

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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Environmental Research 208 (2022) 112761

Contents lists available at ScienceDirect

Environmental Research

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

Impact analysis of environmental and social factors on early-stage COVID-19 transmission in China by machine learning

Yifei Han [a,b], Jinliang Huang [a], Rendong Li [a], Qihui Shao [a,b], Dongfeng Han [a,b], Xiyue Luo [c], Juan Qiu [a,*]

[a] Key Laboratory of Monitoring and Estimate for Environment and Disaster of Hubei Province, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, Wuhan, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Faculty of Resources and Environmental Science, Hubei University, Wuhan, China

ARTICLE INFO

Keywords:
COVID-19
Machine learning
Air pollutants
Social data
Meteorology
Non-pharmaceutical interventions

ABSTRACT

As a highly contagious disease, COVID-19 caused a worldwide pandemic and it is still ongoing. However, the infection in China has been successfully controlled although its initial transmission was also nationwide and has caused a serious public health crisis. The analysis on the early-stage COVID-19 transmission in China is worth investigating for its guiding significance on prevention to other countries and regions.

In this study, we conducted the experiments from the perspectives of COVID-19 occurrence and intensity. We eliminated unimportant factors from 113 variables and applied four machine learning-based classification and regression models to predict COVID-19 occurrence and intensity, respectively. The influence of each important factor was analysed when applicable.

Our optimal model on COVID-19 occurrence prediction presented an accuracy of 91.91% and the best R^2 of intensity prediction reached 0.778. Linear regression-based model was identified as unable to fit and predict the intensity, and thus only the variable influence on COVID-19 occurrence can be explained.

We found that (1) CO VID-19 was more likely to occur in prosperous cities closer to the epicentre and located on higher altitudes, (2) and the occurrence was higher under extreme weather and high minimum relative humidity. (3) Most air pollutants increased the risk of COVID-19 occurrence except NO_2 and O_3, and there existed a lag effect of 6–7 days. (4) NPIs (non-pharmaceutical interventions) did not show apparent effect until two weeks after.

1. Introduction

The novel coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus-2 (SARS CoV-2) has become a global pandemic. As of June 15, 2021, there have been over 175 million confirmed patients and over 3.8 million deaths due to COVID-19 (WHO, 2021). On the initial stage of COVID-19 transmission, WHO confirmed on January 12, 2020, that the infection in Wuhan belongs to coronaviruses (Bashir et al., 2020). The virus spread to surrounding provinces in China rapidly during January, and thus Wuhan, Hubei province restricted personnel movement on Jan 23 to prevent more severe transmission (Yu et al., 2020; Andersen et al., 2021). All provinces of mainland China were on high alert and responded swiftly (Ren et al., 2021). Comprehensive and stringent epidemic prevention and control measures were taken nationwide (The State Council Information Office of the People's Republic of China, 2020). It was proved that the feedback of these measures was encouraging, only sporadic cases have been reported on mainland China since April 29, 2020 (The State Council Information Office of the People's Republic of China, 2020).

The recovery of COVID-19 in mainland China cost less than half a year and has provided a valuable background for scholars to investigate the correlation between the COVID-19 transmission and meteorological, social, demographic and geographical factors (Liu et al., 2020; Sun et al., 2020b; Wang et al., 2021; Xiao et al., 2021). Currently, it is a popular methodology to statistically analyse how various parameters influence the spreading of COVID-19 from both global and regional perspectives. For example, related studies showed that some environmental and economic factors such as $PM_{2.5}$ and GDP have a positive effect on

* Corresponding author. 340 XuDong Rd. Wuhan, 430077, Hubei, China.
E-mail address: qiujuan@apm.ac.cn (J. Qiu).

https://doi.org/10.1016/j.envres.2022.112761
Received 21 October 2021; Received in revised form 14 December 2021; Accepted 16 January 2022
Available online 21 January 2022
0013-9351/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

preventing COVID-19 infection (Ahmed et al., 2021). A study in the United States verified that meteorological parameters (higher temperature) and social background (physical distancing intervention) together contribute to a lower risk of COVID-19 transmission (Guo et al., 2021b). However, most studies on COVID-19 transmission in early 2020 were focused on the impact of environmental quality rather than socio-economic, governance, transportation and other elements (Sharifi and Khavarian-Garmsir, 2020).

Previous studies on COVID-19 transmission factors were usually conducted via statistical methods. Varies of models have been proved to be significantly effective in revealing the impact of different factors and thus provided recommendations in prevention (Babuna et al., 2021; Coşkun et al., 2021; Lorenzo et al., 2021; Xiao et al., 2021). Apart from conventional statistical analysis, spatial statistics has also shown its reliability and priority in digging out the complex correlation from geographical data (Sun et al., 2020a; Andersen et al., 2021; Dupre et al., 2021). Furthermore, some of these statistical models showed the potential to forecast the trend of the COVID-19 pandemic (Guo et al., 2021a).

As a popular technology, machine learning has also been widely applied and thus brought significant findings when it comes to the analysis of the COVID-19 pandemic. Kuo and Fu (2021) integrated 8 types of basic machine learning techniques to build up a robust model to predict the number of patients based on environmental, demographic and mobility data. Li et al. (2021) built a ridge logistic model and found out how numerous factors may affect COVID-19 transmission and fatality. Nevertheless, most of these machine learning-based studies focused on the influence factors of worldwide transmission instead of the early-stage pandemic in China.

Therefore, we applied several machine learning methods on a wide range of environmental, social, economic, meteorological and geographical data during the early-stage COVID-19 pandemic in China, and thus revealed the impact of all these factors. Meanwhile, we evaluated and compared the machine learning-based models to work out an optimal solution for estimating the transmission of COVID-19.

2. Materials and methods

2.1. Data source

In this study, we collected prefecture-level data in mainland China and all the data obtained were daily data between Jan 10 and Jun 11 when the number of daily new patients decreased to less than 300 (Johns Hopkins University, 2021). Each record represented the parameters of each prefecture per day.

The data of confirmed and newly confirmed patients were obtained from the National Health Commission of the People's Republic of China (2021) and provincial or prefectural health commissions. 13,332 newly confirmed cases of novel coronavirus pneumonia (NCP) were officially released on Feb 12, such a numerous increase was due to the change of diagnosis standard, which was previously relying on nucleic acid testing but then changed to clinical diagnosis (Chinadaily, 2020). We evenly distributed the increment of patients on Feb 12 to each of the previous days to assure our data reflected the true transmission.

