

Mathematical Models for Analyzing COVID-19 Pandemic in Different Regions with Climate factors

Zhibin Shen^{1, †}, Dawei Wang^{2, †, *}

¹Department of Mathematics, College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign, Urbana, IL, United States

²Fu Foundation School of Engineering and Applied Science, Columbia University in the City of New York, New York City, NY, United States

*Corresponding author: dw2987@columbia.edu

[†]These authors contributed equally.

Abstract. Aiming at the Covid-19 pandemic problem, to find out whether the climate factors could affect the development of pandemic, this paper mainly uses mathematical modeling and machine learning to analyze the correlation between climate factors and COVID-19 cases. Firstly, Weather conditions are classified by seasons, and cases differ in regions, then correlation results of different features are conduct to see what are the most important features that affect the pandemic, according to the correlation result, KNN model is used for predicting the future COVID-19 cases including potential danger zones. In order to test the effectiveness of the methods utilized in this paper, real climate data and covid-19 cases data of different regions in USA are deployed, the results show that, the temperature plays an important role in the pandemic, and the KNN method could predict the future development of Covid-19 with R square reached 0.25, which verifies the effectiveness of the machine learning method.

Keywords: Climate factor, COVID-19, Mathematical Modeling, Machine Learning.

1. Introduction

At the end of 2019, Wuhan China found a new COVID virus: SARS-CoV-2. Soon this virus led to a pandemic over China and whole word at early 2020 [1]. As one of the deadliest epidemics in human history, studying the way COVID-19 (Coronavirus disease 2019) transmit is essential in prevention and treatment.

The transmission of COVID-19 relies on people breathe in virus contained in small airborne particles or droplets. Many meteorological features like humidity, temperature, wind speed, rainfall and air quality affect COVID-19 transmission. Scientists held many researches on the potential affect weather factors to virus spread internationally. Bashir's team used Spearman test shows that above factors are significant for new cases. Although Bashir's team has shown the correlation, the data set from New York is limited and unilateral [2,3]. PengShi's team showed that COVID-19 case rate have a biphasic relationship with temperature at impact of temperature on the dynamics of the COVID-19 outbreak in China [4-9]. Penshi's team confirmed the statistically significant association between temperature and relative infected risk during their study period by random-effect meta-analysis 28 provinces in China, but the data used lack the temporal breadth and the summer period with higher temperature [10].

Therefore, this paper mainly focuses on the analysis of the correlation of COVID-19 cases and weather factors throughout a complete seasonal cycle, then prediction will be made by machine learning technique. Factors with high correlation will be identified and used in next phase. The prediction phase will use KNN model to predict the COVID-19 cases based on weather factor given from correlation analysis. The result of this research would be clearer to identify potential danger zone and easier to make prevention advises.

2. Methodology

2.1. correlation analysis

A scatter plot of testing object and features can intuitively express the potential relation wanted to find [10]. Even though the data set used is too complicate to directly find the correlation relation, the scatter plot can provide a reasonable hypothesis for future analysis. While scatter plot provides intuitive picture, correlation

coefficient is a common and simple way to identify correlation relationship among different features in a numerical level. There are three common correlation coefficients in statistics study: Pearson, Kendall tau and Spearman. The analysis is based on Pearson and Spearman coefficient since they are more representative.

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

Pearson coefficient uses the covariance, which requires expectation and mean, and standard deviation of two object wanted to measure. The range of Pearson coefficient is from -1 to 1. Coefficient equal to 1 means two object is exactly linear and Y increases as X increases, and coefficient equal to -1 means they are exactly linear but Y decreases as X increases. It is important to notice that Pearson coefficient performs best on linear correlation, so this is to be expected that the coefficient will be relatively small on nonlinear data.

$$r_s = \rho_{R(X),R(Y)} = \frac{\text{Cov}(R(X),R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \quad (2)$$

While Pearson coefficient is used to evaluate linear correlation, Spearman coefficient uses monotonic function to conduct its assessment. Spearman coefficient convert raw data to ranked data and follows same step as Pearson. The range of Spearman coefficient is the same as Pearson coefficient. Only difference is that, instead of linear related, 1 and -1 represent perfectly monotonically related.

