

COVID 19 Peak Time Prediction via a Gradient Boosting Method [†]

Basak Cetinguc * and Eyup Calik

Industrial Engineering Department, Engineering Faculty, Yalova University, Central Campus, Yalova 77200, Turkey; ecalik@yalova.edu.tr

* Correspondence: basak.cetinguc@yalova.edu.tr; Tel.: +90-226-8155359

[†] Presented at the 7th International Management Information Systems Conference, Online, 9–11 December 2020.

Abstract: The outbreak of COVID-19 has caught humanity off guard. Peak-times differ in countries based on their characteristics and precautions taken by governments. In this study, we aimed to determine relative importance of indicators on the spread and to assist non-peaked countries to estimate their peak-times. Gradient Boosting Method was employed on 82 countries which reached peak-times. The findings indicate that hospital beds per thousand is the main predictor of peak-time estimation. Restrictions on gatherings and closing public transportation have the highest relative importance among governmental precautions. This model can be utilized and employed with various indices and alternative machine-learning algorithms.

Keywords: COVID-19; peak times; gradient boosting method; government; restrictions

1. Introduction

The outbreak of COVID-19 pandemic has been one of the major health issues for human beings came across since decades. COVID-19 is a novel coronavirus disease known as severe acute respiratory syndrome coronavirus 2 appeared in Wuhan, China in December 2019 [1]. By the end of July, almost 16.5 million people is infected all over the world and more than 652 thousand people died because of this virus. Yet, COVID-19 is neither the first pandemic nor the last one. In the history, there have been several pandemics such as Black Death, Spanish flu, Smallpox, HIV etc. [2]. However, in the 21st century, living in a global world have caused COVID-19 to spread in a short time to all over the world due to heavy air traffic routes, international trade actions and tourism habits. Under these circumstances, there have been several essential issues to manage by public health concern such as identifying people at risk, controlling the borders, monitoring tourism, surveillance of active cases [3]. There has been done tremendous research by investigators from various disciplines under the same goal of understanding the pharmaceutical and non-pharmaceutical dimensions of COVID-19. Especially in the first phases of pandemic, exploring COVID-19 characteristics was very crucial to identify the virus and who were at risks. Fu and his colleagues (2020) examined 43 different studies (with 3600 patients) that focused on clinical characteristics of COVID-19, and found that the most common symptoms are fever, cough and fatigue among patients [4]. Furthermore, elder people with comorbidities are associated with the highest risk [5]. Moreover, according to John Hopkins University, mortality rate of COVID-19 for the world is approximately 3.97% and differs between 0.1% and 28.5% among countries by 27 July 2020 [6]. On the other hand, unfortunately there is still not a cure for this virus, medications applied to COVID-19 patients at hospitals are not proved. Although vaccine studies have been undergoing in different countries, yet there is no publicly announced ready vaccination. Under these circumstances, being infected is very vital, if public health strategies are not applied to the community, individuals remain very vulnerable. Therefore, which strategies are applied

Citation: Cetinguc, B.; Calik, E. COVID 19 Peak Time Prediction via a Gradient Boosting Method. *Proceedings* **2021**, *74*, 8. <https://doi.org/10.3390/proceedings2021074008>

Published: 4 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

when makes difference in spread pattern [7]. Moreover, strategies should help to flatten and decrease the COVID-19 cases and mortality curves [8]. Governments take precautions by implementing policies such as external border restrictions, closure of schools, public awareness campaigns, lockdowns, health monitoring and testing etc. [9]. To our knowledge, Government Response Index [10] and Government Policy Activity Index [9] are two indices in the literature that investigating government policies regarding COVID-19 among countries. These indices are valuable research resources for policy makers and scientists to understand the dynamics and effects of implemented policies. In this study, we used some indicators from Government Response Index alongside with the attributes of human resources, economical background and cultural dynamics of each country. These variables were employed in prediction of COVID-19 peak time via examining the curves of COVID-19 cases from the first day of occurrence. The main aims of this study are to examine the importance of the policies and dynamics regarding COVID-19 and to predict peak time of the countries which are not reached their peak yet.

In the next section data and methodology will be introduced. Third chapter includes analysis and results of the study. Then, implications of the study will be mentioned in discussion section. Finally, in the last section limitations and further research will be given.