Since the relation between an independent factor and disease incidence counts may fail to indicate the association with disease transmission (Zhao, 2020). It should be noticed that COVID-19 was not balanced distributed since the COVID-19 transmission varied a lot among different prefectures. Therefore, our study was conducted from two separate perspectives of COVID-19 occurrence (COV_O) and COVID-19 intensity (COV_I). COV_O were binarized to 1 and 0 depending on whether COVID-19 is over 0. COV_I inherited the natural logarithm of COVID-19 to smooth the unbalanced distribution.

The meteorological data we collected included daily minimum temperature (MinT), maximum temperature (MaxT), mean temperature (MeanT), relative humidity (Rh) and minimum relative humidity (MinRh). These data were downloaded from the China Meteorological Data Service Center (2021).

The atmospheric environmental quality data, including daily CO, NO_2, O_3, $PM_{2.5}$, PM_{10}, SO_2 and AQI (air quality index), were downloaded from The Data Center of The Ministry of Ecology And Environment of The People's Republic of China (2021). We applied linear interpolation to replace missing values. Meanwhile, considering the delay effect of atmospheric environmental indicators, we defined CO_1 as the CO content of one day before, and CO_9 was the CO content of 9 days before. The same naming rule was used on all the other atmospheric environmental indicators.

The geographical data used in this study was a 30m digital elevation model (DEM) product called 'ASTER GDEM', which was downloaded from Geospatial Data Cloud (2021). The mean DEM (MeanDEM) of each prefecture was calculated.

The socio-economic data we used consisted of three parts. The first part is human-related economic data. We collected the household population (Pop), gross domestic product (GDP), and population density (PD) of the year 2019, which were the latest accessible version. The first two were collected from all provincial statistical yearbooks of China. The population density was downloaded from Ministry Of Housing and Urban-Rural Development of the People's Republic of China (2021).

The second part is migration-related data. We collected destination proportion in population flow from Wuhan (WH) and migration scale (MS) from Jan 1 to Jan 23 (Wuhan lockdown), 2020 (Baidu, 2020). Then we calculated the destination migration scale flow from Wuhan (Popmob) according to the following equation:

$$Popmob = WH \times MS \tag{1}$$

Meanwhile, considering the delay effect, we defined Popmob1 as Popmob of one day before, and Popmob9 was defined as Popmob of 9 days before, and so on naming Popmob0 to Popmob9. Besides, the cumulative Popmob (Popmobsum) from Jan 1 to each corresponding date was also calculated. The distances from Wuhan to the geometric centre of each prefecture (DisWH) were calculated.

The third part is the level of the adopted control measures (Reslevel) in 366 prefectures. These data were collected from the National Health Commission of the People's Republic of China (2021) and provincial or municipal health commissions. China's public health alert system is categorized into four levels in terms of the nature of the incidents, the extent of harm and scope: Level-I (extremely significant), Level-II (significant), Level-III (major) and Level-IV (normal) (The Central People's Government of the People's Republic of China, 2020). We quantified these response levels that no response, Level-IV, Level-III, Level-II and Level-I were ranked 0, 1, 2, 3 and 4, respectively. Meanwhile, considering the delay effect, Reslevel1 was defined as the response level one day before, Reslevel20 was defined as the response level 20 days before, and so was the others.

Days from Jan 10 (Time) of each record was also extracted and regarded as a potential influence parameter considering that the temporal aspect may influence a lot in the early-stage transmission.

More specific information of all the aforementioned data was shown in Appendix A.

2.2. Factor filtering

In this study, we collected 113 factors. Such high dimensional data would massively increase the difficulty of machine learning. Besides, some factors may not have a strong correlation to COVID-19 occurrence or COVID-19 intensity where such interruption might bring unexpected noise to our analysis. Therefore, a preprocess of factor filtering was necessary.

2.2.1. COVID-19 occurrence

We built Gradient Boosting Decision Tree (GBDT) and Random Forest (RF) classification models on COV_O to find out the importance of

each factor. These two machine learning-based classification models were built and modified based on the Python module 'Scikit-learn' (Pedregosa et al., 2012).

RF fits a certain number of decision trees trained on randomly selected subsamples of the training data by the bagging approach and the output of the RF classification model is the class selected by most trees (Kulkarni and Lowe, 2016; Kuo and Fu, 2021). GBDT is also in the form of ensemble decision trees but uses boosting approach and it usually overperforms RF (Piryonesi and El-Diraby, 2020; Kuo and Fu, 2021). In this preprocessing, we set the number of trees to 1000 for both two models.

Each type of data was normalized to the interval of [0, 1] based on the maximum and the minimum. The influence of each factor was calculated by both these two models. The influence was computed via Gini importance, which was derived from the Gini index (Breiman, 2001). Since our classification was binary according to two categories of COV_O, the Gini index was then computed by the following equation (Qi, 2012):

$$G_k = 2p(1-p) \tag{2}$$

where G_k is the Gini index of node k, and p represents the fraction of the positive example assigned to this node.

The Gini importance of each feature in a tree was the sum of the Gini index reduction and the feature importance was the sum of Gini importance in all trees (Qi, 2012). The higher feature importance means a stronger correlation.

For each model, we recorded 50 factors the influence of which ranked the least and deleted the factors which were listed in both these two results. In this way, the least related variables were all removed to avoid noise interruption in machine learning.

2.2.2. COVID-19 intensity

Similar to the preprocess on COVID-19 occurrence, GBDT and RF were also adopted to build regression models to explore the importance of each factor for COVID-19 intensity and we also applied the same normalization to each type of data. The number of trees was set to 1100 and 1500 for GBDT and RF regression models respectively. The last 50 influence factors were also recorded and the repetitive factors were dropped as well.

2.3. Estimation models

2.3.1. COVID-19 occurrence

Since we have 51,776 records the COV_O of which were 0 and only 4588 records were 1, we randomly select 10% from the former COV_O records and thus the distribution of these two categories was more balanced to avoid overfitting problem. Then we randomly split the training and test data for the following classification models by the proportion of 70% and 30%.

Four machine learning models were adopted to estimate the COVID-19 occurrence. We retained identical GBDT and RF classification applied in factor selection considering they were among the most popular realizations of ensemble learning (Piryonesi and El-Diraby, 2020). Moreover, we have built a ridge classification model and a 3-layer artificial neural network (ANN), then compared them with the aforementioned two models to find the optimal solution.

The ridge classifier is referred to a least-squares support vector machine (LSSVM) with a linear kernel that transforms the binary targets into (−1, 1) and then does a regression task (Scikit-learn, 2021). To increase the reliability of COV_O estimation, we applied 10-fold cross-validation to our ridge classifier.

