Can a Meteorological Variable Be Considered As a Predictor of COVID-19 Cases in Urban Agglomerations of Indian Cities?

Asha B. Chelani
NEERI: National Environmental Engineering Research Institute CSIR
Sneha Gautam (snehagautam@karunya.edu)
Karunya Institute of technology and Science: Karunya Institute of Technology and Sciences
https://orcid.org/0000-0002-2978-844X

Short Report

Keywords: COVID – 19, Lockdown, Random forest model, Hybrid model, Meteorological parameters, India

Posted Date: October 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-921666/v1

Can a meteorological variable be considered as a predictor of COVID-19 cases in urban agglomerations of Indian cities?

Asha B. Chelani[1], Sneha Gautam[2†]

[1]Air Pollution Control Division
National Environmental Engineering Research Institute (CSIR-NEERI), Nehru Marg, Nagpur, India-440020

[2] Department of Civil Engineering, Karunya Institute of Technology and Sciences, Coimbatore – 641114, Tamil Nadu, India

[†] corresponding author: gautamsneha@gmail.com / snehagautam@karunya.edu

Abstract

Coronavirus has been identified as one of the deadliest diseases and WHO has declared it as pandemic and global health crisis. It has become a massive challenge for humanity. India is also being facing its fierceness as it is highly infectious and mutating at a rapid rate. Many interventions have been applied in India since the first reported cases i.e. on January 30, 2020. Several studies have been conducted to assess the impact of climatic and weather conditions on its spread in the last year span. As it is a well-established fact that temperature and humidity could trigger the onset of diseases such as influenza and respiratory disorders, the association of several meteorological variables has been studied in the past with the COVID-19 related number of cases. The conclusions in those studies were based on the data obtained at the early stage and it was too early to draw any inference. This study attempted to assess the influence of temperature, humidity, wind speed, dew point, previous day's number of deaths, and government intervention's effect on the number of COVID-19 confirmed cases in 18 districts of India. It is also attempted to identify the important predictors of number of confirmed COVID-19 cases in those districts. The random forest model and the hybrid model obtained by modelling the random forest model's residuals are used to predict the response variable. It is observed that meteorological variables are useful only to some extent that too when used with

the data on number of the previous day's deaths and lockdown information in predicting the COVID-19 cases. Partial lockdown is more important than the complete or no lockdown in predicting the number of confirmed COVID-19 cases. The information is useful to policy makers in balancing the restriction activities and economic losses to individuals and the government.

Keywords: COVID – 19; Lockdown; Random forest model; Hybrid model; Meteorological parameters; India

1. Introduction

Novel Coronavirus (COVID-19), has been spread in almost all the countries for the last one and a half years. The nations are facing the wrath of the disease with outbreaks, during which, the seasonal variations have been observed. In India, over 32,474,773 cases and 435,050 deaths have been reported (Worldometer, 2021). The temporal variations in the number of COVID-19 cases observed during January 2020 till date across the regions indicate the seasonal variations (Gautam et al., 2020; Chelani and Gautam 2021). It is a well-established fact that temperature and humidity could trigger the onset of diseases such as influenza and respiratory disorders (Shaman and Kohn, 2009; Golakota et al. 2021). The association of several meteorological variables with the COVID-19 related number of cases has therefore been assessed in various studies (Zhu et al., 2020; Yao et al., 2020; Cole et al., 2020; Gautam 2020a&b; Gautam et al., 2021; Ambade et al., 2021). In winter, a high number of flu or influenza cases are usually witnessed due to low temperature, whereas during summer, fewer cases are generally observed (Damette et al., 2021; Chen et al., 2021). It has been observed that the cities with average temperature less than 10^0C and lower humidity have more chances of spread of the virus than with higher temperature (Sajadi et al., 2020; Araujo and Naimi, 2020). It is observed that the rise of 1^0C in temperature may cause approximately 4.861% increase in the daily confirmed COVID-19 cases in China (Xie & Zhu, 2020). Few studies, however have established the insignificant or negative effect of meteorological factors on COVID-19 confirmed cases (Yao et al., 2020; Liu et al., 2020; Shi et al., 2020; Méndez-Arriaga, 2020; Wu et al., 2020). The studies on the association between the meteorological parameters and COVID-19 have provided mixed results and do not provide an empirical evidence or confirmed statistical significance of the association. In India, the warmer temperature is usually observed in most regions for many days, which was the reason for the hope that the disease would not spread to the tune of other areas having cooler climates (Chen et al., 2021). However, the havoc

created by the virus in India is not unknown (BBC, 2021). Even in the warmer period, large number of cases have been reported (https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus).