2. Data and Methodology

2.1. Data Sources

In this study, our data comprised of two main dimensions; the first one is constituted of restriction policies, the other one is characteristics of each country based on human capital, economics and habits. The restriction policies are selected based on containment and closures as school closing, workplace closing, cancel public events, restrictions on internal movements, international travel controls, restrictions on gatherings, stay at home requirements and closing public transportation, and derived from Government Response Index [10]. The remaining characteristics: hospital beds per thousand, GDP per capita, population density, handwashing facilities, life expectancy, aged 65 older and median age are obtained from Coronavirus Source Data [11]. Totally 173 countries are included in the analysis and distributed in terms of region and income groups as seen in Table 1.

Table 1. Country Distribution.

Regions	High Income	Lower Middle Income	Upper Middle Income	Low Income	Grand Total
East Asia & Pacific	8	9	4		21
Europe & Central Asia	32	4	11	1	48
Latin America & Caribbean	7	4	17	1	29
Middle East & North Africa	7	5	5	2	19
North America	3				3
South Asia		4	1	2	7
Sub-Saharan Africa	1	15	5	24	45
Grand Total	58	41	43	30	173

Peak times is the target variable which is evaluated by the authors based on daily COVID-19 case curves. Out of 173 countries, 83 country has been peaked while 53 country has not reached its peaked yet. The data of remained 37 countries has not enough to evaluate whether peaked or not. Descriptions and scales of variables are given in detail with their sources in Table A1.

2.2. Gradient Boosting Method

In this study, a gradient boosting regression tree method (GBM) was performed to estimate COVID-19 peak time and to determine which factors affect the peak time by giving their relative importance. This method provides more prediction accuracy and model interpretability while comparing to the single decision tree model. GBM is the generalization of tree reinforcement that tries to reduce accuracy and interpretability problems to provide a precise and operative way for data mining [12]. By emphasizing the training data which are difficult to predict, multiple models are developed sequentially, and the accuracy of the prediction is increased. In the training data during the boosting process, while using previous base models examples that are estimated difficultly are much more common than correctly estimated ones. Mistakes made by previous base models are handled to corrected by each additional base model [13].

In GBM, the model is constituted gradually and updated by minimizing the expected value of loss function via the number of iterations (i.e., the number of trees). Fitted model may accomplish illogically small training error by adding many trees to the model. However, this can cause an overfitting problem because of depending on the training data and a lack of generalizability. The number of gradient boosting iterations can be controlled to prevent overfitting [13]. Alongside with the number of iterations, the other parameters as learning rate and tree complexity that directly affect the performance of the algorithm should be handled.

3. Analysis and Results

Firstly, after data cleaning process, one country is eliminated from data set due to missing data of some predictor variables. Considering the beforementioned explanations, after trying the combinations of the parameters (tree complexity, learning rate and number of iterations), we decided to employ 400 trees with 0.01 learning rate to avoid overfitting. Furthermore, a leave-one-out strategy was performed to validate the GBT models because of limited sample size ($n = 82$ countries). In each iteration, the algorithm sequentially operated 81 data points for training and remaining datum point for testing. Rapid Miner academic version is selected to conduct analysis. According to the results, some variables have strong effects while others have no importance on prediction of COVID-19 peak time. GBT algorithm provides relative score of predictor variables to indicate the importance in building of the trees and ignores some variables completely [12]. In our model, school closing, workplace closing, cancel public events, and restrictions on internal movement variables are not importance in prediction of COVID-19 peak time. The remaining variables, hospital beds per thousand, close public transport, GDP per capita, population density, handwashing facilities, life expectancy, aged 65 older, median age, international travel controls, restrictions on gatherings, income group, stay at home requirements are ordered decently in terms of score can be seen in Figure 1. At the same time our determination coefficient is 0.832 that means our model explains 83.2% of variance in prediction of COVID-19 peak time. Moreover, root mean squared error is 23.901 in the model.

Non-peaked countries peak times were predicted via the algorithm and predicted times are given in Figure A1 in Appendix A. According to our results, the highest time to peak days are estimated in Nigeria, Indonesia, India and Philippines. On the other hand, Uruguay, Bosna Hercegovina and Poland are expected to reach peak day in a short time relatively to other non-peaked countries.

