Forecasting COVID-19 Cases and Mortality using SIR Model with Time-Varying Parameters Forecasted via Ensemble Learning

CSE-8803-EPI Fall 2021 Project Final Report

Whitaker Chu*
Georgia Institute of Technology
Atlanta, Georgia, USA
mchu31@gatech.edu

Kai Ouyang*
Georgia Institute of Technology
Atlanta, Georgia, USA
ko40@gatech.edu

Nicholas Saney*
Georgia Institute of Technology
Atlanta, Georgia, USA
nsaney3@gatech.edu

KEYWORDS

COVID-19, SARS-CoV-2, Coronavirus, epidemiology, data science, SIR, compartmental, time-varying, regression, machine learning, ensemble learning

ACM Reference Format:
Whitaker Chu, Kai Ouyang, and Nicholas Saney. 2021. Forecasting COVID-19 Cases and Mortality using SIR Model with Time-Varying Parameters Forecasted via Ensemble Learning: CSE-8803-EPI Fall 2021 Project Final Report. In *Special Topics in Computational Science and Engineering: Data Science for Epidemiology, 2nd Edition.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*All authors contributed equally to this project, and are listed alphabetically by surname.

CSE-8803-EPI Fall 2021, December 2, 2021, Atlanta, GA

ACM ISBN 978-x-xxxx-xxxx-x/21/11.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The world is facing a pandemic. Coronavirus disease 2019 (COVID-19) is a rapidly spreading disease spreading across the globe and is a significant public health threat to humanity, infecting millions worldwide. To combat this, multiple models are used to predict future infection and death rates under various circumstances. These models are loosely divided into two classes: network-based models and deep-learning based models. Network-based models like SIS and SIR provide a classical solution that relies on constant variables to account for the spread of infection between connected portions of undirected graphs, providing an excellent baseline estimator, but an ultimately flawed ultimate predictor. Combating this are modern deep-learning based models that take in a wide breadth of features that are believed to correlate with new infections and deaths. These far-more-modern approaches are incredibly computationally expensive and face a all-too-familiar struggle: the spread of disease is incredibly difficult to learn to predict, regardless of hardware or methodology. To combat the potential flaws of these two methodologies, we propose a new system that uses modern machine learning to solve an easier problem: predicting the "constant" values associated with network-based models. Ideally, this would allow our system to maintain the classical approach, while challenging the common assertion that the associated values are truly constant as the infection spreads and evolves. Doing so would

be far less computationally expensive and theoretically exhaustive than a modern CNN or deep-learning solution, while being more accurate than using constant values in a network-based model to extrapolate values in the distant future.

RESPONSE TO MILESTONE COMMENTS

In response to our milestone findings, as well as input from Professor Prakash, we have decided to keep our desired output (predicted infection rates at time T), but greatly change the method in which our predictions are made. In keeping with our SIR-based title, we chose to use a new mixture of features to instead predict future beta and gamma values at a future time T1, and then use these predicted values as inputs into a graph-based network model to predict infections at a second time T2.

2 PROBLEM STATEMENT

Given the complexities associated with predicting deaths/infections using deep learning, and the general lack of precision associated with network-based models, we seek to combine both methodologies to form a compromise that will allow for faster, less-complex predictions by using ML to predict ODE inputs rather than the deaths/infections that are output.

3 RELATED WORK AND SURVEY

3.1 SIR Model Foundations

Compartmental models of epidemiology have been in use for about a century, building on the foundation laid out by Kermack and McKendrick [9]. These models break up a population into two or more sub-populations – compartments – and give rules for how the size of each compartment changes over time. The sizes of all compartments in such models always sum up to the size of the total population N.

In the common SIR compartmental model, the spread of a disease is described using three compartments: S for those who are "susceptible" to the disease, I for those who are currently "infectious" to others, and R for those who have been "removed" from these interactions, whether from recovery, death, or long-term isolation. The model poses that each compartment changes in size over time according to the following collection of differential equations, where β and γ (the model parameters) encode the rate of disease transmission and the rate of removal, respectively:

$$\begin{cases} S'(t) &= -\beta \cdot \frac{S(t)}{N} \cdot I(t) \\ I'(t) &= \beta \cdot \frac{S(t)}{N} \cdot I(t) - \gamma \cdot I(t) \\ R'(t) &= \gamma \cdot I(t) \end{cases}$$

Of particular importance is the ratio of the rates of transmission and removal, $\mathcal{R}_0 = \frac{\beta}{\gamma}$, called the basic reproductive number of a disease. An epidemic occurs when $\mathcal{R}_0 > 1$, or equivalently when $\beta > \gamma$, because if we begin with $\frac{S(t)}{N} \approx 1$, then $I'(t) \approx \beta \cdot I(t) - \gamma \cdot I(t)$ will put $I(t)$ in a positive feedback loop, only slowing when $S(t)$ is so depleted that the $\frac{S(t)}{N}$ term begins to bring $I'(t)$ back down. Consequently, a higher $\mathcal{R}_0$ indicates a steeper $I(t)$.

