A Stacking Ensemble Machine Learning Strategy for COVID-19 Seroprevalence Estimations in the USA based on Genetic Programming

Gontzal Sagastabeitia*, Josu Doncel*, Antonio Fernández Anta†, Jose Aguilar†‡ and Juan Marcos Ramirez†
*University of the Basque Country UPV/EHU, Leioa, Biscay
Email: gontzal.sagastabeitia@ehu.eus
†IMDEA Networks Institute, Madrid, Spain
‡CEMISID, University of the Andes, Merida, Venezuela

***Abstract*—The COVID-19 pandemic exposed the importance of research on the spread of epidemic diseases. In the case of COVID-19, official data about infection prevalence was based on PCR and antigen tests reports, which can be unreliable. In our work, we construct prediction models based on Genetic Programming to estimate the SARS-CoV-2 seroprevalence of a given population from multiple estimates of the COVID-19 prevalence (official prevalence data, estimates derived from wastewater data, and estimates obtained from massive surveys with different rules and ML methods). To do that, we propose the use of stacking techniques based on Genetic Programming to obtain Machine Learning Ensemble Methods. Our approach produces more accurate prediction models than conventional stacking techniques based on Linear Regression.**

I. Introduction

The Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) [1], has raised public interest in epidemics. During the pandemic, media outlets mainly reported daily updates on the number of COVID-19 infections, hospitalisations, and deaths to provide information about the spread of the disease. COVID-19 estimated number of cases was primarily obtained from large-scale screening using PCR and antigen tests [2]. However, this method may not be the most reliable source of information when attempting to understand the full scope of the pandemic and accurately determine the percentage of the population affected, since the accuracy of the information obtained from test screening is affected by various factors such as the limited availability of test kits [3] (especially at the beginning of the pandemic), the time between infection and the test timing [4], and the high number of asymptomatic infected individuals [5].

The traditional approach for estimating the proportion of previously infected individuals within a population relies on the measurement of seroprevalence. Specifically, seroprevalence refers to the proportion of individuals who test positive for a specific antibody in their blood [6]. In the case of COVID-19, a seropositive individual is a person who has SARS-CoV-2 antibodies in their blood. The presence of antibodies is considered sufficient evidence to confirm past infection, even without a positive test result. Multiple seroprevalence studies were conducted during the COVID-19 pandemic in different countries, which required blood analyses of thousands of individuals along multiple rounds [7], [8]. These campaigns required substantial resources for logistical and organisational purposes.

On the other hand, many approaches have been proposed during the COVID-19 pandemic that rely on data analysis and artificial intelligence to estimate the number of daily cases accurately [3], [9], [10]. These methods exploit the ability of online tools to track health indicators in almost real-time by collecting vast amounts of data from self-reported information. Estimating the seroprevalence from this data type is required to provide healthcare systems with a less expensive method for tracking the spread of diseases. In this context, it is necessary to analyse Ensemble Methods that allow combining the different estimation approaches (regardless of whether they are based on machine learning techniques or not). In particular, we are interested in stacking techniques as an ensemble learning strategy, since it allows learning how to combine the estimates of numerous machine learning models to obtain a final estimate [11], [12].

Although the COVID-19 pandemic is losing its relevance, it has revived the fear of the spread of infectious diseases and made the population aware of the threat that pandemics can pose. Therefore, even though COVID-19 may not be as important anymore, our work can provide a general framework based on data for tracking viral spread. Furthermore, this work is not confined to epidemic tracking or even medical applications. The methods and procedures used for this problem can be applied to any data-set to predict or estimate other aspects of a population. An example of an application in other areas is to use surveys and assemble different models and data to evaluate the voting intention in future elections.

A. State of the Art

Many approaches have been proposed that rely on data analysis and artificial intelligence to estimate the number of COVID-19 cases [3], [10], [13]. In particular, we are interested in stacking techniques as an ensemble learning strategy, since

they allow learning to combine the estimates from numerous machine learning models to obtain a final estimate [11], [12].

There have been extensive studies carried on the spread of the COVID-19 pandemic from a statistical point of view, as can be seen from works on literature reviews on the subject [14], [15], [16]. These studies have mainly focused in implementing machine learning-based models, as seen in [17], [18], [19]; and ensemble approaches have been widely studied, like those of [20], [21], [22], [23]. Most of these works have taken COVID-19 test positives as reference, although there are some interesting works that have tried to predict other metrics, such as COVID-19 mortality in [24]. Overall, not a lot of research has been conducted around SARS-CoV-2 seroprevalence, which is the metric we focus on with our models. There has also been some work on the use of Genetic Programming to estimate COVID-19 prevalence in the United States, like [25]; but there is very little research on the application of ensemble approaches using non-machine learning-based models.

B. Contributions

In our work, we use Genetic Programming (GP) to estimate the seroprevalence of SARS-CoV-2 in the United States of America (USA) based on data of daily infections obtained from multiple sources. Notice that our approach does not result from mass screenings using PCR or antigen tests. Specifically, we use GP as a stacking machine learning strategy which learns to combine the estimations of several base models to obtain a final prediction. GP was chosen because the models it constructs explicitly show how the different explanatory variables are combined to calculate the seroprevalence rate, so the explainability of the model is high. We consider as a performance metric the mean absolute relative error (MARE) and we use GP to find the model that best combines the estimations of seroprevalence rates, i.e., that minimises the MARE. We consider two different settings: state by state (statewide models), where data from a single state is considered for the model; and the whole USA (nationwide models), where the data from all states is considered for a single model.

We also present two approaches to dealing with the available data: cumulative and non-cumulative aggregation. Cumulative aggregation uses the first available seropositivity data value as a reference, while non-cumulative aggregation uses the latest available seropositivity data value. We conclude that cumulative aggregation is better when working with individual states within the USA, but non-cumulative aggregation more accurately fits the data when multiple states are considered together. We find that GP obtains much more accurate prediction estimations than those yielded by a simple Linear Regression (LR) least-square model on average and that, as expected, the complexity of the models obtained using GP is usually inversely correlated with its MARE.