Although the data set researched required more advance method than these simple coefficients designed for linear or monotonic correlation, a general overview of features is essential in following study.

2.2. KNN

K Nearest Neighbors (K-NN) classification can be used to categorize COVID-19 severity under different weather conditions. Every new instance is allocated to the category where K (a number) of its nearest neighbor instances belong to. The Euclidean distance formula is used for finding a neighbor instance of the new instance:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^D (x_{im} - x_{jm})^2} = \sqrt{\|x_i\|^2 + \|x_j\|^2 - 2x_i^T x_j} \quad (3)$$

This algorithm will be executed repeatedly until all instances have been classified or a predetermined end. The number of “K” affects a lot on the final classification result. A small “K” value risks overfitting, leading to a high accuracy on training set but a low accuracy on testing set. A large number of “K” brings low estimate error but high approximate error. “Cross-Validation Method” can be used to find the optimal value of “K”.

However, the above categorizing method doesn’t take the effect of distance on classification into consideration. There exists an improved algorithm: Weighted K-Nearest Neighbors. For those K nearest neighbors, the closer a neighbor instance to the new instance, the higher weight it’s category counts. In this case, the measure of distance is as follows:

$$d(x_i, x_j) = \sum_{m=1}^D w_m d(x_{im}, x_{jm}) \quad (4)$$

2.3. Process of Modeling

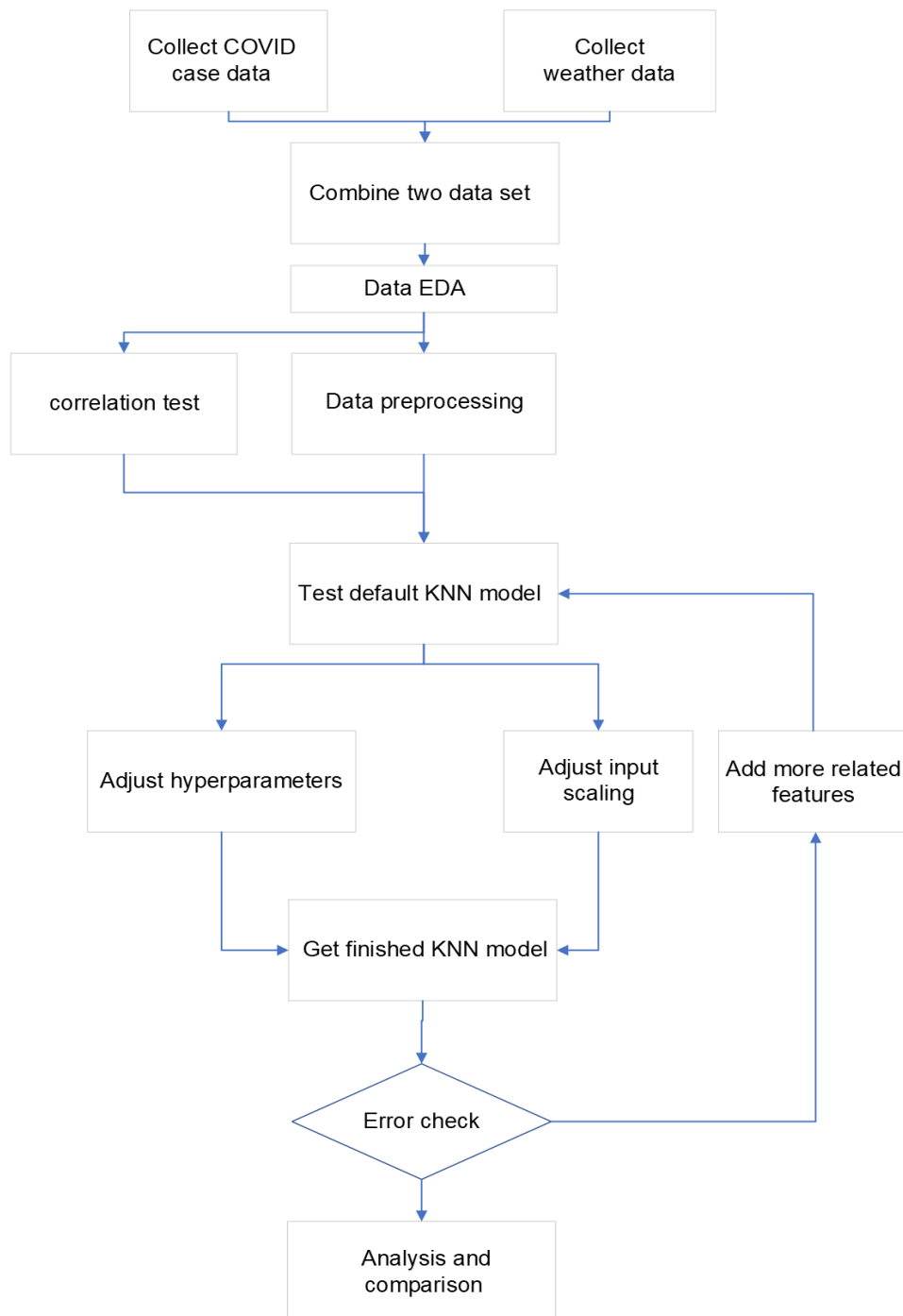


Figure 1. flow chart of the modeling process

Figure 1 shows the flow chart of the modeling process, the detailed steps are described below:

Step 1: Collect data from reliable data source. Both COVID case data and weather data is easy to find but since they usually are collected separately so it is necessary to combine separate data set for following study.

Step 2: Choose potential features in data for more detailed analysis, like correlation test.

Step 3: Before construct the model, the data prepossessing is made to prevent nan term within data set interrupt model processing.

Step 4: Depending on the performance of test run, the parameters of model will be modified and input features will be added and removed.