This formulation is undoubtedly familiar to those who have studied epidemiological models. It must be noted, however, that Kermack and McKendrick introduced the described model in their paper under the section "Special Cases", sub-section "B. *Constant Rates*", which considers a situation in which the rates of infectious spread and removal remain unchanged over the course of the modeled epidemic. Even with this clear simplification, the resulting system may be sufficient to describe the dynamics of how some diseases spread, particularly if the time span under consideration is relatively limited.

But in order to model the ongoing COVID-19 pandemic, with population-level countermeasures being updated in response to changing factors and healthcare systems experiencing fluctuations in capacity and treatment capability, the model parameters may be better described by functions whose values change over time.

And of course, each instance of this kind of model can only describe an individual region where the model parameters can reasonably represent the region's population as a whole – the world does not have uniform population density and the governments of the world certainly do not act in unison.

3.2 SIR Models with Time-Varying Transmission Rate

Even before this pandemic, researchers have investigated various SIR models that incorporate an transmission rate $\beta(t)$ that changes with time – while continuing to describe removal rate as a constant value γ.

One such model is explored in [4], which focuses on approximating the transmission rates of seasonal or otherwise re-emergent diseases using a sinusoidal function that is itself parameterized to account for amplitude, frequency, and phase. As their model factors in births as well as deaths, the susceptible population is able to increase when $\beta(t)$ is low enough. Their findings are interesting in their own right but their only consideration of non-sinusoidal transmission rates is restricted so as to discuss systems that stabilize in the long term into a cyclical trade-off between $S(t)$ and $I(t)$.

Regarding studies of COVID-19 behavior, several compartmental models have been researched that include time-varying transmission rate. These typically have a primary goal of analyzing the effects of large-scale countermeasures imposed by regional and national governments, showing the degree to which those interventions helped to slow the spread of the disease. As with most models, these are employed for predictions as well.

In [3], a collection of models are considered, but one specifically addresses strict countermeasures, such as lockdowns, that reduce social contact enough to make the transmission rate asymptotically approach 0 over time. That model is then applied to data from Senegal and France, showing an improvement in closeness of fit over simply using constant β.

In [12], a model with broader potential (which the authors call "varying-coefficient SIR" or "vSIR") is proposed that estimates a constant value for γ and a time series for $\beta(t)$ using daily infected and removed counts for $I(t)$ and $R(t)$, respectively. The model is then applied to data from mainland provinces in China as well as more granular data from 15 cities within the Chinese province of Hubei (where Wuhan is located), showing lockdown dates correlated with steep reductions in $\mathcal{R}_0$ and achieving decently accurate 3-day forecasts.

In [11], a similar model is described (which the authors call "adaptive SIR" or "aSIR") that uses a fixed $\gamma = 1/6$ estimates a time series for $\beta(t)$ using a sliding window over daily confirmed case counts as $R(t)$, working with the premise that quarantine and hospitalization remove confirmed cases from the general population. Here, the value for $I(t)$ is part of the estimation. The model is then applied to data from every state and county in the US, showing lockdown dates correlated with drops in $\mathcal{R}_0$ and achieving good 1-day forecasts.

3.3 SIR Models with Time-Varying Transmission and Removal Rates

Some SIR-based models take the next step and also consider not just a time-varying transmission rate, but a time-varying removal rate $\gamma(t)$ as well.

In [1], a model is developed using a piecewise function for $\beta(t)$ that breaks up the time span being studied into multiple phases, where each phase has its own constant value for β. This particular formulation splits the removal rate into a time-varying death rate $d(t)$, which is piecewise in the same way as the transmission rate, alongside a constant recovery rate r. The removal rate is implied to be a function $\gamma(t) = d(t) + r$. The model is applied to data from the US states of New York and New Jersey, providing insight into hospitalizations required during the first wave in those states.

