Does Climate Help Modeling COVID-19 Risk and to What Extent?

Giovanni Scabbia
Hamad Bin Khalifa University
Antonio Sanfilippo (asanfilippo@hbku.edu.qa)
Hamad Bin Khalifa University https://orcid.org/0000-0001-7097-4562
Annamaria Mazzoni
Hamad Bin Khalifa University
Dunia Bachour
Hamad Bin Khalifa University
Daniel Perez-Astudillo
Hamad Bin Khalifa University
Veronica Bermudez Benito
Hamad Bin Khalifa University
Etienne Wey
Transvalor S.A.
Mathilde Marchand-Lasserre
Transvalor S.A.
Laurent Saboret
Transvalor S.A.

Article

Keywords: COVID-19 transmission, spread, meteorological variables

Posted Date: November 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108398/v1

Abstract

A growing number of studies has suggested potential impacts of meteorological variables on the spread of the COVID-19 pandemic. These impacts are supported by data from similar viral contagions, such as SARS and the 1918 Flu Pandemic, and corroborated by US influenza data relative to the last decade. However, there is still limited understanding about the extent to which meteorology affects COVID-19 transmission rates and how meteorology can be a relevant element to anticipate social and politic measures. This study demonstrates that such an understanding is attainable through the development of regression models that verify the contribution of meteorology to the modeling of COVID-19 transmission, and the use of feature importance techniques assessing the relative weight of meteorological variables compared to epidemiological, socioeconomic, environmental, and global health indicator factors. The study results show that meteorological factors play an important role in regression models of COVID-19 transmission that have low error rates (R2 0.964). These results are corroborated by a panel data fixed-effect model showing that meteorological coefficients are often significantly correlated with COVID-19 transmission rates (R2 0.691-0.746, $p<0.01$).

Introduction

Since the early stages of the COVID-19 pandemic, climate has provided an important reference point to explain the spread of the virus. Just three months after the first outbreak in Wuhan, China, Bukhari and Jameel (2020) reported that 90% of COVID-19 cases were recorded through 22 March 2020 in world areas with colder and less humid late winters and early springs (temperatures of 3-17 °C, and absolute humidity of 4-9 g/m^3). As recognized by the authors, these initial data were likely to be biased by minimal testing per capita in tropical countries. Ten months into the pandemic, 217 countries across the globe have reported a total of 51,820,423 confirmed COVID-19 cases[1]. We now have a clearer picture of the global distribution of this disease. There are numerous examples indicating that the cooler season in the northern hemisphere may have favored the spread of the disease, while warmer and more humid weather in late spring and summer has seen a substantial and rapid decline in transmission numbers, once the different containment strategies adopted worldwide are taken into consideration[2,3]. Prior work on betacoronavirus genus shows that viruses similar to SARS-CoV-2 such as HCoV-HKU1 and HCoV-OC43 also display seasonal peak trends[4]. Previous coronavirus epidemics, such as SARS-related and Middle East respiratory syndrome (MERS)-related coronaviruses have also displayed correlation between their spread and seasonal weather changes, although with different behaviors[5-9]. In general, outbreaks of respiratory virus infections are commonly associated with seasonality, with peaks often occurring with low temperatures during the winter months. Data from the 1918-19 Flu Pandemic support this correlation. Peak infection/mortality during the 1918-19 Flu Pandemic occurred in the winter months and waned as solar radiation and absolute humidity increased from late March onward. The resurgence of mortality that took place during the 1918 Flu Pandemic in the winter of 1919 and its decline with the arrival of spring confirm this trend. These observations are corroborated by US influenza data relative to the last

decade where the percentage of patient visits for influenza-like illness consistently grows in the winter months.

A comprehensive understanding of the underlying elements involved in the contagion spread of this virus is crucial to plan and execute a timely response to the challenges that people are facing globally. These challenges transcend health aspects as they also affect the global markets and political institutions of our society. Further complexities arose from the growing spread of COVID-19 in countries that were going through the austral winter in the southern hemisphere (e.g. Australia, Brazil, South Africa) in the first phase of the pandemic, and have more recently developed with a second wave of infections, especially in Europe and the Northern hemisphere, bolstered by the flu season and the easing of restrictions during the summer. Despite the unprecedented efforts to obtain an effective vaccine in the shortest possible time, its distribution will probably differ considerably across countries and population sectors. Learning to cope and coexist with the COVID-19 pandemic is becoming an essential aspect of daily life. For these reasons, the ability of the scientific community to continue providing evidence on the evolution and dynamics of the pandemic is indispensable to enhance institutional preparedness to mitigate the spread of the pandemic.

Enhancing the understanding of the impact that climate factors have in the contagion spread is necessary to adjust to the COVID-19 “new normal” and promote the most appropriate policies of containment and mitigation to avoid overcrowding hospitals and impair health management systems among others. Regression analyses in the growing literature correlating meteorological factors with the spread of SARS-CoV-2 have reported contrasting results. A number of studies found an inverse (negative) correlation between temperature and the spread of COVID-19 in China and the US[10-12], Italy[13], Mexico[14], Latin America and the Caribbean[15], and worldwide[16]. By contrast, Xie and Zhu[17] suggest a direct (positive) relationship between temperature and the spread of COVID-19 (up to 3°C), and city-level data from Oslo, Jakarta and five Brazilian cities show a direct correlation between COVID-19 transmission and higher temperatures, and inverse correlation with precipitations[18-20]. Studies on the impact of humidity has also given conflicting results between reports of direct correlations, e.g. Jiang et al.[10] for China, and inverse correlations, e.g. Ward et al.[21] for Australia, Qi et al.[11] and Wang et al.[12] for China, and Jüni et al.[22] worldwide. Overall, there is still lack of conclusive evidence[23,24]. This is probably due to the fact that studies on the impact of climate on COVID-19 transmission have been piecemeal (e.g. limited to country or administrative division-level data), have only taken into account a few climatic parameters, and have not considered the impact of socioeconomic factor, as highlighted in Mecenas et al.[3]

The present study aims to investigate the relationship between meteorological and detected COVID-19 cases at the global scale to determine the relative impact of climate factors on the spread of the virus, using socioeconomic, environmental, and global health factors as control variables. It provides a comprehensive cohort study of COVID-19 transmission including 180 countries over an 11-month period spanning from 22 January 2020 to 11 November 2020. The study employs two complementary approaches to measure the correlation between the growth of the COVID-19 pandemic as reported by

epidemiological data and the selected meteorological, socioeconomic, environmental, and global health factors. The first approach uses machine learning modelling with gradient boosted decision trees in combination with feature importance analysis. The second approach is based on econometric analysis with panel-data fixed-effect regression models.

Results

We designed a prospective cohort study using global data of confirmed daily cases of COVID-19 to examine the association between the COVID-19 pandemic growth and climatic conditions, using several socioeconomic, environmental, and global health factors as control variables. We first apply machine learning to the selected data to understand the relative importance that climatic and control variables have on the spread of COVID-19. We then use econometrics analysis to verify the statistical significance of the correlation between climatic conditions and COVID-19 transmission.

Feature importance analysis

Relative feature impact is computed by applying the Tree SHAP (SHapley Additive explanation for tree-based machine learning models) algorithm to a Gradient Boosted Regression Tree (GBRT) model (see Methods section for details).