The studies above have used a small sample size due to the non-availability of the COVID-19 confirmed cases. The outcome based on the small datasets may lead to erroneous conclusions. Over a period of time, enormous dataset has been obtained. With the availability of large sample size, a rigorous study can be conducted on the relationship between meteorological factors and number of COVID-19 confirmed cases to assess the role of the former in governing the disease spread. In addition, the effect of policy interventions such as complete or partial lockdowns implemented to prevent the spread of the virus can also be assessed. In India, the complete lockdown was initiated from March 24, 2020, which was later relaxed from April 14, 2020 in phase-wise manner. During the complete lockdown, all the activities except the essential services were completely halted. Partial lockdowns were imposed after April 14, 2020 with relaxation in phases (Wikipedia, 2021). It is interesting to know the influence of complete and partial lockdowns in the cities in the containment of spread of the disease. The number of deaths incurred due to COVID-19 on previous day may influence the spreading of the virus because people gather for condolence meets, although in fewer numbers due to government restrictions. The number of deaths incurred on the previous day was included in the model as it contains the effect of earlier days due to autocorrelation. In this study, with the data on meteorological factors such as daily mean temperature, wind speed, relative humidity, dew point temperature, number of previous day's deaths and interventions i.e. lockdown variables in 18 districts across India, a random forest model was applied with number of confirmed COVID-19 cases as a response variable and others as predictors during March 24, 2020 to June 15, 2021. The study shall be useful in decision making and speculating the onset of high number of cases.

2. Method

2.1 Data collection and model application

Fig 1 shows the districts which are included in the study. The total number of cases reported for each district divided by the corresponding population since the onset of the virus till June 15, 2020 are plotted in Fig. 2. The COVID-19 data were obtained from COVID-19-India (2021). Although the lockdown was imposed since March, 2020, the number of confirmed cases across India is not available. Based on the availability of the data, the number of confirmed cases from 18 districts in different states of India during April 26, 2020 to June 15, 2021 are considered.

The time series of number of confirmed cases for each district is normalized by dividing with the population of the corresponding district. The newly formed time series is considered as the response variable, which is modelled as a function of meteorological variables such as temperature, wind speed, relative humidity and dew point, lockdown variable which is a categorical dummy variable, and number of deaths reported on previous day. The meteorological data are obtained from Wunderground (2021). The policy interventions such as complete and partial lockdown are incorporated in the model as dummy variable with the following descriptors as;

Complete lockdown - The complete lockdown was initiated in India starting with the 'Janta Curfew' on March 22, 2020. The complete lockdown was announced on March 24, 2020 till May 31, 2020. During this phase, the non-essential services and factories were suspended except essential services such as grocery stores and vegetable sellers, which were allowed to remain open for a particular duration of time. The traffic movement was restricted due to strict compliance by the police and local governments.

Partial lockdown - The essential and non-essential activities were allowed for a few hours till 2PM or 4PM or 8PM.

Unlock - On June 1, 2020, central government announced the unlock 1.0 with ease on restrictions followed by series of unlock phases like unlock 2.0, unlock 3.0, unlock 4.0 and unlock 5.0 which were extended till November 30, 2020. All the phases of unlock are incorporated in the model as one dummy variable.

Complete upliftment of the lockdown or no lockdown - From December 1, 2020, the restrictions were uplifted till the beginning of the second wave of COVID 19, which initiated on April 5, 2021.

During April 5, 2021 to April 30, 2021, partial lockdown was imposed which became stricter from May 1, 2021 to May 31, 2021. The lockdown phases were followed in this study as imposed by Central government only. The model incorporates the lockdown variables as a categorical dummy variable with complete lockdown as Lk1, unlock as Lk2, partial lockdown as Lk3 and no lockdown as Lk4. The four dummy variables were therefore introduced in the model code with binary values as described above. The other variables such as mean daily temperature (^{0}C), humidity (%), wind speed (m s^{-1}), dew point temperature (^{0}C) for the respective district are used in the model as Temp, RH, WS, Dew. The number of deaths on previous few days may have influence on the spread of the virus. The number of deaths occurred on previous day only is included in the model as it is assumed that the previous day's number of deaths effect is included in the current day's number of deaths. It is denoted as Death_1 in the model.

The reported cases may be dependent on the previous number of cases due to lagged behaviour. The other lagged variables of confirmed cases are however, not included to avoid the tautological effect on the model. The inclusion of the previous day's number of confirmed cases also overpowers the model and shall have more importance than other features. Therefore, the model is developed only with meteorological variables, number of previous day's deaths and intervention exogenous variables. The meteorological variables may also be

auto-correlated, however the lags of meteorological variables are not included in the model as it is assumed that the values observed on a particular day are inclusive of the effect of previous day's values, so any effect on the response variable of those previous day's values shall be taken care off by the present day's observations.

2.2 The random forest model

Random forest (RF) is an ensemble method involving the random forest of decision trees (Breiman, 2001). It is a supervised learning algorithm that constructs decision trees based on the training data set. The combination of decision trees minimizes the out of bag error based on the training data sets. The studies have shown its applicability even in the presence of noise in the data (Kontschieder et al., 2011). Many nonlinear and high-dimensional complex classification and regression problems have been solved by applying RF (Yu et al., 2016). In case of regression problem, the predictions are obtained by the random selection of number of predictor variables in the decision tree. The best solution is determined based on the number of nodes and variables in the nodes in the fully grown tree. Every training set is fed to each decision trees (Breiman, 2001; Breiman, 2002). The number of variables or features to construct the model are selected with a specified value. For many regression problems, $1/3^{rd}$ of the total number of variables is often used (Liaw and Wiener, 2002). The usual practice to select the number of trees is randomly selecting the initial value from 10 to 1000 or higher and choosing one based on model performance criteria. Higher number of trees however slows the learning process. A discussion on other selection criteria based on cross-validation and tuning are given in Stone (1974). The decision tree is constructed for the training cases with the specified nodes of the tree. The training set is selected and trained whereas the remaining cases are used to estimate the error. So for each case, out of bag error estimates are provided. Random forest model has advantage over other supervised learning models for classification and regression problems as it avoids over-fitting samples and provides the solution based on