In [10], a more detailed model (which the authors call simply a "time-varying beta and gamma SEIR model", or "TVBG-SEIR") is explored that describes $\beta(t)$ and $\gamma(t)$ with non-polynomial (exponential) splines across a time span that includes two known dates of population-wide countermeasure implementation. These two dates are considered interior nodes for a total of four nodes (start, first measure, second measure, end) through which the values of $\beta(t)$ and $\gamma(t)$ are to be interpolated. Several restrictive assumptions are made, such as $\beta(t)$ and $\gamma(t)$ needing to be respectively decreasing and increasing in a monotone fashion, along with both rates being held constant prior to the date of the first countermeasure. To describe the SEIR model's transition from the "exposed" population $E(t)$ to $I(t)$, the "latency rate" is considered to be a constant value $\sigma = 1/5.2$, based on prior research into the incubation period of COVID-19. The model is applied to data from Bulgaria and Germany with successful fits, though application to data from Italy required some adjustments.

Lastly, we mention [7], in which the time series for both $\beta(t)$ and $\gamma(t)$ are calculated using available daily infected and removed counts for $I(t)$ and $R(t)$, respectively. But it is not just past changes in transmission and removal rate parameters that are described – the analysis is taken a step further by modeling the changes in these parameters themselves as Finite Impulse Response (FIR) filters. This secondary model, which has its own parameters fit via ridge regression, is then used to predict future values of $\beta(t)$ and $\gamma(t)$, and those predicted values are fed back to the primary compartmental model in order to predict future values of $I(t)$ and $R(t)$. The paper also addresses other topics such as undetectable infectious individuals, herd immunity, and network modeling, applying the model to data from China as well as the US, the UK, France, Iran, and Spain in order to track and estimate $\mathcal{R}_0(t)$ in each. However, it is the concept of using a secondary model to predict future values of the variable parameters for the primary model that we build upon in our present work.

4 METHODOLOGY

Our overall goal can be summed up as follows: we aim to utilize one model to predict future values of a time series, where the parameterization of that primary model is itself a separate time series whose future values are predicted with a second model.

Specifically in the present work, we aim to utilize a discrete time-varying SIR model to predict future case and mortality counts, where the model's own parameters are time-variant $\beta(t)$ and $\gamma(t)$ and are predicted with a machine learning model trained not only on prior $\beta(t)$ and $\gamma(t)$ time-series data, but also on contemporaneous mobility information.

Whereas there is significant research devoted towards the prediction of case and mortality counts directly, we find that indirectly forecasting case and mortality counts, by way of predicting a time-varying SIR model's $\beta(t)$ and $\gamma(t)$, to be novel.

This section details the steps we took toward that goal.

4.1 Data Procurement

We acquired mobility data provided by Google [8] and Apple [2] together with case count and mortality data from the CDC [6] and Johns Hopkins [5] (JH). We used the parts of the data relevant to the US in the date range of January to September 2021, cutting out data for other countries and outside of that date range.

The mobility data from Google came as a single file with details for all US counties from 2021-01-01 to 2021-08-31. The original data had 706568 rows with 73 columns, and was 71MB in size.

The mobility data from Apple came as a single file with details for all US counties from 2020-01-01 to present. The original data had 4792 rows with 675 columns, and was 19MB in size.

The data from the CDC came as a single file that included confirmed case counts as well as total and probable counts at the state and territory level in the US from 2020-01-30 to 2021-08-30. The original data had 39061 rows with 15 columns, and was 3.2MB in size.

The data from JH came in 466 files that included daily confirmed case counts and recovered counts at the state and territory level in the US from 2020-01-01 to present. The original data had 27435 rows with 17 columns, and was 14MB in size.

For population sizes of the 50 individual states, we utilized data from the US Census Bureau [13] for 2019, because the release of data from the 2020 decennial census have been delayed until at least March 2022, partly due to the pandemic.

4.2 Data Preparation

After limiting our scope to US-only data, we split each of the described data sets into 50 parts, one for each US state. We focused exclusively on US states and did not include data for territories or special districts. We then assembled data points from each data set for each state on a per-date basis.

4.2.1 Selection of Mobility Time-Series Data. From the Apple mobility data set, we chose to use data points for 2 included features: activity levels of driving and walking. We did not use Apple's data for other transit usage, since that data was missing for several

states. From the Google mobility data set, we chose to use data points for 6 included features: activity levels at retail/recreation, grocery/pharmacy, parks, transit stations, workplaces, and residences. This gave us a total of 8 mobility features. For all mobility data points that we chose, the time span we selected data for was 2021-Jan-01 thru 2021-Aug-31.