The GBRT model was trained on a time series dataset with epidemiological, meteorological, socioeconomic, environmental, and global health indicator variables, where the number of Covid-19 daily cases is the dependent variable. The model is trained on 80% of the dataset and evaluated on the remaining 20% using the Root Mean Square Error (RMSE) as performance metrics for model selection. Due to the highly skewed distribution of COVID-19 daily cases reported for the different locations under study during the evolution of the pandemic, a single training-testing partition would lead to a biased analysis. To avoid such bias, the evaluation is performed in over 200 iterations with randomly selected training-testing partitions. Results across all iterations are then averaged to provide unbiased evaluation results. As detailed in Table 5, the GBRT model achieved an overall RMSE of 952 88 on the test set, which is 25% better than the prediction performance of the baseline persistence model, where the value of the predicted dependent variable is assumed to be the same as the previous day. The GBRT model also outperforms (in terms of RMSE) the prediction obtained by using the 7-day moving average by 11%. The GBRT model provides a reliable and accurate account of the correlation between the dependent variable (COVID-19 daily cases) and the independent model variables, as shown by the close fit between the global aggregate of the confirmed and predicted COVID-19 daily cases in Figure 1 (R^2 of 0.964 on the test set). The model also offers a high degree of interpretability through the associated feature importance analysis computed with the Tree SHAP algorithm.

The results of feature importance analysis suggest that climate plays a meaningful role in modulating the dynamics of the COVID-19 pandemic, as indicated in Figure 2 where feature importance is ranked in

terms of logarithmic mean absolute SHAP-values. The analysis is based on a 7-day moving average for all the time-variant explanatory inputs of the regression. All the climatic and air quality factors score at similar levels of importance, showing that there is no dominant meteorological or air quality predictor. In line with results reported in the current literature on COVID-19 transmission studies and other coronaviruses, population susceptibility in terms of number of infected subjects (here estimated by the average number of previous daily cases) and the number of days elapsed from the first detected case are the most impactful features[25]. Socioeconomic, environmental, and global health indicator variables all show minor impact, including the Oxford COVID-19 Government Response Tracker (OxCGRT) stringency index which provides information on several common policy responses by governments to the pandemic such as school closures and travel restrictions.

Figure 3 shows the intensity and polarity of specific SHAP variable values for each data point (a dot in the plot) with reference to the predicted dependent variable value (high/low daily COVID-19 cases). The red-to-blue color scale indicates magnitude (high/red vs. low/blue). Position on the horizontal axis signals polarity (negative vs. positive). The resulting analysis suggests that temperature is expected to be significantly negatively correlated with the target, similarly to UV-index. Wind speed does not show a specific polarity correlation with the target/dependent variable, despite being among the most important features for the model prediction. Rainfall appears to be negatively correlated with the target/dependent variable, but there are too few observations for high rainfall values in the dataset to properly confirm this result. PM 2.5 levels show a weak positive correlation with COVID-19 spread, while PM 10 levels show a weak negative correlation with COVID-19 spread, although there likely external cofounding factors that may play a decisive role in this result. Absolute humidity displays a weak positive correlation, despite the expected impact according to previous studies (see Introduction). SHAP values for the epidemiological terms have been omitted from the plot to focus on the effect of the other variables.

Econometric analysis

The econometric analysis was carried out using Panel data Fixed Effect Models with ordinary least-square (OLS) regression analysis. Confirmed daily cases of COVID-19 in log-scale were regressed against climate and air-quality factors, with reference to cross-sectional and time fixed effects. The main focus of the analysis is on the significance of the coefficients since the nominal magnitude of the single variable is potentially correlated with the error term (endogeneity) due to undetermined confounding effects. Since the effect of variables that behave as time-invariant factors for the period of focus (e.g. socioeconomic, environmental, and some global health indicator variables) would be absorbed in the intercept due to the use of time fixed effect regression, these variables were removed. For additional comments on the removal of OxCGRT stringency index, see final paragraph below. As a caveat, the statistical significance of correlation coefficients does not necessarily imply causality. This is because it is not possible to test for causality by creating an appropriate treatment and control groups with balanced baseline characteristics due to the rapidly changing nature of the epidemiological data.

The statistical significance of correlation coefficients is computed by clustering the regression standard error at the country/regional level, to account for error correlation within the geographical areas where our unit of observation was collected. For each location, we select days with a minimum of 10 confirmed COVID-19 cases, and we limit our analysis to the period following the closure of the country international borders. This strategy is intended to limit the inclusion of imported cases in the daily COVID-19 infections count of each location. The results are reported in Table 1. The regression has an R^2 of about 72% over 36,084 observations. COVID-19 spread shows a strong negative correlation with both UV radiation (UV-index) and absolute humidity, statistically significant at .01 level ($P < 0.01$), and a mild negative correlation with temperature at a lower statistically significant level ($P < 0.10$). For other climatic factors, the econometric analysis is congruent with the feature importance analysis but does not yield statistically significant correlation coefficients with the dependent variable (daily COVID-19 cases). Also, different moving averages (5, 7, 10, 12 and 14 days), which relate climatic variables to incubation periods of diverse duration, do not seem to influence the overall result of the econometric analysis.

The removal of the OxCGRT stringency index, which serves as a proxy indicator for health intervention policies, was needed to avoid biased changes in the polarity of its coefficient due to the skewed distribution of the OxCGRT stringency index across country data. When the OxCGRT stringency index is included in the econometric analysis, the absolute values of the correlations coefficients of the climatic factors are unperturbed, but the polarity of the stringency index results positive (Table 2). This is probably due to the fact that the temporal lag between the enactment of restrictions and the pandemic peak tends to vary from country to country due to the diversity, severity, and enforcement capacity of the containment policies implemented. This observation is corroborated by the fact that correlations coefficients for climate variables have statistically significant negative polarity (Table 3), as expected, when the analysis focuses on countries with similar socioeconomic characteristics, such as Italy, Spain, Germany, France and the UK, that have implemented similar containment policies from the inception of the pandemic to the time of lifting restrictions (March-May 2020). Due to these discrepancies, it is best to omit the OxCGRT stringency index in the econometric analysis. The inclusion of cross sectional and time fixed effects ensures that this omission does not compromise the robustness of the analysis.

Discussion

Overall, disease susceptibility is the main factor driving the pandemic growth. Compliance with lockdown and restrictions policies and regulations and increased testing are the most effective strategies for disease control and COVID-19 spread prevention. For example, various studies have reported that interventions such as restrictions on mass gatherings, school closures, and social distancing measures are strongly associated with a decrease in the COVID-19 transmission growth rate[22,26,27]. The correlation of COVID-19 transmission with climate factors provides a valuable complementary diagnostic that sheds light on the seasonal characterization of the pandemic and helps refine measures containing and preventing the spread of COVID-19. More specifically, weather forecasts could help predict new cycles of the pandemic and future outbreaks and thus contribute to the definition of ad-hoc measures that limit the

economic impact of complete lockdowns. This study also extends the reach of previous studies on the relationship between COVID-19 transmission and climate factors by assessing how climate helps modeling COVID-19 through a systematic validation through feature importance and econometric analyses. Such a validation is crucial in establishing which are the contributing factors and their relative magnitude and direction of change.

Limitations and Assumptions

As for other data driven studies on COVID-19 transmission, the present analysis relies on records whose quality varies across sources, due to heterogeneous collection and reporting practices worldwide. Furthermore, reports on confirmed COVID-19 cases tend to underestimate the actual number of infections because of asymptomatic patient and undetected COVID-19 deaths. For the purpose of this study, we assume that the number of confirmed COVID-19 cases is monotonically related to the true number of infections.