The ANN included three hidden layers where the neural nodes were 64, 48, and 16. All records were segmented into training, validation, and test sets by 60%, 20%, and 20%, respectively.

The results of all four classification models were evaluated based on their accuracy, precision and recall calculated by the following equations:

$$recall = c_i/n_i \tag{3}$$

$$precision = c_i/m_i \tag{4}$$

$$accuracy = (c_1 + c_0)/s \tag{5}$$

where c_i refers to the number of correct estimations of class i, n_i refers to the true record number of the class i, m_i is the number of class i records estimated by the model, and s is the total amount of all records.

Considering the algorithm of the ridge classifier is based on linear regression, the coefficients of all selected factors were finally computed. The factors with an absolute coefficient over 0.1 were then listed and analysed.

The overall workflow on COVID-19 occurrence is shown in Fig. 1.

2.3.2. COVID-19 intensity

Based on 4588 records that presented non-zero COVID-19 intensity, we kept them all and also split them into training and test sets by 70% and 30%.

Consistently, we employed four machine learning techniques to estimate the COVID-19 intensity. GBDT and RF were still preserved due to their good performances. However, instead of ridge regression, in this experiment, we used elastic net (EN) with 10-fold cross-validation. Compared to ridge regression, EN can discard unimportant variables but keeps the stability, and it usually gives a significantly accurate prediction (Zou and Hastie, 2005). Besides, we also constructed a 2-layer ANN which had two hidden layers individually consisting of 96 and 10 nodes.

The results were finally assessed by the coefficient of determination (R^2) and mean squared error (MSE) of all estimations for both training and test sets. R^2 and MSE were calculated by the following equations:

$$R^2 = \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{6}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{7}$$

where m refers to the records of all non-zero COV_I, y_i refers to the true COV_I of the i^{th} record, $\hat{y}_i$ refers to the estimated COV_I of the i^{th} record, and $\bar{y}$ is the mean COV_I of all records.

The EN model is also a linear regression-based model and thus its coefficient analysis of all input variables was then conducted to explain

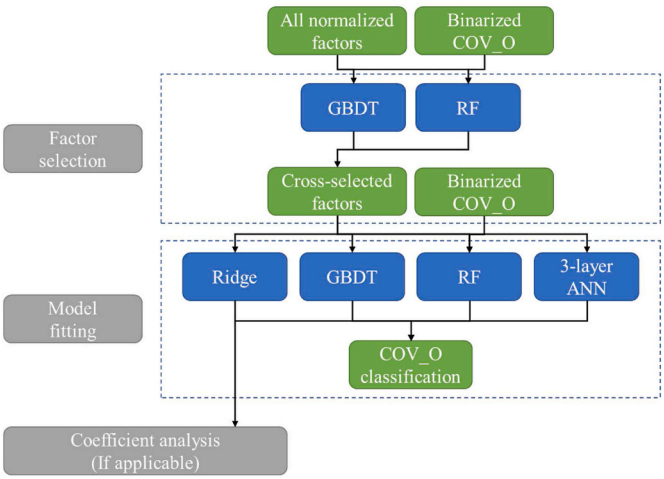


Fig. 1. Overall workflow on COVID-19 occurrence classification.

how each factor would impact the COVID-19 intensity.

The overall workflow on COVID-19 intensity is shown in Fig. 2.

3. Results

The last 50 COV_O influence factors calculated by GBDT and RF classification are shown in Table 1. Based on these results, 26 factors (Table 2) were finally cross-selected and dropped before fitting the estimation model.

Apart from the AQI with a lag of 7 days, all the other AQI factors were proved to be unrelated to the number of patients and deleted. Some other air pollutants such as SO_2, CO, PM_{10}, $PM_{2.5}$ and also the migration scale flow from Wuhan with the lag of various days were identified lack of influence and then removed.

As for the last 50 COV_I influence factors (Table 3), some Air pollutants such as AQI, CO, NO_2, $PM_{2.5}$ and SO_2 were removed based on the cross-selection results (Table 4). Besides, two social factors – the migration scale from Wuhan and the response level – with the lag of certain days were also removed.

The final performances of four machine learning models on the test sets for COVID-19 occurrence classification are shown in Fig. 3 as the format of confusion matrixes. The accuracy, precision, and recall were also calculated and shown in Table 5.

All these four techniques showed spectacular performances on the classification of test sets that their accuracies were identically over 88%, and GBDT presented the highest accuracy of 91.91%. Meanwhile, all the high precisions and recalls also indicated that these four classifiers were similarly convincing.

Considering that our Ridge classifier presented excellent performance, 38 factors that showed high influence with the absolute coefficient over 0.1 were retained and listed in Table 6.

As for the prediction on COVID-19 intensity, the results of all four regression models are presented in Fig. 4 and their assessment metrics are shown in Table 7.

The R^2 of the EN model was below 0.4 and its MSE reached over 1.0, which was not feasible enough when the dependent variable (natural logarithm of COV_I) ranged from −4 to 6. In this case, the coefficient analysis was not applicable. The other three regression models (GBDT, RF, and 2-layer ANN) all performed well that their R^2 on both training and test sets were over 0.75.

4. Discussion

Most of our proposed machine learning classification or regression models were tested to be reliable. All four classification models provided convincing accuracies and GBDT was tested to be the optimal solution. GBDT performed the best on all three assessment indices compared to the other models.

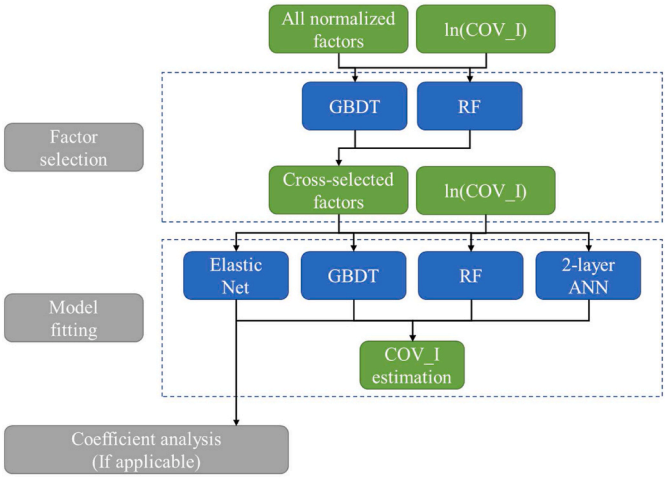


Fig. 2. Overall workflow on COVID-19 intensity estimation.