II. Our approach

We have access to multiple data sources that provide us with useful information with respect to the epidemic, including estimated COVID-19 prevalence rates via different prediction methods based on COVID-19 prevalence surveys. To reduce the biases each method may have, we will use all of them to construct an ensemble model that will use the estimations of all those methods, as well as some extra explanatory variables (mainly, official prevalence data and estimates from wastewater SARS-CoV-2 concentration studies). Our problem consists of finding an appropriate prediction model that combines said estimations and variables to predict the seropositivity of a certain population on a given date.

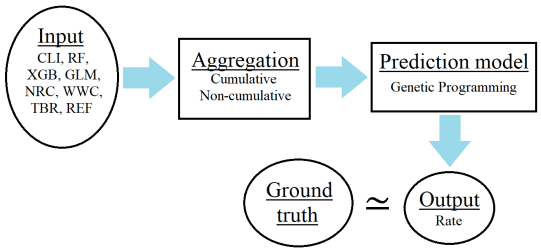


Figure 1: Diagram showing our stacking machine learning strategy.

A. Our Ensemble Method

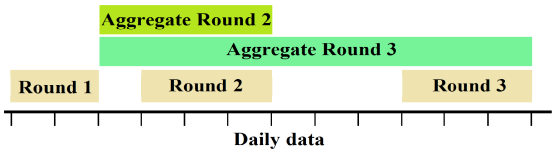
We require a stacking machine learning strategy, which learns to combine the estimates from these methods with the extra explanatory variables, to obtain a final estimate. The seropositivity values we use as ground truth for our work are the seroprevalence measurements made by the Centers for Disease Control and Prevention (CDC), the national public health agency of the USA [26].

In Figure 1, a diagram schematically explains our stacking machine learning strategy to obtain the seroprevalence rate estimations. As the figure shows, we have various input variables from multiple data sources that need to be aggregated before we can use them to build our GP-based models. It is important to note that the values of these variables come from estimation/prediction methods based on machine learning techniques (Random Forest - RF, Extreme Gradient-Boosting - XGB, etc.) or are extra explanatory variables (New reported cases - NRC, Wastewater cases - WWC, etc.).

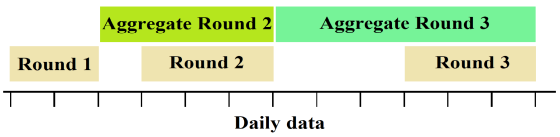
On the other hand, the aggregation is necessary because there is an inconsistency between the number of input data points and the ground truth (seropositivity). The latter comprises at most 30 measurements per US state, while most input variables have daily values. Therefore, we aggregate the input data into the same number of data points as the ground truth.

After aggregation, the variables are combined into a prediction model using GP as the stacking ensemble strategy that outputs estimated seroprevalence rates. We then compare the output of the models with the seroprevalence ground truth to evaluate the accuracy of the constructed model.

In order to have a baseline model to compare our results to, we also build another stacking ensemble strategy based on a least-square Linear Regression (LR) model. Least-square regression is the most widely used type of regression [27]. It



(a) Cumulative aggregation



(b) Non-cumulative aggregation

Figure 2: Diagrams showing how daily data points are aggregated to transform daily data into sporadic data.

minimises Sum of Squared Residuals (SSR) to find the best linear model to fit the data.

B. Aggregation

The data can be classified into two groups based on its frequency: sporadic and daily data. The ground truth has one measurement per seropositivity survey round for a total of at most 30 data points per state (sporadic data), while every explanatory variable we are going to use has daily values (daily data). Note that each seropositivity survey round spans several days of data collection. Therefore, we need to establish some criteria for how we are going to unify these two types of data. We have to choose how to aggregate daily data so that it aligns with the sporadic ground truth.

We have defined two different approaches to this aggregation problem, namely, "cumulative" and "non-cumulative" aggregation, both of which add the daily values of the explanatory variables. A graphical representation of how the daily data is aggregated with each approach can be seen in Figure 2. The cumulative aggregation approach (Figure 2a) adds up the daily data into Aggregate Round n, for the n-th survey round, starting at the end-date of the first round, up to the end of the current n-th round, so that the data aggregated for each round is a subset of the data aggregated for the next round. On the other hand, the non-cumulative approach (Figure 2b) adds up the daily data into Aggregate Round n from the end-date of the round $n-1$ up to the end-date of round n, so that the aggregates of each round are disjoint.

C. Dataset

1) Input: Our main source of data has been the US COVID-19 Trends and Impact Survey (CTIS). This project, operated by the Delphi Group at Carnegie Mellon University in collaboration with Facebook, has continuously operated surveys between the 6th of April 2020 and the 25th of June 2022, and has collected over 20 million responses [28], [9]. Every day for the duration of the project, a random sample of Facebook users were invited to complete a questionnaire about the COVID-19 pandemic: symptoms, COVID testing, social distancing, vaccination, mental health and economic security.

In this work, we have not used directly the raw data obtained by the CTIS. Instead, we have used daily prevalence estimates obtained from the responses to the CTIS using various methods and individual features, as described in [29]. Moreover, in addition to the estimates obtained from the CTIS, we have also chosen four other input variables for our models, from three different sources: official daily reported new COVID cases, wastewater SARS-CoV-2 concentration [30], previous seroprevalence measurement, and normalised time since the previous seroprevalence measurement.

