Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Environmental Management 302 (2022) 114085

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

The impact of weather condition and social activity on COVID-19 transmission in the United States

Xinxuan Zhang [a,*], Viviana Maggioni [a], Paul Houser [a], Yuan Xue [a], Yiwen Mei [b]

[a] George Mason University, Fairfax, VA, 22030, USA
[b] University of Michigan, Ann Arbor, MI, 48109, USA

ARTICLE INFO

Keywords:
COVID-19 transmission
Weather condition
Social activity factor
Machine learning
Random forest regression model

ABSTRACT

The coronavirus disease 2019 (COVID-19) has been first reported in December 2019 and rapidly spread worldwide. As other severe acute respiratory syndromes, it is a widely discussed topic whether seasonality affects the COVID-19 infection spreading. This study presents two different approaches to analyse the impact of social activity factors and weather variables on daily COVID-19 cases at county level over the Continental U.S. (CONUS). The first one is a traditional statistical method, i.e., Pearson correlation coefficient, whereas the second one is a machine learning algorithm, i.e., random forest regression model. The Pearson correlation is analysed to roughly test the relationship between COVID-19 cases and the weather variables or the social activity factor (i.e. social distance index). The random forest regression model investigates the feasibility of estimating the number of county-level daily confirmed COVID-19 cases by using different combinations of eight factors (county population, county population density, county social distance index, air temperature, specific humidity, shortwave radiation, precipitation, and wind speed). Results show that the number of daily confirmed COVID-19 cases is weakly correlated with the social distance index, air temperature and specific humidity through the Pearson correlation method. The random forest model shows that the estimation of COVID-19 cases is more accurate with adding weather variables as input data. Specifically, the most important factors for estimating daily COVID-19 cases are the population and population density, followed by the social distance index and the five weather variables, with temperature and specific humidity being more critical than shortwave radiation, wind speed, and precipitation. The validation process shows that the general values of correlation coefficients between the daily COVID-19 cases estimated by the random forest model and the observed ones are around 0.85.

1. Introduction

The coronavirus disease 2019 (COVID-19) has been first reported in December 2019 and rapidly spread worldwide. According to the World Health Organization (WHO), there are more than 174 million COVID-19 cases that have been confirmed across 219 countries, areas or territories globally as of early June 2021, and number of COVID-19 related deaths are over 3.7 million. The global pandemic has affected our society in many aspects. To better understand this challenging situation, there has been a significant number of studies investigating the dynamics of COVID-19 transmission (Kucharski et al., 2020; Davahli et al., 2021, Sapkota et al., 2021) and the short and long term impacts of COVID-19 on people's life, health condition, and social activities (Goodell, 2020; Melo-Oliveira et al., 2021; Fiok et al., 2021; Sonza et al., 2021, Joseph, 2021).

The first COVID-19 case in the United States (U.S.) was identified in Washington state in January 2020 and remained at a relatively slow rate of transmission throughout February of the same year. The daily number of U.S. confirmed cases started to increase dramatically in March until hitting its first peak in early April 2020. Then, the spread of COVID-19 in the U.S. slowed down due to the stay-at-home orders issued by most of the states. However, the daily number of confirmed cases began to rise again in mid-June since the states reopened gradually. The number of daily confirmed cases started to reduce again in August through October 2020, but a new extreme peak came right after (November 2020 to early January 2021) with over of 200,000 daily confirmed cases during the winter time. As a plausible consequence of the effective COVID-19 vaccines that became available towards the end of 2020, the number of daily confirmed cases started to decrease in mid-January 2021. However, the sharp decreasing curve of daily confirmed cases became

* Corresponding author.
E-mail address: xzhang37@gmu.edu (X. Zhang).

https://doi.org/10.1016/j.jenvman.2021.114085
Received 28 June 2021; Received in revised form 14 October 2021; Accepted 7 November 2021
Available online 11 November 2021

flat since mid-March 2021, while the number of administered vaccine doses are increasing rapidly. The pandemic is still ongoing. As of early June 2021, there are about 63% of the adults in the U.S. had received at least one dose of vaccine, but the average number of daily confirmed COVID-19 cases was still around 20,000 and the cumulative confirmed cases already reached 33 million. More than 607,000 deaths happened in the U.S. indicating an overall COVID-19 death rate of 1.8% while the death rates of the seasonal influenza is usually below 0.1% according to a recent report from World Health Organization (WHO, 2020).

The COVID-19 case rate (number of cases per million people) varies dramatically across the U.S. Despite the socioeconomic differences among different states and counties, it is speculated that the transmission mechanism of the COVID-19 might be related with local meteorological conditions as other respiratory viruses. Several studies have investigated the relationship between the weather condition and COVID-19 transmission. For instance, Wu et al. (2020) used a log-linear generalized additive model to explore the effect of temperature and humidity on the COVID-19 transmission in 166 countries and found that temperature and relative humidity were negatively related to the COVID-19 cases. Chen et al. (2020) established a statistical model to estimate the number of COVID-19 cases with four weather variables (temperature, relative humidity, wind speed, and visibility) during the January to March 2020 time frame. The model-estimated case counts showed an acceptable correlation with the real counts based on the data of 54 countries around the world. Wang et al. (2021) applied a Fama-Macbeth Regression and found effective reproductive number of COVID-19 to decrease with increasing air temperature and relative humidity based on the data of 100 Chinese cities and 1005 U.S. counties from January to April 2020. Haque and Rahman (2020) focused on Bangladesh in the March-to-May 2020 period using a linear regression framework to conclude that high temperature and humidity significantly reduce the COVID-19 transmission. Another study (Mofijur et al., 2020) in Bangladesh using the Spearman rank correlation test showed different results, i.e., only minimum and average temperatures had a significant relationship with the number of COVID-19 cases. More recently, He et al. (2021) studied 9 major Asian cities with generalized additive modeling (GAM) and Pearson correlation. The GAM analysis showed the number of daily COVID-19 cases to be positively associated with the weather variables (i.e., temperature and relative humidity), while the Pearson correlation showed the relationships between COVID-19 cases the weather variables can be either negative or positive depending on different cities.