Methods

Data sources and processing

The data used in this study include epidemiological, socioeconomic, environmental, global health indicator, and meteorological variables. All population-related variables are converted to percentages of total population per country. Parameters that behave as time-invariant variables during the period of focus for this study, e.g. socioeconomic variables, were used as control variables.

Epidemiological data on cumulative number of confirmed COVID-19 cases were retrieved for the period from 22 January 2020 to 11 November 2020 from two main sources: the data repository by the Johns Hopkins Center for Systems Science and Engineering[28], and the Corona Data Scraper online data service that pulls COVID-19 Coronavirus case data from verified sources worldwide and adds population data on a daily basis (coronadatascraper.com). Data from these two sources are merged to create the initial starting dataset. When available, data at the regional or state level were included, in addition to country-level records. Country, region, and state aggregations were selected so as to achieve location records with comparable population size. We derive the number of daily registered COVID-19 cases by differencing entries in the initial dataset. Inconsistent data points (e.g. negative values) and records reporting less than 10 cases were removed. The time of exposure to the pandemic for each country was calculated as the cumulative temporal distance from the first registered case in the country.

Socioeconomic data include demographic information, technology adoption rates, and Gross Domestic Product per-capita (GDPP). Socioeconomic parameters were used as control variables as they change during the time period considered for this study. Demographic, population density and population age data were derived from the 2019 Population Division dataset compiled by the Department of Economic

and Social Affairs of the United Nation (UNDESA)[29] and the World Bank indicators database[30]. Information for locations not included in the UNDESA dataset were retrieved online directly from national official sources. Rates of internet users, subscribers to mobile telephony services, and the number of secure Internet servers were retrieved from the World Bank indicators database. These technology adoption variables are used as proxies for the capacity of different countries to provide smart-work environments under lockdown restrictions in terms of ease of implementation of smart-working initiatives and keep the national population informed about the development of the pandemic. GDPP data at constant price purchasing-power parity were sourced from the International Monetary Fund's World Economic Outlook Database[31].

Environmental indicators retrieved from the World Bank indicators database include population-weighted exposure to ambient PM2.5 pollution, carbon dioxide, methane, nitrous oxide emissions, and greenhouse gas emissions. These variables were used as indicators of the degree of pollution of each country, on the assumption that long-term exposure to pollutants may increase the risk of contracting COVID-19. Environmental indicators were also used as control variables.

Health indicators include the Global Health Security (GHS) index, diabetes prevalence and number of hospital beds for both acute and chronic care, and the Oxford COVID-19 Government Response Tracker (OxCGRT) stringency index. GHS providing a country-level score of health security, was used as proxy variable for a country's capability to prevent and mitigate infectious diseases. For the purpose of this study, only the "detect" category was used, which focuses on the country readiness to promptly identify and report disease outbreaks of potential international concern[32]. Health indicators relative to diabetes prevalence and number of hospital beds for both acute and chronic care were retrieved from the World Bank indicators database. These variables serve as proxies for population health status and public health preparedness. Differences in intervention responses by governments to mitigate the pandemic are accounted for in terms of the Oxford COVID-19 Government Response Tracker (OxCGRT) stringency index[33]. The OxCGRT stringency index includes information on containment and closure policies, such as school closures and restrictions in movement, limits on private gatherings, and full lockdowns.

Meteorological variables were obtained from two main sources: The Global Forecast System (GFS), a weather forecast model produced by the National Centers for Environmental Prediction, and the Copernicus Atmosphere Monitoring Service (CAMS) with specific reference to the McClear Clear-Sky Irradiation model. Data obtained through GFS include daily averages, at 0.25° spatial resolution, of temperature and relative humidity at 2 m above ground, pressure at ground level, wind speed at 10 m above ground, and rainfall. Temperature is combined with relative humidity to derive measures of absolute humidity[34]. Data obtained from CAMS include the UV biologically effective dose rate (UVBED) in W/m^2, and Particulate Matter (PM) concentrations (PM2.5 and PM10) for 3-hour periods at a spatial resolution of 40 km. The UVBED values are then averaged to obtain a daily estimate and divided by 0.025 W/m^2 to derive the resulting dimensionless global solar UV index forecast. UV exposure can have a sterilizing effect[35]. UVB, which is present in small amounts in natural sunlight, is known to rapidly

inactivate SARS-CoV-2 on surfaces[36]. Data on particulate matter[37], originally in kg/m^3, is converted to micrograms/m^3, provide preliminary evidence on the relation between air quality and the chronicity of exposure to the viral infection in northern Italy. Coccia[38] suggests instead that air pollution may have accelerated the transmission rate of COVID-19, even though the viability of infectious virus embedded on suspended aerosol particles is still under debate[39]. Streaming access to GFS and CAMS was provided by Transvalor. For each reference point, we derived the geographical centroid and used the corresponding latitude and longitude coordinates to query the climatic information from GFS and CAMS through Transvalor's SoDa data service (http://www.soda-pro.com/).

After merging all the different sources, the resulting dataset includes data on 180 countries covering a total population of 7,505,800,179 (96% of word population) and 46,906,240 confirmed COVID-19 cases (97.7% of worldwide cases). Data for 28 of these countries are detailed in the dataset at a state or regional level for the available periods (see additional information in appendix for the detailed list). We consider only country-level epidemiological data for the remaining 152 countries, even if regional-level data are available from coronadatascraper.com, to maintain a certain level of minimum comparison between the locations under study in terms of overall population size.

Information on the enforcement of policy responses to COVID-19 regarding the closure of international borders and the implementation of lockdown or curfew measures (or similar policies restricting the movement of the population within the country borders) were collected from a variety of in-country data sources, including government public health websites, regional newspaper articles and crowd-sourced information on Wikipedia when verifiable sources were available.

Moving Averages for Climatic Data

The COVID-19 mean incubation period, defined as the time period ranging between the exposure to the virus and the onset of the illness, is estimated by WHO at 5-6 days (median 5.1 days, 95% CI 4.5 to 5.8 days)[40]. According to Lauer et al.[40], 97.5% of those who develop symptoms will do so within 11.5 days (CI, 8.2 to 15.6 days) from the day of infection. Furthermore, on average there is a delay of few days (3.6 days according to Cereda et al.[41]) to receive the answer from the COVID-19 PCR test. For these reasons, the number of new COVID-19 cases that are publicly announced each day correspond to a time-window of infection that spans from few days up to potentially two weeks earlier. To account for this timeframe uncertainty and test the robustness of our results, the analysis is carried out with moving averages for the climatic variables of different duration: 5, 7, 10, 12, and 14 days (minimum length of 3 days).

Machine learning modelling strategy

The objective of machine learning modeling is to compute the relative feature impact of factors contributing to COVID-19. Feature impact is computed by applying the Tree SHAP algorithm to a Gradient

Boosted Regression Tree (GBRT) model.

GBRT is an additive stochastic model that combines multiple sequentially connected weak learners (regression trees) in way that each new learner fits to the residuals from the previous step to optimize the overall predictive performance[42]. The resulting model can describe multiple nonlinear interaction and partial dependency with sufficient flexibility, very high predictive accuracy, and robustness to missing data and outliers.