In summary, our models consist of eight explanatory variables, of which three are predicted by machine learning or statistical models. Those eight variables are the following:

- **COVID-like illness** (*CLI*). Daily rates for reported COVID compatible symptoms, from the CTIS, aggregated to 30 data points by addition.
- **Random Forest** (*RF*). Daily estimated prevalence rate via a RF model from CTIS data, aggregated to 30 data points by addition.
- **Extreme Gradient-Boosting** (*XGB*). Daily estimated prevalence rate via an XGB model from CTIS data, aggregated to 30 data points by addition.
- **Generalised Linear Model** (*GLM*). Daily estimated prevalence rate via a Generalised Linear Model from CTIS data, aggregated to 30 data points by addition.
- **New reported cases** (*NRC*). Official total number of daily reported SARS-CoV-2 test positives, aggregated to 30 data points by addition. The resulting data has been divided by its maximum value so that the scale lines up with the other variables.
- **Wastewater cases** (*WWC*). Daily estimated total active COVID-19 cases via wastewater virus concentrations, aggregated to 30 data points by addition. The resulting data is divided by its maximum value so that the scale lines up with the other variables.
- **Time between rounds** (*TBR*). Number of days of the time interval we are aggregating (days from the end of the reference value's round). The resulting number is divided by the maximum value of the variable so that the scale lines up with the other variables.
- **Reference value** (*REF*). The official rate of seropositivity in the round from which we are aggregating the daily data (the first round for cumulative and previous round for non-cumulative).

2) Output and Ground Truth: We have chosen to work with seroprevalence data from the United States of America (USA), because data is available for each one of its states. The US CDC has collected extensive data with respect to SARS-CoV-2 seroprevalence in their Nationwide Commercial Laboratory Seroprevalence Survey, which can be found in [26]. For that survey, the CDC conducted 30 rounds of seroprevalence testing among the population of separate states within the USA, between July 2020 and February 2022. Unfortunately, even though the CDC conducted 30 rounds in most of the US states, there are 13 states with less than 30 rounds: Arizona, Indiana,

Maryland, Montana, Nevada, New Hampshire, New Jersey, Utah and Virginia have 29 rounds; Hawaii has 27 rounds; Wyoming has 26 rounds; South Dakota has 21 rounds; and North Dakota has 4 rounds. Note that the 10 most populous states all have had 30 rounds conducted in them.

Therefore, the CDC seroprevalence measurements are the ground truth of this problem, and our models' output will be predictions of these values. In order to measure the accuracy of our prediction models, we are going to use their MARE when compared to the ground truth. The formula for a data-set of n observations is presented in Eq. (1), where y_i is the real rate for the i-th observation and $\hat{y}_i$ is its predicted value.

$$\text{MARE}(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{1}$$

D. Genetic Programming

GP is a method inspired by natural genetic processes that tries to find the best solution to a problem by evolving a set of equations. In our specific case, the GP-based models are mathematical formulas that combine the eight input variables presented above, and its output is a value that represents the estimation of the seroprevalence rate. The aim is to minimise the error between the values provided by these equations considering the available data set and the ground truth.

In our work, when building the prediction models, we minimise the SSR, which reduces the variance of the resultant residuals. Its formula is shown in 2, where y_i is the real rate for the i-th observation and $\hat{y}_i$ is its predicted value.

$$\text{SSR}(\hat{\mathbf{y}}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2}$$

We minimise SSR because we compare the GP-based models to the baseline least-square LR models, which minimise the SSR. Therefore, we minimise the same error in both cases. However, as the SSR is a relatively abstract measurement of error, we use the MARE as the error metric to compare the models. The MARE represents the relative deviation from the observed data on average and is more explicit and easily interpretable than the SSR. This error metric allows for a more comprehensible reading of the accuracy of the models.

GP uses operations based on natural genetic evolution to evolve and update a set of given equations (prediction models) so that they get better over time. GP works with tree structures to manipulate the equations. This tree structure allows the algorithm to change and swap parts of an equation by manipulating nodes or subtrees in a given tree.

For our work, we use the following operators: addition, subtraction, multiplication, division, negative sign, exponential, and natural logarithm. We have also added a set of constants to the algorithm's pool of resources to make it easier to get constant terms and factors in the equations. The set of chosen constants is the following set of powers of ten: $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$.

The GP algorithm can be considered to have three phases: initialisation, selection, and reproduction. The algorithm starts by generating a random initial population of trees of a predetermined size using the available operators and variables. The randomness of the initial population allows the algorithm to start with a wide range of possibilities to cover enough of the search space.

Once the population is initialised, the algorithm uses a selection method to pick several individuals. In our case, the evaluation criteria for selection is SSR. Therefore, our algorithm uses SSR to select a population subset, using tournament selection of size three, which consist on randomly taking three individuals to then pick the best individual among those three, and repeating until the desired number of individuals are selected. Then, the algorithm manipulates the selected individuals ("parents") to create a new generation ("children"). For the reproduction of the parents, our algorithm uses three operations:

- Crossover, which picks pairs of parents and uses one-point crossover to generate one or two children.
- Mutation, which picks a single child and use subtree replacement mutation to randomly change it.
- Replication, when the child is just a copy of its parent.

Crossover is usually applied before mutation, and in our algorithm both operations have assigned probabilities and are applied to the parents based on those probabilities: p_c, p_m.

After the children have been created, a new selection takes place to form the new generation of the same size, which in our case includes the parent population.

With the new generation, the algorithm repeats the process of selection and reproduction, until the stopping criteria is met. As stopping criteria, we have set a maximum number of generations, but if the fitness (SSR) of the best individual in each generation does not improve beyond a specified threshold δ for a total of m_s generations, the algorithm is stopped. δ and m_s are predefined hyper-parameters.

Another hyper-parameter of the algorithm is the maximum depth of the GP-based models. This hyper-parameter prevents the model's complexity to grow too much, and we do not need an excessive complexity to obtain good prediction models, as we see below.

In summary, the pseudo-code of the constructed GP algorithm for our problem is presented below on Algorithm 1. The variables introduced to the algorithm are as follows:

- P the initial population, a set of models.
- $p_c \in [0, 1]$ the probability of crossover.
- $p_m \in [0, 1]$ the probability of mutation.
- $\delta \in \mathbb{R}^+$ and $m_s \in \mathbb{N}$, the previously mentioned hyper-parameters for the stopping criteria.
- $g_{max} \in \mathbb{N}$ the maximum number generations of the algorithm.
- $d_{max} \in \mathbb{N}$ the maximum allowed depth of the model.