The results from previous studies all highlight the impact of weather variables such as temperature and humidity on the spread of COVID-19, though the conclusions may vary significantly, e.g., Wu et al. (2020) found a negative relationship between temperature and COVID-19 cases whereas He et al. (2021) reported an increasing in temperature may yield an increase in daily COVID-19 cases in some cities. Moreover, these studies are based either on large (county) scale analyses or several distinct cities. Most research methods proposed in such studies did not consider public health interventions, such as mask wearing and social distancing, which can largely influence the transmission dynamics of COVID-19. In this study, we aim to provide a comprehensive analysis of the relationship between weather conditions and COVID-19 transmission at the county level across the Continental U.S. (CONUS) on a daily basis. A machine learning algorithm is thus developed, by considering not only weather variables, but also the impact of public health interventions. Although the COVID-19 has been spreading across US since January 2020 and the pandemic is still ongoing as of April 2021, for this work we focused on the period from January to September 1st, 2020 because we considered the virus transmission was less impacted by vaccination conditions during this period. Thus, it should be easier to isolate and analyse the influence of weather factors.

2. Study area and Data

2.1. Study area

The study was carried out in 48 states across CONUS, including 3142 counties and independent cities. The population density varies widely across counties (Fig. 1 top). For example, the population density of New York County (Manhattan) is close to 27,000/km^2 while the population density in some rural midwestern counties can be as low as 0.1/km^2. The climate over CONUS varies due to the different terrain features and the wide range of latitudes (Fig. 1 bottom). Major climate types include humid continental, humid subtropical, semi-arid, desert, and Mediterranean, which are all characterized by extremely different temporal and spatial patterns of temperature and humidity.

2.2. Data collection

2.2.1. COVID-19 data

The nationwide county level COVID-19 data were provided by the University of Maryland COVID-19 Impact Analysis Platform (https://data.covid.umd.edu) that was originally developed by Zhang et al. (2020, preprint) at the Maryland Transportation Institute (MTI) in partnership with the Center for Advanced Transportation Technology Laboratory (CATT Lab).

The variables used in this study include the number of new COVID-19 cases (NewC) and the social distance index (SDI) at the county level. Both variables were extracted for the period of March 23rd to September 1st, 2020. NewC represents the number of daily confirmed cases that tested positive to coronavirus detection. The SDI is an integer that ranges from 0 to 100 and represents the extent residents and visitors are practicing social distancing. A value of zero indicates that no social distancing is observed in the community, while 100 indicates that all residents are staying at home and no visitors are entering the county (Zhang et al., 2020, preprint). Specifically, Zhang et al. (2020 preprint) defined the SDI as a combination of six mobility metrics, according to the following equation:

$$SDI = 0.8[SH + 0.01(100 - SH)(0.1RAT + 0.2RBT + 0.4RNT + 0.3RDT)] + 0.2RTO \quad (1)$$

where SH stands for Staying Home, which is the percentage of residents staying at home; RAT is the percentage of Reduction of All Trips compared to a pre-COVID-19 benchmark; RBT is the reduction of business trips (%), RNT is the reduction of non-business trips (%), RDT is the percent Reduction of Travel Distance; and ROT is the Reduction of Out-of-county Trips (%). The weights are chosen based on shared residents and visitor trips (e.g., about 20% of all trips are out-of-county trips, which led to the selection of a weight of 0.8 for resident trips and 0.2 for out-of-county trips); what trips are considered more essential (e.g., business trips more essential than non-work-related trips); and the principle that higher SDI scores should correspond to fewer chances for close-distance human interactions and virus transmissions.

2.2.2. County attributes

The county attributes used in this study, including boundary, area,

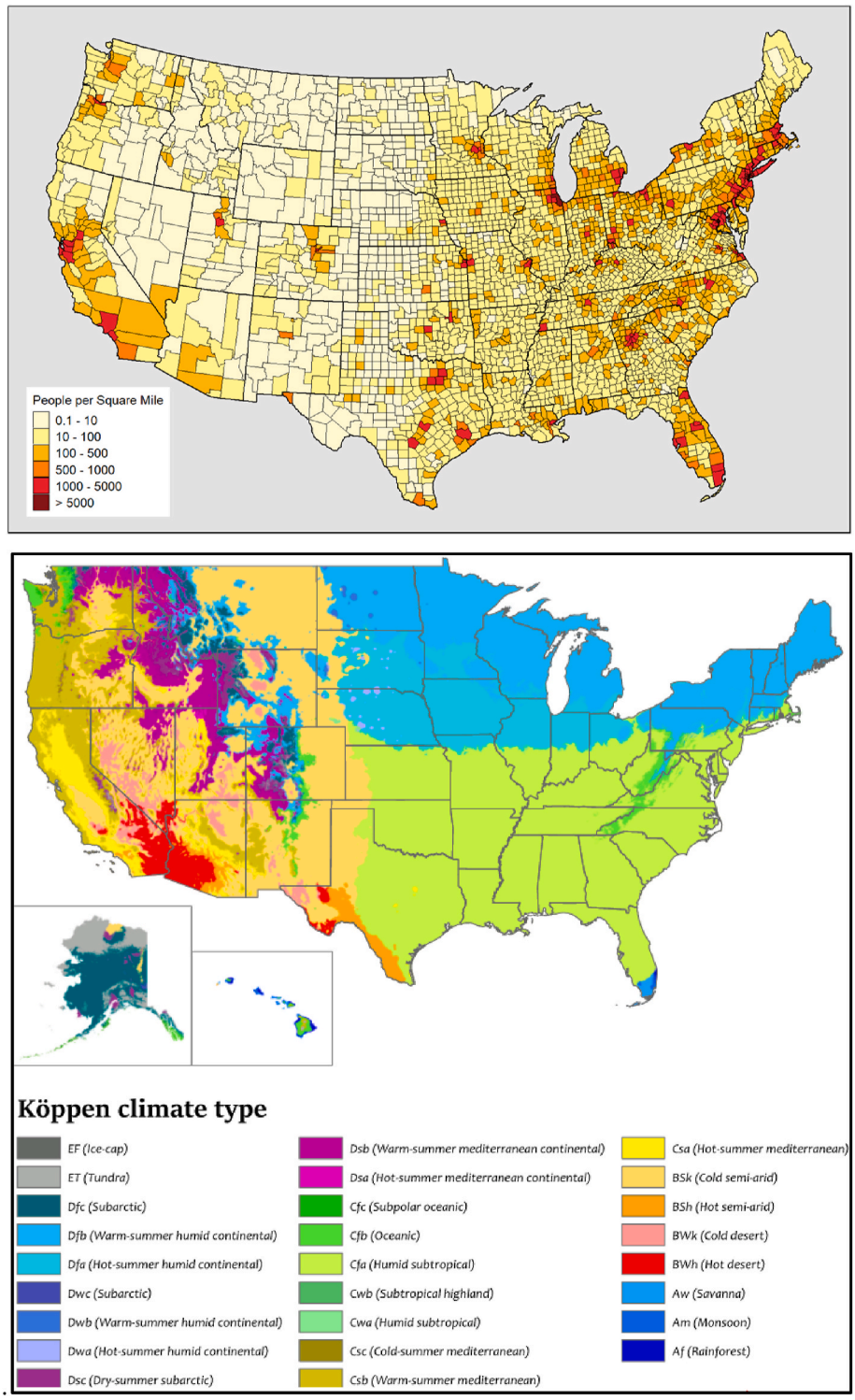

Fig. 1. Top: Population Density by County over CONUS. Bottom: Köppen climate types of the U.S. (Map source https://en.wikipedia.org/wiki/Climate_of_the_United_States. Data source: Köppen types calculated based on data from PRISM Climate Group https://prism.oregonstate.edu/explorer/).

and population, are collected from the U.S. Census Bureau (https://www.census.gov/). The boundary and area data are based on the files in year 2015, while the population data are based on the 2019 estimates.