The study uses the open source *xgboost* Python library as which provides a highly efficient, flexible, and portable implementation of GBRT[43]. The xgboost algorithm provides several ways to control and reduce overfitting, i.e. when the model fits the training data so precisely that it ends up learning the noise instead of the target signal, thus failing to generalize well when used with new data. The maximum depth of the individual trees used in the boosting process is one aspect that allows to modulate the degree of feature interactions that the model can fit. Another option to control model complexity is the implementation a lower bound on the minimum number of samples that each leaf can contain. The third most important regularization parameter is the learning rate that scales newly added weights by a factor after each step of tree boosting round. Overfitting is reduced by introducing randomization into the tree building process by enabling the creation a subsample of the training set before deriving each tree, and the subsampling of the columns features before searching for the best node split.

The GBRT model was trained on a randomly selected 80% portion of the whole data, using the remaining 20% for testing the model prediction accuracy. We further optimized the model hyperparameters using grid-search approach combined with a 5-fold cross-validation technique on the training set as preventative measure against overfitting and to reduce the model variance error. This step led to a 1% further improvement on the final prediction performance over the same algorithm initialized with the default hyperparameters values. Once the GBRT model is trained, the relative ranking of the model parameters is obtained through the SHAP method (see below). Table 4 summarizes the set of hyperparameters leading to the best evaluations results.

Tree SHAP is an algorithm that computes SHAP values for Decision Trees models such as GBRT. SHAP (SHapley Additive exPlanation)[44,45] uses a game theoretic approach to explain the prediction for any instance as a sum of contributions from its individual feature values. This type of analysis does not identify causal correlation, but it is still a useful metric to capture relative feature importance.

Econometric analysis approach

The econometric analysis of the association between the daily number of confirmed COVID-19 infections and climatic factors at a global scale is carried out using the multivariate equation in (1)[46]:

$$ln\_daily\_cases_{i,t} = \beta_0 + \beta_1\, temperature_{i,t} + \beta_2\, absolute\_humidity_{i,t} + \beta_3\, wind\_speed_{i,t} + \beta_4\, rainfall_{i,t} + \beta_5\, UV\_index_{i,t} + \beta_6\, PM10_{i,t} + \beta_6\, PM2.5_{i,t} + c_i + \lambda_t + u_{i,t} \quad (1)$$

The dependent variable $Ln\_daily\text{-}cases_{it}$ expresses the number of daily cases of COVID-19 cases for country/region *i* and time index *t*. β_0 is the regression intercept, while $\beta_{1\ldots6}$ are the different regression slope coefficients of the respective climatic factors. Further details on the construction of all the variables is given in the previous sections. Equation (1) is computed using Ordinary Least Squares (OLS) regression on the panel dataset. We include a vector of cross-sectional unit fixed effects c_i to account for all time-invariant factors that affect the local growth rate of infections, such as differences in demographics, socioeconomic status, culture and health systems. We also include a vector of (daily) time fixed λ_t effects to absorb the autoregressive component specific of the COVID-19 spread growth, and to account for the presence of any potential seasonal bias. Finally, we cluster the standard errors u_{it} at entity-level to account for error correlation within each location. Table 1 lists the values of the intercept (constant) and the β coefficients with their respective standards of error for each climatic factors resulting from our panel regression analysis under the different hypotheses of duration for the moving average window (T). Table 2 and Table 3 present the estimation results of a modified version of equation (1) with different explanatory variables where we try to single out the sole contribution effect of the containment policies on the spread of COVID-19.

Declarations

Code availability

All machine learning models, and data processing steps presented in this work were developed using Python and the open-source scikit-learn and the xgboost software library.

Acknowledgements

This publication was made possible by QNRF grant RCC-2-044 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

Competing interests

The authors declare no competing interests. This study did not require research ethics approval, as publicly available, anonymized aggregate data were used for all analyses

Additional Information

Supplementary information is available for this paper in the Appendix

Correspondence and requests for materials should be addressed to Giovanni Scabbia gscabbia@hbku.edu.qa, and/or to Antonio Sanfilippo asanfilippo@hbku.edu.qa

References

1 Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* **20**, 533-534 (2020).

2 Briz-Redón, Á. & Serrano-Aroca, Á. The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques. *Progress in Physical Geography: Earth and Environment* **44**, 591-604 (2020).

3 Mecenas, P., Bastos, R., Vallinoto, A. & Normando, D. Effects of temperature and humidity on the spread of COVID-19: A systematic review. *medRxiv* (2020).

4 Moriyama, M., Hugentobler, W. J. & Iwasaki, A. Seasonality of respiratory viral infections. *Annual review of virology* **7** (2020).

5 Al-Ahmadi, K., Alahmadi, S. & Al-Zahrani, A. Spatiotemporal clustering of Middle East respiratory syndrome coronavirus (MERS-CoV) incidence in Saudi Arabia, 2012–2019. *International journal of environmental research and public health* **16**, 2520 (2019).

6 Altamimi, A. & Ahmed, A. E. Climate factors and incidence of Middle East respiratory syndrome coronavirus. *Journal of Infection and Public Health* **13**, 704-708 (2020).

7 Gardner, E. G. *et al.* A case-crossover analysis of the impact of weather on primary cases of Middle East respiratory syndrome. *BMC infectious diseases* **19**, 1-10 (2019).

8 Lin, K., Fong, D. Y.-T., Zhu, B. & Karlberg, J. Environmental factors on the SARS epidemic: air temperature, passage of time and multiplicative effect of hospital infection. *Epidemiology & Infection* **134**, 223-230 (2006).

9 Yuan, J. *et al.* A climatologic investigation of the SARS-CoV outbreak in Beijing, China. *American journal of infection control* **34**, 234-236 (2006).

10 Jiang, Y., Wu, X.-J. & Guan, Y.-J. Effect of ambient air pollutants and meteorological variables on COVID-19 incidence. *Infection Control & Hospital Epidemiology*, 1-11 (2020).

11 Qi, H. *et al.* COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Science of the Total Environment*, 138778 (2020).

12 Wang, J., Tang, K., Feng, K. & Lv, W. High temperature and high humidity reduce the transmission of COVID-19. *Available at SSRN 3551767* (2020).

13 Coro, G. A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate. *Ecological Modelling* **431**, 109187 (2020).

14 Méndez-Arriaga, F. The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1. *Science of The Total Environment*, 139560 (2020).

15 Bolaño-Ortiz, T. R. *et al.* Spread of SARS-CoV-2 through Latin America and the Caribbean region: a look from its economic conditions, climate and air pollution indicators. *Environmental research* **191**, 109938 (2020).

16 Sobral, M. F. F., Duarte, G. B., da Penha Sobral, A. I. G., Marinho, M. L. M. & de Souza Melo, A. Association between climate variables and global transmission oF SARS-CoV-2. *Science of The Total Environment* **729**, 138997 (2020).

17 Xie, J. & Zhu, Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment* **724**, 138201 (2020).

18 Auler, A., Cássaro, F., da Silva, V. & Pires, L. Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: A case study for the most affected Brazilian cities. *Science of The Total Environment*, 139090 (2020).

19 Menebo, M. M. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. *Science of The Total Environment*, 139659 (2020).

20 Tosepu, R. *et al.* Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of The Total Environment*, 138436 (2020).

21 Ward, M., Xiao, S. & Zhang, Z. The Role of Climate During the COVID-19 epidemic in New South Wales, Australia. *Authorea*, doi:10.22541/au.158879258.84484606 (2020).

22 Jüni, P. *et al.* Impact of climate and public health interventions on the COVID-19 pandemic: a prospective cohort study. *Cmaj* **192**, E566-E573 (2020).