Considering the good performance of our ridge classification model, the coefficient analysis of all the selected factors on COVID-19 occurrence, therefore, brought reasonable and inspiring findings:

The analysis of our meteorological data showed that the maximum temperature had a positive impact on the occurrence of COVID-19 cases while the minimum temperature showed a negative impact. Although it was believed that the higher temperature would reduce the possibility of COVID-19 transmission (Liu et al., 2020), our findings indicated that the COVID-19 was more likely to occur where the temperature was extreme and such connection was stronger where the maximum temperature is higher. However, the meteorological impacts could be very complex considering that human interruptions such as the frequencies of outdoor activities were highly related to the weather. Our study also indicated that the minimum relative humidity had a positive impact on COVID-19 occurrence and this could be supported by current studies (Qiu et al., 2021).

The analysis of the atmospheric environmental quality data explained the correlation from different perspectives. It was believed that CO, NO_2, O_3, PM_{10}, and $PM_{2.5}$ synergistically impact the COVID-19 occurrence with a lag of 8–9 days (Qiu et al., 2021). Our study verified that such lag did exist but such duration should be 6–7 days, except CO and O_3. CO at 4–5 days in advance was found to be the inflection point that its lag effect was shorter than the other pollutants. Besides, O_3 did not show a significant inflection point and thus its lag effect was still yet to be explored. Another surprising finding was that NO_2 and O_3 both showed a negative correlation with COVID-19 occurrence, which was not aligned with the conclusion that most scholars thought they were supposed to increase the severity of transmission (Barnett-Itzhaki and Levi, 2021). However, we applied the lag effect to avoid the possible impacts of COVID-19 on the air pollutants which can be confusing when fitting the prediction model. Therefore, our study should be more reliable compared to the former researches.

In this study, it was found that COVID-19 patients were more likely to occur in the prefectures with higher GDP, denser population, more intensive population mobility and closer to the epicentre, which was aligned with some studies (Coccia, 2020; Qiu et al., 2021). Nevertheless, for population density, some studies stated that it was not significantly associated with the spreading of COVID-19 (Sun et al., 2020b). Considering this inconsistency, an explanation for our result was that the transportation capacity in prosperous cities was usually larger and thus increased the frequency of human interaction, which made it vulnerable to the COVID-19 epidemic. Whereas the medical facilities were usually more accessible and sufficient in big cities to prevent a further internal epidemic, and thus these factors might negatively influence the COVID-19 intensity instead of occurrence. However, this needed to be proved by further exploring.

Our result also presented that the advance control measure response level was not positively influencing the nationwide spreading until the lag of 15 days, where it was found to be the turning point from the negative coefficient to the positive. This finding was consistent with the widely acknowledged viewpoint that the NPIs (non-pharmaceutical interventions) continuing for two weeks can reduce the spreading effectively (Vardavas et al., 2021).

In our study, the finding on the geographical data revealed that COVID-19 occurrence is higher where the altitude is lower, which was consistent with the finding of Sun et al. (2020b). The landscape and the prefectural development status in China mainland, the prefectures located on higher elevation were closely associated with both the meteorological data as well as the socio-economic status. Hence, such a negative correlation was assumed partly due to the synergy of these two causes.

As for the prediction on the number of COVID-19 patients, RF presented the lowest MSE and the best R^2 on the training set but GBDT

Table 1
The last 50 COV_O influence factors calculated by GBDT and RF.

GBDT				RF			
Factor	Influence	Factor	Influence	Factor	Influence	Factor	Influence
Reslevel17	<0.0001	AQI_2	<0.0001	Popmob1	0.0011	SO2	0.0026
Reslevel6	<0.0001	SO2_6	<0.0001	Popmob0	0.0013	PM10_8	0.0026
Reslevel14	<0.0001	CO_1	0.0001	Popmob2	0.0015	AQI_2	0.0026
Reslevel7	<0.0001	Popmob3	0.0001	Popmob3	0.0015	PM25	0.0026
Reslevel8	<0.0001	PM10_2	0.0001	CO_1	0.0020	Popmob4	0.0026
Reslevel11	<0.0001	AQI_3	0.0001	Popmob5	0.0020	AQI_6	0.0027
Reslevel9	<0.0001	PM25_4	0.0001	Popmob7	0.0020	AQI_3	0.0027
AQI_4	<0.0001	Reslevel20	0.0001	CO_3	0.0021	SO2_1	0.0027
AQI_9	<0.0001	CO_4	0.0001	CO_9	0.0021	AQI_9	0.0027
Popmob6	<0.0001	PM25_6	0.0001	CO	0.0021	PM10_7	0.0027
Popmob5	<0.0001	AQI_1	0.0001	SO2_3	0.0022	Popmob9	0.0027
Popmob2	<0.0001	PM10	0.0001	CO_6	0.0022	PM10_2	0.0027
Reslevel3	<0.0001	AQI	0.0001	CO_2	0.0023	PM10_3	0.0027
Reslevel13	<0.0001	PM10_1	0.0001	CO_5	0.0023	AQI_7	0.0027
Reslevel5	<0.0001	PM10_5	0.0001	SO2_5	0.0023	PM10_6	0.0027
Reslevel12	<0.0001	SO2	0.0001	SO2_6	0.0024	SO2_9	0.0028
Reslevel16	<0.0001	CO_9	0.0001	SO2_2	0.0024	AQI_4	0.0028
Reslevel10	<0.0001	NO2_9	0.0001	SO2_7	0.0024	AQI_1	0.0028
CO_8	<0.0001	O3_8	0.0001	CO_8	0.0024	PM10_9	0.0028
Reslevel4	<0.0001	NO2_1	0.0001	SO2_4	0.0024	PM10_1	0.0028
Rh	<0.0001	PM25_8	0.0001	Popmob6	0.0025	Popmob8	0.0028
PM10_8	<0.0001	Popmob1	0.0001	CO_4	0.0025	PM10_5	0.0028
AQI_6	<0.0001	AQI_5	0.0001	AQI_8	0.0025	PM25_6	0.0029
AQI_8	<0.0001	PM25_9	0.0001	CO_7	0.0026	AQI_5	0.0029
Popmob8	<0.0001	CO_3	0.0001	SO2_8	0.0026	PM10_4	0.0029

Table 2
26 cross-selected COV_O influence factors to be deleted.

AQI_1	AQI_8	CO_9	Popmob1	SO2
AQI_2	AQI_9	PM10_1	Popmob2	SO2_6
AQI_3	CO_1	PM10_2	Popmob3	
AQI_4	CO_3	PM10_5	Popmob5	
AQI_5	CO_4	PM10_8	Popmob6	
AQI_6	CO_8	PM25_6	Popmob8	

Table 4
30 cross-selected COV_I influence factors to be deleted.