Step 5: In the end there should be a KNN model that is good enough to make prediction based on selected features.

Step 6: The performance of this model will be analyzed and compared with other types of models.

3. Experiments and Result

3.1. Data

The information about COVID-19 cases from a website of Centers for Disease Control and Prevention (CDC) [11]. Specifically, data about the number and type of COVID-19 cases by States over time were downloaded. CDC get those data from the voluntary report of about 60 jurisdictions. The data set may exist error because of an absence/incompleteness of report from some jurisdictions, or the delay between the symptom's emergence on individuals and testing.

Additional information about climate condition from “visual crossing”, a leading provider of weather data are utilized. Specifically, the historical weather information by States over January 01, 2022 and September 30, 2022 were downloaded. The weather factors include temperature, precipitation, wind speed, cloud cover, snow depth, solar radiation etc.

3.2. Experiment design

3.2.1 Evaluation index

Mean Square Error (MSE) and Root Mean Square Error (RMSE) are two evaluation index for assessing the performance of models precisely. RMSE and MSE values too small means over-fitting, and too large means under-fitting. Therefore, a good machine learning model should expect relative low value for both.

$$MSE(y_{true}, y_{pred}) = \frac{1}{n_{samples}} \sum (y_{true} - y_{pred})^2 \quad (5)$$

Explained variance is a common way to measure the variety of a machine learning model. It reports the variance of the error. This method is used on regression models, the more its value approach to 1 the better the prediction. Explained variance does not consider the systematic offsets in the prediction, so usually R^2 score will take its job. In this research both of them will be used to see the difference.

$$EV(y_{true}, y_{pred}) = 1 - \frac{Var(y_{true} - y_{pred})}{y_{true}} \quad (6)$$

$$R^2(y_{true}, y_{pred}) = 1 - \frac{\sum (y_{true} - y_{pred})^2}{\sum (y_{true} - \bar{y})^2} \quad (7)$$

$$\bar{y} = \frac{1}{n_{samples}} \sum y_{true} \quad (8)$$

3.2.2 Experiment Setting

Most model parameters are default. The training and testing data will be separated via time by 0.8 train 0.2 test ratio.