23 Neher, R. A., Dyrdak, R., Druelle, V., Hodcroft, E. B. & Albert, J. Potential impact of seasonal forcing on a SARS-CoV-2 pandemic. *Swiss medical weekly* **150** (2020).

24 Sajadi, M. M. *et al.* Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19). *JAMA Network Open* **3**, e2011834-e2011834 (2020).

25 Weitz, J. S. *et al.* Modeling shield immunity to reduce COVID-19 epidemic spread. *Nature Medicine* **26**, 849-854, doi:10.1038/s41591-020-0895-3 (2020).

26 Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395-400 (2020).

27 Hsiang, S. *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262-267 (2020).

28 CSSE, J. (ed JHU CSSE) (2020).

29 Affairs), U. U. N. D. o. E. a. S. *World Population Prospects 2019* (ed UNDESA (United Nations Department of Economic and Social Affairs)) (2019).

30 Bank, W. (ed World Bank) (2020).

31 Fund), I. I. M. (ed IMF (International Monetary Fund)) (Washington, DC, 2019).

32 Ravi, S. J. *et al.* The value proposition of the Global Health Security Index. *BMJ global health* **5**, e003648 (2020).

33 Hale, T., Petherick, A., Phillips, T. & Webster, S. Variation in government responses to COVID-19. *Blavatnik school of government working paper* **31** (2020).

34 Oyj, V. Humidity conversion formulas. Calculation formulas for humidity. (Helsinki, Finland, 2013).

35 Hockberger, P. E. The discovery of the damaging effect of sunlight on bacteria. *Journal of Photochemistry and Photobiology B: Biology* **58**, 185-191 (2000).

36 Ratnesar-Shumate, S. *et al.* Simulated sunlight rapidly inactivates SARS-CoV-2 on surfaces. *The Journal of Infectious Diseases* (2020).

37 Fattorini, D. & Regoli, F. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environmental Pollution*, 114732 (2020).

38 Coccia, M. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. *Science of the Total Environment*, 138474 (2020).

39 Lewis, D. Is the coronavirus airborne? Experts can't agree. *Nature* **580**, 175 (2020).

40 Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine* **172**, 577-582 (2020).

41 Cereda, D. *et al.* (Arxiv, 2020).

42 Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232 (2001).

43 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

44 Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).

45 Lundberg, S. M. & Lee, S.-I. in *Advances in neural information processing systems.* 4765-4774.

46 Wooldridge, J. M. *Econometric analysis of cross section and panel data*. (MIT press, 2010).

Tables

Table 1. Panel data Fixed Effects model. T-days moving average

Dependent variable: Daily cases (log)	(1) T = 5	(2) T = 7	(3) T = 10	(4) T = 12	(5) T = 14
PM2.5	-0.001	-0.003	-0.005	-0.007	-0.008
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
PM10	0.002	0.002	0.002	0.003	0.003
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Temperature	-0.022*	-0.024*	-0.025*	-0.026*	-0.026*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Absolute Humidity	-0.057***	-0.063***	-0.068***	-0.070***	-0.072***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Wind speed	-0.009	-0.013	-0.003	0.006	0.017
	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Rainfall	-0.005	-0.005	-0.003	-0.001	0.001
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
UV index	-0.012***	-0.011***	-0.010**	-0.010**	-0.009**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	2.581***	2.541***	2.449***	2.415***	2.361***
	(0.14)	(0.15)	(0.19)	(0.20)	(0.22)
Observations	36084	36084	36084	36084	36084
R^2	0.725	0.725	0.726	0.727	0.727
Adjusted R^2	0.718	0.719	0.720	0.720	0.721

Standard errors in parentheses are clustered at location (country/region) level,

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2. Panel data Fixed Effects model. Testing the effect of restrictions.

Dependent variable: Daily cases (log)	(1)	(2)	(3)	(4)
PM2.5	-0.001	-0.001	-0.003	-0.004
	(0.01)	(0.01)	(0.01)	(0.00)
PM10	0.002	0.002	0.002	0.002
	(0.00)	(0.00)	(0.00)	(0.00)
Temperature	-0.028**	-0.031**	-0.023*	-0.022*
	(0.01)	(0.01)	(0.01)	(0.01)
Absolute Humidity	-0.045*	-0.046*	-0.054**	-0.053**
	(0.02)	(0.02)	(0.02)	(0.02)
Wind speed	-0.026	-0.025	-0.001	-0.001
	(0.03)	(0.02)	(0.02)	(0.02)
Rainfall	-0.009	-0.008	-0.002	-0.001
	(0.01)	(0.01)	(0.01)	(0.01)
UV index	-0.008**	-0.007*	-0.008*	-0.007*
	(0.00)	(0.00)	(0.00)	(0.00)
Lockdown (7 days shift)	0.571***			
	(0.10)			
Lockdown (12 days shift)		0.456***		
		(0.10)		
Stringency index (7 days shift)			0.034***	
			(0.00)	
Stringency index (12 days shift)				0.035***
				(0.00)
Constant	2.798***	3.132***	0.520**	0.570***
	(0.21)	(0.19)	(0.22)	(0.21)
Observations	31872	29263	35866	35873
R^2	0.757	0.775	0.751	0.753
Adjusted R^2	0.752	0.769	0.746	0.747

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3. Panel data Fixed Effects model. Result only for the regions of Germany, Italy, Spain, France, and the United Kingdom.

Dependent variable: Daily cases (log)	(1)	(2)	(3)	(4)
Lockdown (7 days shift)	0.194**			
	(0.09)			
Lockdown (10 days shift)		-0.118		
		(0.07)		
Lockdown (12 days shift)			-0.333***	
			(0.07)	
Lockdown (14 days shift)				-0.542***
				(0.06)
Constant	7.574***	7.936***	8.126***	8.378***
	(0.07)	(0.05)	(0.05)	(0.04)
Observations	2158	1981	1864	1748
R^2	0.846	0.856	0.862	0.877
Adjusted R^2	0.841	0.852	0.857	0.873

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4. Xgboost hyperparameter tuning result

Hyperparameter	Tuning range	Tuned value
Learning rate (eta)	0.01 - 0.3	0.1
Maximum depth	3 - 10	4
Minimum sum of instance weight (hessian)	1 - 10	5
Gamma	0 - 0.4	0
Subsample ratio of the training instances	0.5 - 1	0.9
Subsample ratio of columns when constructing each tree	0.3 - 1	1
Lambda	1	1
Alpha	0	0

Table 5. Regression modelling performance comparison (best model – lowest mean RMSE)*

Model	Mean RMSE Standard Deviation	Difference from best-model performance (higher RMSE in %)
GBRT (with hyperparameter optimization) * - test score	952 88	-
GBRT (no optimization) - test score	959 92	1%
GBRT (no optimization) - train score	268 8	
Constant: 7-days moving average	1052 247	11%
Persistence (1-day ahead)	1189 247	25%

Appendix

Data available at regional-, county-, or province-level for the following countries: Czechia, Denmark, Estonia, France, Ireland, Italy, Japan, Latvia, Lithuania, Netherlands, Panama, Poland, South Africa, South Korea, Spain, Sweden, Ukraine.

Data available at state-level for the following countries: Australia, Austria, Brazil, Canada, China, Germany, India, Russia, Switzerland, United States, United Kingdom.

