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

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

Joseph Galasso\(^a\(^b\), Duy M. Cao\(^b\), Robert Hochberg\(^c\)

\(^a\) Department of Biology \#11, University of Dallas, Irving, TX 75062, USA
\(^b\) Department of Computer Science \#134, University of Dallas, Irving, TX 75062, USA
\(^c\) Department of Computer Science \#50, University of Dallas, Dallas, TX 75062, USA

**Abstract**

During the COVID-19 pandemic, predicting case spikes at the local level is important for a precise, targeted public health response and is generally done with compartmental models. The performance of compartmental models is highly dependent on the accuracy of their assumptions about disease dynamics within a population; thus, such models are susceptible to human error, unexpected events, or unknown characteristics of a novel infectious agent like COVID-19. We present a relatively non-parametric random forest model that forecasts the number of COVID-19 cases at the U.S. county level. Its most prioritized training features are derived from easily accessible, standard epidemiological data (i.e., regional test positivity rate) and the effective reproduction number \((R_s)\) from compartmental models. A novel input training feature is case projections generated by aligning estimated effective reproduction number (pre-computed by COVIDActNow.org) with real-time testing data until maximally correlated, helping our model fit better to the epidemic’s trajectory as ascertained by traditional models. Poor reliability of \(R_s\) is partially mitigated with dynamic population mobility and prevalence and mortality of non-COVID-19 diseases to gauge population disease susceptibility. The model was used to generate forecasts for 1, 2, 3, and 4 weeks into the future for each reference week within 11/01/2020 - 01/10/2021 for 3068 counties. Over this time period, it maintained a mean absolute error (MAE) of less than 300 weekly cases/100,000 and consistently outperformed or performed comparably with gold-standard compartmental models. Furthermore, it holds great potential in ensemble modeling due to its potential for a more expansive training feature set while maintaining good performance and limited resource utilization.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

The COVID-19 epidemic in the United States proved devastating economically as it is projected to cause 3.2 to 4.8 trillion USD in net U.S. GDP loss [1]. The epidemic has had a devastating toll on life, particularly among the elderly and members of ethnic minorities such as African Americans [2]. Throughout the COVID-19 pandemic, it has been necessary to forecast the progression of the epidemic at the U.S. county/county-equivalent (CCE) level so that local authorities can effectively implement public health measures such as social distancing or quarantines [3]. This need is particularly great due to findings that there is low reliability between state-wide and county-specific reported data, requiring that the pandemic is tracked at the most granular level possible [4].

Forecasts have been traditionally done with compartmental models such as Susceptible-Exposed-Infected-Recovered (SEIR) [5,6]. These models segment the population into various compartments for stages in progression of the disease of interest. Transitions between these compartments represent epidemic dynamics. These models can be used to solve for the time-path of the \(R_s\), which is the estimated ratio of new infections caused by each currently infected individual, over the course of an epidemic; this in turn can be used to forecast future epidemic progression [6]. However, these models involve numerous assumptions in their design about disease spread dynamics and their interpretability and usability is limited by the quality of these assumptions [5–7].

Thus, relatively non-parametric deep or machine-learning models, such as Random Forests (RFs), are attractive alternatives to compartmental models, as they avoid assumptions about the dis-
tation of input data and generate forecasts based on observed empirical trends in this data [8,9]. In addition to being non-parametric, RF-regressors are highly effective at extracting non-linear relationships from input data while being time efficient [10,11]. First described by Breiman, RF-regressors are ensemble models of regression trees each trained on different subsets of input data, reducing the variance of predictors and minimizing overfitting [12]. In addition, RFs enable relatively easy estimations of variable importance [13], which can serve as an assessment of model performance. RFs have been successfully utilized in prior studies for predicting diarrheal infections, Dengue, H5N1 influenza, and West Nile virus [10,14–17]. They also have been applied for short-term forecasting of COVID-19 infections by Ribeiro et al. [11].

We similarly apply an RF to project COVID-19 infections at the more granular CCE level up to 4 weeks in the future. Unlike prior RF epidemiology forecasting models, we use a unique and extensive combination of features including population health and mobility, demographic, and COVID-19 testing data. We develop a novel method of computing case and Rt forecasts with an algorithm that is based on our observation that COVID-19 cases and SEIR-generated Rt have a similar trajectory over the course of an epidemic, but Rt lags behind cases, making it predictive of cases in the lag period. These forecasts are then added as a training feature for our RF model, which produces the final forecasts.

