see that the curve first increases and then decreases a little before rising exponentially. In the research publications[3] there has been particular attention to the association between weather variables in the affected regions and the spread of COVID-19. The most affected areas include epicenters of outbreaks such as parts of the Northeastern United States, Hubei Central Province of China, South Korea, Japan, Iran, Italy, Spain, Germany, and England, all of which share an average temperature of 5C to 11C in January and February 2020, and 47 percent to 79 percent humidity. In the case of Italy, regions with temperatures greater than 15 degrees Celsius and 75 percent humidity have a lower COVID-19 spread. Therefore, we can believe that the spread of the virus in areas with higher temperatures and humidity would decrease relative to areas with average records.

II. Motivation

A. Need for Awareness: To support our environment, we carry out this prediction through a rate of infection from corona viruses based on the climate of the current infected region. For instance, if the weather forecast predicts it is raining for the next two days, and then it is gloomy for the next two days and further on sunny for the next three days. Based on these forecast, on a day by day basis we try to find that, how does the rate of infection change?
B. Design Predictive model: Building a model that has more readiness in healthcare systems and predicting the future using sophisticated deep learning algorithms is the best way to find out regular verified cases. People can better aware, what is the optimal weather and how to take proper precautions.

III. Related Work

Several research studies on the weather effect on COVID-19 spread and distribution have been performed. Dangi et al.[4] proposed a system to predict the forthcoming COVID-19 outbreak in 35 major cities in India (March and April 2020) by correlating the temperature factor of five major cities worldwide.The results showing that 27 cities were showing temperature correlation.

The role of ambient temperature in the survival and transmission of viruses is highlighted in several studies, including laboratory, epidemiological analysis, and mathematical modeling[5]. This research was inspired by a huge number of studies supporting both ambient temperature and humidity in the role of transmission and infection to investigate the influential factors for COVID-19.

In spite of this interest, we are trying to connect knowledge of weather conditions and other external factors, such as air population and COVID-19 spread, to find a connection between temperature and humidity and other transmission-related external factors.

IV. Methodology

Our study follows the Data Life cycle process which involves Data Collection, Data Scraping, Map reduce function Data Planning, Data Visualization, Data Access, and the implementation of deep learning models.
We use ridge regression model, the main aim of using this model is it is seems to be identical to linear regression which we try to implement previously. As predictive model in my Deep learning experiments it was observed that linear regression model was over fitting. Further we predicts time series, and measuring accuracy. In future we can try to implement this model on other Deep learning algorithms, if we found them more feasible.
Before we begin to do data analysis, we first conduct Data Collection and Data Preparation.

A. Data Collection and scrapping: We collect data from the National Center for Environmental Information and USA Facts[13] and National Weather Service[12] which is a reliable data provider with numerous useful and meaningful datasets. After searching for the keyword "COVID", there are numbers of results coming out and we select the one with more than 1 GB dataset in CSV because CSV is more suitable for big data analysis on both Machine learning and Pyspark.

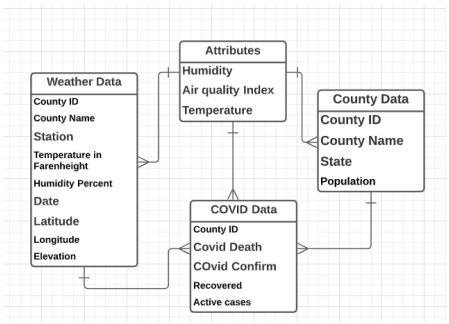

Fig. 1. ER Diagram

B. Data Preparation and Data Understanding: The dataset includes Weather data, COVID data, and County Data. To identify necessary information, we develop an ER diagram and try to understand the data structure and relationship. According to the ER diagram illustrated in Fig. 1. We find attributes that are useful for our project since we only need to discover the relationship between weather and its dependency on COVID. Meanwhile, the County table contains each US state county and is distinguished by ID representing the population.