In this research the correlation test is essential. Scatter plots of COVID total cases and weather factors will provide intuitive result and the team will analyze the behavior of these plots with a general conclusion. A numerical result will be given by correlation coefficients mentioned before. Consider the limitation of chosen coefficients, the correlation level should be relatively low.

3.3. Results

3.3.1 Correlation result

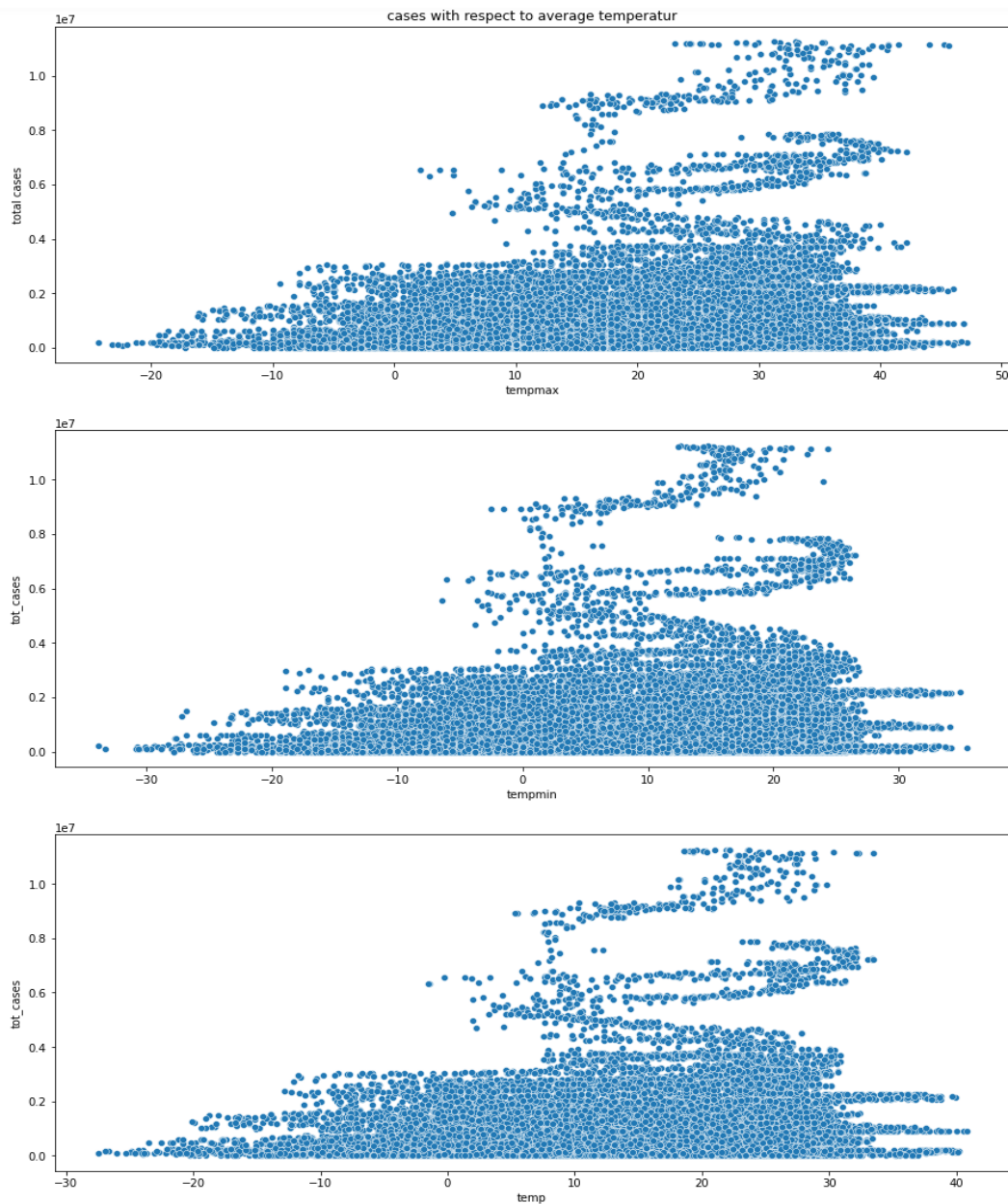


Figure 2. Scatter plots of COVID-19 cases and temperature

According to the COVID-temperature shown in figure 2, it is obvious that high amount of COVID cases lie on [10,40] maximum temperate, [-5,28] minimum temperature and [0,33] average temperature. This shows that the virus has a preferred temperature range and cannot adapt extreme climate.