2.2.3. Weather data

The weather variables are extracted from the North American Land Data Assimilation System – second phase (NLDAS-2) dataset (Xia et al., 2012). NLDAS-2 is an upgraded version of the first phase of the multi-institution NLDAS-1 (Mitchell, 2004) project, which was initiated to provide coupled atmosphere–ocean-land models with reliable initial land surface states for improving weather predictions (Xia et al., 2012). The original NLDAS-2 is available at 1/8° and hourly spatial and temporal resolution, respectively. We processed the gridded hourly NLDAS-2 data from March 23rd to September 1st, 2020 and obtained the county level daily mean values of temperature, specific humid, wind speed,

shortwave radiation, and precipitation over CONUS.

3. Methodology

The study is organized in two parts. The first part presents a traditional statistical analysis to explore the impact of each weather variable on the COVID-19 transmission. In the second part of the study, we develop a machine learning algorithm. The overarching goal is to investigate which weather variable(s) can explain most of the COVID-19 transmission variability across the U.S.

3.1. Traditional statistical analysis

This analysis assesses Pearson correlation coefficients of each weather variable and SDI versus the number of COVID-19 cases per 1000 people (NewC1000) at the county level. The NewC1000 is used here instead of NewC for the sake of eliminating the influence of the county population on the results. In addition, to minimize the uncertainty of COVID-19 spreading due to population density, the analysis is applied to five separate county groups with population density: i) greater than 10,000 people per mile^2, ii) around 1000 people per mile^2, iii) around 500 people per mile^2, iv) around 100 people per mile^2, and v) around 10 people per mile^2. Moreover, as the COVID-19 disease transmission might be largely influenced by human activities, policies, and social distancing, the study time frame is divided into three periods according to different public health intervention levels: i) Period A: non-intervened COVID-19 transmission period (January 1st to March 22nd, 2020), ii) Period B: intervened period (most states issued mandatory stay-at-home orders, March 23rd to May 10th, 2020), and iii) Period C: reopening period (May 11th to September 1st, 2020). The correlation coefficients between each weather variable and COVID-19 case numbers are studied for each group of counties during each period of analysis.

3.2. Machine learning algorithm

Correlation coefficients used above are valid metrics when investigating linear or slightly non-linear problems. Given the complicated nature and highly non-linear response of the COVID-19 transmission mechanisms, we hypothesized that the signal-to-noise ratio between the weather variables and the NewC might not be high enough to be detected due to the complex interactions coupled with human-related impacts. Thus, a nonparametric random forest (RF) regression algorithm (Breiman, 2001; Liaw and Wiener, 2002) was adopted here to further investigate the impact of weather data on COVID-19 transmission. An RF model is a collection of decision trees trained on a random subset of data, using a random subset of predictors. For training the RF model in this study, we used the NewC data, population and population density, SDI, and different selections of weather variables. Specifically, the RF model allows to predict daily NewC values at county level based on all available predictors. Theoretically, the RF model could also be used to predict future values of NewC by using weather forecasts, population data, and predicted SDI values. However, this study did not include the testing of future cases due to the lack of SDI predictions.

The RF model was trained with three versions of predictor list (Table 1). The first version estimated NewC values utilizing eight predictors: county population, county population density, county SDI, and the five weather variables (daily maximum surface air temperature, daily maximum specific humid, daily maximum shortwave radiation, daily total precipitation, and daily maximum wind speed). The second version used the same list of predictors but removed three weather variables (shortwave radiation, precipitation, and wind speed). For the third version, all weather variables were removed in model training, leaving only three predictors (county population, population density, and SDI).

Table 1
RF model predictors.

Predictors	RF model version 1 (8-predictor)	RF model version 2 (5-predictor)	RF model version 3 (3-predictor)
Population	✓	✓	✓
Population density	✓	✓	✓
SDI	✓	✓	✓
Temperature (daily maximum)	✓	✓	–
Specific humidity (daily maximum)	✓	✓	–
Shortwave radiation (daily maximum)	✓	–	–
Precipitation (daily accumulation)	✓	–	–
Wind speed (daily maximum)	✓	-	-

The RF regression algorithm is applied for the intervened and reopening periods (Period B and Period C, March 23rd to September 1st, 2020, 163 days in total) when the reported COVID-19 cases started to increase rapidly. The non-intervened period (Period A, January 1st to March 22nd, 2020) is eliminated in the RF model because of the limited records of nationwide COVID-19 cases. As there are 3112 counties and county equivalent administrative units across CONUS and each region has 163 days of weather, population, and COVID-19-related data, a total of over half a million data records are included in this study. First, we randomly selected 70% of the data records for the RF model training. The remaining 30% of the data records were used for independent model validation. In the RF model training procedure, for each tree, a different two-third subset was randomly taken from the training data records, while the remaining one-third of the training data records served as out-of-bag samples for model evaluation. The number of subset predictors for each tree in the forest was set to be 5 (out of 8) for the first version of the predictor list; 3 (out of 5) for the second version; and 2 (out of 3) for the third version. This procedure was carried out to quantify the importance of predictors for the NewC estimation in the RF model. Different numbers of trees were also investigated in the RF model training in order to identify an optimal number of trees (which would balance model accuracy with model efficiency). Specifically, the RF model performance using 10, 20, 30, 50, 60, and 100 trees was assessed, and results are shown in Section 4.