C. Data Prepossessing: The data is processed using jupyter notebook and pandas in spark.The dataset is already broken into pieces so that we can obtain sample information from it. As there are more than 3,000 counties that are massive. We have split this dataset during the data analysis, they can be randomly selected as sample test datasets. We also considered dummy variable the combination of the county and state regulations name emergency and mapped it with regular covid data.
One approach for creating a dummy variable is it is range from -1 to 1 such that 1 could be assigned to the most strict state let's say to California, where the government is very strict in response to COVID, 0 can be given to the states like North Carolina where masks were voluntary and -1 could be indexed to the states where no policies were designed such as Arizona. In the resulting document(fig-2), we have daily data of temperature and the new COVID cases.

```
# checking the training data
trainData.select(predictionCol, *featuresCols).show(5)
```

```
+------------------+------+---------+----------+---------+---------------+
|         COVID_LOG|TEMP_F|HUMID_PCT|POPULATION|EMERGENCY|EMERGENCY_STATE|
+------------------+------+---------+----------+---------+---------------+
|  9.63652272167307|  65.1|     86.0|  943332.0|      1.0|            1.0|
|0.6931471805599453|  47.3|     56.0|    8910.0|      1.0|            1.0|
|0.6931471805599453|  47.3|     56.0|    8910.0|      1.0|            1.0|
| 4.663439094112067|  56.4|     73.0|  169509.0|      1.0|            1.0|
| 4.394449154672439|  79.3|     80.0|    8630.0|      1.0|            0.0|
+------------------+------+---------+----------+---------+---------------+
only showing top 5 rows
```

Fig. 2. Train Data set

D. Data analysis: We analyze the weather and temperature data of the respective states of the US. We have composed a dataset than load the cleaned data from COVID 19 Dataset[9][10]. The file contains the cumulative count of population, temperature, and covid cases from different states of the United States, we scrape it for the different US counties from 22nd January 2020. As described in Fig-2 the assumption here is that, there is a correlation between certain weather metrics and government laws and orders and the speed of the number of infections/deaths.[3]
Hence, the main approach should be a prediction, a classic method of Deep learning. First, find out the temperature data for various counties, count the humidity percent and air quality index, and connect them to COVID data. Second, applying a model on the dataset to investigate and understand the real effect of temperature and humidity on the spread of COVID-19.

V. Algorithm and Implementation

The methods presented shall be trained in the enormous volume of the dataset in order to carry out the forecast of the distribution of COVID-19. In the training phase, the amount of dataset plays a vital role and affects the efficiency of the algorithms proposed. As for the initial stage, we are considering the dataset mainly for temperature and humidity, so we have to perform an individual operation for these variables:

Month of D..	Confirmed	Deaths
January	38	0
February	378	1
March	1,091,068	26,160
April	19,552,582	1,035,013
May	45,407,574	2,728,676
June	64,933,835	3,518,537
July	93,360,473	3,703,024

TABLE I
MONTHLY STATISTICS OF CONFIRMED AND DEATH CASES

We have analyzed the correlation between weather variables and the spread of COVID-19 in the case of log-transforming the number of reported COVID-19 cases (dependent variable) to make it follow a normal distribution as per the assumption of statistical analysis.
Since the original data is highly skewed in selected states. We have utilized the standard ridge regression model to estimate the relationship.
Equation 1 estimates the humidity in the weather
Equation 2 estimates the temperature of different counties in the US

A. Calculation of humidity The relative humidity is the ratio of the actual water vapor pressure to the saturation water vapor pressure at the prevailing temperature, calculated using the following equation:

Relative humidity=Ew×100 - Equation 1

where E (hPa) donates the vapor pressure of air at temperature t(°C). Ew (hPa) donates the saturated vapor pressure of the pure horizontal liquid surface at dry-bulb temperature t (°C).

The dew point is the temperature at which air must be cooled to become saturated with water vapor. The actual vapor pressure E can be calculated using the dew point temperature and the saturated vapor pressure can be calculated from the actual temperature using the Magnus formula for saturation water vapor pressure

E=E0×eAtB+t - Equation-2

where E0 denotes saturation vapor pressure at a reference temperature T0 (273.15 K) which equals 6.11 MB. A is a constant of 17.43 and B is a constant of 240.73. t (°C) is the actual temperature or dew point.[5]