After a hyper-parameter optimisation process, the final values for the hyper-parameters were: a population size of $|P| = 300$, crossover and mutation probabilities of $p_c = 0.8$ and $p_m = 0.3$, the stopping parameters $\delta = 0.005$ and $m_s = 100$, a maximum number of generations of $g_{max} = 1000$, and maximum depths $d_{max} \in \{4, 6, 8, 10\}$.

Algorithm 1 The GP algorithm created

Require: P, $p_c \in [0,1]$, $p_m \in [0,1]$, $\delta \in \mathbb{R}^+$, $m_s \in \mathbb{N}$, $g_{max} \in \mathbb{N}$
  $best \leftarrow \arg\min_{f \in P}\{\text{Eval}(f)\}$
  $locked\_eval \leftarrow \text{Eval}(best)$
  $m \leftarrow 0$
  $p_m^* \leftarrow \min\{1, 2p_m\}$
  for $g = 1, \ldots, g_{max}$ **do**
    $C \leftarrow \text{Select}(P, |P|)$
    if $g = 100$ **then**
      $p_m^* \leftarrow p_m$
    end if
    Crossover(C, p_c)
    Mutate(C, p_m^*)
    $P \leftarrow \text{BestOf}(P \cup C, |P|)$
    $best \leftarrow \arg\min_{f \in P}\{\text{Eval}(f)\}$
    if $locked\_eval - \text{Eval}(best) \leq \delta$ **then**
      $m \leftarrow m + 1$
      if $m = m_s$ **then**
        break for
      end if
    else
      $locked\_eval \leftarrow \text{Eval}(best)$
      $m \leftarrow 0$
    end if
  end for
  return $(best,\ g)$

We want to analyse the performance of GP-based models with this particular problem. The GP-based models are known to provide a more general framework than LR. Therefore, one of the objectives of this work is to investigate the performance improvement of GP-based models with respect to the LR models.

The GP algorithm is stochastic: it has an element of randomness that can cause the results of each iteration to be different from each other. Therefore, one execution is not enough to see how good the GP algorithm is at finding accurate prediction models. For that reason, we will test the algorithm by executing it 20 times for each combination of state, aggregation, and maximum depth. After those 20 executions, we average the MARE of all the resultant prediction models, in order to have a better approximation of the accuracy of our GP algorithm.

III. Result Analysis

A. Results by state

First, we build a model for each state and aggregation method. All hyper-parameters are fixed, except d_{max}, the maximum depth of the GP algorithm. Therefore, for each state-aggregation combination, we obtain multiple models depending on the maximum depth picked.

In particular, we use four maximum depths: 4, 6, 8 and 10. We observe that a tree with less than 4 levels is too simple to represent the observed data accurately, and as we will see below, and that 10 levels are enough to get a relatively low MARE (higher values may result in over-fitting the data). In Table I, we display the mean MARE of three example states per maximum depth for both aggregation methods and the MARE obtained with the LR models. These three examples are representative of most statewide models. On the other hand, the box plots of the MARE per maximum depth for each state aggregation are also displayed in Figure 3.

		By maximum depth			
		4	6	8	10
California	Non-cum.	0.109	0.100	0.078	0.083
	Cum.	0.097	0.086	0.077	0.067
Texas	Non-cum.	0.122	0.108	0.097	0.087
	Cum.	0.090	0.083	0.076	0.067
Pennsylvania	Non-cum.	0.103	0.090	0.077	0.065
	Cum.	0.103	0.083	0.063	0.056

Table I: Table with the mean MARE of 20 executions of the GP algorithm for different maximum tree depths.

As we can see in the box plots, the larger depth they are allowed to have, the more precise GP-based models get (Texas, Pennsylvania, and cumulative California), even though there are a few cases where more depth beyond a certain point is shown to produce higher MARE (non-cumulative California).

When observing the behaviour of the MARE, the cumulative approach results in lower MARE than the non-cumulative on average for the GP statewide models, as seen in all three examples. If we compare the GP-based models' MARE to that of the LR models, we see that for all three examples (and all states studied beyond these examples), GP achieves a lower MARE than LR, especially with non-cumulative aggregation (even the 4-level model is below the linear MARE for all executions of California and Texas). The cumulative aggregation model usually needs more depth than the non-cumulative to better its linear counterpart, and in California there are some executions where the GP-based model was worse, but it is better than LR on average.

With these statewide GP-based models, we are achieving very low mean MAREs, below a 10% deviation from the observed data on average. This low MARE looks like the models are working extremely well, and could lead us to think that allowing even more depth would be desirable, as we may be able to reduce the MARE even more. However, there are two main reasons why that may not be a good idea. On the one hand, the more levels the model has, the more complex and confusing it becomes. Hence, if we want to understand the internal workings of the model, more complex trees could be a problem. Besides, a small reduction of the MARE may not be worth the great growth in complexity. On the other hand, when building a prediction model, reducing the error of the training data to a minimum (the observed data per state in our case) runs the risk of over-fitting the model to said training data and including the noise of the observations into the model, which gravely reduces the usefulness of the model outside the small data-set used.

Therefore, we decided that the small MARE obtained with maximum depths of up to 10 levels is a good enough result and that it is unnecessary to try to lower it even more by raising the maximum depth further.

In order to see couple of examples from all the models generated, we are going to show some estimations for both the minimum depth allowed (4) and the maximum depth (10).