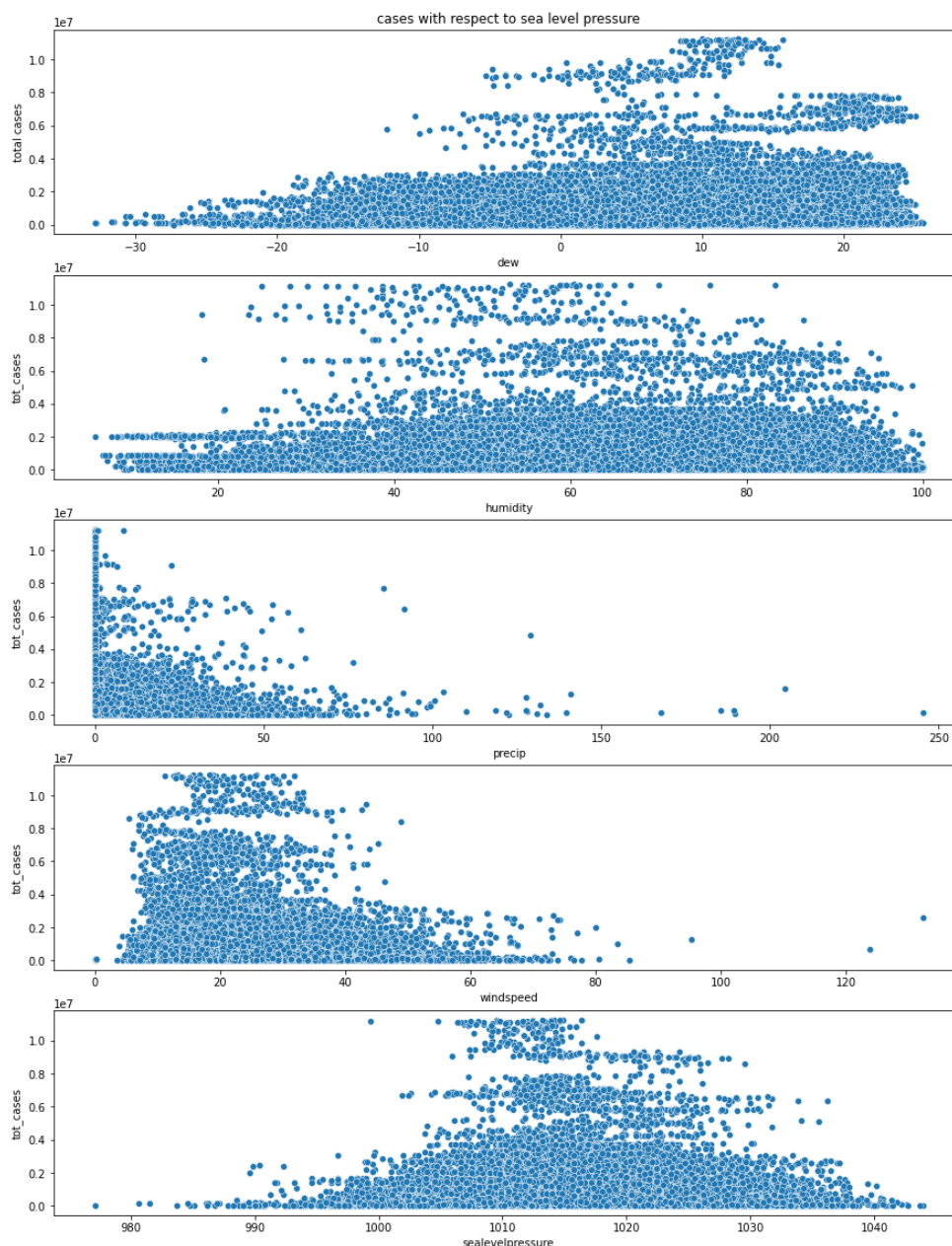


Figure 3. Scatter plots of COVID-19 cases and other climate factors

Figure 3 is scatter plots the COVID cases and other climate factors, also shows some intriguing information. The dew plot shows that virus might not adapt environment with low dew amount. The humidity gives average distribution, but few case-prone area lies on low humidity zone. The precipitation plot indicates that more rain might interrupt case transmission. The windspeed plot reveals the virus might spread more rapidly in less wind area. The sea level pressure plot demonstrates that more cases gather at medium altitude. These results evidence that COVID virus prefer moist weather, but excessive rainfall and windy will obstruct virus transmission. COVID-19 transmission method mentioned before back up this outcome. Virus could stay active in damp air but rain and wind will flush them away.