4.2.2 Selection of Case and Mortality Time-Series Data. From the CDC and JH data sets, we chose to use data points for daily confirmed COVID-19 cases and mortality counts, selecting values in the time span from 2020-Dec-02 thru 2021-Aug-31. The inclusion of the last 30 days of 2020 for these data points is required for the way that we calculated time-series data for β and γ.

4.2.3 Calculation of Beta and Gamma Time-Series Data. Following the premise that β and γ vary with time, we calculated their time-series values using the prepared data for confirmed cases and mortality. This method calculates results for each date using a window of some number of preceding days. The largest preceding window that we used had a size of 30 days, which is why we included the last 30 days of 2020 when preparing the data points from the CDC and JH.

All of the β and γ time-series values were calculated according to the following algorithm:

(1) Select a US state s from the set of all 50 states
(2) Select a date t from the range 2021-Jan-01 thru 2021-Aug-31
(3) Select a data source d from the values $\{\text{CDC}, \text{JH}\}$
(4) Select a window size w from the values $\{7, 14, 30\}$
(5) Let N be the population for s in 2019
(6) Let $I_0 = c_0/N$
    - where c_0 is the confirmed case count that d reported in s on the wth day before t
(7) Let $R_0 = m_0/N$
    - where m_0 is the mortality count that d reported in s on the wth day before t
(8) Let $S_0 = 1.0 - I_0 - R_0$
(9) Calculate the ground-truth removed series $R_g(u) = \{m_u/N\}$
    - where the date u ranges through the w days prior to t
    - and m_u is the mortality count that d reported in s on u
(10) Find the best fit for the β and γ parameters on a time-invariant ODE SIR model initialized with (S_0, I_0, R_0) so as to minimize the mean squared difference between the ground-truth removed values $R_g(u)$ and the model's predicted removed values $R_p(u)$
(11) Save the fitted β and γ values in a data store, keyed under the combination (s, t, Z, w)
(12) Repeat steps 1 to 11 for every combination (s, t, Z, w) under consideration

The result is that for each of the 50 states, on each of the dates being analyzed, we have six different versions, from $\{\text{CDC}, \text{JH}\} \times \{7, 14, 30\}$, of time-series data for both β and γ for the time period of January 1 2021 to August 31 2021.

4.3 Forecasting Beta and Gamma via Ensemble Learning

Once we assembled both types of time-series data (mobility features and SIR parameter versions) for each state, we then worked on building a forecaster for $\beta(t)$ and $\gamma(t)$. For this task, we turned to machine learning techniques, specifically an ensemble learning model. This ensemble consisted of the following models:

- Auto-Regressive Integrated Moving Average (ARIMA)
- Ordinary Least-Squares (OLS)
- Decision Tree (DT)
- Random Forest (RF)
- Bagged DT
- Bagged RT

We chose these particular models because they are inherently different prediction algorithms, each with their own strengths and weaknesses. ARIMA is feature agnostic, only taking into consideration a series of data and forecasts temporally. ARIMA is good at predicting patternistic data. Given that we don't have seasonal time frame epidemiological patterns as one would expect with disease such as influenza, this model may fall short in producing accurate forecasts. We acknowledge that there are "waves" of infection and deaths, but since these "waves" will only marginally change beta and gamma, intuitively, it is possible that ARIMA produces forecasts which can be generally described as linear in nature.

OLS is a simple least square forecast - in essence, it is a simple regression model - there is no randomness, variability, variability. This model serves the purpose of being our discrete prediction model.

The basic decision tree model in its base form is also discrete and deterministic. The tree is built based on the 'gini' entropy (information gain) of each split, and as such, produces the same tree given the same data and hyper parameters each time. The problem with this model is that it tends to over fit. We could adjust the depth and number of leafs to try to mitigate this by limiting the number of leafs or the depth; but in our case, we choose a depth of 20 just for standardization. Further, we could tune a variety of parameters, but the forecast is in the base state computationally expensive (6+ hrs on i5 laptop 8gb memory, 1.5hrs+ on a AMD 5600 32gb memory), and as such, we choose not to tune hyper parameters.

We justified this decision because in addition to the base decision tree, we also had a bootstrapped forest of decision tree, which randomly selected a subset of data for forecasts, adding some degree of variability. In addition the random forest would reduce over fitting by randomly selecting some of the feature values which leaves are split on. Further, our random forest also has the bootstrap feature activated, so not only are we randomly selecting feature values for splitting, but we are also randomly selecting a subset of data to predict upon. We set n estimators for these bagged bootstrapped trees to 10 - as such, we are building 10 bootstrapped decision trees and 10 bootstrapped random forest trees and consolidating their votes.