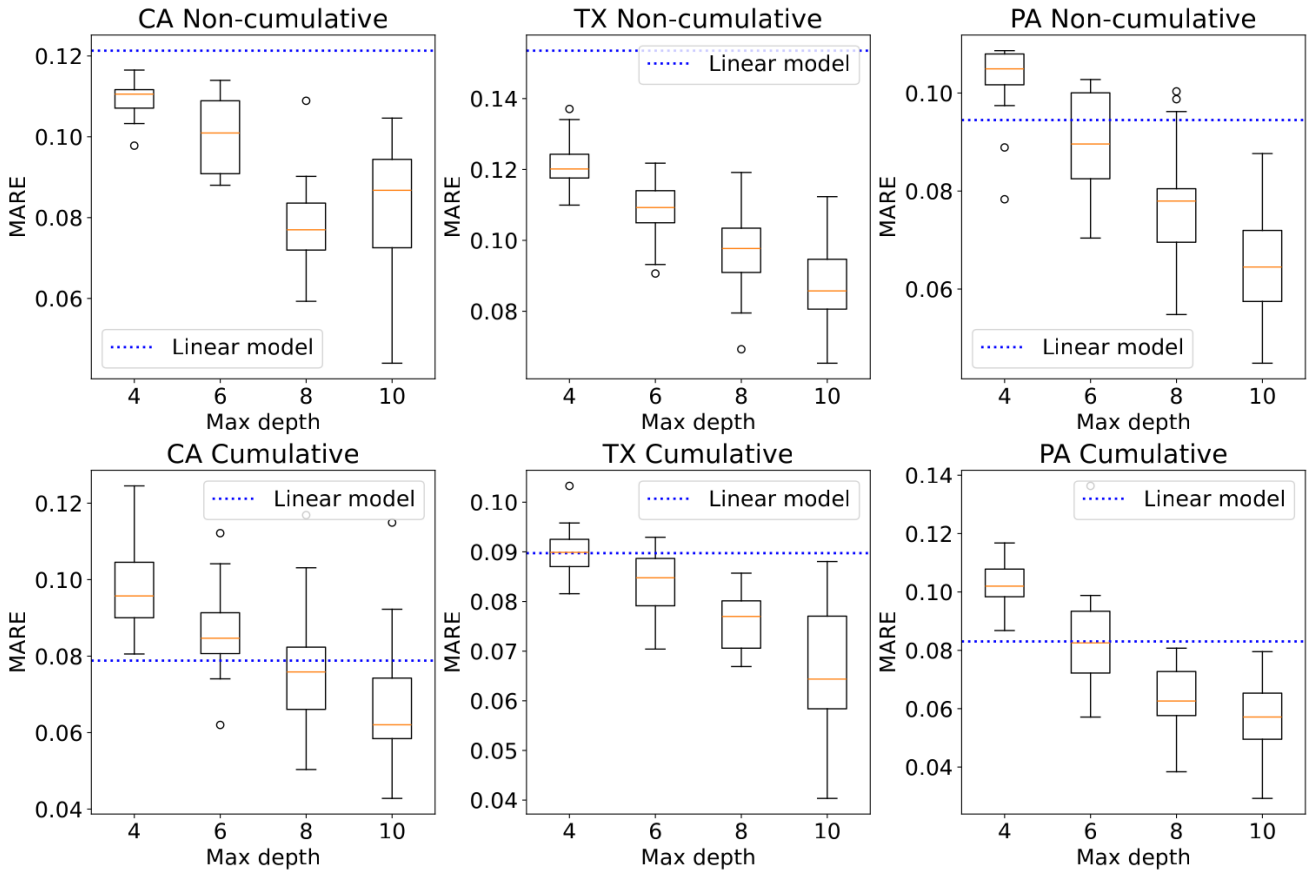


Figure 3: Box-plots of the MARE of 20 executions of the GP algorithm with different maximum depths for California (CA), Texas (TX) and Pennsylvania (PA).

We have picked the models that are closest to the mean MARE of all 20 executions as examples, for the most populous state (CA). The model of depth four that the algorithm returned for CA with non-cumulative aggregation, is the following:

$$\left(REF + 0.01RF + \frac{0.01}{TBR}(RF - 0.1)\right) e^{REF(WWC-TBR)}.$$

And with cumulative aggregation, the result is:

$$(NRC - TBR - \ln(CLI))\frac{CLI + 0.1}{100RF} + \frac{RF}{(REF + 10)e^{TBR}}.$$

It becomes evident at first glance that these models are more complex than a simple LR model, even if these are the GP-based models with the least depth. They are also clearly non-linear. These models have the MARE and R^2 values shown on Table II. The resultant estimated seropositivity rates can be seen in Figure 4.

We have done the same with the maximum depths of ten. The MARE and R^2 values can be seen on Table II, and the estimations on Figure 4.

B. Results for all the USA

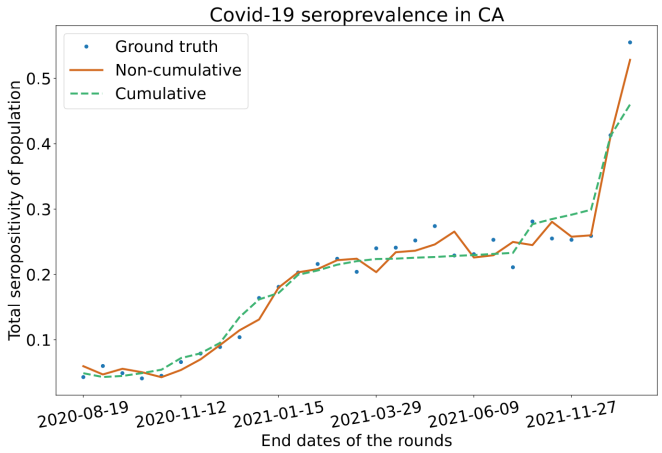
We have also used GP to obtain nationwide prediction models, aggregating all available data from all the USA. As previously mentioned, there are some states that have had less than 30 rounds of the CDC survey conducted on them, which may indicate that the accuracy of those measurements is lower. In order to see whether the accuracy of the nationwide models works better with some states, we have built three nationwide models using three sets of states: all states, only the states with 29 or 30 rounds surveyed (all but HI, ND, SD and WY), and the top 10 most populous states (CA, TX, FL, GA, NY, PA, IL, OH, MI and NC).