B. Calculation For temperature: To better understand the impact of temperature on the COVID-19 epidemic, the temperature was considered based on a ridge regression model[12]. The equation of regression line is represented as:

$$\widetilde{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

The basic criteria for using Ridge regression can be said as dependent feature variables, over fitting, biasing importance of feature variables. An L2 parameter called ridge estimator is multiplied to an identity matrix which then added to the X transpose * X. It has been proven that the ridge estimator can lie between 0 to positive infinity, it cannot be negative or infinity[13].
We have combined all our features into a single column by using vector assembler. The generated data-frame is then fed into linear regression model with regularization parameter of 0.3 with a maximum of 10 iteration refer fig-3.

```
+--------------------+-------+
|            features|  COVID|
+--------------------+-------+
|[65.1,86.0,943332...|15314.0|
|[47.3,56.0,8910.0...|    2.0|
|[47.3,56.0,8910.0...|    2.0|
+--------------------+-------+
only showing top 3 rows
```

Fig. 3. Generated Data frame after applying Vector Assembler on Features

We have performed R-Square estimation to evaluate the model developed. R-Square value tells us the percent accuracy of the model, the bigger it is the better the model between 0-1

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

```
+--------------------+
only showing top 20 rows

RMSE: 5356.625072
r2: 0.500164
```

Fig. 4. Generated R square Value

VI. Data Visualization

We used Tableau, pair visualization from matplotlib for data visualization which is an effective tool to quickly create interactive data visualizations, Due to its simple and user-friendly interface.

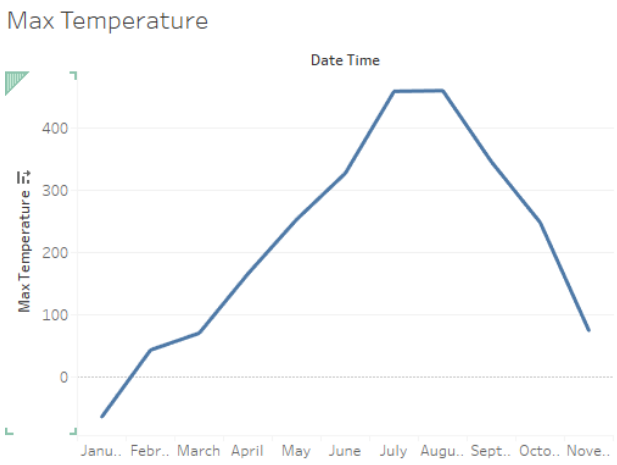

Fig. 5. Relation between Temperature Change and COVID Cases

After the Data collection, we review temperature, confirm COVID cases, and population. We further predict them on the basis of the temperature change.

The Fig-5 represent the change of temperature, where there were no cases in Feb, March. Here, We observed that lines lie in a different curve path, which suggests both temperature and confirmed COVID cases are related. We also wanted to assess how better our Analysis predicts the data as we increase the number of training instances.

The experimental results for the comparison of the confirmed covid cases models for random dates of months. We visualize for different counties of and evaluate them on a test set, shown in Fig-6.

Confirm COVID Cases

County Name	1/22/20	2/1/20	3/11/20	4/10/20	5/1/20	6/1/20	7/1/20	8/10/20	9/11/20	10/10/20	11/15/20
Audrain County	0	0	0	0	1	85	138	203	334	483	865
Audubon County	0	0	0	1	2	12	17	28	42	115	252
Auglaize County	0	0	0	9	33	72	99	256	504	759	1,867
Augusta County	0	0	0	15	44	112	184	280	414	561	861
Aurora County	0	0	0	1	1	26	34	38	42	126	297
Austin County	0	0	0	8	13	27	72	249	456	503	544
Autauga County	0	0	0	17	42	233	553	1,215	1,551	1,898	2,456
Avery County	0	0	0	0	0	2	12	95	228	457	730
Avoyelles Parish	0	0	0	56	72	130	341	1,070	1,398	1,617	1,851
Baca County	0	0	0	9	10	12	14	15	17	19	42
Bacon County	0	0	0	14	25	129	248	438	552	614	649
Bailey County	0	0	0	0	0	17	100	171	203	241	408
Baker County	0	0	0	27	45	64	119	1,035	1,462	1,800	2,218
Baldwin County	0	0	0	94	401	640	1,234	4,765	6,763	8,383	10,069
Ballard County	0	0	0	1	8	12	14	35	46	60	112
Baltimore City	0	0	0	689	2,162	5,604	7,605	12,704	15,085	16,390	20,737
Baltimore County	0	0	1	1,072	3,013	6,299	8,040	13,374	16,520	18,887	24,285
Bamberg County	0	0	0	6	10	22	117	479	571	657	711
Bandera County	0	0	0	1	6	6	21	91	112	177	208
Banks County	0	0	0	7	25	85	139	271	423	523	643
Banner County	0	0	0	0	0	0	1	2	2	2	7
Bannock County	0	0	0	5	11	32	104	400	737	1,452	3,426