Data available at country-level for the following countries: Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh,

Barbados, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Burkina Faso, Burma, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo (Brazzaville), Congo (Kinshasa), Costa Rica, "Cote dIvoire", Croatia, Cuba, Cyprus, Czechia, Denmark, Djibouti, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Eswatini, Ethiopia, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guam, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Korea, South, Kosovo, Kuwait, Kyrgyzstan, Latvia, Lebanon, Lesotho, Liberia, Libya, Liechtenstein, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Moldova, Monaco, Mongolia, Montenegro, Morocco, Mozambique, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, North Macedonia, Norway, Oman, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russia, Rwanda, San Marino, Sao Tome and Principe, Saudi Arabia, Senegal, Serbia, Seychelles, Sierra Leone, Singapore, Slovakia, Slovenia, Somalia, South Africa, South Sudan, Spain, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, US, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States Virgin Islands, Uruguay, Uzbekistan, Venezuela, Vietnam, West Bank and Gaza, Yemen, Zambia, Zimbabwe.

Figures

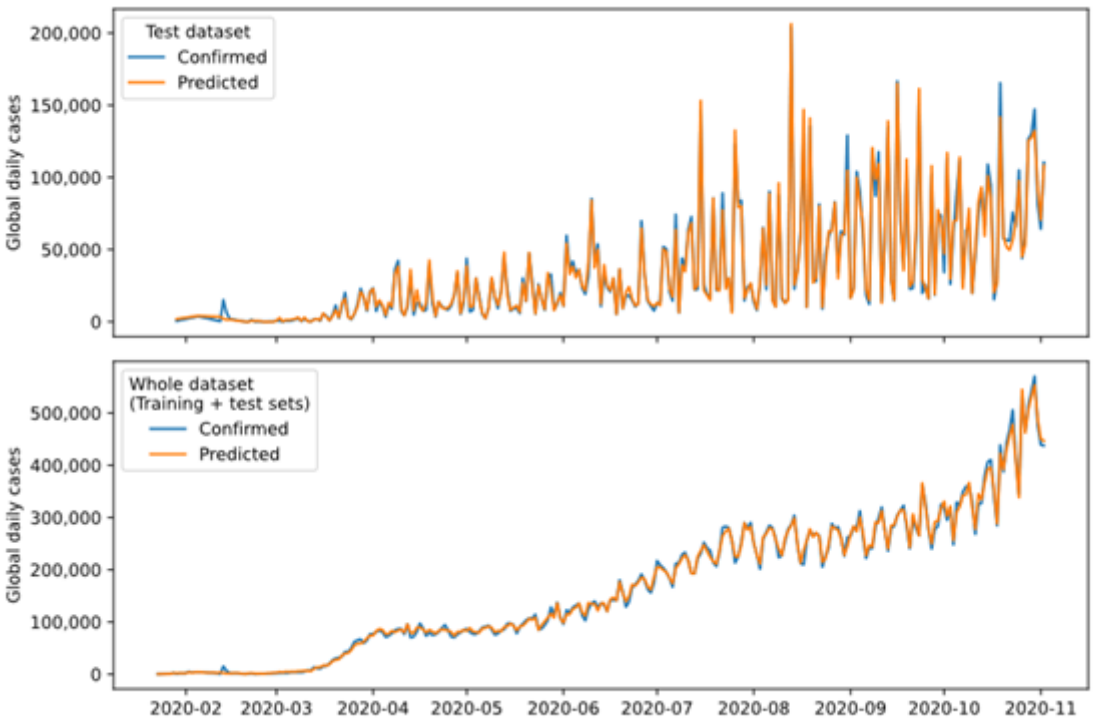

Figure 1

Comparison between the confirmed and predicted COVID-19 daily cases at the global level. Top image shows the prediction accuracy of our methods when tested on a unseen portion of the dataset. Bottom figure depicts the overall model performance when considering the whole dataset (including both the training and testing sets).

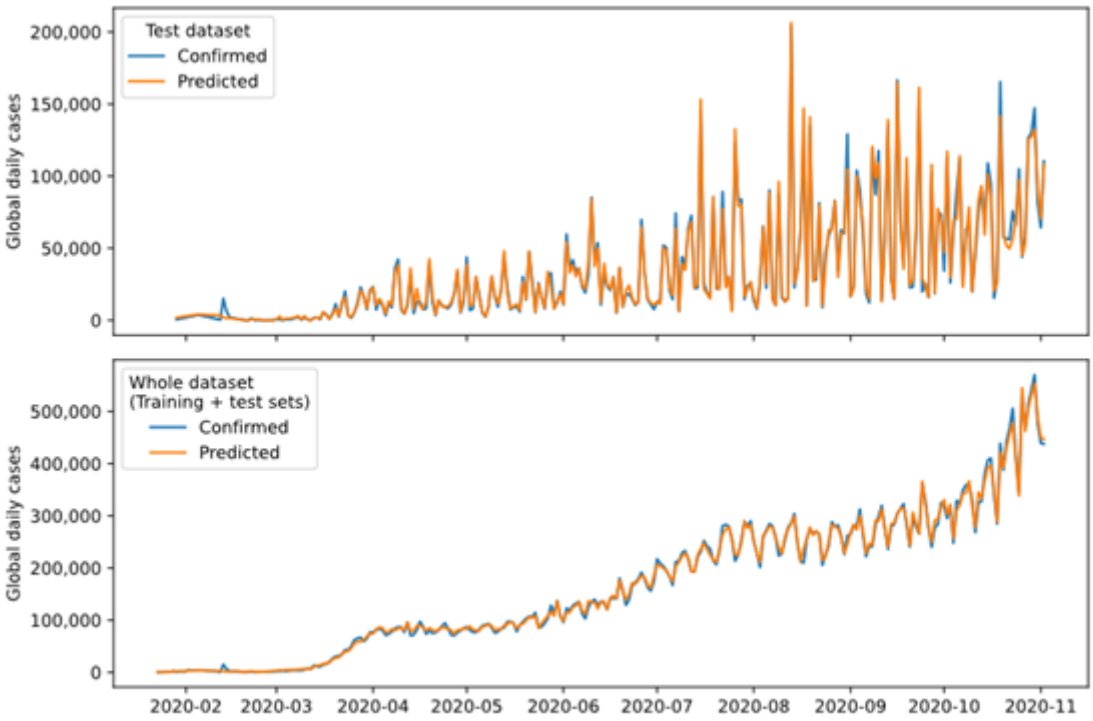

Figure 1

Comparison between the confirmed and predicted COVID-19 daily cases at the global level. Top image shows the prediction accuracy of our methods when tested on a unseen portion of the dataset. Bottom figure depicts the overall model performance when considering the whole dataset (including both the training and testing sets).