AQI_1	NO2_9	Popmob7	Reslevel15	SO2_2
AQI_2	PM25_1	Reslevel1	Reslevel18	SO2_3
AQI_3	PM25_2	Reslevel10	Reslevel20	SO2_4
CO_5	PM25_6	Reslevel11	Reslevel5	SO2_5
CO_6	Popmob5	Reslevel12	Reslevel6	SO2_6
CO_9	Popmob6	Reslevel13	Reslevel7	SO2_8

Table 3
The last 50 COV_I influence factors calculated by GBDT and RF.

GBDT				RF			
Factor	Influence	Factor	Influence	Factor	Influence	Factor	Influence
Reslevel6	<0.0001	SO2	0.0002	Reslevel12	0.0000	CO_3	0.0008
PM25_9	<0.0001	Reslevel7	0.0002	Reslevel14	0.0000	SO2_5	0.0008
Reslevel20	<0.0001	Popmob5	0.0002	Reslevel10	0.0001	SO2_4	0.0008
Reslevel5	<0.0001	NO2_2	0.0002	Reslevel13	0.0001	SO2_3	0.0008
Reslevel13	<0.0001	SO2_6	0.0002	Reslevel11	0.0001	CO_6	0.0009
SO2_2	<0.0001	Popmob7	0.0002	Reslevel9	0.0001	PM25_6	0.0009
CO_2	<0.0001	PM10_1	0.0002	Reslevel6	0.0001	PM25_1	0.0009
CO_5	<0.0001	Reslevel11	0.0002	Reslevel15	0.0001	AQI_1	0.0009
Reslevel12	<0.0001	PM25_6	0.0003	Reslevel8	0.0002	PM25	0.0009
CO	<0.0001	AQI_8	0.0003	Reslevel17	0.0002	SO2_9	0.0009
CO_8	<0.0001	PM10_9	0.0003	Reslevel7	0.0002	Reslevel20	0.0010
Reslevel1	<0.0001	PM10	0.0003	Reslevel5	0.0002	CO_9	0.0010
SO2_8	<0.0001	SO2_4	0.0003	Reslevel18	0.0002	PM25_7	0.0010
CO_1	0.0001	AQI_2	0.0003	Popmob8	0.0003	AQI_3	0.0010
AQI_3	0.0001	NO2_8	0.0003	Popmob9	0.0003	PM25_2	0.0010
NO2_4	0.0001	PM25_8	0.0003	Reslevel16	0.0003	SO2_8	0.0010
Reslevel15	0.0001	PM25_2	0.0003	Popmob7	0.0004	SO2_2	0.0010
Reslevel10	0.0001	Reslevel18	0.0003	Popmob6	0.0004	NO2_9	0.0011
AQI_1	0.0001	NO2_3	0.0003	Popmob5	0.0005	PM25_5	0.0011
SO2_3	0.0001	O3_6	0.0003	CO_4	0.0006	PM25_3	0.0011
Popmob6	0.0002	SO2_5	0.0003	CO_5	0.0006	AQI_4	0.0012
Reslevel0	0.0002	NO2_1	0.0003	SO2_6	0.0007	Reslevel1	0.0012
CO_7	0.0002	CO_9	0.0004	AQI_2	0.0007	PM25_4	0.0012
CO_6	0.0002	PM25_1	0.0004	SO2_7	0.0007	AQI_6	0.0012
NO2_9	0.0002	MaxT	0.0004	AQI_5	0.0008	PM10_2	0.0013

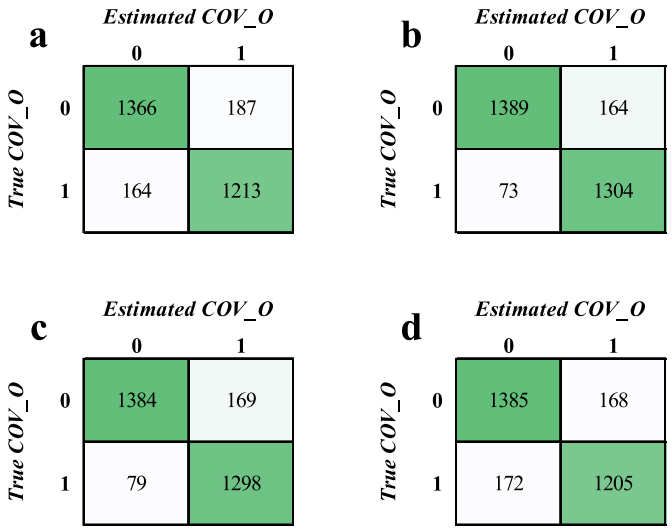


Fig. 3. Confusion matrixes of COV_O classification results by Ridge classifier (a), Gradient Boosting Decision Tree (b), Random Forest (c), and 3-layer Artificial Neural network.

Table 5
Assessment indexes of COV_O classification results by four models.

	Ridge	GBDT	RF	3-layer ANN
Accuracy	88.02%	91.91%	91.54%	88.40%
Precision	89.28%	95.01%	94.60%	88.95%
Recall	87.96%	89.44%	89.12%	89.18%

Table 6
38 COVID-19 occurrence factors that have high coefficients.

Factor	Coefficient	Factor	Coefficient
AQI_7	0.5675	PM25_1	0.8735
CO_2	0.1655	PM25_2	0.1621
CO_5	−0.3688	PM25_3	0.7964
DEM	−0.1584	PM25_5	0.9568
DisWH	−0.4398	PM25_7	−0.7619
GDP	1.1081	PD	0.3344
NO2_1	−0.8857	Popmob4	0.6645
NO2_4	−0.1488	Popmobsum	0.5297
NO2_5	−0.1183	Reslevel1	0.7163
NO2_6	−0.2154	Reslevel2	0.1711
NO2_8	0.1998	Reslevel15	−0.1343
O3_1	−0.1249	Reslevel19	−0.2020
O3_6	−0.1668	Reslevel20	−0.4461
O3_7	−0.1183	MinRh	0.1623
O3_8	−0.1001	SO2_5	0.5294
PM10_4	0.3215	SO2_7	−0.1303
PM10_6	0.2028	MaxT	0.3439
PM10_7	−1.0103	MinT	−0.1128
PM10_9	−0.2273	TIME	−1.1932

outperformed the others with its best estimation results on the test set. The performance of ANN on COVID-19 intensity estimation was brilliant even though it was not as feasible as GBDT. This could be explained by two reasons: (1) Given that all records were at the prefectural level, data of the same prefecture more or less had a homogeneity problem that most variables were similar values, and thus the neural network failed to train on enough effective data. (2) The distribution of COVID-19 intensity records was still a little skewed even after the transformation of the natural logarithm. Although these three models were slightly overfitted, such results were still good enough considering the extremely unbalanced distribution of COVID-19 intensity. However, the poor fitting performance of the EN model indicated that this prediction on the number of patients is very complex rather than a single linear regression model can express.