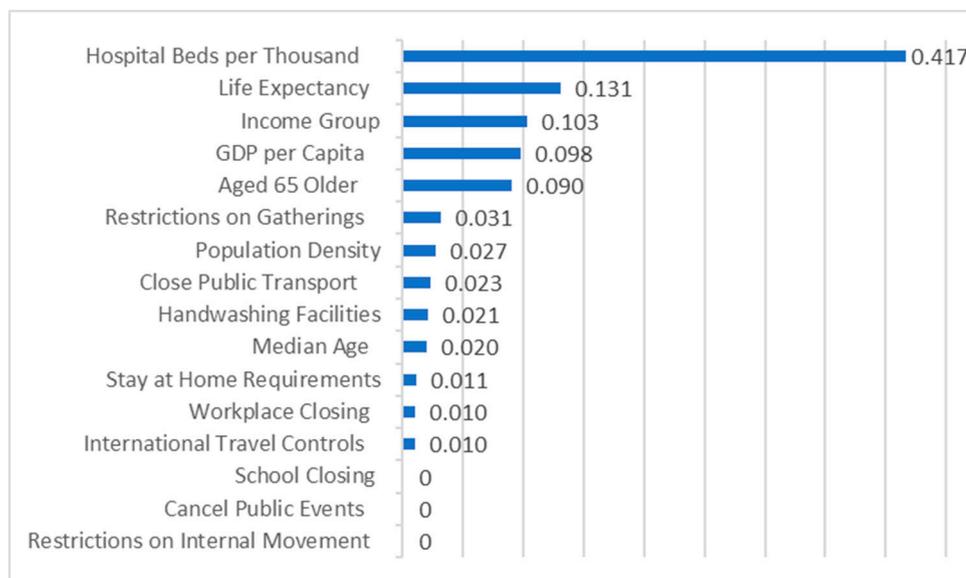


Figure 1. The relative importance of variables.

4. Discussion

Among 82 countries, relative importance of government policies and characteristics have been examined regarding COVID-19 peak times. Most of the characteristics of countries have been more effective than government policies to constitute peak times. The highest relative importance was found as hospital beds for thousand that means the main determinator to predict the countries peak time duration. If there are not enough hospital beds for patients, they will not be the place where they are supposed to be and continue to spread the virus. Respectively, life expectancy, income group, GDP per capita and percentage of aged 65 older have similar scores while population density and median age have relatively lower importance on to identify peak times of COVID-19. Moreover, contrary to our expectations the only indicator of cultural dynamics of countries having handwashing facilities has mild importance to estimate peak times. Although one of the contagion ways of COVID-19 is hand contact, finding mild importance of hand washing facilities show that transmission rate by hand contact might not be higher as scientists clarified. Among restrictions, parallel to our expectations the highest score is restrictions on gatherings and followed by closing of public transport because the areas, where there is high human density, are effective to increase infection rate due to high circulation and close contact of people. Furthermore, staying at home requirements, closing of workplaces and control of international travels have also slightest importance on estimation of peak times. On the other hand, based on our analysis, the importance of closing schools, cancelling public events and restrictions on internal movements indicators are found as zero in our model. When investigating the data closely, most of the countries take similar strict actions based on these topics. Furthermore, these results can be used by policy makers and government administrators.

5. Limitations and Further Research

This study make contribution to the evaluations of COVID-19 precautions; however, it has several limitations. Firstly, COVID-19 literature has been changing every day, even in hours. All humanity tries to contribute to prevent and remediation of COVID-19. Therefore, updated data should use to conduct analysis while using the proposed models. Secondly, by the nature of this study, number of countries are limited and while employing machine learning techniques the limited number of observations narrows method choices

and the prediction accuracy. Moreover, especially restriction data is scaled by perceptually might affect the performance of the model, hence different indices might be employed for further research.

Author Contributions: Introduction and data, B.C.; methodology, analysis and results E.C.; discussion, writing, review and editing, B.C and E.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Variable Descriptions and Sources.

Indicators	Scale	How to Measure/Description	Source
Peak Times	Numerical	Duration until the peak time	Authors' examination
Income Group	Polynomial	1—High income 2—Upper middle income 3—Lower middle income 4—Low income	World Bank https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed on 27 July 2020)
School Closing	Polynomial	0—No measures 1—recommend closing 2—Require closing 3—Require closing all levels	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
Workplace closing	Polynomial	0—No measures 1—recommend closing 2—require closing for some sectors or categories of workers 3—require closing all-but-essential workplaces (e.g., grocery stores, doctors)	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
Cancel public events	Polynomial	0- No measures 1—Recommend cancelling 2—Require cancelling	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
Restrictions on gatherings	Polynomial	0—No restrictions 1—Restrictions on very large gatherings (the limit is above 1000 people) 2—Restrictions on gatherings between 101–1000 people 3—Restrictions on gatherings between 11–100 people 4—Restrictions on gatherings of 10 people or less	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
Close public transport	Polynomial	0—No measures 1—Recommend closing 2—Require closing	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)