Fig. 6. Confirmed covid cases for random dates of months

A. Performance visualization: From the experimental results, in the case of death rate, it is experimentally state that the weather variables are more important when compared to other factors such as population, and urban percentage.[Fig-5]

B. Result Visualization : With the use of Matplotlib

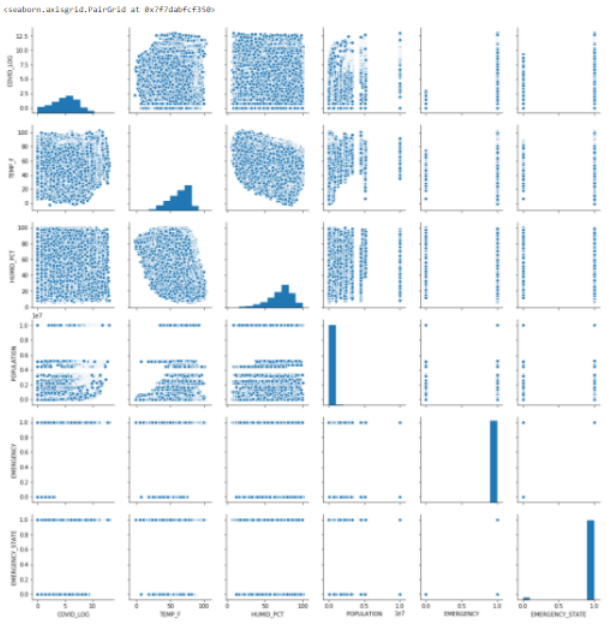

Fig. 7. Pair plot Features

to pair-plot all the features We compare COVID cases, temperature, Humidity and Population. In the fig-7 we take these four features from the month January to November and compared with the corresponding covid data.

We differentiate new COVID cases according to the state with the help of Matplotlib hue feature. The points are scattered all over the graph, hence a plane can best fits the data points.It is also visible in the graph that the facts of delayed COVID reporting significantly affected the best fit plane. We used ridge regression in order to avoid data points which are too low or too high, after applying shrinkage in the data it improves our model and hence the r-squared value which is 0.51 proved our hypothesis to be true.

VII. Actionable knowledge

Our COVID-19 spread analysis approach begins with researching how the novel COVID-19 spread over the year and how it came to be what it is now. We find the relationship between temperature and COVID spread which can be denoted by the rate of change in new cases of COVID data very interesting. Since it is a form of respiratory disease, it tends to be more affected by atmospheric constraints such as temperature, humidity, and air quality, analysing the relation could be quite interesting and could lead to some deeper insights.

To conduct the survey we needed the weather data and COVID-19 daily case data in different counties, we considered the average temperature of each county and mapped it with new COVID-19 cases of that county in a particular day. Since We cannot find a single source for our data, we need to scrape it from National Oceanic and Atmospheric Administration data. The plugin for World Weather Online historical weather data API came in handy for achieving this task, and COVID-19 Data is taken from NY Times COVID-19 repository.
Capturing weather data and mapping it to COVID data according to the date and county-state unique identifier was achieved using Python libraries such as pandas.
A lot of undesired data came along with temperature and COVID-19 data which then shredded using pandas library, and to provide asynchronous behaviour to our data processing, we created milestones and saved them as CSV file, for instance, mapping the data from one source to another can be counted as a milestone and the output was saved in file storage.
In order to truly understand the data, we visualized it on Tableau and pair visualization from matplotlib. We experimented with the visualization by taking different variables over different axis, and tends to find our hypothesis. To make our model run over a distributed environment, we used spark mllib to execute our ridge regression algorithm. Transforming the data and extracting the features are some of the key steps in implementing our model in Apache Spark.