		By depth	
		4	10
MARE	Non-cum.	0.1053	0.0799
	Cum.	0.0980	0.0637
R^2	Non-cum.	0.9708	0.9854
	Cum.	0.9483	0.9872

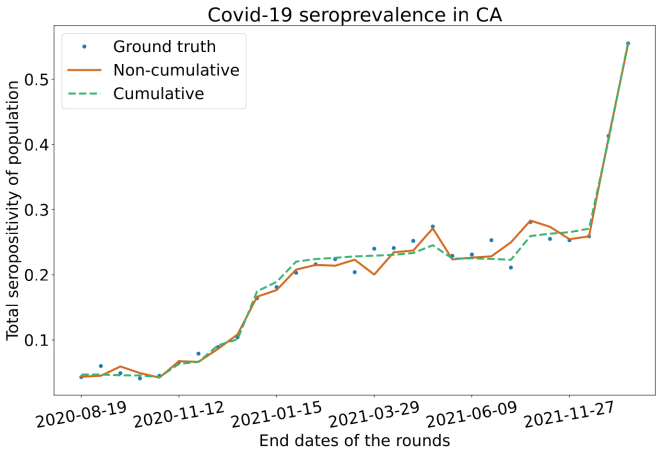
Table II: Table with the MARE and R^2 of the California GP-based models with a maximum tree depth of 4 and 10.

After running the GP algorithm for those three sets of states 20 times, we computed the mean MARE for each maximum depth, just like we did with the statewide models. The resultant mean MARE values are on Table III; and we have also displayed the MAREs of all 20 executions with box-plots on Figure 5.

Looking at the box-plots, we can see that the MARE of



(a) Maximum depth of 4



(b) Maximum depth of 10

Figure 4: GP estimations of seropositivity rates for California, with a maximum depth of 4 and 10.

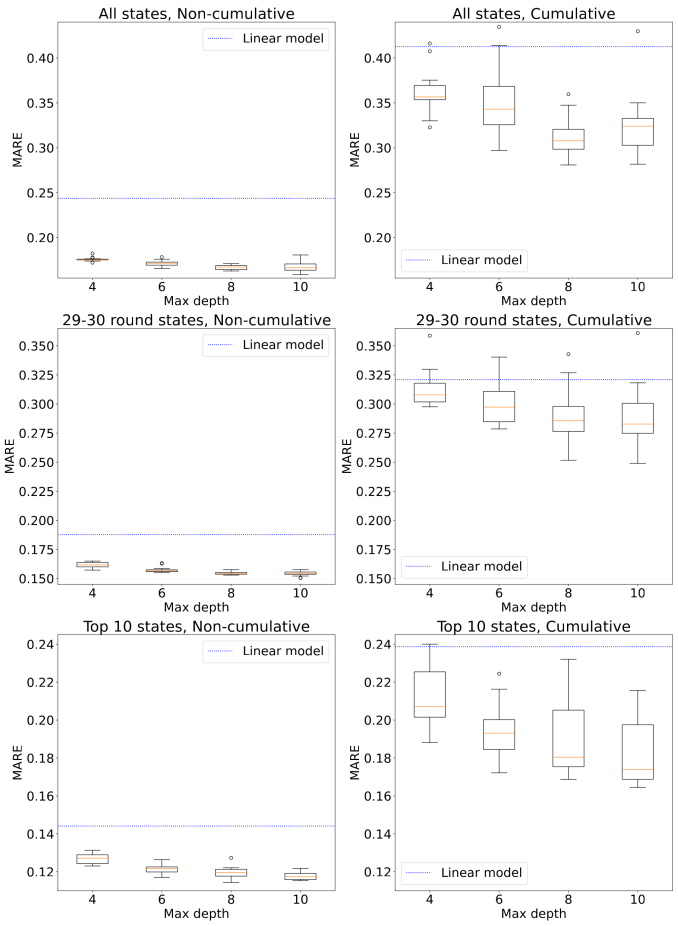


Figure 5: Box-plots of the MARE of 20 executions of the nationwide GP algorithm with different maximum depths for five sets of states: all states, 29-30 round states, and the top 10 most populous.

		By maximum depth			
		4	6	8	10
All	Non-cum.	0.176	0.171	0.167	0.168
	Cum.	0.362	0.351	0.311	0.325
29-30	Non-cum.	0.162	0.157	0.155	0.154
	Cum.	0.313	0.302	0.291	0.291
Top 10	Non-cum.	0.127	0.122	0.120	0.118
	Cum.	0.211	0.193	0.191	0.183

Table III: Table with the mean MARE of 20 executions of the GP nationwide algorithm with all states, states with 29-30 rounds and the top 10 states; for different maximum tree depths.

the nationwide GP-based models is higher on average than the MARE of the statewide models. However, the GP-based models greatly over-perform the linear nationwide models, specially with non-cumulative aggregation. Furthermore, the GP nationwide models seems to suggest that the nationwide model performs poorly with smaller states, driving the mean MARE up, because the GP-based models without the states with less than 29 rounds shows better results, and we get even better MARE if we only account for the ten most populous states. This behaviour is also observed with LR.

Besides, just like the statewide GP-based models, the larger the maximum depth of the models, the more accurate they get. However, there is barely any improvement from maximum depth 8 to 10 for non-cumulative aggregation when states with 29-30 rounds are considered, and for both aggregations with all states. That seems to indicate that a maximal depth beyond 8 levels does not result in a big improvement in accuracy. This leads us to think that sacrificing simplicity for relatively minuscule improvements beyond depth 8 is not worth it.