Stay at home requirements	Polynomial	0—No measures 1—recommend not leaving house 2—require not leaving house with exceptions for daily exercise, grocery shopping, and essential trips 3—Require not leaving house with minimal exceptions	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
Restrictions on internal movement	Polynomial	0—No measures 1—Recommend not to travel between regions/cities 2—internal movement restrictions in place	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
International travel controls	Polynomial	0—No measures 1—Screening 2—Quarantine arrivals from high-risk regions 3—Ban on arrivals from some regions 4—Ban on all regions or total border closure	https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker (accessed on 27 July 2020)
Handwashing facilities	Numerical	Share of the population with basic handwashing facilities on premises, most recent year available	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)
Hospital beds per thousand	Numerical	Hospital beds per 1000 people, most recent year available since 2010	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)
Life expectancy	Numerical	Life expectancy at birth in 2019	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)
Population density	Numerical	Number of people divided by land area, measured in square kilometers, most recent year available	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)
Median age	Numerical	Median age of the population, UN projection for 2020	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)
Aged 65 older	Numerical	Share of the population that is 65 years and older, most recent year available	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)
Gdp per capita	Numerical	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available	https://ourworldindata.org/coronavirus-source-data (accessed on 27 July 2020)

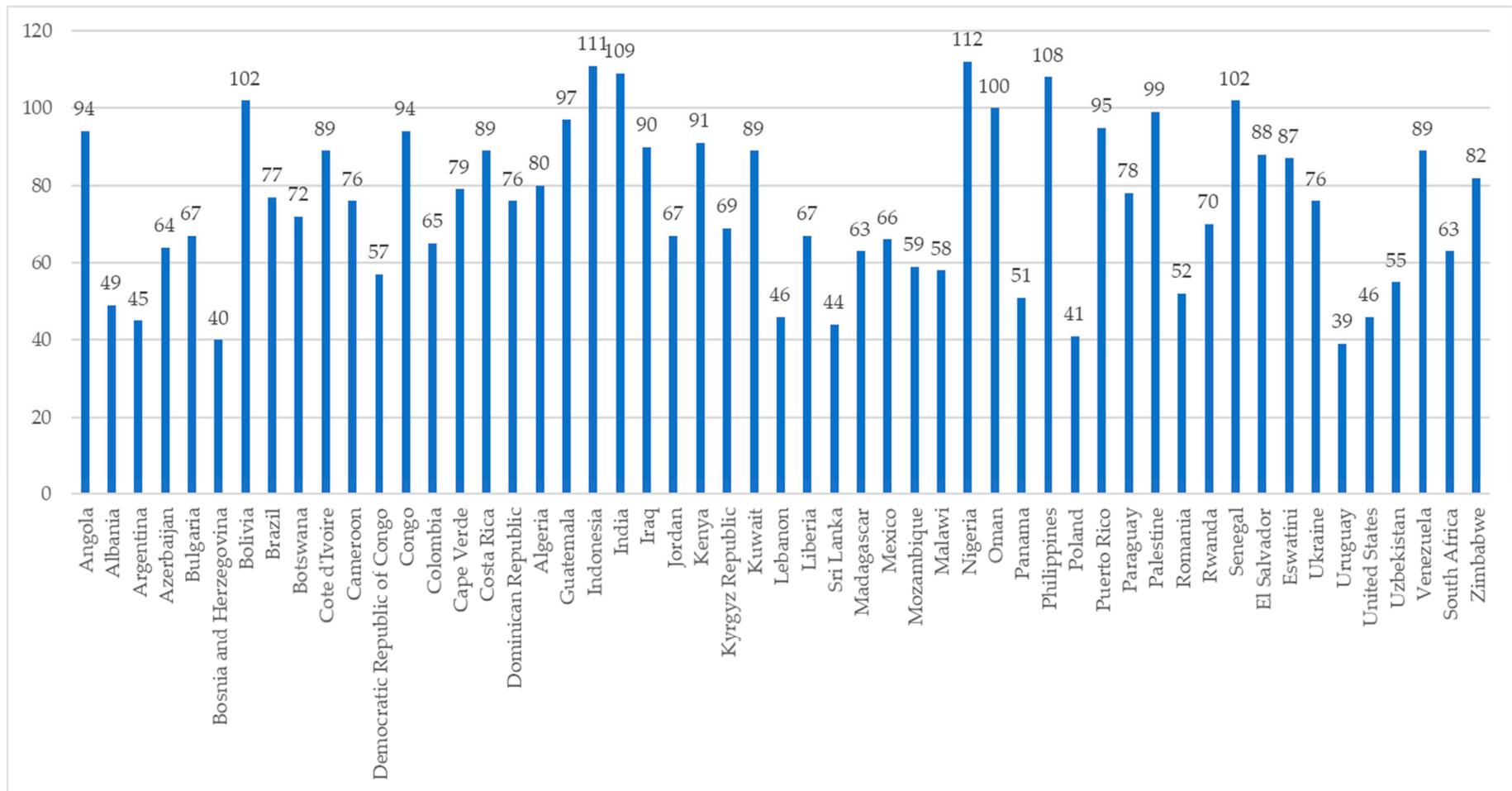


Figure A1. Predicted Peak Times of Non-Peaked Countries.

References

1. Li, H.; Liu, S.M.; Yu, X.H.; Tang, S.L.; Tang, C.K. Coronavirus disease 2019 (COVID-19): Current status and future perspectives. *Int. J. Antimicrob. Agents* **2020**, *55*, 105951.
2. Turner, J.A. Pandemics and Epidemics through History: This Too Shall Pass Pandemics and Epidemics through History: This Too Shall. *J. Hosp. Librariansh.* **2020**, *20*, 280–287.