averaging. Each variable has its importance in the overall model performance, which can be shown based on the Gene index for classification and mean square error (MSE) for regression problems. RF is applied using R4.0.0 (R Development Core Team, 2010). For variable importance, %IncMSE is computed in the R package, which is the increase in MSE of predictions of out of bag samples as a result of permutations of the variables.

2.3 AR1 model

Sometimes, the fitted model shows heteroscedasticity in the residuals and accepting the model is not advisable as some patterns in the datasets may not have been appropriately captured. Hence further modelling of the residuals may explain the existing temporal relationships and may provide more reliable model performance than just relying on the single model fitted to the datasets. The autoregressive model of order 1 (AR1) model is therefore fitted to the residuals of the RF model. The details of the model are given below.

Let r_t represents the residuals obtained by fitting RF model to number of confirmed cases for the training set.

$$r_t = y_t - \widehat{RF}_t \qquad \text{- - - - (1)}$$

Where $\widehat{RF}_t$ is the forecast value of RF model for time *t*. AR1 model can be fitted as in equation (2);

$$r_t = \varphi_0 + \varphi_1 r_{t-1} + \varepsilon_t \qquad \text{- - - - (2)}$$

Like RF model, the coefficients are estimated for residuals of the training set and the predictions are obtained for residuals of the testing set.

2.4 Hybrid model

The predictions obtained by RF model and AR1 model are combined. The hybrid methodology is based on the combination of linear autocorrelation model and random forest model (Chelani and Devotta, 2006), which can be given as;

$y_t = r_t + RF_t$ - - - - (3)

Where r_t denotes the AR1 model fitted to residuals of RF model as given in Equation (2) and and RF_t denotes the RF model of number of confirmed cases. First the RF model is fitted to the data and the residuals are obtained, which were then modelled as AR1 process.

Let the forecast from the AR1 model be denoted as $\hat{r}_t$. The new forecasts can therefore be obtained as,

$\hat{y}_t = \widehat{RF}_t + \hat{r}_t$ - - - - (4)

To evaluate the performance of the models, the error statistics such as correlation between observed and predicted cases, root mean square error (RMSE), normalized root mean square error (NRMSE) are utilized. These test statistics can be obtained as,

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)} \quad \text{- - - - (5)}$$

$$NRMSE = \frac{RMSE}{{}^{1}/_{n}\sum_{t=1}^{n} y_t} \quad \text{- - - - (6)}$$

Where y_t is the observed and $\hat{y}_t$ is the predicted data and *n* is the total number of data points.

2.5 Results and Discussion

The descriptive statistics of number of confirmed cases and meteorological variables is given in Table 1. The mean daily temperature, wind speed, relative humidity and dew point temperature ranges from 23.3±6.7^{0}C to 35.8±12.5^{0}C, 0.4±0.2 ms^{-1} to 3.1±1 ms^{-1}, 44.7±24.3% to 72.9±13.5%, 7.08±3.82^{0}C to 21.69±5.02^{0}C. The number of confirmed cases range between 0 to 2036 per million, whereas the number of deaths range between 0 to 184 per million in all the districts during the study period. Further the correlation analysis of meteorological variables, number of deaths and number of confirmed cases is given in Table 2.

The monthly variations in confirmed cases divided by the corresponding population of the district are given in Fig.2, which shows that April and May have witnessed outbreak due to second wave in all the districts. In Nagpur, Patiala, Thane, Mumbai, Ludhiana and Pune, high number of confirmed cases have also been observed in March. Few spikes are also observed in other months. The correlation between number of deaths and confirmed cases is statistically significant in all the districts except in Mumbai and Patna. The correlation of number of confirmed cases with temperature is significant at p=0.05 in all the districts except in Chandrapur, Thane and Ujjain. The correlation of relative humidity and number of confirmed cases is mostly negative and significant at p=0.05 except at Kolkata and Nagpur. In Mumbai and Thane, the correlation with relative humidity is positive. The correlation of number of confirmed cases with wind speed is mostly negative and insignificant. Wind speed has been characterized as one of the factors in defining the ventilation coefficient of an area (Goyal and ChalapatiRao, 2007). It has been observed in the past that high wind speed and good ventilation are associated with less number of COVID-19 cases. The relationship in the study is however negative and not significant. The relationship of dew point and confirmed cases is sporadic with negative correlation in Chandrapur, Dewas, Kanpur and Ujjain and positive correlation in Chennai, Kolkata, Mumbai, Nagpur, Pune and Varanasi. When one looks at the scatter plot of confirmed cases and meteorological parameters, it can be seen that there is an inverse parabolic relation of confirmed cases with temperature, relative humidity and dew point. The number of cases increase with these factors and beyond a certain point, number of confirmed cases starts decreasing. The number of confirmed cases start decreasing with temperature at 31 ± 0.5^0C, relative humidity at 51.7± 0.6% and dew point at 21 ± 0.4^0C, respectively. The relationship of the number of confirmed cases with wind speed was exponentially decreasing with the initial increase in the number of confirmed cases.