While most COVID-19 spreading prediction models using machine learning focused on the number of patients-only statistics instead of taking multifactor as the parameters (Ardabili et al., 2020; Pinter et al., 2020; Rustam et al., 2020), the models proposed in our study counted a series of variables to ensure convincingness. Besides, the brilliant performances of these models verified that a reliable forecast on the regional early-stage COVID-19 transmission could be accessible and it is worth further researches. Moreover, the early-stage infection of more contagious diseases rather than COVID-19 is also believed to be achievable based on our study.

Potential improvement on the analysis of early-stage COVID-19 transmission is worth more endeavour from different perspectives: (1) A wider range of variables such as vaccination may have indirect impacts which have not been fully understood, and they might contribute to a more comprehensive model, especially for complex COVID-19 intensity prediction; (2) The quantification of the control measures in other countries is yet to be studied for better use of our optimal model; (3) More specific correlation and the synergistic effect of different factors are worth further exploration.

5. Conclusion

Various factors synergistically influence the early-stage COVID-19 transmission in China: (1) COVID-19 was more likely to occur in prosperous cities closer to the epicentre and located on higher altitudes; (2) The extreme weather and higher minimum relative humidity increased the vulnerability to COVID-19 occurrence; (3) More air pollutants increased the risk of COVID-19 occurrence except for NO_2 and O_3, and most of the pollutants did not show the impact until 6–7 days after; (4) The control effects of NPIs on COVID-19 occurrence became significant after two weeks. The accuracy of our optimal model used to predict the COVID-19 occurrence reached 91.91%. Besides, our optimal regression model presented the best R^2 of 0.778 on the prediction. However, the specific connections between the number of COVID-19 patients and each factor were complex and work further exploration. Overall, this is the first study that analysed multifactor early-stage COVID-19 transmission in China and precisely estimated the pandemic trend by machine learning. Our study brings guiding significance to the forecast of the early-stage regional COVID-19 epidemic and enables quicker response to prevent regional outbreaks from becoming a nationwide pandemic.

Funding

This work was supported by Multidisciplinary Cross-cultivation project of Innovation Academy of Precision Measurement Science and Technology, CAS [grant number S21S3202], the Hubei Provincial Natural Science Foundation of China [grant number 2020CFA048], and the National Natural Science Foundation of China [grant numbers 62071457, 81803297].

Authors' contributions

Juan Qiu and Rendong Li conceptualized the study, Yifei Han directed the study's implementation, drafted and finished the manuscript, Jinliang Huang review the drafts and supervised the study, Qihui Shao, Dongfeng Han and Xiyue Luo collected resources and helped to interpret the findings.

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.

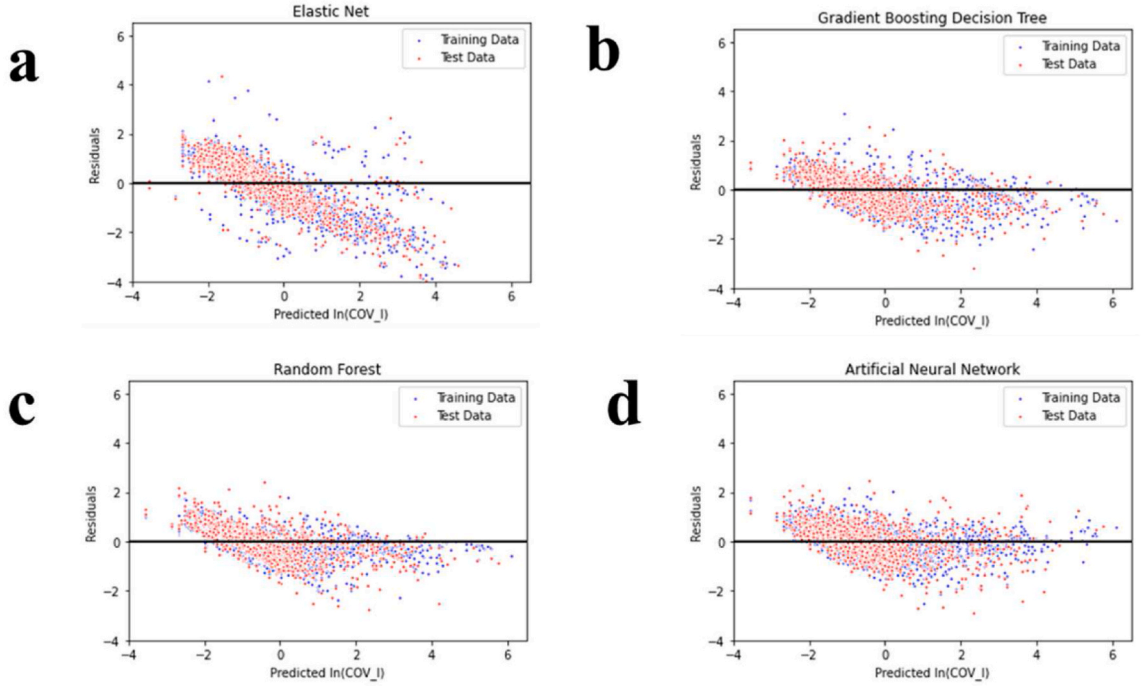


Fig. 4. Estimated values and residuals of COV_I by Elastic Net (a), Gradient Boosting Decision Tree (b), Random Forest (c), and 2-layer Artificial Neural Network (d).

Table 7
Assessment metrics of COV_I regression model on training and test set.

	EN		GBDT		RF		2-layer ANN	
	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
MSE	1.182	1.126	0.332	0.415	0.295	0.431	0.307	0.454
R^2	0.340	0.397	0.815	0.778	0.835	0.770	0.829	0.757

Appendix A

Data type, variables and data source.