We attempted to expedite this prediction process using Python's pooling multiprocessor functionality - initially with much fanfare and processing time reduction. However, we soon ran into problems with any model which required a random generator - spooling up additional cores produced exactly the same results. We discovered that because we couldn't not stagger the time in which each iteration is executed, and due to the initiation of set random number pools being created by by random states generated by the time a

core is initiated, parallel processing of any model requiring random selection such as random forest, or bootstrapping, used the same random seed and yielded the exact same results. There is a python feature which is supposed to force each iteration to pick a new random state, but after significant testing, we discovered that due to a known bug in version 3.8 of python, theoretically, each new random state is generated from the same random pool of each core's existing state, and as such, switched to exactly the same new random state. As such, we gave up on attempting to parallel process models.

As we pooled a variety of models, each with their own respective strengths and weaknesses, we then created an weighted ensemble forecast of the aforementioned 6 models based on the weighted root squared root mean squared deviation of each of the model - some better performing models would be weighted more. In theory, this should be a more viable model that would outperform most of the individual models, and has the potential to outperform all of the models.

The 6 versions of the SIR parameter time series were each utilized to train their own separate instances of the ensemble learner. All instances were also given all 8 mobility features as inputs.

The date range used for training data was 2021-Jan-01 thru 2021-June 30. Those we did have the data through August, we choose to train up to the end of June because we are forecasting 7, 14, 30 day data for 30 days. As such, the first day's forecast is for 2021-August-01, forecasting the 30 day beta, gamma for 2021-July-31. Given that we are producing 30 days of forecasts, the last day's forecast is for the day 2021-August-1, and the 30 day forecast is for the day of 2021-Aug-31.

4.4 Forecasting Cases and Mortality via Discrete Time-Varying SIR

We used each of the versions of predicted values for $\beta(t)$ and $\gamma(t)$ as the parameter settings for different instances of a discrete time-varying SIR (DTV-SIR) model. Taking known case and mortality counts on a particular start date as I_0 and R_0, respectively, along with total population, we set up the initial state of the model. Then we step forward one day at a time, using the following formulation:

$$\begin{cases} S(t+1) & = -\beta(t) \cdot \frac{S(t)}{N} \cdot I(t) \\ I(t+1) & = \beta(t) \cdot \frac{S(t)}{N} \cdot I(t) - \gamma(t) \cdot I(t) \\ R(t+1) & = \gamma(t) \cdot I(t) \end{cases}$$

The result of this is time series data that forecasts cases via $I(t)$ and mortality via $R(t)$. We produced this prediction over the same time range in which we made predictions for $\beta(t)$ and $\gamma(t)$.

5 RESULTS

5.1 Questions to be answered:

In this section, we list the questions our experiments are designed to answer:

(1) Can we use forecasted beta and gamma values to predict infections and deaths?
(2) Do certain predictive models out perform others?
(3) Do forecasts decrease in accuracy as we predict further out?
(4) Are certain states produce significantly better forecasts than others?
(5) Do ensemble predictors outperform other predictors?
(6) Are there any surprise findings?

5.2 Details of experiments:

We took the 2019 census data (latest available) to calculate the infection and death ratio with respect to the daily infection and casualty numbers published by John Hopkins and the CDC. We then ran the time-invariant SIR ODE model to produce the JH and CDC beta and gamma values for a look back window of 7, 14, and 30 days.

We then took these calculated beta and gamma values for each state for each day for the time period of Jan 1 2021 to August 31, 2021, and combined the data with mobility data from apple and google for each respective day. This combined dataset was our full training data set.

We then iteratively removed the feature we are each training models when predicting the feature (beta or gamma for a certain lookback window), except for ARIMA which exclusively predicts using the only the feature data we were predicting.

We used walk forward with ARIMA and OLS, appending the predicted results each time in forecasting future future results.

To ensure we don't taint the training data with unknowable data from the day of, the beta and gamma values in the training data are already shifted one day.

To forecast 7, 14, or 30 days out, we need only to shift using panda data frames by the forecast period. As all of the result generation is dynamic - at the start of the code execution, we can set any forecast period, predict range, or look back period in the initial settings and rerun the experiment to produce new results without changing any of the functions.

We outputted two files at the end of this stage: The predicted look back window beta and gamma for each state for each time period for each data source, and the RMSE of each forecast. After this output is complete for all states, we then take the RMSE for each each model and use it to create a weighted ensemble prediction, and calculated the RMSE of the ensemble learner and appended these results into each state's respective prediction file as well as each state's RMSE result file.