Random forest model was developed using library 'randomForest' in R. The data was divided into training and testing set with a ratio of 85:15 for each district. The number of trees and the random selection of number of variables were adjusted according to the mean square error as a cost function. The bootstrapping with a sample size of 500 was done to arrive at an optimum model. The 'Çaret' library in R was used to carryout bootstrapping of the training samples. The optimum number of trees and number of variables with minimum mean square error is observed to be 80 and 7.

RF model helps in ranking the relative importance of the predictors in modelling the response variable. %IncMSE is used to rank the important features for the response variable. The model is run each time when a split on one predictor is conducted. A large change in the MSE is usually observed with the important predictor. The variable importance ranking shown in Table 3 suggests that the major governing factors of confirmed cases are the number of deaths occurred on previous day. Temperature is the second most important factor influencing the confirmed cases closely followed by wind speed. Dew point and relative humidity on the other hand have relatively less influence on confirmed cases predictability. The influence of lockdown variables is quite low as compared to meteorological variables and the number of previous day's deaths. Partial lockdown effective only for few hours is seen to be highly effective in confining the cases as compared to the complete lockdown and no lockdown. The unlock period is also shown to be the second influencing factor among the lockdown measures governing the number of confirmed cases. This finding is very useful to policy makers and economic growth point of view. Policymakers can opt for the partial lockdowns instead of complete or no lockdown et al to sustain the economic activities *viz.* a *viz.* confine the spread of the virus. Moreover, it can be seen from the importance matrix that meteorological variables alone cannot be used as predictors to predict or estimate the number of confirmed cases for any district. Instead the number of previous day's deaths and lockdown details are also required.

The results of RF models are then further improved by fitting AR1 model on the residuals of the model. AR1 model is observed to be with coefficient $\varphi_0 = -0.000127$ (p>0.05) and φ_1= 0.46 (0.44±0.48) with p<0.05. The estimations obtained by AR1 model i.e. $\hat{r}_t$ are added to the estimations of RF model ($\widehat{RF}_t$) as in Equation (4). The performance of the final results of hybrid model are given Table 4. The prediction results for training and testing set in terms of box plot is shown in Fig. 3a and 3b, respectively. It can be seen that the hybrid model performs better than RF model as NRMSE is lower for the training and testing set. The model can obtain the number of confirmed cases based on the exogenous meteorological variables for a particular day along with the previous day's deaths.

2.6 Conclusion

Considering the fierceness of COVID-19, WHO has declared it as a pandemic and global health crisis. In India, many interventions have been applied and in the last year span, several studies have been conducted to assess the impact of climatic and weather conditions on its spread. The earlier studies on the linkage between meteorological variables and COVID-19 related number of cases were based on the data obtained at the early stage and it was too early to draw any inference. This study attempts to study the influence of temperature, humidity, wind speed, dew point, previous day's number of deaths, and government intervention's effect on the number of COVID-19 confirmed cases in 18 districts of India with the data observed during April 26, 2020 to June 15, 2021. It is also attempted to identify the important predictors of the number of cases in those districts using the random forest model and the hybrid model obtained by modelling the residuals of the random forest model. It is observed that meteorological variables are useful in predicting the number of COVID-19 related cases only to some extent. The data on the number of previous day's deaths is more important factor in governing the number of COVID-19 cases. Comparing the degrees of restrictions in terms of complete, partial, unlock phases and no lockdown, partial lockdown is observed to be more important

than the complete lockdown and no lockdown in predicting the number of confirmed COVID-19 cases. The information is useful to policy makers as instead of restricting completely to contain the spread of the virus or allowing all the activities, government can opt for partial lockdowns to minimize the economic losses to individuals and government.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Ambade, B., Sankar, T.K., Kumar, A., Gautam, A.S., Gautam, S., 2021. COVID-19 lockdowns reduce the Black carbon and polycyclic aromatic hydrocarbons of the Asian atmosphere: source apportionment and health hazard evaluation. Environ. Develop. Sustain. DOI:org/10.1007/s10668-020-01167-1.
- Araujo, M. B., Naimi, B., 2020. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. https://doi.org/10.1101/2020.03.12.20034728.
- BBC, 2021. Covid-19: India in a 'delicate phase' of its coronavirus battle as cases surge. https://www.bbc.com/news/world-asia-india-56206004.
- Breiman, L., 2001. Random Forests, Machine Learn. 45(1), 5-32.
- Breiman, L., 2002. Manual on setting up, using and understanding Random Forests V3.1. https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.
- Chen, S., Prettner, K., Kuhn, M., Geldsetzer, P., Wang, C., Barnighausen, T., Bloom, D.E., 2021. Climate and the spread of COVID-19. Sci. Rep. 11, 9042.
- Chelani, A.B., Devotta, S., 2006. Air quality modeling using a hybrid autoregressive and nonlinear model. Atm. Env. 40, 1774-1780.
- Cole, M. A., Ozgen, C., and Strobl, E., 2020. Air pollution exposure and Covid-19 in Dutch municipalities. Environ. Resour. Econ. 76(4), 581-610. https://doi.org/10.1007/s10640-020-00491-4.
- Conticini, E., Frediani, B., and Caro, D., 2020. Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?. Environ. Pollut. 261, 114465. 10.1016/j.envpol.2020.114465.
- COVID-19-India, 2021. *Accessed via* COVID-19india.org.