Data type	Variable	Abbreviation	Units	Scale	Data source
COVID-19 data	Case incidence ratios (cases per 10,000,000 persons)	CIR	0.00001%	Daily	National Health Commission of the People's Republic of China (http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml) and provincial or municipal health commissions
Meteorological data	Minimum temperature	MinT	0.1 °C	Daily	http://data.cma.cn
	Maximum temperature	MaxT	0.1 °C	Daily	
	Mean temperature	MeanT	0.1 °C	Daily	
	Relative humidity	Rh	%	Daily	
	Minimum relative humidity	MinRh	%	Daily	
Atmospheric environmental quality data	Carbon monoxide, CO	CO, CO_1, …, CO_9	mg/m^3	Daily	https://datacenter.mee.gov.cn/
	Nitrogen dioxide, NO2	NO2, NO2_1, …, NO2_9	μg/m^3	Daily	
	Ozone, O3	O3, O3_1, …, O3_9	μg/m^3	Daily	
	Fine particles, PM2.5	PM25, PM25_1, …, PM25_9	μg/m^3	Daily	
	Inhalable coarse particles, PM10	PM10, PM10_1, …, PM10_9	μg/m^3	Daily	
	Sulfur dioxide, SO2	SO2, SO2_1, …, SO2_9	μg/m^3	Daily	
	Air Quality Index	AQI, AQI_1, …, AQI_9		Daily	
Geographical data	Mean DEM	MeanDEM	m	Daily	http://www.gscloud.cn
Socio-economic data	Household population	Pop	10,000 Person	2019, Fixed value	http://tjj.shandong.gov.cn/tjnj/nj2020/zk/indexch.htm, etc.

(*continued on next page*)

(*continued*)

Data type	Variable	Abbreviation	Units	Scale	Data source
	Population density	PD	Person/km^2	2019, Fixed value	http://www.mohurd.gov.cn/xytj/index.html
	GDP per capita	GDP	100 million RMB	Yearly	http://tjj.shandong.gov.cn/tjnj/nj2020/zk/indexch.htm, etc.
	Destination migration scale flow from Wuhan = destination proportion in population flow from Wuhan * migration scale	Popmob, Popmob1, …, Popmob9, Popmobsum		Daily	http://qianxi.baidu.com/
	The distance of each city from Wuhan	DisWH	km	Fixed value	Distance measurement based on GIS
	National emergency response	Reslevel, Reslevel1, …, Reslevel20		Daily	National Health Commission of the People's Republic of China (http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml) and provincial or municipal health commissions
Temporal data	Days from Jan 10, 2020	Time		Daily	Corresponding to each record

References

Ahmed, J., Jaman, M.H., Saha, G., Ghosh, P., 2021. Effect of environmental and socio-economic factors on the spreading of COVID-19 at 70 cities/provinces. Heliyon 7, e06979. https://doi.org/10.1016/j.heliyon.2021.e06979.

Andersen, L.M., Harden, S.R., Sugg, M.M., Runkle, J.D., Lundquist, T.E., 2021. Analyzing the spatial determinants of local Covid-19 transmission in the United States. Sci. Total Environ. 754, 142396. https://doi.org/10.1016/j.scitotenv.2020.142396.

Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T., Atkinson, P.M., 2020. Covid-19 outbreak prediction with machine learning. Algorithms 13, 249. https://doi.org/10.2139/ssrn.3580188.

Babuna, P., Han, C., Li, M., Gyilbag, A., Dehui, B., Awudi, D.A., Supe Tulcan, R.X., Yang, S., Yang, X., 2021. The effect of human settlement temperature and humidity on the growth rules of infected and recovered cases of COVID-19. Environ. Res. 197, 111106. https://doi.org/10.1016/j.envres.2021.111106.

Baidu, 2020. Overall Migration Trend in China [Online]. Available:. Accessed Jun 10 2021. http://qianxi.baidu.com/.

Barnett-Itzhaki, Z., Levi, A., 2021. Effects of chronic exposure to ambient air pollutants on COVID-19 morbidity and mortality - a lesson from OECD countries. Environ. Res. 195, 110723. https://doi.org/10.1016/j.envres.2021.110723.

Bashir, M.F., Ma, B.J., BilalKomal, B., Bashir, M.A., Farooq, T.H., Iqbal, N., Bashir, M., 2020. Correlation between environmental pollution indicators and COVID-19 pandemic: a brief study in Californian context. Environ. Res. 187, 109652. https://doi.org/10.1016/j.envres.2020.109652.

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

China Meteorological Data Service Center, 2021. China Meteorological Data Service Center [Online]. Available:. Accessed Jun 10 2021. http://data.cma.cn.

Chinadaily, 2020. Reasons for the surge of confirmed NCP cases. China Dail. Feb 13 http://global.chinadaily.com.cn/a/202002/13/WS5e461369a31012821727796e.html.

Coccia, M., 2020. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Sci. Total Environ. 729, 138474. https://doi.org/10.1016/j.scitotenv.2020.138474.

Coşkun, H., Yildirim, N., Gündüz, S., 2021. The spread of COVID-19 virus through population density and wind in Turkey cities. Sci. Total Environ. 751, 141663. https://doi.org/10.1016/j.scitotenv.2020.141663.

Dupre, N.C., Karimi, S., Zhang, C.H., Blair, L., Gupta, A., Alharbi, L.M.A., Alluhibi, M., Mitra, R., Mckinney, W.P., Little, B., 2021. County-level demographic, social, economic, and lifestyle correlates of COVID-19 infection and death trajectories during the first wave of the pandemic in the United States. Sci. Total Environ. 786, 147495. https://doi.org/10.1016/j.scitotenv.2021.147495.

Geospatial Data Cloud, 2021. ASTER GDEM [Online]. Available:. Accessed Jun 10 2021. http://www.gscloud.cn.

Guo, C., Bo, Y., Lin, C., Li, H.B., Zeng, Y., Zhang, Y., Hossain, M.S., Chan, J.W.M., Yeung, D.W., Kwok, K.-O., Wong, S.Y.S., Lau, A.K.H., Lao, X.Q., 2021a. Meteorological factors and COVID-19 incidence in 190 countries: an observational study. Sci. Total Environ. 757, 143783. https://doi.org/10.1016/j.scitotenv.2020.143783.

Guo, C., Chan, S.H.T., Lin, C., Zeng, Y., Bo, Y., Zhang, Y., Hossain, S., Chan, J.W.M., Yeung, D.W., Lau, A.K.H., Lao, X.Q., 2021b. Physical distancing implementation, ambient temperature and Covid-19 containment: an observational study in the United States. Sci. Total Environ. 789, 147876. https://doi.org/10.1016/j.scitotenv.2021.147876.

Johns Hopkins University, 2021. COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University (JHU) [Online]. Available:. Accessed Jun 21 2021. https://origin-coronavirus.jhu.edu/map.html.