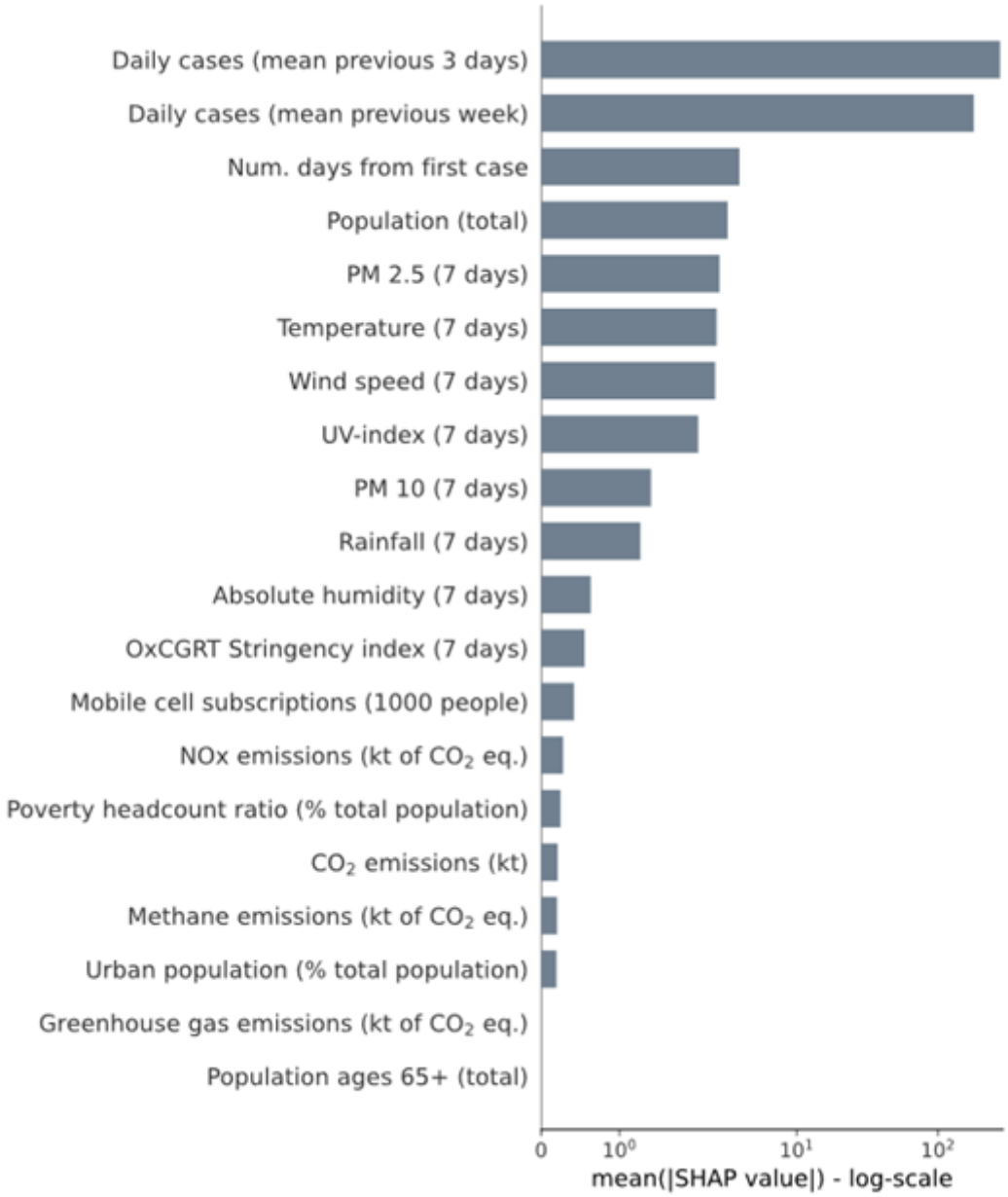

Figure 2

Feature importance plot. Mean absolute SHAP value (in log scale) of each variable showing the average impact on the model output magnitude

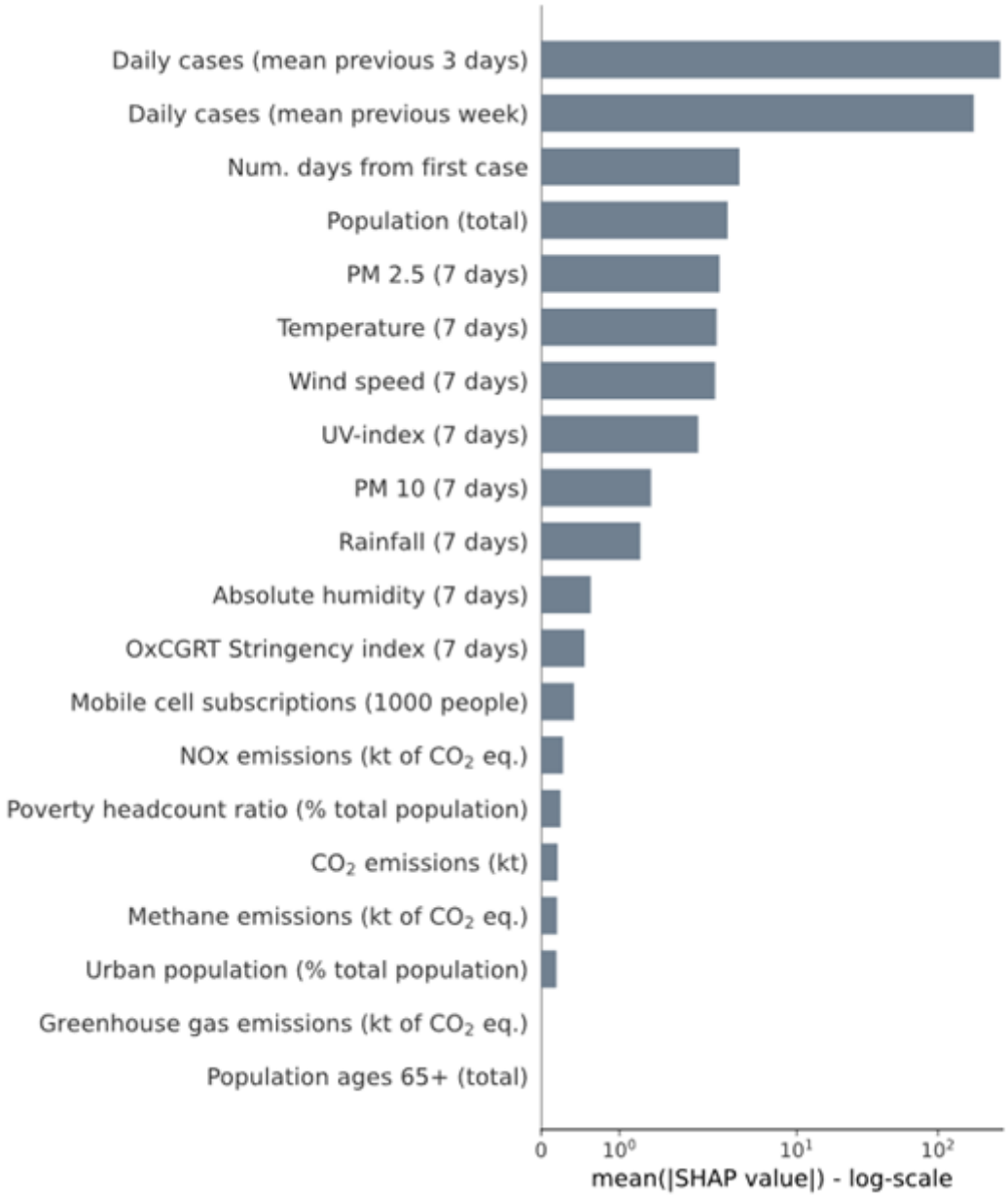

Figure 2

Feature importance plot. Mean absolute SHAP value (in log scale) of each variable showing the average impact on the model output magnitude