It is also worth noting that when multiple states are considered, the aggregation that results in the best MARE is the non-cumulative, opposite to what was observed with the statewide models. Furthermore, the accuracy of the GP-based

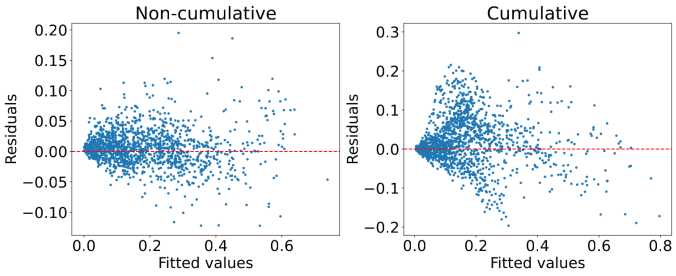


Figure 6: Residuals vs. fitted values plot for the nationwide model with all states.

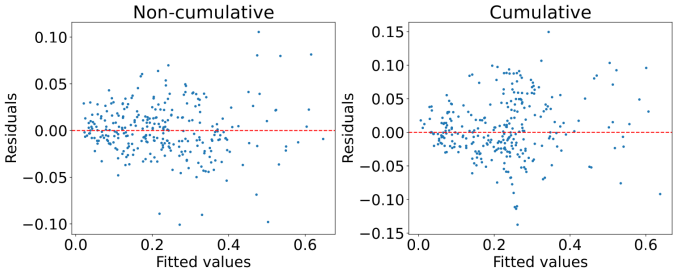


Figure 7: Residuals vs. fitted values plot for the nationwide model with the Top 10 states.

models found with cumulative aggregation vary much more than those found using non-cumulative aggregation. The box-plots of the cumulative models show that the MARE values are more spread out. This suggests that the non-cumulative approach results in more deterministic or predictable behaviour for the GP algorithm, while cumulative aggregation is more random and variable.

Lastly, we have checked that the residuals resultant from the GP-based models are symmetric and do not contain any kind of discernible pattern. For that, we first plotted the fitted values of the model versus their residuals, and then we checked the quantiles of said residuals. On Figures 6 and 7, we show the residual plots of the GP nationwide models with all states and with the top 10 states, respectively (with a maximum depth of 10). By a quick look at the plots, one cannot see any clear pattern or trend, and the points seem to be randomly distributed and centred around zero. The only two possible anomalies are: there are many more points with small fitted values, and the cumulative model with all states seems to have a downward protuberance around 0.3. The former is simply due to the nature of the data observed; and the latter is only a problem for the model with all states, so the model without the problematic states (the smallest ones) doesn't have any clear problems.

In Table IV we can see the quantiles of the residuals of the same two GP-based models, and except for the cumulative all states model, which is a bit skewed towards the positives, they are symmetric and centred around zero, specially the top 10 states model.

It is worth mentioning that a big disadvantage of GP-based models is their execution time. While the classical LR model

	All states		Top 10 states	
	Non-cum.	Cum.	Non-cum.	Cum.
min	-0.122	-0.196	-0.095	-0.126
25%	-0.011	-0.015	-0.014	-0.023
50%	0.005	0.011	-0.001	-0.001
75%	0.020	0.056	0.015	0.030
max	0.195	0.297	0.085	0.202

Table IV: Table with the quantiles of the residuals from the GP nationwide models with all states and with the top 10 states.

is built almost instantaneously (less than one second), the GP equivalent requires much more time to be built (orders of magnitude greater execution time). In some cases, the GP algorithm may need multiple hours to converge. Besides, as GP is a stochastic modelling approach the execution time is not fixed and can fluctuate significantly.

IV. Conclusions and Future Work

During the COVID-19 pandemic, many approaches based on machine learning and statistics have been proposed to estimate its behaviour. In this context, we have proposed a stacking ensemble machine learning approach which learns to combine the estimates from these methods, even incorporating other explanatory variables, to obtain a final estimate.

In our work, we have tested how well can GP work for this problem, and the obtained result show that the GP-based models used can very accurately estimate the seroprevalence rates of SARS-CoV-2. When comparing the MARE of the GP-based models to that of a baseline LR model, we clearly see that GP can obtain a lower MARE. Furthermore, the more depth and complexity is allowed when running the GP algorithm, the better the models get, but beyond a model depth of 8 the MARE does not improve considerably.

Overall, we see that statewide models are more accurate than nationwide models, with multiple states, and that the smaller states deviate from the trend of the bigger states, resulting in a higher MARE. Between the two aggregation methods used, we noticed that the cumulative approach is more appropriate for statewide models, but when working with multiple states the non-cumulative approach is more accurate.

Possible future work includes the following research lines: (i) studying GP more in depth by, for example: trying a larger set of operators beyond the ones used in this paper; allowing the GP algorithm to run for more time or using a bigger population size; trying to minimise the MARE instead of the SSR, (ii) checking whether adding new explanatory variables such as number of deaths by COVID-19 or vaccination rates can improve the models, (iii) trying to use the models to forecast future seroprevalence rates, (iv) researching the application of Neural Networks and other machine learning techniques to this problem, and (v) exploring new aggregation approaches other than cumulative and non-cumulative.

Acknowledgements

This work was done while G. Sagastabeitia was in an internship at IMDEA Networks Institute. The authors would like to thank Alexander Brodbelt and Mohamed Kacem for

their contribution to pre-processing the data used in this work. J. Aguilar and J.M. Ramírez work has been supported by project TED2021-131264B-I00 (SocialProbing), funded by MICIU/AEI /10.13039/501100011033 and the European Union-NextGenerationEU/PRTR. The work was also partially supported by project PID2022-140560OB-I00 (DRONAC) funded by MICIU/AEI /10.13039/501100011033 and ERDF, EU, by the Department of Education of the Basque Government, Spain, through the Consolidated Research Group MATHMODE (IT1456-22) and by the Marie Sklodowska-Curie grant agreement N. 777778.

REFERENCES

[1] R. Wölfel *et al.*, "Virological assessment of hospitalized patients with COVID-2019," *Nature*, vol. 581, p. 465–469, 2020.

[2] M. P. Cheng *et al.*, "Diagnostic testing for severe acute respiratory syndrome–related coronavirus 2: a narrative review," *Annals of internal medicine*, vol. 172, p. 726–734, 2020.