VIII. Future Work

Improving upon the current model, need some more work on mapping the data according to the county. In my research I got to know that each county is given a unique identifier number valid within a country. In future work we can use that instead of the combination of county and state name. Adding more features could increase accuracy of current models and would highly encourage some research by adding other environmental constrains and seasonal diseases as a dummy variable into the model. It would be very interesting to know how the novel COVID-19 is shaped by the strictness of government policies applied and how seriously people are following them.

I would also like to put more pressure on the Government policy measurement and trend of the population to be included in the model since they can impact COVID spread a lot. I found a handy quantified data on government policies strictness all over the world, it contains the indexed data of how stringent the governments were while responding back on COVID-19. Since the scope of the project is limited to the United States of America. I believe that would be huge research topics in itself but combining them with our model as a would really allow us to analyse the COVID-19 spread and provide us the opportunity to consider probable solutions and make a quick response to future pandemics.

We also believe that results of using other predicting algorithms would be very interesting to analyze, for instance, the Time Series method could be a good predictor of the seasonality of COVID-19 spread with the variation in temperature and could be very helpful in predicting the daily new cases. The findings would be interesting because it can reveal more information about the COVID-19 spread pattern and in combination with this model we can more accurately determine the time, temperature and location of new COVID cases, although these are some hypothesis that we believe quite firmly but the before analysing the temporal data with all the variables, it is hard to say anything with certainty.

IX. Conclusion

We used Prediction Analysis; It is a promising field in Machine Learning. It provides a strong platform and the performance state of the art Prediction analysis technology. We believe the importance of this technology will be more pronounced as user-generated content gets bigger and more prevalent.
With the current dataset, we can not say that the COVID-19 spread does depend over only on Temperature and Humidity.Since there are many other factors. As this is a huge pandemic of this era, there is still a lot more to study and implement various other factors increasing these Viruses. AWS Cluster Implementation: It was also a big procedure but AWS EC2 Cloud Platform made it easier.

X. References

1. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, and B. Cao, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," The Lancet, 24-Jan-2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673620301835?via=ihub. [Accessed: 04-Jan-2021].

2."Coronavirus," World Health Organization. [Online].Available:https://www.who.int/health-topics/coronavirus. [Accessed: 04-Jan-2021].

3. Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, "Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches," Chaos, solutions, and fractals, Sep-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7367008/. [Accessed: 04-Jan-2021].

4.M. M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, and A. Amoroso, "Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19)," JAMA network open, 01-Jun-2020. [Online]. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7290414/. [Accessed: 04-Jan-2021].

5.Y. Wu, W. Jing, J. Liu, Q. Ma, J. Yuan, Y. Wang, M. Du, and M. Liu, "Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries," Science of The Total Environment, 28-Apr-2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S004896972032568[Accessed: 04-Jan-2021].

6."COVID-19: Government Response Stringency Index," Our World in Data. [Online]. Available: https://ourworldindata.org/grapher/covid-stringency-index. [Accessed: 04-Jan-2021].

7."Modeling COVID-19 scenarios for the United States," Nature News, 23-Oct-2020. [Online]. Available: https://www.nature.com/articles/s41591-020-1132-9. [Accessed: 04-Jan-2021].

8. Y. Qiu, X. Chen, and W. Shi, "Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China," Journal of population economics, 09-May-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7210464/. [Accessed: 04-Jan-2021].

9. N. O. A. A. US Department of Commerce, National Weather Service. [Online]. Available: https://www.weather.gov/. [Accessed: 04-Jan-2021].

10. "National Centers for Environmental Information," National Centers for Environmental Information (NCEI). [Online]. Available: https://www.ncei.noaa.gov/. [Accessed: 04-Jan-2021].

11. "Overview," Overview - Matplotlib 3.3.3 documentation. [Online]. Available: https://matplotlib.org/3.3.3/contents.html. [Accessed: 04-Jan-2021].

12. Stephanie, "Ridge Regression: Simple Definition," Statistics How To, 14-Oct-2018. [Online]. Available: https://www.statisticshowto.com/ridge-regression/. [Accessed: 04-Jan-2021].

13. NCSS Statistical Software, Ridge regression. [Online]. Available:https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge$_Regression.pdf$ $[Accessed : 04 - Jan - 2021]$.