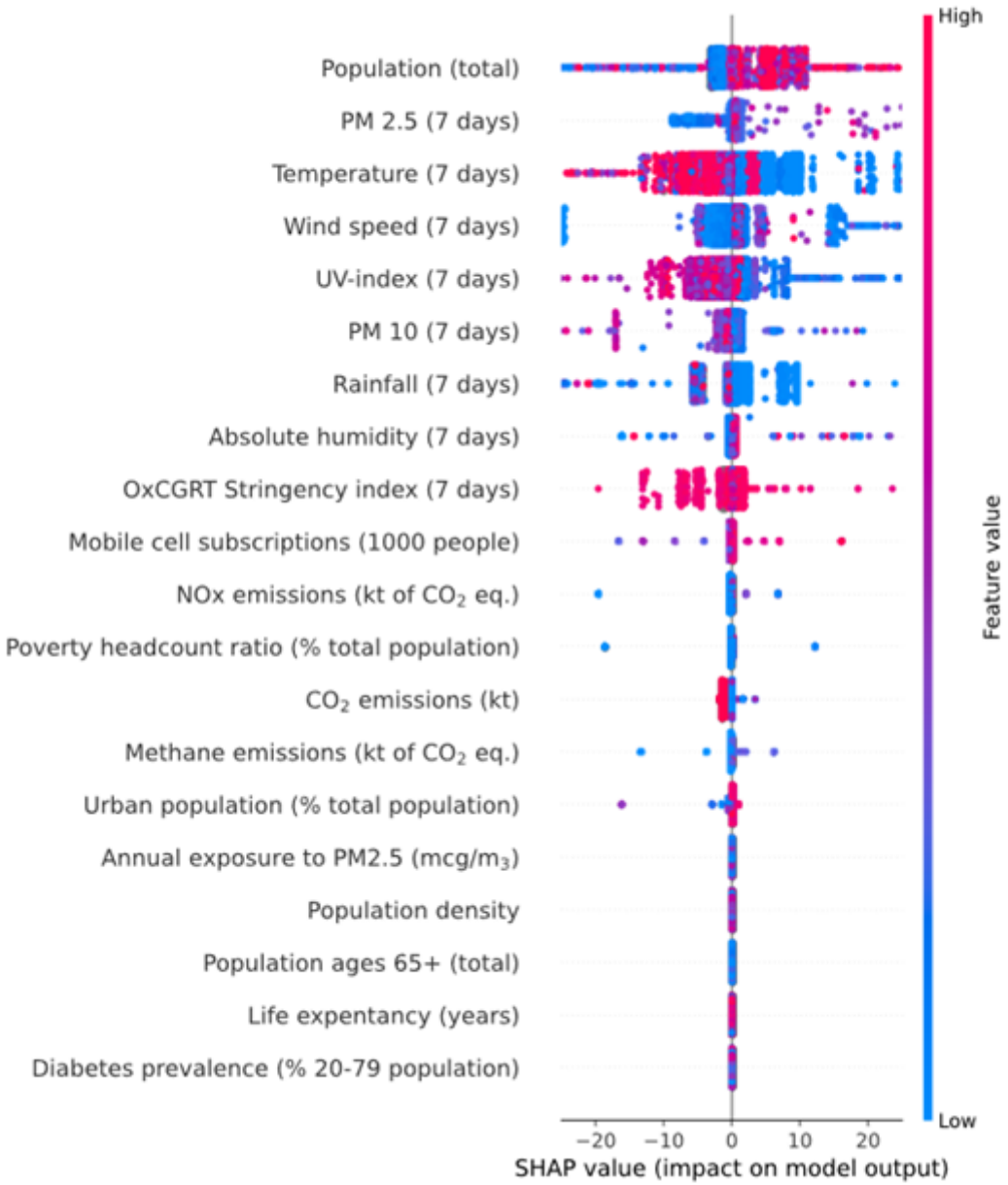

Figure 3

Feature impact plot. SHAP value of each variable for all the single observations as a function of their relative value.

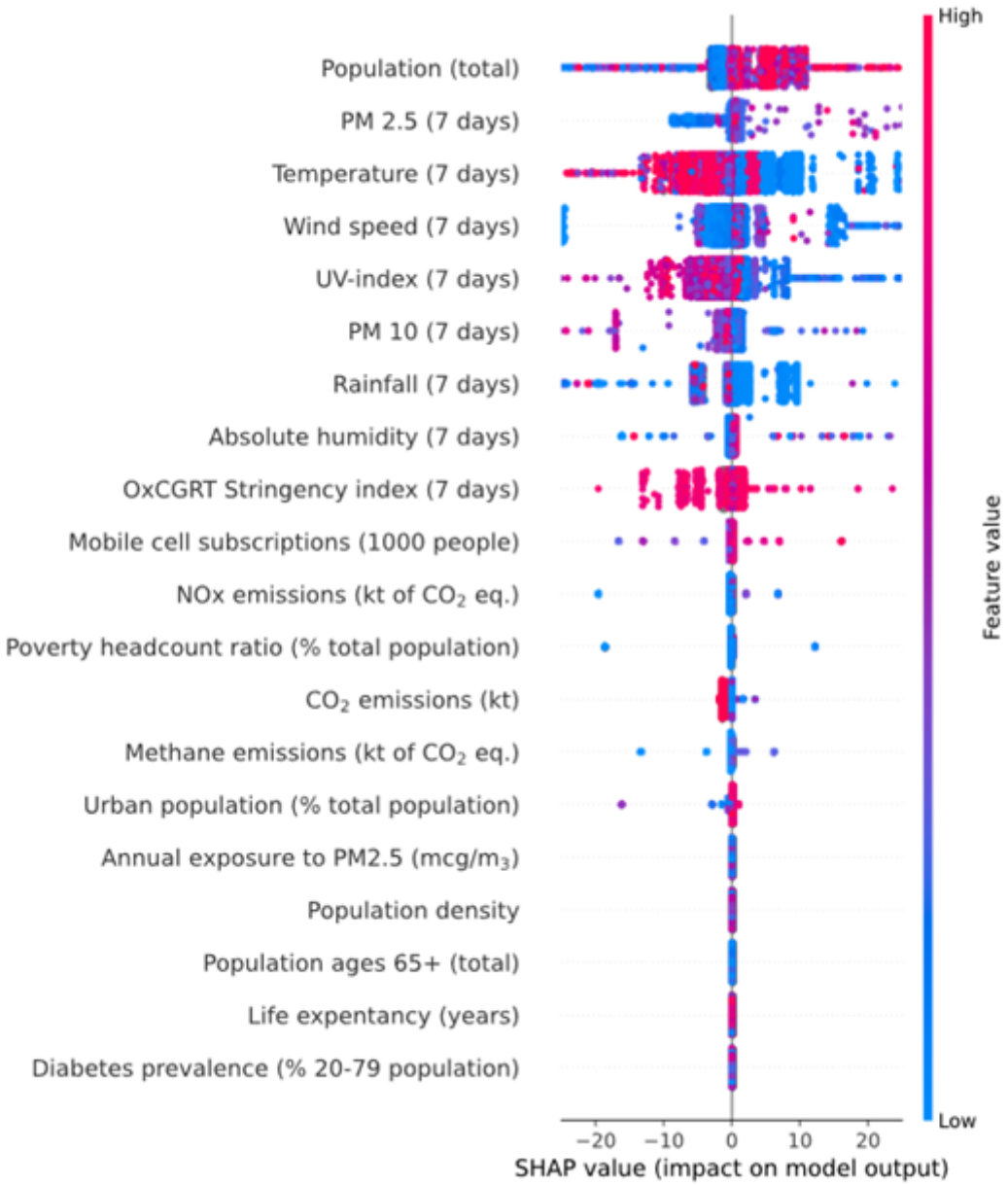

Figure 3

Feature impact plot. SHAP value of each variable for all the single observations as a function of their relative value.

Supplementary Files

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

- COVID19v6.720201111.zip
- COVID19v6.720201111.zip