### 2. Methods

#### 2.1. Data acquisition

Our time-series datasets begin on 03/31/2020 for 3068 U.S. CCEs with complete, consistent data that could be processed without error in our downstream pipelines. To normalize testing and case counts by population, these metrics were converted to incidence rate (IR) using the 2018 U.S. Census CCE-level population estimates in Eq. (1).

\[
IR = \frac{\text{Number of incident cases, tests, etc. during time period}}{\text{County population, state population}} \times 100,000
\]

(1)

#### 2.2. Generation of Rt and case alignment forecast features

For any given date in every U.S. CCE, we forecast Rt and testing-normalized cases for 1, 2, 3, and 4 weeks into the future. First, a testing-normalized COVID-19 cases time-series was generated by dividing the “Weekly case increase” feature by the “Daily tests increase” feature (see Table 1). The normalized cases time-series and the “Daily estimated Rt” feature were used to generate Rt and case-prediction features, as shown in Fig. 1. All linear regression models were implemented with the `linear_model.LinearRegression` module [28].

#### 2.3. RF training and forecasting

The dataset used to train and validate the RF forecasting models includes all features in Table 1 for each U.S. CCE; however, “Weekly case increase” and “Daily estimated Rt” features were replaced with the Fig. 1F features for 7-, 14-, 21-, and 28-day forecasts. Thus, the preliminary forecasts were used as features to generate the final forecasts. All features were normalized by removing their mean and scaling to unit variance. For each Sunday from 11/01/2020 to 01/10/2021, we trained and validated RF forecasting models via incremental learning. Thus, we filtered out feature data that occurs on or after the Sunday of interest, but later Sundays will have more feature data than earlier Sundays. Then, we randomly split the feature data for a given Sunday into a training subset and a validation subset with a 9:1 ratio, respectively.
The training subset is used to train a RF regression model for each forecast timepoint (7, 14, 21, or 28 days into the future); the target outcome was the “Weekly case increase” from Table 1 shifted backwards to represent future cases on the forecast timepoint. The validation subset is used to ensure that trained RF models do not overfit the training subset. Finally, data for the Sunday of interest is input to the trained RF models to generate forecasts for each forecast timepoint.

RF regression models are ensemble machine learning algorithms that were first described by Breiman [12]. They create multiple regression trees trained on unique bootstrap samples of the training dataset with a random subset of the input features. The output of all trees is averaged to create the final projection. We used the Scikit-learn (version 0.23.2) implementation with default hyperparameters and 20 estimators [28].
2.4. Model evaluation

The permutation importance of all features input into the RF models was calculated as described by Breiman as the decrease in mean squared error of the model when a feature of interest is randomly shuffled [12]. We used two metrics, MAE and R-squared ($R^2$), to evaluate the accuracy of our model’s forecasts relative to actual case counts for both the training dataset and forecasts outside of the training dataset. These metrics are calculated as follows:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$  \hspace{1cm} (2)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$  \hspace{1cm} (3)

We also used these metrics to calculate the error for equivalent forecasts by the JHU Infectious Disease Dynamics group (IDD), the JHU Applied Physics Lab (APL), and One Quiet Night (OQN) forecasting models [29–31]. These were selected as they are parts of the U.S. Centers for Disease Control and Prevention (U.S. CDC) ensemble model, have relatively high performance, and forecast for a large breadth (2349) of U.S. counties along with the RF model [32].