- Damette, O., Mathonnat, C., Goutte, S., 2021. Meteorological factors against COVID-19 and the role of human mobility. PLoS ONE 16(6): e0252405. https://doi.org/10.1371/journal.pone.0252405.
- Gautam, A.S., Dilwaliya, N., Srivastava, A., Kumar, S., Baudh, K., Singh, D., Gautam, S., 2020. Temporary reduction in air pollution due to anthropogenic activity switch-of during COVID-19 lockdown in northern parts of India. Environ. Develop. Sustain. DOI: 10.1007/s10668-020-00994-6.
- Gautam, S., Sammuel, C., Gautam, A.S., Kumar, S., 2021. Strong link between coronavirus count and bad air: A case study of India. Environ. Develop. Sustain. doi.org/10.1007/s10668-021-01366-4.
- Gollakota, A.R.K., Gautam, S., Santosh, M., Sudan, H.A., Gandhi, R., Jebadurai, V.S., Shu, C.M., 2021. Bioaerosols: characterization, pathways, sampling strategies, and challenges to geo-environment and health. Gondwana Res. 99, 178 – 203. 10.1016/j.gr.2021.07.003.
- Goyal, S.K., ChalapatiRao, C.V., 2007. Assessment of atmospheric assimilation potential for industrial development in an urban environment: Kochi (India). Sci. Total Environ. 376 (1-3), 27-39.
- Izquierdo-Verdiguier, E., Zurita-Milla, R., 2020. An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. Int. J. Appl. Earth Obs. Geoinf. 88, 102051. https://doi.org/10.1016/j.jag.2020.102051.
- Kontschieder, P., Bulò, S.R., Bischof, H., Pelillo, M., 2011. Structured class-labels in random forests for semantic image labelling. Inter. Conf. Computer Vision, 2190-2197. https://doi.org/10.1109/ICCV.2011.6126496.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News, 2 (3), 18-22.

- 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.
- Méndez-Arriaga, F., 2020. The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1. Sci. Total Environ.735. 139560. https://doi.org/10.1016/j.scitotenv.2020.139560.
- Chelani, A., Gautam, S., 2021. Lockdown during COVID-19 pandemic: A case study from Indian cities shows insignificant effects on persistent property of urban air quality. Geoscience Frontiers. (doi.org/10.1016/j.gsf.2021.101284).
- Ogen, Y., 2020. Assessing nitrogen dioxide (NO_2) levels as a contributing factor to coronavirus (COVID-19) fatality. Sci. Total Environ. 726, 138605. 10.1016/j.scitotenv.2020.138605.
- R Development Core Team, 2010. A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing.
- Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., Amoroso, A., 2020. Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. JAMA. Network Open. 3(6): e2011834, pmid:32525550.
- Setti, L., Passarini, F., De Gennaro, G., P., Perrone, M. G., Borelli, M., Palmisani, J., Di Gilio, A., Torboli, V., Fontana, F., Clemente, L., Pallavicini, A., Ruscio, M., Piscitelli, P., and Miani, A., 2020. SARS-Cov-2RNA found on particulate matter of Bergamo in Northern Italy: first evidence. Environ. Res. 188, 109754. https://doi.org/10.1016/j.envres.2020.109754.
- Shaman, J., Kohn, M., 2009. Absolute humidity modulates influenza survival, transmission, and seasonality. Proc. Natl. Acad. Sci. USA 106(9), 3243–3248.

- Shi, P., Dong, Y., Yan, H., Zhao. H., Li, X., Liu, W., He, M., Tang, S., Xi, S., 2020. Impact of temperature on the dynamics of the COVID-19 outbreak in China. Sci. Total Environ.728, 138890. https://doi.org/10.1016/j.scitotenv.2020.138890.
- Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions J. R. Stat. Soc.: Ser. B (Methodol.), 36 (2), 111-133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.
- Wikipedia, 2021. *Accessed via* https://en.wikipedia.org/wiki/COVID-19_lockdown_in_India#cite_note-81.
- Worldometer, 2021. *Accessed via* https://www.worldometers.info/coronavirus/country/india/, accessed on 24/8/2021.
- Wunderground, 2021. *Accessed via* www.wunderground.com, accessed on 28/7/2021.
- Wu Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., et al., 2020. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Sci. of Total Environ.729, 139051.
- Xie, J., and Zhu, Y., 2020. Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci. Total Environ. 724, 138201. https://doi.org/10.1016/j.scitotenv.2020.138201.
- Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., and Wang, W., 2020. No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. Eur. Respir. 55(5), 2000517. 10.1183/13993003.00517-2020.
- Yu, R., Yang, Y., Yang, L., Han, G., Move, O.A., RAQ- A random forest approach for predicting air quality in urban sensing systems. Sensors 16, 86.
- Yuan, J., Wu, Y., Jing, W., Liu, J., Du, M., Wang, Y., and Liu, M., 2021. Association between meteorological factors and daily new cases of COVID-19 in 188 countries: A time series analysis. Sci. Total Environ. 780, 146538. 10.1016/j.scitotenv.2021.146538.