3. Azarpazhooh, M.R.; Morovatdar, N.; Avan, A.; Phan, T.G.; Divani, A.A.; Yassi, N.; Stranges, S.; Silver, B.; Biller, J.; Belasi, M.T.; et al. COVID-19 pandemic and burden of non-communicable diseases: An ecological study on data of 185 countries. *J. Stroke Cerebrovasc. Dis.* **2020**, *29*, 105089.
4. Fu, L.; Wang, B.; Yuan, T.; Chen, X.; Ao, Y.; Fitzpatrick, T.; Li, P.; Zhou, Y.; Lin, Y.F.; Duan, Q.; et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. *J. Infect.* **2020**, *80*, 656–665.
5. Cai, Q.; Huang, D.; Ou, P.; Yu, H.; Zhu, Z.; Xia, Z.; Su, Y.; Ma, Z.; Zhang, Y.; Li, Z.; et al. COVID-19 in a designated infectious diseases hospital outside Hubei Province, China. *Allergy Eur. J. Allergy Clin. Immunol.* **2020**, *75*, 1742–1752.
6. John Hopkins University Coronavirus Research Center. Covid-19 Dashboard by the Center for Systems Science and Engineering at John Hopkins University. 2020. Available online: <https://coronavirus.jhu.edu/map.html> (accessed on 27 July 2020).
7. Iyanda, A.E.; Adeleke, R.; Lu, Y.; Adaralegbe, A.; Lasode, M.; Ngozi, J. A retrospective cross-national examination of COVID-19 outbreak in 175 countries: A multiscale geographically weighted regression analysis (January 11–June 28, 2020). *J. Infect. Public Health* **2020**, *13*, 1438–1445.
8. Hennekens, C.H.; George, S.; Adirim, T.A. The Emerging Pandemic of Coronavirus: The Urgent Need for Public Health Leadership. *Am. J. Med.* **2020**, *133*, 648–650.
9. Cheng, C.; Barceló, J.; Hartnett, A.S.; Kubinec, R.; Messerschmidt, L. COVID-19 Government Response Event Dataset (CoronaNet v.1.0). *Nat. Hum. Behav.* **2020**, *4*, 756–768.
10. Hale, T.; Hale, A.J.; Kira, B.; Petherick, A.; Phillips, T.; Sridhar, D.; Thompson, R.; Webster, S.; Angrist, N. Global Assessment of the Relationship between Government Response Measures and COVID-19 Deaths. *medRxiv* **2020**, 1–23. doi:10.1101/2020.07.04.20145334.
11. Ritchie, H. Coronavirus Source Data. 2020. Available online: <https://ourworldindata.org/coronavirus-source-data> (accessed on 27 July 2020).
12. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
13. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 308–324.