Kulkarni, A.D., Lowe, B., 2016. Random forest algorithm for land cover classification. Int. J. Recent Innovat. Trends Cpmut. Commun. 4, 58–63.

Kuo, C.-P., Fu, J.S., 2021. Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. Sci. Total Environ. 758, 144151. https://doi.org/10.1016/j.scitotenv.2020.144151.

Li, M., Zhang, Z., Cao, W., Liu, Y., Du, B., Chen, C., Liu, Q., Uddin, M.N., Jiang, S., Chen, C., Zhang, Y., Wang, X., 2021. Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. Sci. Total Environ. 764, 142810. https://doi.org/10.1016/j.scitotenv.2020.142810.

Liu, J., Zhou, J., Yao, J., Zhang, X., Li, L., Xu, X., He, X., Wang, B., Fu, S., Niu, T., Yan, J., Shi, Y., Ren, X., Niu, J., Zhu, W., Li, S., Luo, B., Zhang, K., 2020. Impact of meteorological factors on the COVID-19 transmission: a multi-city study in China. Sci. Total Environ. 726, 138513. https://doi.org/10.1016/j.scitotenv.2020.138513.

Lorenzo, J.S.L., Tam, W.W.S., Seow, W.J., 2021. Association between air quality, meteorological factors and COVID-19 infection case numbers. Environ. Res. 197, 111024. https://doi.org/10.1016/j.envres.2021.111024.

Ministry Of Housing and Urban-Rural Development of the People's Republic of China, 2021. Statistical Yearbook of Urban Construction [Online]. Available:. Accessed Jun 10 2021. http://www.mohurd.gov.cn/xytj/index.html.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., 2012. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., Gloaguen, R., 2020. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics 8, 890. https://doi.org/10.3390/math8060890.

Piryonesi, S.M., El-Diraby, T.E., 2020. Data analytics in asset management: cost-effective prediction of the pavement condition index. J. Infrastruct. Syst. 26 https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512, 04019036.

Qi, Y., 2012. Random Forest for Bioinformatics. *Ensemble Machine Learning*. Springer. https://doi.org/10.1007/978-1-4419-9326-7_11.

Qiu, J., Li, R., Han, D., Shao, Q., Han, Y., Luo, X., Wu, Y., 2021. A multiplicity of environmental, economic and social factor analyses to understand COVID-19 diffusion. One Health 13, 100335. https://doi.org/10.1016/j.onehlt.2021.100335.

Ren, X., Li, Y., Yang, X., Li, Z., Cui, J., Zhu, A., Zhao, H., Yu, J., Nie, T., Ren, M., Dong, S., Cheng, Y., Chen, Q., Chang, Z., Sun, J., Wang, L., Feng, L., Gao, G.F., Feng, Z., Li, Z., 2021. Evidence for pre-symptomatic transmission of coronavirus disease 2019 (COVID-19) in China. Influenza and Other Respiratory Viruses 15, 19–26. https://doi.org/10.1111/irv.12787.

Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.-W., Aslam, W., Choi, G.S., 2020. COVID-19 future forecasting using supervised machine learning models. IEEE access 8, 101489–101499. https://doi.org/10.1109/ACCESS.2020.2997311.

Scikit-Learn, 2021. 1.1. Linear Models [Online]. Available:. Accessed Jun 25 2021. https://scikit-learn.org/stable/modules/linear_model.html.

Sharifi, A., Khavarian-Garmsir, A.R., 2020. The COVID-19 pandemic: impacts on cities and major lessons for urban planning, design, and management. Sci. Total Environ. 749, 142391. https://doi.org/10.1016/j.scitotenv.2020.142391.

Sun, F., Matthews, S.A., Yang, T.-C., Hu, M.-H., 2020a. A spatial analysis of the COVID-19 period prevalence in U.S. counties through June 28, 2020: where geography matters? Ann. Epidemiol. 52, 54–59. https://doi.org/10.1016/j.annepidem.2020.07.014 e1.

Sun, Z., Zhang, H., Yang, Y., Wan, H., Wang, Y., 2020b. Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China. Sci. Total Environ. 746, 141347. https://doi.org/10.1016/j.scitotenv.2020.141347.

The Central People's Government of the People's Republic of China, 2020. National Emergency Plan for Public Health Emergencies [Online]. Available:. Accessed Jun 10 2021. http://www.gov.cn/yjgl/2006-02/26/content_211654.htm.

The Data Center of The Ministry of Ecology And Environment of The People's Republic of China, 2021. Air Quality [Online]. Available:. Accessed Jun 10 2021. https://datacenter.mee.gov.cn/.

The National Health Commission of The People's Republic of China, 2021. Outbreak Notification [Online]. Available:. Accessed Jun 10 2021. http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.

The State Council Information Office of The People's Republic of China, 2020. Fighting COVID-19: China in Action [Online]. Available:. Accessed 15 June 2021. http://www.china.org.cn/chinese/2020-08/06/content_76173252.htm.

Vardavas, R., De Lima, P.N., Baker, L., 2021. Modeling COVID-19 nonpharmaceutical interventions: exploring periodic NPI strategies. medRxiv. https://doi.org/10.1101/2021.02.28.21252642.

Wang, Q., Dong, W., Yang, K., Ren, Z., Huang, D., Zhang, P., Wang, J., 2021. Temporal and spatial analysis of COVID-19 transmission in China and its influencing factors. Int. J. Infect. Dis. 105, 675–685. https://doi.org/10.1016/j.ijid.2021.03.014.

WHO, 2021. WHO Coronavirus (COVID-19) Dashboard [Online]. Available:. World Health Organization. Accessed 15 June 2021. https://covid19.who.int/.

Xiao, S., Qi, H., Ward, M.P., Wang, W., Zhang, J., Chen, Y., Bergquist, R., Tu, W., Shi, R., Hong, J., Su, Q., Zhao, Z., Ba, J., Qin, Y., Zhang, Z., 2021. Meteorological conditions are heterogeneous factors for COVID-19 risk in China. Environ. Res. 198, 111182. https://doi.org/10.1016/j.envres.2021.111182.

Yu, A., Wang, Z., Ren, W., Wu, Z., Hu, Z., Li, L., Ruan, Y., Hu, R., Shi, F., 2020. Epidemic analysis of COVID-19 in China after Wuhan was restricted. Res. Square. https://doi.org/10.21203/rs.2.24289/v1.

Zhao, S., 2020. To avoid the noncausal association between environmental factor and COVID-19 when using aggregated data: simulation-based counterexamples for demonstration. Sci. Total Environ. 748, 141590. https://doi.org/10.1016/j.scitotenv.2020.141590.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.