- Zhu, Y., Xie, J., Huang, F., Cao, L., 2020. Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. Sci. Total Environ. 727, 138704. https://doi.org/10.1016/j.scitotenv.2020.138704.

List of Figures

List of Tables

Table 1. Descriptive statistics of number of confirmed cases and meteorological variables

S. No.	District	Temp Mean	±	SD	RH Mean	±	SD	WS Mean	±	SD	Dew Mean	±	SD	Death_1 Mean	±	SD	Cases* Mean	±	SD
1	Chandrapur	28.4	±	2.2	49.1	±	12.4	0.4	±	0.2	18.24	±	3.56	1.7	±	4.5	95.8	±	172.2
2	Chennai	32.0	±	3.5	72.8	±	8.3	2.1	±	1.4	26.0	±	3.31	2.7	±	4.3	178.5	±	211.7
3	Delhi	26.0	±	5.9	56.4	±	14.5	0.8	±	0.3	17.3	±	5.67	3.2	±	4.7	181.5	±	284.8
4	Dewas	29.3	±	1.3	51.0	±	23.1	3.1	±	1.0	19.45	±	4.9	0.1	±	0.3	11.9	±	19.2
5	Faridabad	26.6	±	5.8	53.0	±	18.7	0.7	±	0.3	17.16	±	5.95	0.9	±	1.3	132.6	±	202.4
6	Jodhpur	31.6	±	5.3	44.7	±	24.3	0.8	±	0.4	20.57	±	8.02	0.7	±	1.5	73.2	±	122.1
7	Kanpur	35.7	±	2.5	60.4	±	14.7	1.7	±	0.8	27.8	±	3.34	1.0	±	1.5	43.5	±	84.6
8	Kolkata	26.9	±	4.4	71.1	±	14.7	0.8	±	0.7	21.14	±	5.79	0.8	±	0.7	49.3	±	65.3
9	Ludhiana	24.6	±	7.5	65.6	±	20.5	0.5	±	0.2	17.74	±	6.43	1.4	±	1.7	59.4	±	88.6
10	Mumbai	25.9	±	2.0	72.9	±	13.5	0.5	±	0.1	20.49	±	3.47	3.0	±	4.1	138.6	±	161.7
11	Nagpur	24.2	±	6.4	49.1	±	11.9	0.5	±	0.1	14.14	±	7.11	4.1	±	7.0	255.1	±	388.6
12	Patiala	23.3	±	6.7	69.7	±	21.4	0.5	±	0.2	17.29	±	6.34	1.7	±	2.3	60.9	±	77.8
13	Patna	25.6	±	5.8	70.7	±	16.7	0.5	±	0.2	19.7	±	5.74	1.0	±	9.0	60.1	±	103.7
14	Pune	35.8	±	12.5	48.6	±	14.8	0.5	±	0.2	21.52	±	8.46	4.0	±	6.3	264.9	±	313.0
15	Rohtak	25.5	±	7.9	51.7	±	9.9	1.1	±	0.6	16.37	±	7.68	1.1	±	2.9	58.5	±	86.1
16	Thane	26.8	±	4.2	71.8	±	12.5	0.5	±	0.7	7.08	±	3.82	2.2	±	3.9	123.8	±	131.3
17	Ujjain	31.4	±	1.4	51.6	±	21.1	2.8	±	1.3	21.69	±	5.02	0.2	±	0.6	22.8	±	40.1
18	Varanasi	25.3	±	6.7	57.1	±	23.5	1.1	±	0.9	18.16	±	6.18	0.6	±	0.9	55.7	±	118.7

Cases* refers the number of confirmed cases per million

Death_1: number of deaths due to COVID-19 on previous day, WS: wind speed, Temp: temperature, RH: relative humidity, Dew: dew point temperature, Lk1: no lock down, Lk2: unlock, Lk3: partial lockdown, Lk4: complete lockdown

Table 2. Correlation between meteorological variables and number of confirmed cases