3. Results and discussion

3.1. Analysis of $R_t$ and case alignment forecasts

When the case time-series for a CCE was normalized by population and state-level testing data, it often showed a very similar shape to the CCE and/or state $R_t$ time-series, as shown in Fig. 1. This should be expected, as both time-series indicate new infection load over the course of an epidemic, although $R_t$ lags behind cases, which we suspect is because $R_t$ represents infection load in the present moment, but these infections will not be detected via testing until much later due to viral incubation periods of 2–14 days and test result reporting delays [33]. In the 01/10/2021 dataset, the average optimal shift for the most highly correlated $R_t$ time-series, whether state or CCE, was 34 days forward.
However, there were many CCEs where this relationship was weak. We observed that 1782 CCEs selected, optimally shifted $R_t$ time-series have a Pearson correlation with the case time-series $< 0.5$. They also have a mean population density of 89 people/mi$^2$, vs. the 273 people/mi$^2$ mean over the entire set of 3068 CCEs. Thus, we attribute the low correlations to poor and/or inconsistent testing efforts and data quality in rural CCEs. Further supporting this is our observation that in CCEs where we select their state’s $R_t$ time-series as opposed to their own, the mean population density is also relatively low at 210 people/mi$^2$.

In concurrence with Omori et al., we found that normalization of the case time-series with testing data is critical to expose underlying changes in COVID-19 progression, as seen in Fig. 2 [34]. However, our approach is limited by use of state-level testing data as opposed to CCE-level testing data, which was inaccessible to us. However, state-level testing data has the advantage of including individuals who are not tested in their CCE of residence due to inequity in regional testing access.

### 3.2. RF feature prioritization and training performance

On average over the 11 epi weeks, RFs always prioritized population-normalized, state-level test numbers, test positivity, and case and $R_t$ alignment forecasts as seen in Fig. 3. This validates the model, as these features most directly measure COVID-19 progression. Importantly, of all 8 alignment forecasts, those for the desired forecast target were always prioritized highest. Conversely, demographic, population health, and population mobility features are relatively inconsequential; on average across Fig. 3A-D, these 46 features share of the total sum of all 56 features permutation scores is just 14.33%.

On the training datasets for all 4 forecast targets and all 11 epi weeks, $R^2$ was very stable, averaging 0.97 with a standard deviation of 0.00. For the corresponding validation datasets, $R^2$ fell to 0.92 with a standard deviation of 0.02. Thus, overfitting is not extreme and, considering the RFs relatively high $R^2$ on actual case data in comparison to other models as seen in Figs. 4 and 5, is not detrimental to our objectives.

### 3.3. RF forecasting performance vs. other models

As may be observed in Figs. 4 and 5, over the 11 reference weeks for which forecasts were generated by the RF models, MAE and $R^2$ remain relatively consistent compared to the OQN and JHU models. The RF’s $R^2$ is competitive with and often higher than the $R^2$ for other models, notably November 2020. Cases in November were harder to model accurately, as November saw a 40% increase in cases in the 3068 counties in its fourth week relative to its first week, whereas December only saw an 8% decrease in cases for its equivalent comparison period. Periods of relatively low performance are generally shared by all models, indicating that case load changes during these weeks are simply less predictable.

For each epi week, case projections from each model were collected and compared to actual cases that occurred during these weeks as reported by JHU [18]. The results were calculated and visualized in Figs. 4 and 5.

The JHU IDD and APL models are SEIR models [29,30], whereas the OQN model applies a linear regression model to each CCE [31].
The distribution of output from a model exclusive to a CCE will be skewed towards the distribution of its training dataset labels, which may be a factor explaining OQN’s low MAE. On the other hand, the RF model forecasts for all counties, affording it a larger dataset which possibly contributes to its relatively high, stable fit to the actual cases as indicated by its \( R^2 \).

4. Conclusion

We present a unique method to project COVID-19 cases for CCEs by using their or their state’s \( R_t \) time-series to predict cases, taking advantage of the backward lag of regional \( R_t \) time-series from the case time-series despite their similar trajectory. These predictions are input into a RF regression models with regional testing, demographic, population mobility and population health data for final case forecasts. Our approach is computationally inexpensive while remaining very effective, as our model achieves consistently high \( R^2 \) and low MAE relative to gold standard models used in the U.S. CDC ensemble model during a highly dynamic case spike period in November 2020 early January 2021. This model is limited to forecasting cases detected by testing as opposed to latent, asymptomatic cases, which can be estimated by compartmental models [35]. Thus, its best use scenario is for public health officials to identify potential outbreaks in their community to help them optimize their response. It shows evidence of good consistency in its current iteration but has some potential for improvement via addition of new features to its training dataset, particularly case and \( R_t \) forecasts from other compartmental models. Such ensemble forecasting approaches have improved model performance significantly [36].