3.3. Model performance measures

The estimated number of the new COVID-19 cases by all three models for the validation data set are compared with the observed number of cases to determine which model gives the best prediction. As described in section 3.2, there are more than 151,000 county level data records included in the validation procedure. We adopt the normalized root-mean-square-error (NRMSE) and the Pearson correlation coefficient (CORR) to evaluate the model performances. The NRMSE can be calculated using

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(NewC_{model,\ i} - NewC_{observed,i}\right)^2}{n}}}{\overline{NewC_{observed}}} \quad (2)$$

The CORR is obtained by

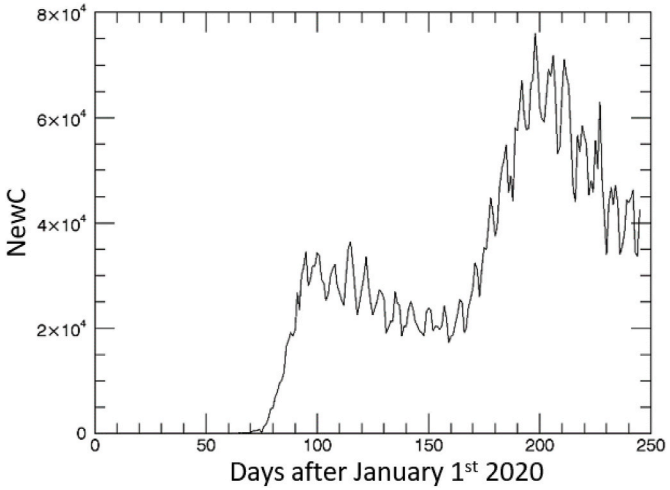

Fig. 2. Time series of CONUS total NewC from January 1st to September 1st, 2020.

$$CORR = \frac{\sum_{i=1}^{n}\left(NewC_{model,\,i} - \overline{NewC_{model}}\right)\left(NewC_{observed,i} - \overline{NewC_{observed}}\right)}{\sqrt{\sum_{i=1}^{n}\left(NewC_{model,\,i} - \overline{NewC_{model}}\right)^2 \sum_{i=1}^{n}\left(NewC_{observed,i} - \overline{NewC_{observed}}\right)^2}} \quad (3)$$

where n is the number of county level validation data records; $\overline{NewC_{observed}}$ is the mean of the NewC values in the validation dataset, and $\overline{NewC_{model}}$ is the model estimated ones.

4. Results

4.1. Temporal change of the COVID-19 and weather condition

The CONUS total COVID-19 NewC values reached two peaks from January to September 2020 (Fig. 2). The NewC showed two peaks during this period. One was in late March and the other one was in mid-July. The spreading of COVID-19 mildly slowed down after the first peak mainly due to the mandatory stay-at-home orders issued in late March in many states.

This is confirmed in the time series of county averaged SDI (Fig. 3f) that the SDI started to go up in late March because of the stay-at-home orders. With the stay-at-home orders gradually expiring in May and June, the SDI decreased immediately and the NewC value went up accordingly. The SDI is definitely a crucial factor that largely impact on COVID-19 transmission, while the impact of weather on COVID-19 is much more complicated. Unsurprisingly, the variation of CONUS averaged weather condition (Fig. 3 a-e) reveals increasing trend for temperature, specific humidity, and shortwave radiation, and there is no clear temporal pattern for precipitation and wind speed. None of the weather variables represents clear relationship with COVID-19 transmission at CONUS scale. The quantitative relationship between the two needs to be further analysed at finer special scale such as county level.

4.2. Traditional statistics

The correlation coefficients between NewC1000 and each of the weather variables and SDI are shown for different group of counties in three periods at different public health intervention levels (Fig. 4). In the first period (January 1st to March 22nd, 2020), most of the weather variables show near-zero values in terms of correlation with NewC1000. This was because the COVID-19 had just started to spread across U.S. at that time. The number of COVID-19 cases was small and the health departments in most states were experiencing difficulty to collect the real-time data. In the second period (March 23rd to May 10th, 2020), although the stay-at-home orders began to be effective, the confirmed COVID-19 cases maintained at a relative high level in many regions (Fig. 2). The correlations between the weather variables and the

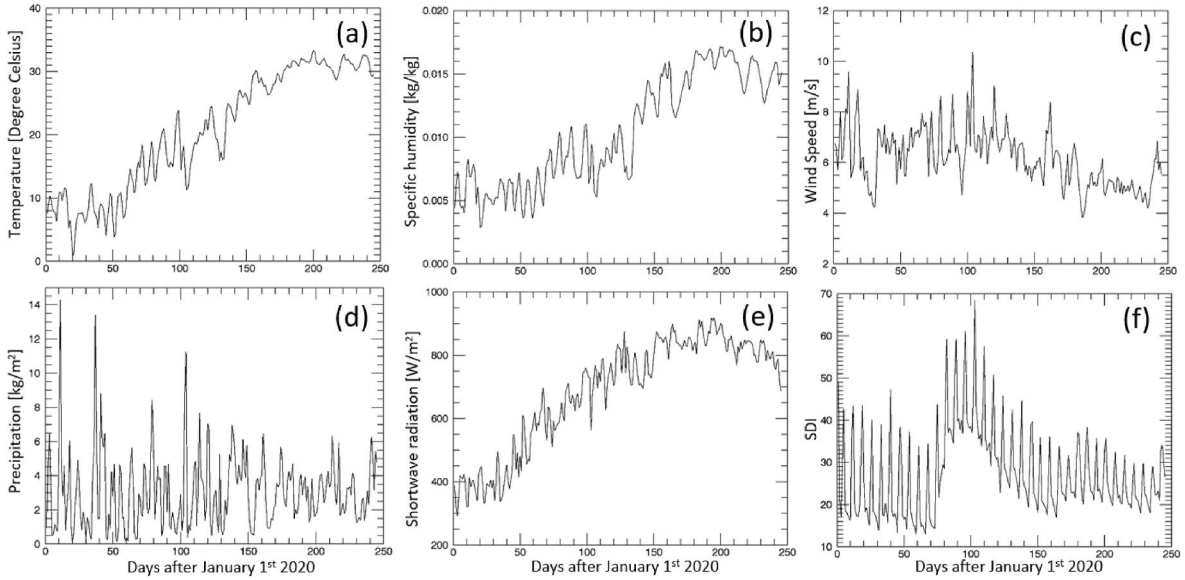

Fig. 3. The series of CONUS averaged a) temperature, b) specific humid, c) wind speed, d) precipitation, e) shortwave radiation, and f) SDI from January 1st to September 1st, 2020.