S. No.	District	Death_1		Temp		RH		WS		Dew	
1	Chandrapur	0.44	*	-0.01		-0.18	*	-0.03		-0.13	*
2	Chennai	0.55	*	0.13	*	-0.22	*	0.05		0.11	*
3	Delhi	0.80	*	0.12	*	-0.34	*	-0.03		-0.05	
4	Dewas	0.17	*	0.23	*	-0.3	*	-0.02		-0.22	*
5	Faridabad	0.72	*	0.14	*	-0.29	*	0		-0.04	
6	Jodhpur	0.88	*	0.21	*	-0.23	*	-0.04		0	
7	Kanpur	0.60	*	0.11	*	-0.47	*	0.15	*	-0.34	*
8	Kolkata	0.69	*	0.27	*	-0.06		-0.03		0.17	*
9	Ludhiana	0.77	*	0.25	*	-0.4	*	0.02		0.04	
10	Mumbai	0.08		0.42	*	0.23	*	0.23	*	0.42	*
11	Nagpur	0.31	*	0.24	*	0.05		0.02		0.22	*
12	Patiala	0.79	*	0.26	*	-0.41	*	0.06		0	
13	Patna	0.06		0.29	*	-0.48	*	0.21	*	0.01	
14	Pune	0.16	*	0.15	*	-0.43	*	0.09		0.13	*
15	Rohtak	0.71	*	0.12	*	-0.24	*	0.13	*	0.03	
16	Thane	0.11	*	0.10		0.13	*	-0.13	*	0.05	
17	Ujjain	0.29	*	-0.06		-0.44	*	0.00		-0.39	*
18	Varanasi	0.67	*	0.40	*	-0.17		0.16	*	0.26	*

Death_1: number of deaths due to COVID-19 on previous day, WS: wind speed, Temp: temperature, RH: relative humidity, Dew: dew point temperature, Lk1: no lock down, Lk2: unlock, Lk3: partial lockdown, Lk4: complete lockdown

Table 3. Variable importance matrix

Variable	**%IncMSE**
Death_1	39.6
Temp	22.7
WS	20.3
Dew	16.8
RH	14.8
Lk3	11.3
Lk2	9.3
Lk1	6.9
Lk4	0

Death_1: number of deaths due to COVID-19 on previous day, WS: wind speed, Temp: temperature, RH: relative humidity, Dew: dew point temperature, Lk1: no lock down, Lk2: unlock, Lk3: partial lockdown, Lk4: complete lockdown

Table 4. Model evaluation for training and testing set

Statistics	**RF**		**Hybrid model**	
	Training	Testing	Training	Testing
R^2	0.96	0.76	0.96	0.79
RMSE	0.000037	0.000286	0.000033	0.000205
NRMSE	0.36	0.55	0.29	0.38