For the last step we fed the predicted beta and gamma time series data into our implementation of a discrete time-varying SIR model in order to predict case and mortality counts. These results and their respective RMSEs were then output to files for each state.

The beta gamma predicted results for each state is a file with 86 time periods columns (date + truth + (6 models + ensemble) * 3 (time periods) * 2 (sources) * 2 (beta, gamma) and 30 rows (days). The RMSE results for each state includes the RMSE results of 84 models.

In total, we calculated 145,800 beta and gamma values, we ran 5,100 models, we calculated the RMSE for 4,200 models, and we forecasted 129,00 beta and gamma values.

5.3 Observations:

There are certain results that we expect: We expect that as we forecast further out, our overall model accuracy for all models

time range	cdc_beta_ARIMA	cdc_beta_OLS	cdc_beta_DT	cdc_beta_RF	cdc_beta_Bagged_DT	cdc_beta_Bagged_RT	cdc_beta_ensemble
7	1.73E-06	2.64E-09	6.80E-10	6.67E-10	1.38E-09	6.53E-10	3.82E-10
14	1.50E-06	5.61E-09	1.81E-09	1.79E-09	1.83E-09	1.73E-09	1.32E-09
30	2.10E-06	1.98E-09	1.91E-09	1.89E-09	1.91E-09	1.90E-09	1.30E-09
avg	1.78E-06	3.41E-09	1.47E-09	1.45E-09	1.71E-09	1.43E-09	1.00E-09

time range	jh_beta_ARIMA	jh_beta_OLS	jh_beta_DT	jh_beta_RF	jh_beta_Bagged_DT	jh_beta_Bagged_RT	jh_beta_ensemble
7	1.34E-06	5.28E-09	8.28E-10	7.95E-10	3.12E-09	9.58E-10	4.52E-10
14	2.38E-06	6.98E-09	1.75E-09	1.74E-09	1.64E-09	1.58E-09	1.28E-09
30	1.85E-06	1.85E-09	2.19E-09	2.17E-09	2.18E-09	2.17E-09	1.46E-09
avg	1.86E-06	4.70E-09	1.59E-09	1.57E-09	2.31E-09	1.57E-09	1.06E-09

Figure 1: CDC model performance

time range	cdc_gamma_ARIMA	cdc_gamma_OLS	cdc_gamma_DT	cdc_gamma_RF	cdc_gamma_Bagged_DT	cdc_gamma_Bagged_RT	cdc_gamma_ensemble
7	6.16E-06	1.11E-05	3.12E-06	2.75E-06	3.23E-06	2.74E-06	8.89E-07
14	1.63E-05	4.87E-05	4.17E-06	4.40E-06	4.93E-06	4.40E-06	1.72E-06
30	1.90E-05	1.13E-05	5.25E-06	5.07E-06	5.23E-06	5.18E-06	2.29E-06
avg	1.38E-05	2.37E-05	4.18E-06	4.07E-06	4.46E-06	4.11E-06	1.63E-06

time range	jh_gamma_ARIMA	jh_gamma_OLS	jh_gamma_DT	jh_gamma_RF	jh_gamma_Bagged_DT	jh_gamma_Bagged_RT	jh_gamma_ensemble
7	7.16E-06	1.22E-05	3.36E-06	3.43E-06	6.53E-06	3.42E-06	1.34E-06
14	1.97E-05	3.20E-05	5.33E-06	3.36E-06	5.83E-06	3.63E-06	1.65E-06
30	1.87E-05	9.61E-06	4.11E-06	4.26E-06	4.16E-06	4.16E-06	2.04E-06
avg	1.52E-05	1.79E-05	4.26E-06	3.68E-06	5.51E-06	3.74E-06	1.68E-06

Figure 2: JH model performance

would decrease - This was true with the exception of ARIMA for Gamma. ARIMA for gamma outperformed all models except the ensemble models for the 7 day range. We would expect that the further out we go, the greater the loss of accuracy. Interestingly, the further out we predict, the lower the percent decrease in accuracy. This unexpected phenomena is more prevalent with gamma relative to beta predictions. Whether this is an anomaly or a feature requires more smaller incremental date ranges to determine.

We expect that of the regression trees, the base decision tree regressor should perform the best for reasons previously explained. In actually, the results show that for almost all date ranges, the random forest outperformed the decision tree for both beta and gamma. Perhaps this is indicative that the decision tree is overfit and the random forest is more robust for predicting on test data.