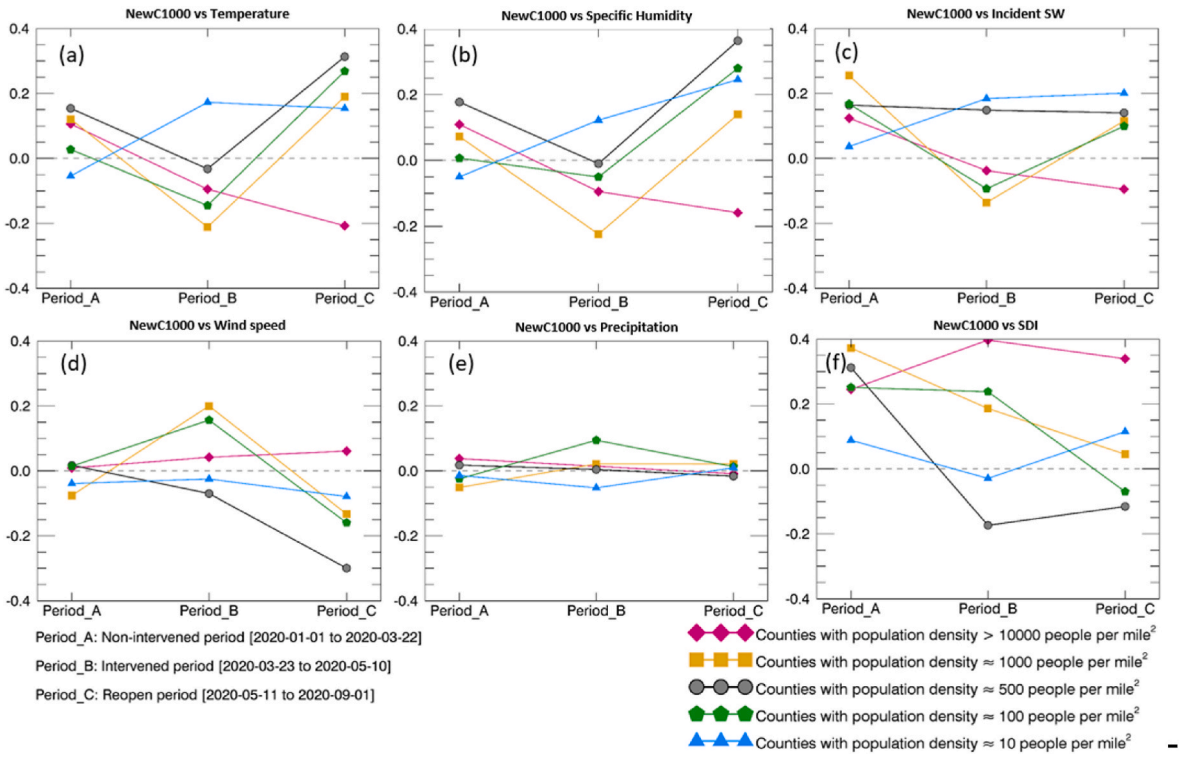

Fig. 4. Correlation coefficients between NewC1000 and weather variables or SDI of selected county groups.

NewC1000 became more obvious except for precipitation which are still close to zero. In the third period, the correlation between temperature/specific humidity and the NewC1000 became even stronger, while wind speed and shortwave radiation keep showing low values as the correlations in the second period. The precipitation has no impact on the NewC1000 in all three periods. In addition, Fig. 4f reveals that the NewC1000 is affected by SDI, especially in densely populated regions. For the areas with low population density (i.e. 10 people per mile^2), SDI rarely impact the NewC1000.

Overall, SDI has a more obvious impact on COVID-19 transmission in densely populated regions. Among the five weather variables, temperature and specific humidity have higher influences on the COVID-19 transmission comparing to wind speed and shortwave radiation. The precipitation can be considered as noninfluential to the COVID-19. These findings are consistent with the results of the RF regression model in section 4.3.

Nevertheless, all the correlation coefficients (even the relative high values) shown in Fig. 4 are not high enough to demonstrate a convincing relationship between the weather variables or SDI and the number of COVID-19 cases. None of the graphs provides a strong and clear trend. Some of them even showed contradictory results. This phenomenon indicates that the traditional linear statistical analysis is not able to

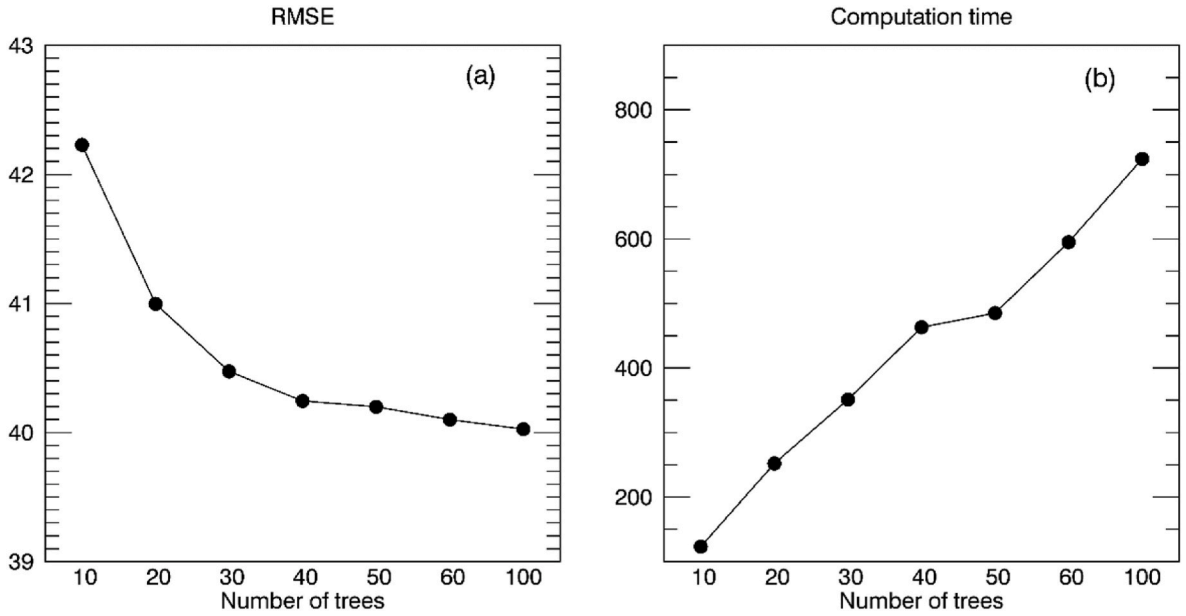

Fig. 5. Model performance in terms of (a) RMSE and (b) computation time as a function of the number of trees in the RF algorithm.