**Fig. 4.** Performance evaluation using MAE. The errors between projections and real number of incident cases were calculated using Eq. (3). The y-axes of the graphs have been limited so that all models can be visually compared.
Fig. 5. Performance evaluation using $R^2$. As with Fig. 4, the y-axes of the graphs have been limited. The proportions of variance between projections and observed values were evaluated using Eq. (2). We notice that there are large anomalies in the weekly $R^2$ of the Google_Harvard and JHUAPL_Bucky models after epi week 202051; however, for the sake of complete comparison, all weeks for all models are shown.

**Availability of software and other materials**

All code, programs, and data supporting the conclusions made in this article are publicly available on GitHub, located at https://github.com/solveforj/pandemic-central.

**Declaration of Competing Interest**

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

**CRediT authorship contribution statement**

**Joseph Galasso:** Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Duy M. Cao:** Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Robert Hochberg:** Writing – review & editing, Supervision.

**Acknowledgments**

We would like to thank Dr. Richard P. Olenick (University of Dallas) for providing us with insightful criticism of the manuscript. We also wish to thank all organizations that curate our input datasets and make them freely available for researchers.

**References**

[1] Walmsley T, Rose A, Wei D. The impacts of the coronavirus on the economy of the United States. SSRN Electron J 2020. doi:10.2139/ssrn.3678835.
[2]ancy CW. COVID-19 and African Americans. JAMA 2020;323(19):1891. doi:10.1001/jama.2020.6548.

[3] Bertozzo AT, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proc Natl Acad Sci 2020;117(29):16732–8. doi:10.1073/pnas.2005625117.

[4] Messner W, Payson S. Variation in COVID-19 outbreaks at the US state and county levels. Public Health 2020;187:15–18. doi:10.1016/j.puhe.2020.07.035.

[5] Paiva HM, Afonso RJM, de Oliveira IL, Garcia GF. A data-driven model to describe and forecast the dynamics of COVID-19 transmission. PLoS One 2020;15(7):e0236386. doi:10.1371/journal.pone.0236386.

[6] Buckman SR, Glick R, Lanning KJ, Petrosky-Nadeau N, Seitelman LM. Replicating and projecting the path of COVID-19 with a model-impacted reproduction number. Infect Dis Model 2020;5:653–51. doi:10.1016/j.idm.2020.08.007.

[7] Ozmen O, Nutaro J, Pullum LL, Ramanathan A. Analyzing the impact of modeling choices and assumptions in compartmental epidemiological models. Simulation 2016;92(5):459–72. doi:10.1177/0037549716648077.

[8] Zeroual A, Harrou F, Dairi A, Sun Y. Deep learning methods for forecasting COVID-19 time-series data: a comparative study. Chaos, Solitons Fractals 2020;140:110121. doi:10.1016/j.chaos.2020.110121.

[9] Shang AC, Calow KE, Calow GC. Regional forecasting of COVID-19 caseload by non-parametric regression: a VAR epidemiological model. AIMS Public Health 2021;8(1):124–36. doi:10.3934/publichealth.202010.

[10] Fang X, Liu W, Ai J, He M, Wu Y, Shi Y, et al. Forecasting incidence of infectious diarrhea using random forest in Jiangsu province, China. BMC Infect Dis 2020;20(1). doi:10.1186/s12879-020-4930-2.

[11] Ribeiro MHDM, da Silva RG, Mariani VC, dos Santos Coelho L. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos, Solitons Fractals 2020;135:109853. doi:10.1016/j.chaos.2020.109853.

[12] Breiman L. Random forests. Mach Learn 2001;45(1):532. doi:10.1023/A:1010933404324.

[13] Chen Y, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99(6):323–9. doi:10.1016/j.ygeno.2012.04.003.

[14] Machado G, Mendoza MR, Corbellini LG. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Vet Res 2015;46(1). doi:10.1186/s13567-015-0219-7.

[15] Mussumeci E, Coelho FC. Large-scale multivariate forecasting models for dengue - LSTM versus random forest regression. Spat Spatio-Temporal Epidemiol 2020;35:100372. doi:10.1016/j.sste.2020.100372.