We expect that of the regression trees, the bootstrapped random forest to perform the worst because it has the greatest amount of randomness and least amount of data to train on. Unexpectedly, similar to the results with the decision tree vs the random forest, the bagged random forest outperformed the bagged decision trees - likely due to outfitting.

We expect the bagged decision tree to outperform the bagged random forest. Unexpectedly, similar to the results with the decision tree vs the random forest, the bagged random forest outperformed the bagged decision trees - likely due to outfitting.

We expect the ensemble learner to out perform all of the other models as the forecast date ranges increase. This is affirmed by the data. What is interesting is that in longer range predictions for beta, ensemble performed approx .3x better than the second best model, whereas for the gamma, ensemble performed approximately 3x better than the second best model.

The data appears to indicate that we can predict beta values with greater accuracy than gamma values.

Using the TV SIR forecast methodology, we can see that our Infection prediction is off. We believe this is likely due to a combination of infection under-reporting (not everyone sick gets tested) and the perceived impact of quarantine/social distancing.

Using the TV SIR forecast methodology, we can see that our Death prediction is pretty accurate when compared to the ground truth. Further, in instances where the ground truth changed course

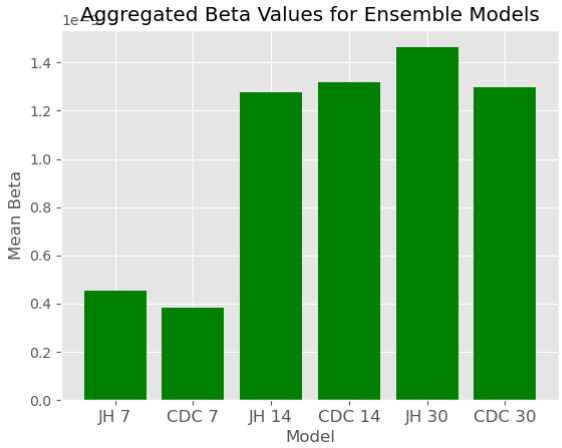

Figure 3: Mean Aggregated Beta

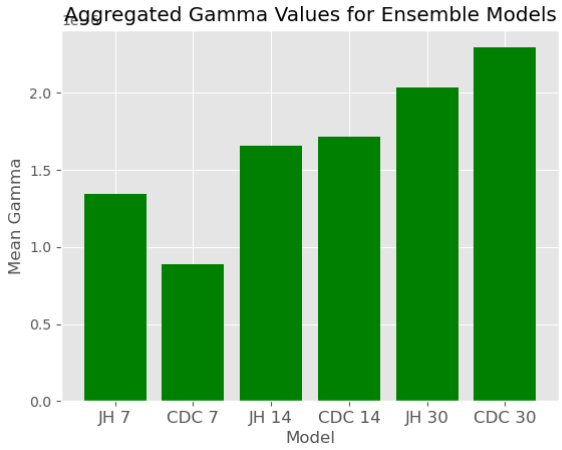

Figure 4: Mean Aggregated Beta

rapidly, our model appears to better estimate death values due to its consideration of dynamic gamma and beta values.

6 CONCLUSION AND DISCUSSION:

There are two main takeaways from this research. The first is that we can use the TV SIR to predict deaths accurately.

The second takeaway is exciting for us. In essence, we have demonstrated that we can accurately forecast beta and gammas a period forward. We also have the variance for these forecasted values (sqrt of RMSE). Thus, in essence, this system also functions as a testing platform for forecasts - as in, one can feed in their forecasted I/R values for a 7, 14, or 30 day period, and our system could tell how likely the forecasts are to be correct - this is of course presuming, that we can validate our models on additional months (currently we only predicted on one month interval).