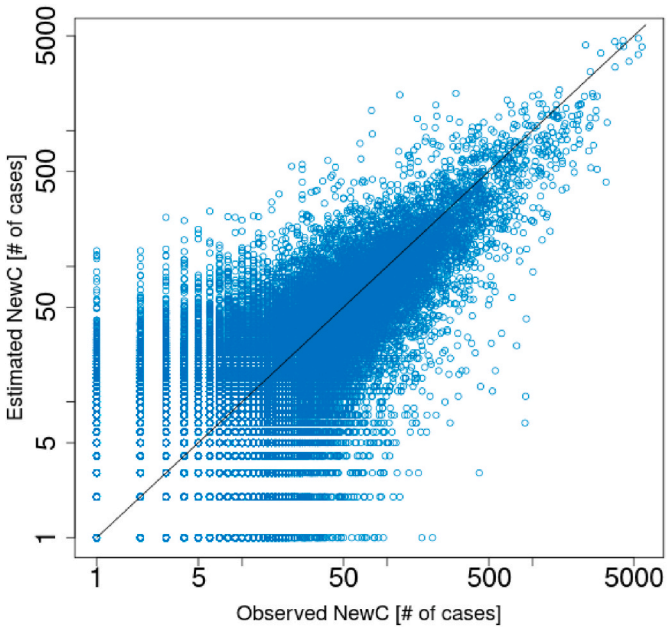

Fig. 6. Scatter plots of estimated and observed NewC at the county level using the 8-predictor version of the RF model.

identify the impact of weather condition and social activity on COVID-19 transmission because of the complicated mechanism of their interactions. The machine learning approach shown in next section is designed to investigate the complicated nature and highly non-linear relationships.

4.3. RF regression model

The RF model aims to estimate the NewC by randomly selecting 5 predictors from the 8 initially identified (temperature, specific humid, wind speed, shortwave radiation, precipitation, county population, county population density, and the county SDI). The model is trained with different numbers of trees to select the optimal number in terms of model accuracy and efficiency. The performance of difference numbers of trees (Fig. 5a) is shown in terms of root mean square errors (RMSEs) of model-simulated NewC versus the observed NewC based on the validation data records (30% of the entire sample pool, around 152,701 county level daily data records). The RMSE values decrease quickly with the increasing of tree numbers from 10 to 40. Then, it shows a mostly flat trend from 40 to 100 trees, though the RMSE still descends slightly. As expected, the model computation time (Fig. 5b) is increasing with adding more trees in the forest, while there is a flat step between 40 and 50 trees. Finally, we chose the 50-tree RF model because it is accurate enough and more efficient than the 60-tree or 100-tree model settings.

As described in section 3.2, the RF regression model was initially trained with 8 predictors, 5 of which are weather variables. Scatter plot based on the validation dataset (Fig. 6) shows that the model performs fairly well for observed NewC above 100, while the bias of the model estimates is evident for values below 100. Although points show a wide spread in the plot, a linear relationship is still distinguishable. Specifically, the correlation coefficient of the data points depicted in Fig. 6 is 0.84.

Table 2
The permutation-based predictor importance for the 8-predictor RF model.

Predictors	Permutation-based importance
Population	24
Population density	18
SDI	10
Temperature (daily maximum)	12
Specific humidity (daily maximum)	12
Shortwave radiation (daily maximum)	8
Precipitation (daily accumulation)	5
Wind speed (daily maximum)	6

Table 3
NRMSE between Estimated and Observed NewC in validation data set.

County population density [People/ Square Mile]	NRMSE		
	RF model version 1 (8-predictor)	RF model version 2 (5-predictor)	RF model version 3 (3-predictor)
0.1–10	4.29	**4.06**	4.88
10–100	3.28	**2.84**	3.08
100–500	**2.74**	3.09	3.31
500–1000	**1.35**	1.43	1.55
1000–5000	1.35	**1.30**	1.57
>5000	**0.98**	1.05	1.09
All testing counties	3.46	3.56	3.96

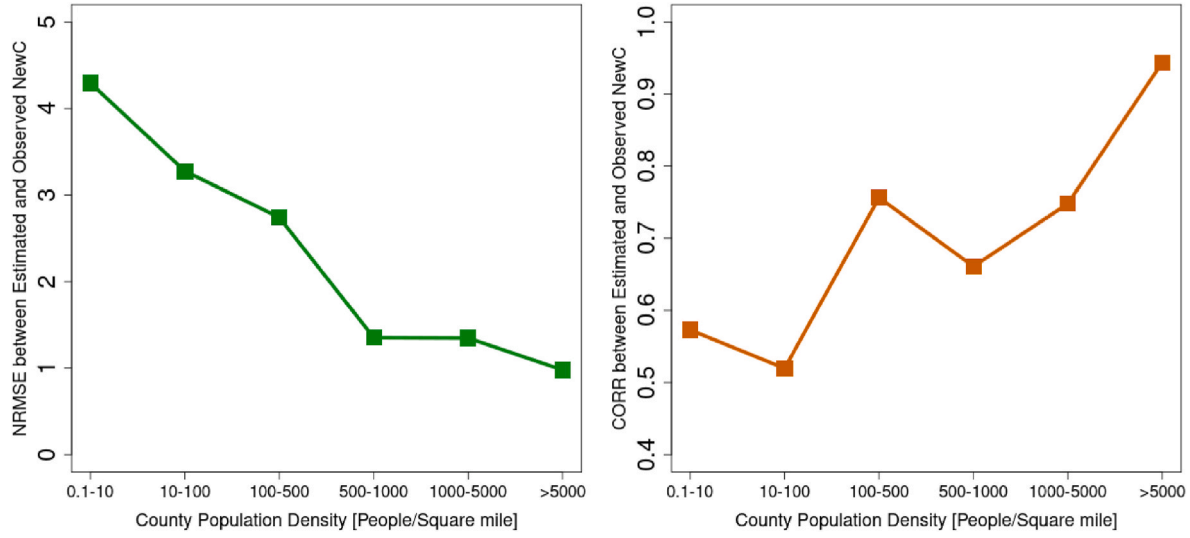

Fig. 7. NRMSE (left) and CORR (right) between the estimated and observed new cases for counties with different population densities (based on RF model 8-predictor version).

Table 4
CORR between Estimated and Observed NewC validation data set.

County population density [People/ Square Mile]	CORR		
	RF model version 1 (8-predictor)	RF model version 2 (5-predictor)	RF model version 3 (3-predictor)
0.1–10	0.57	**0.59**	0.50
10–100	0.52	**0.60**	0.51
100–500	**0.76**	0.75	0.65
500–1000	**0.66**	0.63	0.58
1000–5000	0.75	**0.76**	0.64
>5000	**0.944**	0.938	0.91
All testing counties	0.84	**0.85**	0.77

The model accuracy is affected by county population density as well. There is a clear decreasing trend of the NRMSE between the model estimated and the observed NewC as population density increases (Fig. 7). This performance is substantiated by the higher correlation coefficients observed in more densely populated counties, with the only exception of a small peak value for population densities between 100 and 500 people/mile^2.