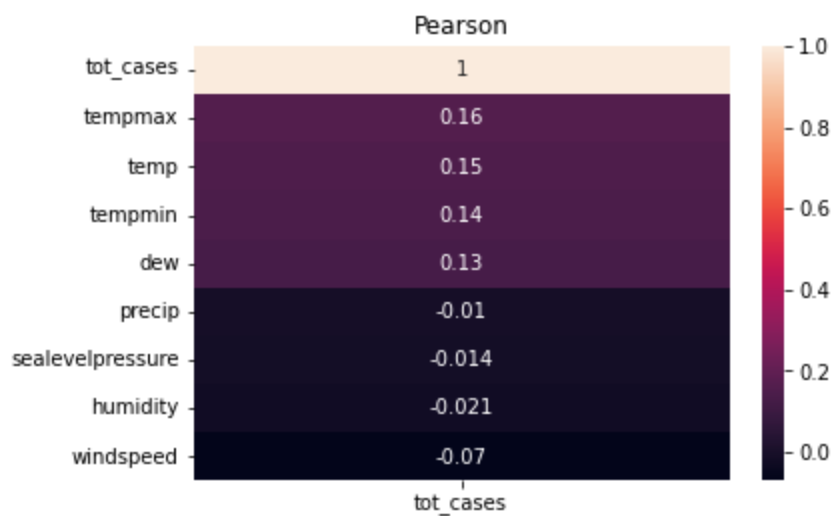


Figure 4. Correlation test from Pearson coefficient

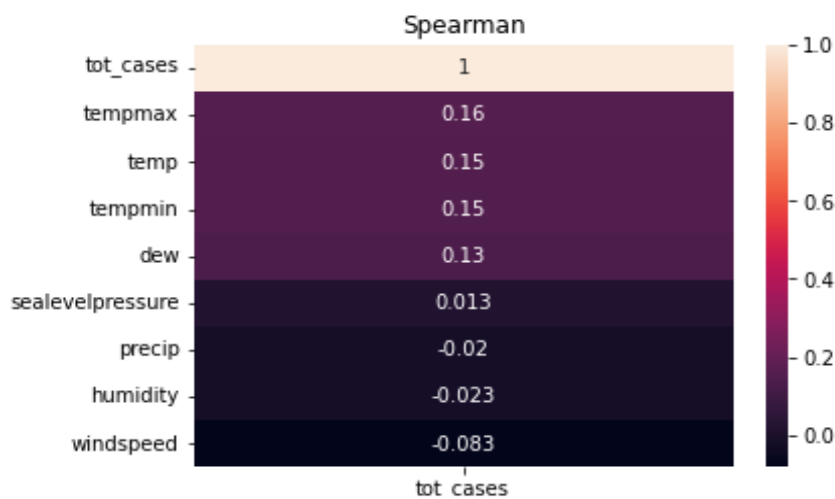


Figure 5. Correlation test from Spearman coefficient

The correlation tests from Pearson and Spearman coefficient gives close result, shown in figure 4 and figure 5. Both result from correlation test give small coefficient. This is predictable because of the limitation and noises in the data. Consider the size of data used in this process, the existence of correlation from COVID cases and selected weather factors is proven. However, currently exhibited correlation also shows that these weather data might not be able to construct a prediction machine learning model alone.