[16] Kane MJ, Price N, Scotch M, Rahinowitz P. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinform 2014;15(1). doi:10.1186/1471-2105-15-276.

[17] Keyel AC, Timm OE, Backenson PB, Prussing C, Quinones S, McDonough KA, et al. Seasonal temperatures and hydrological conditions improve the prediction of west Nile virus infection rates in culex mosquitoes and human case counts in new york and connecticut. PLoS One 2019;14(6):e0217854. doi:10.1371/journal.pone.0217854.

[18] COVID-19 map. Available: https://coronavirus.jhu.edu/map.html

[19] Movement range maps. Available: https://data.humdata.org/dataset/movement-range-maps

[20] The COVID tracking project. Available: https://covidtracking.com

[21] COVID Act Now. Available: https://covidactnow.org/data-api;

[22] United states infectious disease mortality rates by county 1980–2014. Available: http://ghdx.healthdata.org/record/usme-data/united-states-infectious-disease-mortality-rates-county-1980-2014;

[23] United states chronic respiratory disease mortality rates by county 1980–2014. Available: http://ghdx.healthdata.org/record/usme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014;

[24] United states life expectancy and age-specific mortality risk by county 1980–2014. Available: http://ghdx.healthdata.org/record/usme-data/united-states-life-expectancy-and-age-specific-mortality-risk-county-1980-2014;

[25] United states diabetes prevalence by county 1999–2012. Available: http://ghdx.healthdata.org/record/usme-data/united-states-diabetes-prevalence-county-1999-2012;

[26] United states physical activity and obesity prevalence by county 2001–2011. Available: https://doi.org/10.6069/j834-HD26;

[27] County population by characteristics. Available: https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html;

[28] Pedrosga F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Biondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Courneau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825–30. doi:10.1186/s13567-015-109853.

[29] Lemaître JC, Grantz KH, Kaminsky J, Meredith HR, Truelove SA, Lauer SA, et al. A scenario modeling pipeline for COVID-19 emergency planning. Sci Rep 2021;11(1):7534. doi:10.1038/s41598-021-86811-6.

[30] The Johns Hopkins University Applied Physics Laboratory. Bucky's documentation. 2020; Unpublished; https://docs.buckymodel.com/en/latest/index.html.

[31] Jo A, Cho J. OneQuietNight COVID-19 forecast; 2021. Unpublished manuscript; https://one-quiet-night.github.io/vs/static/media/OGN-0316e207.pdf.

[32] Ray E.L., Wattanachit N., Niemi J., Kanji A.H., House K.,ramer E.Y., Bracher J., Zheng A., Yamana T.K., Xiong X., Woody S., Wang Y., Wang L., Walraven R.L., Tomar V., Sherratt K., Sheldon D., Reiner R.C., Prakash B.A., Osthus D., Li M.L., Lee E.C., Koyuoluq U., Keskinocak P., Gu Y., Gu Q., George G.E., España G., Corsetti S., Chhatwal J., Cavany S., Biegel H., Ben-Nun M., Walker J., Slayton R., Lopez V., Biggerstaff M., Johansson M.A., Reich N.C., Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S.; 2020. Unpublished manuscript; https://www.medrxiv.org/content/early/2020/08/22/2020.08.19.20177493.

[33] Clinical questions about COVID-19: Questions and answers. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html;

[34] Omori R, Mizumoto K, Chowell G. Changes in testing rates could mask the novel coronavirus disease (COVID-19) growth rate. Int J Infect Dis 2020;94:116–18. doi:10.1016/j.ijid.2020.04.021.

[35] R’dulescu A, Williams C, Cavanagh K. Management strategies in a SEIR-type model of COVID-19 community spread. Sci Rep 2020;10(1). doi:10.1038/s41598-020-77628-4.

[36] Dean NE, y Piontii AP, Madewell ZZ, Cummings DA, Hitchings MD, Joshi K, Kahn R, Vegusagni A, Halloran ME, Longini IM. Ensemble forecast modeling for the design of COVID-19 vaccine efficacy trials. Vaccine 2020;38(46):7213–16. doi:10.1016/j.vaccine.2020.09.031;

[37] The U.S. COVID Community Vulnerability Index (CCVI). Available: https:// precisionforcovid.org/ccvi;