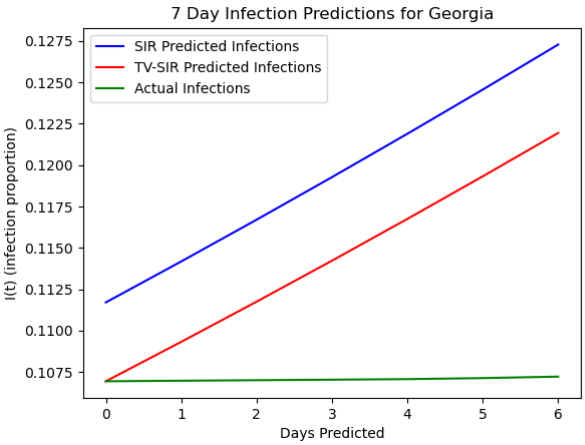

Figure 5: Georgia Infections predicted using TV SIR

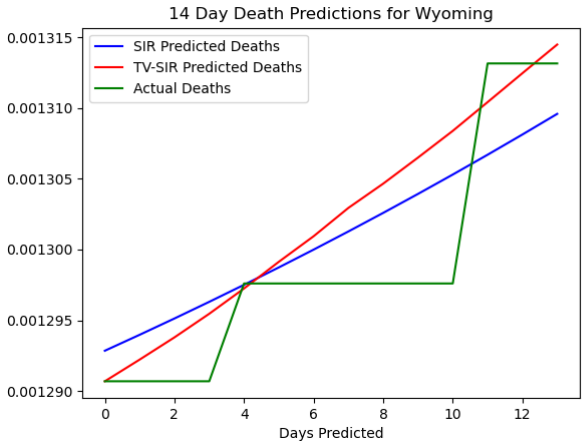

Figure 7: Wyoming Deaths predicted using TV SIR

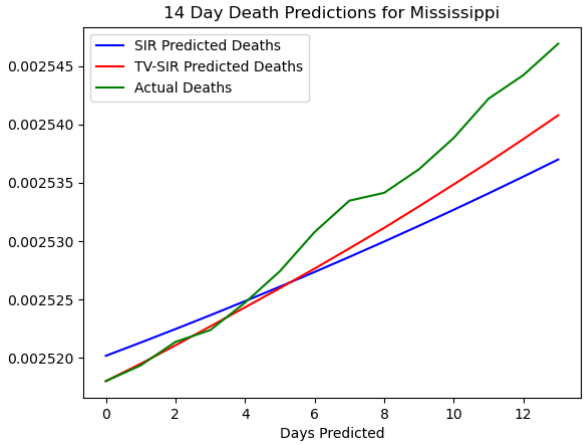

Figure 6: Mississippi Deaths predicted using TV SIR

REFERENCES

[1] Benjamin Ambrosio and MA Aziz-Alaoui. 2020. On a coupled time-dependent SIR models fitting with New York and New-Jersey states COVID-19 data. *Biology* 9, 6 (2020), 135. https://doi.org/10.3390/biology9060135

[2] Apple. 2021. COVID-19 - Mobility Trends Reports. https://covid19.apple.com/mobility/. [Online; accessed 03-October-2021].

[3] Mouhamadou Aliou Mountaga Tall Baldé. 2020. Fitting SIR model to COVID-19 pandemic data and comparative forecasting with machine learning. *medRxiv* (2020). https://doi.org/10.1101/2020.04.26.20081042

[4] Stefanella Boatto, Catherine Bonnet, Bernard Cazelles, and Frédéric Mazenc. 2018. SIR model with time dependent infectivity parameter: approximating the epidemic attractor and the importance of the initial phase. (2018). https://hal.inria.fr/hal-01677886/

[5] Center for Systems Science and Engineering at Johns Hopkins University. 2021. COVID-19 Data Repository. https://github.com/CSSEGISandData/COVID-19. [Online; accessed 02-November-2021].

[6] Centers for Disease Control and Prevention. 2021. United States COVID-19 Cases and Deaths by State over Time. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36. [Online; accessed 02-November-2021].

[7] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. 2020. A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering* 7, 4 (2020), 3279–3294. https://doi.org/10.1109/TNSE.2020.3024723

[8] Google. 2021. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/. [Online; accessed 03-October-2021].

[9] William Ogilvy Kermack and Anderson G McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115, 772 (1927), 700–721. https://doi.org/10.1098/rspa.1927.0118

[10] Ognyan Kounchev, Georgi Simeonov, and Zhana Kuncheva. 2020. The TVBG-SEIR spline model for analysis of COVID-19 spread, and a Tool for prediction scenarios. *arXiv preprint arXiv:2004.11338* (2020). https://arxiv.org/abs/2004.11338

[11] Mark B Shapiro, Fazle Karim, Guido Muscioni, and Abel Saju Augustine. 2020. Are we there yet? An adaptive SIR model for continuous estimation of COVID-19 infection rate and reproduction number in the United States. *medRxiv* (2020). https://doi.org/10.1101/2020.09.13.20193896

[12] Haoxuan Sun, Yumou Qiu, Han Yan, Yaxuan Huang, Yuru Zhu, and Song Xi Chen. 2020. Tracking and predicting COVID-19 epidemic in China mainland. *medRxiv* (2020). https://doi.org/10.1101/2020.02.17.20024257

[13] United States Census Bureau, Population Division. 2019. State Population Totals and Components of Change: 2010-2019. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html. [Online; accessed 06-November-2021].