3.3.2 Modeling

A KNN model with $K=5$ is trained with only weather factors.

Table 1. weather factors only

MSE	RMSE	EV	R^2
1976883094845.6294	1406016.7477116443	0.03170814180424508	0.25287403252618623

The result in table 1 shows that the model failed to perform valid prediction. 1.4 million RMSE is 2 times of the average total cases. EV and R^2 coefficient is too low for the model to provide reasonable result. From the plot it is obvious that model underestimates the pandemic and gives an over optimistic prediction.

An A/B test is made to examine how weather factors could affect a COVID prediction model. First model takes only location factors and total cases day before to stimulate the situation where a model needed that uses today's cases number to predict tomorrow's cases number

	AK	AL	AR	AZ	CA	CO	CT	DE	FL	GA	...	SC	SD	TN	TX	VA	VT	WA	WI	WY	prev_tot_cases
28476	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
16725	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
33644	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
28530	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36898	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35457	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	350117.0
8517	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1197762.0
22252	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3760119.0
9582	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	951283.0
12780	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11247840.0

44848 rows × 47 columns

Second model takes weather factors with higher correlations as additional input for comparison, shown in table 2 and table 3.

Table 2. weather factors as additional input

MSE	RMSE	EV	R ²
49286177.83411371	7020.411514584719	0.9999699328741662	0.9999677889186825

Table 3. without weather factor

MSE	RMSE	EV	R ²
41459366.88205128	6438.894849432726	0.9999765283117747	0.9999746578554388

Model without weather factor gives a good result. 7 thousand RMSE is relatively small compare to mean and median of the total cases. EV and R² coefficient indicate that this is a well preformed model.

Model with weather factor contributes improvement directly reflected on RMSE, but increases the total training time. Such comparison shows that weather factors could improve the accuracy in a small amount with the cost of time efficiency.

A more complex input is introduced to in depth discuss the advantage domain of weather factors. Consider the information delay during the collection of COVID case date, collecting the precise cases number from a day before is a challenge. In order to minimize the potential error caused by data collection, cases number from past 7 days is used to construct the model.

	AK	AL	AR	AZ	CA	CO	CT	DE	FL	GA	...	WA	WI	WY	prev1d_tot_cases	prev2d_tot_cases	prev3d_tot_cases	prev4d_tot_cases
28476	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16725	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
33644	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0
28530	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36898	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35457	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	350117.0	349846.0	349630.0	349522.0
8517	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1197762.0	1197762.0	1196540.0	1196540.0
22252	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3760119.0	3758292.0	3756708.0	3754941.0
9582	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	951283.0	950940.0	950441.0	949896.0
12780	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	11247840.0	11220161.0	11220161.0	11220161.0

	tempmax	tempmin	temp	dew	humidity	precip	windspeed	sealevelpressure	AK	AL	...	WA	WI	WY	prev1d_tot_cases	prev2d_tot_cases
28476	9.5	6.6	8.2	4.0	75.4	0.22	14.1	1019.8	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
16725	7.2	-0.7	3.2	-1.7	71.6	0.00	11.1	1021.4	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
33644	-1.0	-4.2	-2.4	-5.9	77.4	0.00	11.1	1022.8	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0
28530	12.7	2.9	7.7	5.0	83.3	0.00	12.7	1018.9	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
36898	2.8	-0.7	0.7	-2.4	80.0	0.00	11.1	1021.7	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35457	16.2	0.7	7.8	3.6	78.9	0.00	16.4	1027.3	0.0	0.0	...	0.0	0.0	0.0	350117.0	349846.0
8517	27.4	13.2	19.8	6.6	44.6	0.00	25.1	1017.8	0.0	0.0	...	0.0	0.0	0.0	1197762.0	1197762.0
22252	21.7	5.0	12.3	5.9	68.8	0.00	16.5	1023.7	0.0	0.0	...	0.0	0.0	0.0	3760119.0	3758292.0
9582	27.2	11.1	18.4	6.2	48.1	0.00	22.2	1019.8	0.0	0.0	...	0.0	0.0	0.0	951283.0	950940.0
12780	33.0	12.6	22.1	10.0	53.1	0.00	12.9	1009.5	0.0	0.0	...	0.0	0.0	0.0	11247840.0	11220161.0