Figures

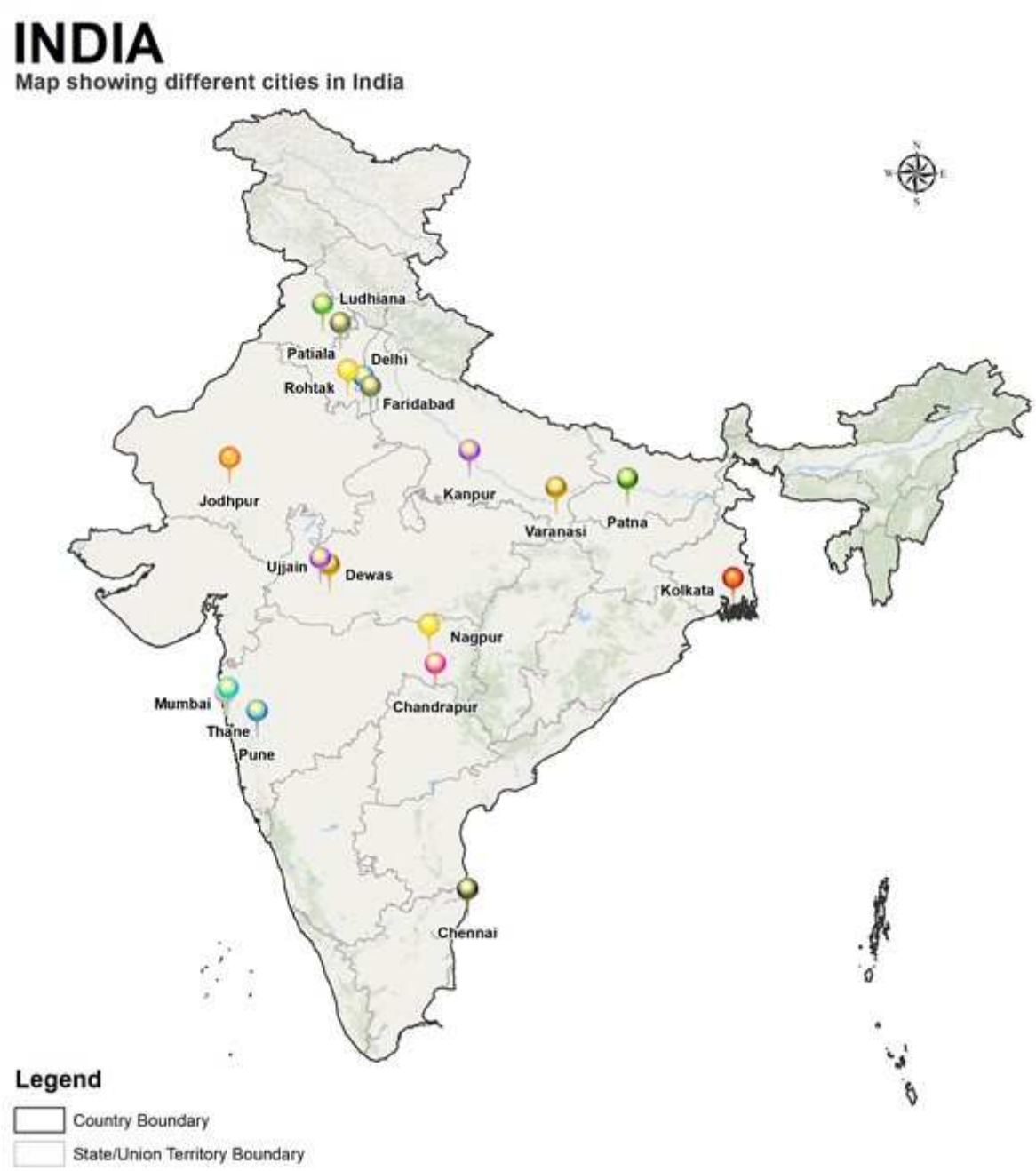

Figure 1

Location of districts in map of India.

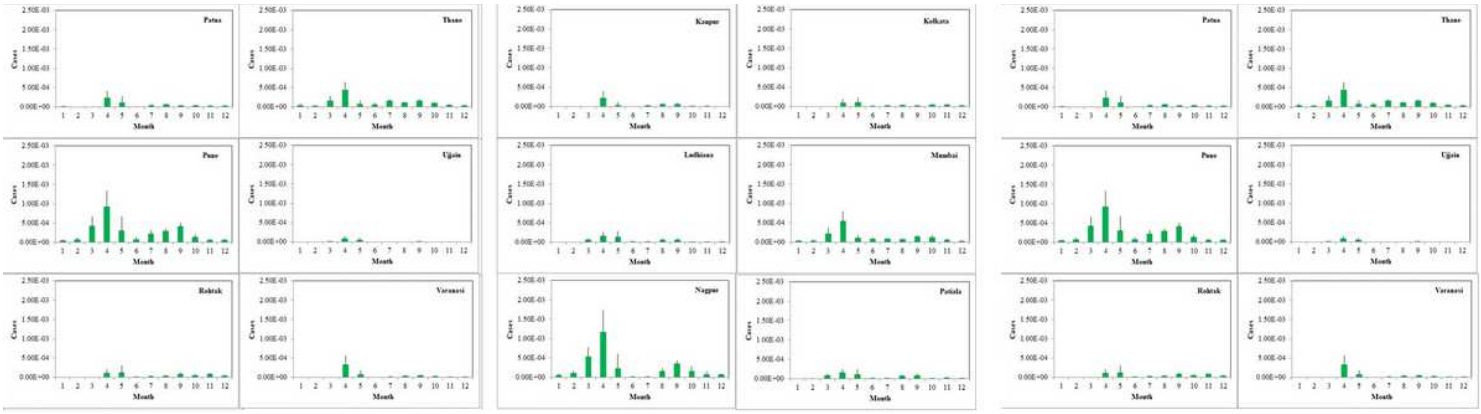

Figure 2

Monthly variations in total number of confirmed COVID-19 cases (divided by population of the district) since April 26, 2020 to June 15, 2021 in various districts of India.

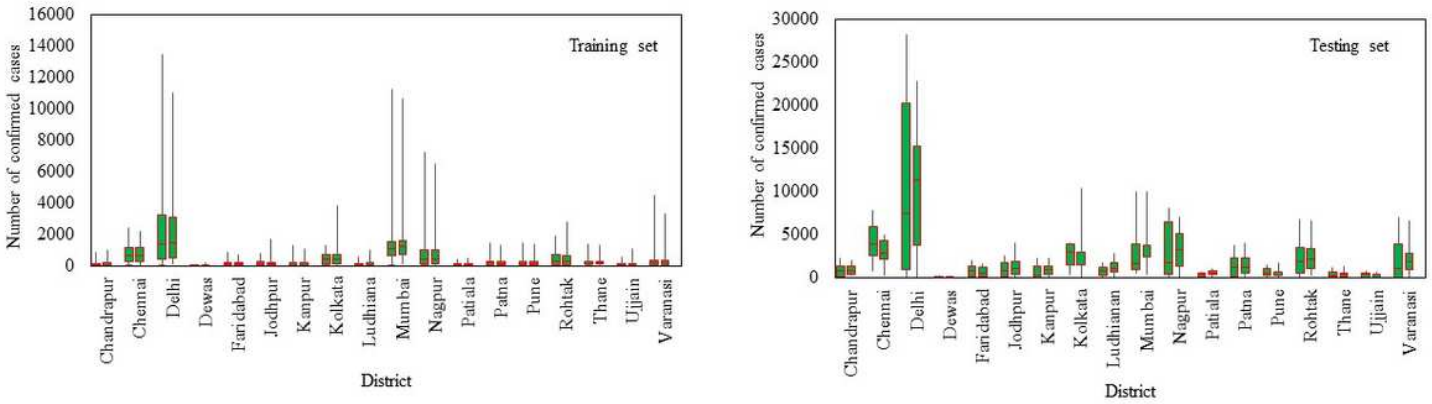

Figure 3

a. Observed and predicted number of cases using hybrid model in 18 districts for training set. b. Observed and predicted number of cases using hybrid model in 18 districts for testing set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CGraphicalAb.jpg