Table 2 lists the predictor importance for the 8-predictor RF model. The importance of a predictor in the RF algorithm is estimated by looking at how much the prediction error increases when the out-of-bag data for that predictor is permuted while all others are left unchanged. Specifically, for each tree in the forest, the model records the mean square error (MSE) of the prediction on the out-of-bag portion of the data, then the same procedure is followed after permuting each predictor. Finally, the difference between the two MSEs is averaged over all trees and normalized by the standard deviation of the differences. The most important predictors in this RF model for estimating NewC are population and population density, followed by the SDI and the 5 weather variables. Temperature and specific humidity are more important than shortwave radiation, wind speed, and precipitation, with precipitation being the least important predictor in the RF model, which is consistent with the results of the traditional statistical analysis presented in section 4.2.

Considering the low importance of shortwave radiation, wind speed, and precipitation shown in the 8-predictor RF model, a 5-predictor RF model was trained without these 3 predictors. Furthermore, a RF model in which all weather variables were eliminated was trained with only 3 predictors (population, population density, and SDI). Normalized RMSEs for the three RF models (Table 3) is shown for county groups with different population density. A similar comparison for correlation coefficients is presented in Table 4. Overall, the performance of the 8 and 5 predictor RF models is comparable. The lowest Normalized RMSE and highest correlation coefficients are shown in either the 8 or the 5 predictor RF model. It is worth noting that the 5-predictor RF model performs the best in low population density regions (less than 100 people/square mile), while the 8-predictor RF model shows best scores in the moderate population density regions (100–1000 people/square mile). The RF model with no weather variable (3-predictor model) is never the winner for any county group, which indicates the effectiveness of using weather variables for predicting COVID-19 cases. To summarize, the addition of weather variables, temperature and specific humidity in particular, improves the RF model performance in the estimation of COVID-19 cases.

4.4. Discussion

In general, the RF models show competitive results in the 8-predictor version and the 5-predictor version (Tables 3 and 4), which indicate that the most important weather variables are temperature and humidity. However, the mechanism for which temperature and humidity impact the COVID-19 transmission is complex. There are a couple of hypotheses that may explain such phenomenon. First, the COVID-19 transmission is highly affected by social activities that are usually sensible to temperature. People tend to gather inside in cold and hot days but are prone to outdoor activities in warm weather. Considering that the virus spreads more efficiently when people share limited indoor spaces, the temperature variation could have indirect impact on COVID-19 transmission. Humidity is another key weather factor on COVID-19 transmission as the aerosol spread of virus is possibly dependent on ambient humidity and temperature. Lowen et al. (2007) investigated 20 experiments performed for different relative humidity values ranging from 20% to 80% and different temperatures (5 °C, 20 °C, or 30 °C), indicating that both cold and dry conditions favor transmission. Their results could partly explain the crucial importance of humidity and temperature that we observe in the RF models we developed.

Table 5
Comparison of RF model performances with daily maximum and daily mean weather variables.

Model	NRMSE		CORR	
	Model with daily maximum weather variables	Model with daily mean weather variables	Model with daily maximum weather variables	Model with daily mean weather variables
RF model version 1 (8-predictor, in which 5 of them are weather related)	**3.46**	3.85	**0.84**	0.82
RF model version 2 (5-predictor, in which 3 of them are weather related)	**3.56**	3.34	**0.85**	0.84

As described in section 3.2, the RF models presented above are all built using daily maximum values of the weather variables (except for precipitation). For a more comprehensive understanding of the effects of weather condition, daily mean weather variables are also tested in the models. However, the RF model based on daily mean weather variables are found to be slightly, but constantly, less accurate than the daily maximum weather variable-based model for estimating NewC (Table 5). A possible explanation is that the maximum value of weather variables (e.g. daily maximum temperature) could have more impact on the people's social activities than average values and therefore have stronger influence on COVID-19 transmission.

Moreover, it is worth to mention that the process of data collection, especially for the daily number of COVID-19 cases, can be influenced by some uncertainties. For example, the number of reported cases tends to decrease during weekends due to the lab working schedule; testing or reporting delays could happen in many circumstances; the case number can also be affected by the accessibility of testing resources in a region, etc. All these factors introduce uncertainties in the original data set thus affect the accuracy of the model.

5. Conclusions

The study presented a traditional statical approach and a machine learning algorithm to analyse the impact of weather, population, and social activity factors on COVID-19 transmission in terms of daily COVID-19 cases (i.e., NewC) for all the counties over CONUS. Specifically, we considered 8 factors: county population, county population density, county SDI (i.e. social distance index), and the 5 daily weather variables (air temperature, specific humidity, shortwave radiation, total

precipitation, and wind speed).

The traditional statistical approach (i.e., correlation coefficients between each weather factor and NewC) shows weak correlations coefficients between NewC values and most of the weather factors, while the precipitation does not show any correlation with NewC.

The machine learning approach adopts a random forest model to estimate county level daily NewC values by the 8 factors aforementioned. Three version of the RF model are tested using different subsets of the 8 factors (Table 1). The validations of the three RF models show that the general value of correlation coefficient between the daily COVID-19 cases estimated by the random forest model and the observed ones is around 0.85. Results also show that the most important predictors in the RF model for the NewC estimation are population and population density, followed by the SDI and the 5 weather variables. Temperature and specific humidity are more important than shortwave radiation, wind speed, and precipitation. Precipitation is the least important predictor in the RF model, which is consistent with the result by the traditional statistical approach.

Most weather variables in the RF models presented in this study are based on their daily maximum value, except for precipitation. The daily mean weather variables are also tested in the models but found be to less accurate than the daily maximum weather variable-based model (Table 5).

To our current knowledge, this paper is among few of the studies presenting COVID-19 transmission at county level and daily scale across the entire U.S. We acknowledge that a single study here is not enough to fully resolve the question of how weather and social activity conditions affect COVID-19 transmission. But we believe the presented paper did successfully demonstrate a systematic and robust framework towards addressing the mechanics in COVID-19 transmission with regards to a variety of weather conditions. In the future work, we will extend the study period and incorporate additional parameters such as the percentage of population fully vaccinated into the model to achieve more comprehensive analysis.