Table 4. more main features

MSE	RMSE	EV	R ²
218688988.67112598	14788.136754545043	0.9998760563033913	0.9998699240853066

Table 5. less main features

MSE	RMSE	EV	R ²
126268592.65707915	11236.929859044201	0.9999232832911351	0.9999170900636078

Table 4 and table 5, show that weather factors still could improve the accuracy of model. Increase of main features leads to more RMSE and processing time, and weather factors no longer reduce time efficiency.

3.3.3 Summary

It is obvious that some weather factors do have correlation with COVID-19 cases both theoretical and on the data. Unfortunately, such correlation is too weak to be directly used in epidemic prediction. There are many more correlated features not taken in consideration such as vaccine rate or overall population, so the information read from weather data could be a reference but shouldn't be used as decisive reason in decision making regards of COVID-19 pandemic. Weather factors presents helpful result when used with other main features with high correlation. Although the improvement provided by weather factors are overall small consider the general model training error, it still gives a good example in taking weather factors into COVID prediction model as reinforcement.

4. Conclusion

The research is created to examine the correlation between COVID-19 cases number and few common weather factors. Moreover, a machine learning model is planned to be built with respect to these weather factors to achieve COVID-19 case prediction.

The correlation analysis proves the consistency of virus transmission theory and real weather data. Although the result of research denies the possibility that weather factors could independently build up prediction model, the potential advantages of these factors be used as support features is confirmed.

References

- [1] Kumar, Anuj and Nishu Ayedee. "An interconnection between COVID-19 and climate change problem." *Journal of Statistics and Management Systems* 24 (2021): 281 - 300.
- [2] Bashir, Muhammad Farhan, Ben-jiang Ma, Bilal, Bushra Komal, Muhammad Adnan Bashir, Duoqiao Tan and Madiha Bashir. "Correlation between climate indicators and COVID-19 pandemic in New York, USA." *The Science of the Total Environment* 728 (2020): 138835 - 138835.

- [3] Mohsinul Hoque, Muhammad, Umme Saima and Sadia Sultana Shoshi. "Correlation of Climate Factors with the COVID-19 Pandemic in USA." (2020).
- [4] Cole, Jennifer and Klaus Dodds. "Unhealthy geopolitics: can the response to COVID-19 reform climate change policy?" *Bulletin of the World Health Organization* 99 (2021): 148 - 154.
- [5] Li, Ziyang, Ya-Han Shen, Lei Huang and Mengzi Yin. "Visual Analysis of Urban Block Risk in Extreme Low Temperature Weather Based on Bayesian Network." *The 2021 3rd International Conference on Big Data Engineering* (2021).
- [6] Lovrić, Mario, Kristina Pavlović, et al. "Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning." *Environmental Pollution (Barking, Essex : 1987)* 274 (2020): 115900 - 115900.
- [7] Magazzino, Cosimo, Marco Mele and Samuel Asumadu Sarkodie. "The nexus between COVID-19 deaths, air pollution and economic growth in New York state: Evidence from Deep Machine Learning." *Journal of Environmental Management* 286 (2021): 112241 - 112241.
- [8] Phillips, Carly A., Astrid Caldas, et al. "Compound climate risks in the COVID-19 pandemic." *Nature Climate Change* 10 (2020): 586-588.
- [9] Shi, Peng, Yinqiao Dong, et al. "Impact of temperature on the dynamics of the COVID-19 outbreak in China." *The Science of the Total Environment* 728 (2020): 138890 - 138890.
- [10] Meyfroidt, Patrick et al. "Ten facts about land systems for sustainability." *Proceedings of the National Academy of Sciences of the United States of America* 119 (2022).
- [11] "United States Covid-19 Cases and Deaths by State over Time - Archived." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, <https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36>.