[3] Y. Zoabi *et al.*, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *NPJ digital medicine*, vol. 4, pp. 1–5, 2021.

[4] L. J. Akinbami *et al.*, "Coronavirus Disease 2019 Symptoms and Severe Acute Respiratory Syndrome Coronavirus 2 Antibody Positivity in a Large Survey of First Responders and Healthcare Personnel, May-July 2020," *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, vol. 73, p. e822–e825, 2021.

[5] M. Klompas, "Coronavirus disease 2019 (COVID-19): protecting hospitals from the invisible," *Annals of internal medicine*, vol. 172, no. 9, pp. 619–620, 2020.

[6] Farlex Partner Medical Dictionary, "Seroprevalence," https://medical-dictionary.thefreedictionary.com/seroprevalence, 2012.

[7] K. Bajema *et al.*, "Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020," *JAMA internal medicine*, vol. 181, no. 4, pp. 450–460, 2021.

[8] M. Pollán *et al.*, "Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study," *The Lancet*, vol. 396, no. 10250, pp. 535–544, 2020.

[9] J. A. Salomon *et al.*, "The US COVID-19 trends and impact survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination," *Proceedings of the National Academy of Sciences*, vol. 118, no. 51, p. e2111454118, 2021.

[10] C. M. Astley *et al.*, "Global monitoring of the impact of the COVID-19 pandemic through online surveys sampled from the Facebook user base," *Proceedings of the National Academy of Sciences*, vol. 118, no. 51, p. e2111455118, 2021.

[11] A. Gupta, V. Jain, and A. Singh, "Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications," *New Gener. Comput.*, vol. 40, pp. 987—-1007, 2022.

[12] T. Zhou and H. Jiao, "Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment," *Educational and Psychological Measurement*, vol. 83, no. 4, pp. 831–854, 2023. [Online]. Available: https://doi.org/10.1177/00131644221117193

[13] Y. Quintero, D. Ardila, E. Camargo, F. Rivas, and J. Aguilar, "Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables," *Computers in Biology and Medicine*, vol. 134, p. 104500, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521002948

[14] M. Jamshidi, S. Roshani, F. Daneshfar, A. Lalbakhsh, S. Roshani, F. Parandin, Z. Malek, J. Talla, Z. Peroutka, A. Jamshidi *et al.*, "Hybrid deep learning techniques for predicting complex phenomena: A review on COVID-19," *AI*, vol. 3, no. 2, pp. 416–433, 2022.

[15] A. H. Elsheikh, A. I. Saba, H. Panchal, S. Shanmugan, N. A. Alsaleh, and M. Ahmadein, "Artificial intelligence for forecasting the prevalence of COVID-19 pandemic: An overview," *Healthcare*, vol. 9, no. 12, p. 1614, 2021.

[16] C. Comito and C. Pizzuti, "Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: A focused review," *Artificial intelligence in medicine*, vol. 128, p. 102286, 2022.

[17] B. Lucas, B. Vahedi, and M. Karimzadeh, "A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA," *International Journal of Data Science and Analytics*, vol. 15, no. 3, pp. 247–266, 2023.

[18] E. Al-Bwana, "Coronavirus (COVID-19) detection using ensemble learning," Ph.D. dissertation, Zarqa University, 2021.

[19] L. Vaughan, M. Zhang, H. Gu, J. B. Rose, C. C. Naughton, G. Medema, V. Allan, A. Roiko, L. Blackall, and A. Zamyadi, "An exploration of challenges associated with machine learning for time series forecasting of COVID-19 community spread using wastewater-based epidemiological data," *Science of The Total Environment*, vol. 858, p. 159748, 2023.

[20] C. Cilgin and M. O. ÖZDEMİR, "Time series forecasting of covid-19 confirmed cases in turkey with stacking ensemble models," *Bingöl Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, vol. 26, pp. 504–520, 2023.

[21] W. Wang, F. Harrou, A. Dairi, and Y. Sun, "Stacked deep learning approach for efficient SARS-CoV-2 detection in blood samples," *Artificial Intelligence in Medicine*, vol. 148, p. 102767, 2024.

[22] S. Sharma, Y. K. Gupta, and A. K. Mishra, "Analysis and prediction of COVID-19 multivariate data using deep ensemble learning methods," *International Journal of Environmental Research and Public Health*, vol. 20, no. 11, p. 5943, 2023.

[23] W. Jin, S. Dong, C. Yu, and Q. Luo, "A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning," *Computers in Biology and Medicine*, vol. 146, p. 105560, 2022.

[24] S. Cui, Y. Wang, D. Wang, Q. Sai, Z. Huang, and T. Cheng, "A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality," *Applied Soft Computing*, vol. 113, p. 107946, 2021.

[25] N. Anđelić, S. B. Šegota, I. Lorencin, Z. Jurilj, T. Šušteršič, A. Blagojević, A. Protić, T. Ćabov, N. Filipović, and Z. Car, "Estimation of covid-19 epidemiology curve of the united states using genetic programming algorithm," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, p. 959, 2021.

[26] Centers for Disease Control and Prevention, "Nationwide Commercial Laboratory Seroprevalence Survey," https://data.cdc.gov/Laboratory-Surveillance, 2023.

[27] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0003267086800289

[28] Delphi Group at Carnegie Mellon University, "Delphi's COVID-19 Trends and Impact Surveys (CTIS)," https://delphi.cmu.edu/covid19/ctis/, 2022.

[29] J. Rufino, J. M. Ramirez, J. Aguilar, C. Baquero, J. P. Champati, D. Frey, R. E. Lillo, and A. F. Anta, "Consistent comparison of symptom-based methods for COVID-19 infection detection," *Int. J. Medical Informatics*, vol. 177, p. 105133, 2023. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2023.105133

[30] A. Srivastava, "The variations of SIkJalpha model for COVID-19 forecasting and scenario projections," *arXiv preprint arXiv:2207.02919*, 2022.