Author contribution

Xinxuan Zhang designed and conducted the model simulations, performed analysis, and wrote the article. Viviana Maggioni and Paul Houser directed the project. Yuan Xue contributed in the research idea. Yiwen Mei assisted with the model tuning procedure. All authors supported the article revision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is sponsored by the George Mason University 2020 College of Science/Volgenau School of Engineering seed grant program. All the RF model computations were done with ARGO cluster, administered by the Office of Research Computing at George Mason University, VA (http://orc.gmu.edu). The authors would like to thank the University of Maryland COVID-19 Impact Analysis Platform (https://data.covid.umd.edu) and the Maryland Transportation Institute (MTI) for their efforts to provide the nationwide county level COVID-19 data.

References

Amankwah-Amoah, Joseph, 2021. COVID-19 pandemic and innovation activities in the global airline industry: a review. Environ. Int. 156, 106719. https://doi.org/10.1016/j.envint.2021.106719.

Breiman, Leo, 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, Biqing, Liang, Hao, Yuan, Xiaomin, Hu, Yingying, Xu, Miao, Zhao, Yating, Zhang, Binfen, Tian, Fang, Zhu, Xuejun, 2020. Roles of meteorological conditions in COVID-19 transmission on a worldwide scale. MedRxiv, preprint. https://doi.org/10.1101/2020.03.16.20037168.

Davahli, Mohammad Reza, Fiok, Krzysztof, Karwowski, Waldemar, Aljuaid, Awad M., Taiar, Redha, 2021. Predicting the dynamics of the COVID-19 pandemic in the United States using graph theory-based neural networks. Int. J. Environ. Res. Publ. Health 18, 3834. https://doi.org/10.3390/ijerph18073834.

Fiok, Krzysztof, Karwowski, Waldemar, Gutierrez, Edgar, Saeidi, Maham, Aljuaid, Awad M., Reza Davahli, Mohammad, Taiar, Redha, Marek, Tadeusz, Sawyer, Ben D., 2021. A study of the effects of the COVID-19 pandemic on the experience of back pain reported on Twitter® in the United States: a natural language processing approach. Int. J. Environ. Res. Publ. Health 18, 4543. https://doi.org/10.3390/ijerph18094543.

Goodell, John W., 2020. COVID-19 and finance: agendas for future research. Finance Res. Lett. 35, 101512. https://doi.org/10.1016/j.frl.2020.101512.

Haque, Syed Emdadul, Rahman, Mosiur, 2020. Association between temperature, humidity, and COVID-19 outbreaks in Bangladesh. Environ. Sci. Pol. 114, 253–255. https://doi.org/10.1016/j.envsci.2020.08.012.

He, Zonglin, Chin, Yiqiao, Yu, Shinning, Huang, Jian, Zhang, Casper JP., Zhu, Ke, Azarakhsh, Nima, Sheng, Jie, Yi, He, Jayavanth, Pallavi, Liu, Qian, O Akinwunmi, Babatunde, Ming, Wai-Kit, 2021. The influence of average temperature and relative humidity on new cases of COVID-19: time-series analysis. JMIR Public Health and Surveillance 7, e20495. https://doi.org/10.2196/20495.

Kucharski, Adam J., Russell, Timothy W., Diamond, Charlie, Liu, Yang, Edmunds, John, Funk, Sebastian, Rosalind, M., Eggo, et al., 2020. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect. Dis. 20, 553–558. https://doi.org/10.1016/S1473-3099(20)30144-4.

Liaw, Andy, Wiener, Matthew, 2002. Classification and regression by randomForest. R. News 2, 18–22.

Lowen, Anice C., Mubareka, Samira, Steel, John, Peter, Palese, 2007. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog. 3, e151. https://doi.org/10.1371/journal.ppat.0030151.

Maryland Transportation Institute, 2020. University of Maryland COVID-19 Impact Analysis Platform. University of Maryland, College Park, USA. https://data.covid.umd.edu.

Melo-Oliveira, Eduarda, Maria, Sá-Caputo, Danúbia, José, Alexandre Bachur, Paineiras-Domingos, Laisa Liane, Sonza, Anelise, Ana, Cristina Lacerda, Mendonça, Vanessa, Seixas, Adérito, Taiar, Redha, Bernardo-Filho, Mario, 2021. Reported quality of life in countries with cases of COVID19: a systematic review. Expet Rev. Respir. Med. 15, 213–220. https://doi.org/10.1080/17476348.2021.1826315.

Mofijur, M., Rizwanul Fattah, I.M., Saiful Islam, A.B.M., Uddin, M.N., Ashrafur Rahman, S.M., Chowdhury, M.A., Alam, Md Asraful, Uddin, Md, 2020. Relationship between weather variables and new daily COVID-19 cases in dhaka, Bangladesh. Sustainability 12, 8319. https://doi.org/10.3390/su12208319.

Sapkota, Nabin, Karwowski, Waldemar, Reza Davahli, Mohammad, Al-Juaid, Awad, Taiar, Redha, Murata, Atsuo, Wrobel, Grzegorz, Marek, Tadeusz, 2021. The chaotic behavior of the spread of infection during the COVID-19 pandemic in the United States and globally. IEEE Access 9, 80692–80702. https://doi.org/10.1109/ACCESS.2021.3085240.

Sonza, Anelise, da Cunha de Sá-Caputo, Danúbia, Sartorio, Alessandro, Tamini, Sofia, Seixas, Adérito, Borja, Sanudo, Süßenbach, Jessica, Marcello, Montillo Provenza, Xavier, Vinicius Layter, Taiar, Redha, Bernardo-Filho, Mario, 2021. COVID-19 lockdown and the behavior change on physical exercise, pain and psychological well-being: an international multicentric study. Int. J. Environ. Res. Publ. Health 18, 3810. https://doi.org/10.3390/ijerph18073810.

Wang, Jingyuan, Tang, Ke, Feng, Kai, Lin, Xin, Lv, Weifeng, Chen, Kun, Wang, Fei, 2021. Impact of temperature and relative humidity on the transmission of COVID-19: a modelling study in China and the United States. BMJ open 11, e043863. https://doi.org/10.2139/ssrn.3551767.

World Health Organization, 2020. Coronavirus Disease (COVID-19): Similarities and Differences with Influenza. https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-similarities-and-differences-with-influenza.

Wu, Yu, Jing, Wenzhan, Liu, Jue, Ma, Qiuyue, Yuan, Jie, Wang, Yaping, Du, Min, Liu, Min, 2020. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.139051, 139051.

Zhang, Lei, Ghader, Sepehr, Pack, Michael L., Xiong, Chenfeng, Darzi, Aref, Yang, Mofeng, Sun, Qianqian, Kabiri, AliAkbar, Hu, Songhua, 2020. An interactive COVID-19 mobility impact and social distancing analysis platform. medRxiv [preprint]. https://doi.org/10.1101/2020.04.29.20085472.