COVID-19 Infection and Mortality Rate Predictions in the US

Christian Brazell, Abdelrahman Kotb, Mahmood Shilleh

Abstract

Ever since the COVID-19 pandemic broke out, massive efforts have been made internationally to improve our understanding of the virus, how it spreads, the impact it makes on populations and societies, and how best to respond in light of these facts. The problem has proved difficult to tackle, bringing in many of the world's leading experts in epidemiology, public health, and data science. This paper presents a method to predict transmission rates and mortality rates in the United States by county, based on a number of demographic and geographic features. The method combines a differential epidemic model, the Susceptible, Infected, Recovered, Dead (SIRD) model, with a Simulated Annealing Algorithm (SAA) to determine which transmission rate and mortality rate best fit the model to real data in each county. A Linear Mixed Effects (LME) model is then coupled with Forward Feature Selection (FFS) to determine which of each counties' features are most important in predicting these parameters. This novel approach aims to bridge the gap between idealistic predictions and the noisy, irregular raw data that is found in national databases, and then to shed light on what kinds of populations are most at risk in the pandemic. While the current methodology is observed to be insufficient to reach the desired conclusions, several useful conclusions are made which should inform future efforts to understand COVID-19 and its impact.

Introduction

COVID-19 has changed the structure of people's lives all over the world. With more than 3.4 million confirmed cases and over 240,000 fatalities worldwide [1], countries are doing what they can to stop the virus. The US is leading the world in the number of cases with more than 1.1 million cases and 66 thousand fatalities. In order to minimize the impact of the virus, a vaccine needs to be developed. However, vaccine development can take over 10 years, which the world can not afford [2]. As a result, alternatives are needed in the fight against COVID-19. In the age of big data, understanding the spread of COVID-19 and the mortality rate can guide legislators to the optimum legislation to help pass the COVID-19 crisis. As a result, the objective of this work will be to use data from different sources to predict the mortality and infection rates in different counties in the US.

Literature Review

In order to understand contemporary epidemiological models, many research papers were analyzed. In order to get a general understanding of machine learning models for epidemiological data, research paper [3] was used. This paper was beneficial in outlining the general process of studying epidemiological models with Machine Learning and Neural Networks. This paper gave a very general explanation in feature engineering, missing data, defining model performance, and etc. However, [3] was not good in the sense that it was too vague and did not show any simulations of the actual models or how to implement them on real problems.

Due to the vagueness of [3], another two papers [4] and [8] that took a more specific approach with machine learning using SVMs and Cellular Automata were used. In [4], the researchers used the SVM based classification of the SARS spatial distribution. The researchers were able to find a relationship between SARS occurrences and other variables such as environmental, societal (gross dependency ratio), and economic data (unemployment rate and tax revenue). Although the approach seemed outside the scope of the class, it gave good ideas in terms of the features needed to construct an epidemiological model, which were incorporated in this research. [8] is a relatively concise conference paper that simulated a model by reflecting the propagation process of infectious disease using Matlab. The main features used in [8] are age structure and population density, further helping with the feature selection in this work.

It was evident that after looking into the research, the more popular methods of epidemiological models are differential equation-based. [5] detailed a general overview in the mathematics of infectious diseases using differential equations, which included SIR and its variants. [5] also detailed many features used in the models, such as age structure, stages of infection, and spatial spread, which were incorporated in this work. The dynamical methods used in [5] became the basis of this work.

After reading [5], a more specific review was done involving the dynamic model approach to learn how this has been incorporated in similar problems to the one this work is trying to solve. [6] and [7] detailed the classic SIR framework to solve the problem. In [6], the seasonal transmission parameters, also known as the contact rate of childhood infectious disease, was calculated with the SIR model. [7] tackles a similar problem to that of [6], but uses a time-series, SIR, as opposed to a regular SIR. It became clear that using such a model on the Covid-19 was feasible and simple.

This literature review helped narrow down the process of model selection for Covid-19. At first, the idea was to use a purely Machine Learning-based approach, but after doing the review, this work shifted to incorporating the dynamical approach while incorporating Machine Learning aspects for the parameters to try to enhance the flaws of the SIR model.

Objectives/Methods

An informed response to epidemical threats can make a critical difference in how a population is affected with regards to health, quality of life, and economic stability. The goal of this project is to determine which communities are likely to experience exceptionally high or exceptionally low transmission rates and mortality rates from the pandemic based on social, geographic, and demographic features. By identifying important characteristics correlated with transmission rates and mortality rates, highly at-risk communities can be identified, and appropriate measures may be taken. Conversely, communities that exhibit low levels of risk may be allowed to continue important societal functions without interruption and without contributing to the spread of the pandemic. The true transmission rate and mortality rate of the virus must be known to develop these correlations.

The true transmission rate and mortality rate of a pandemic, as experienced by a population, is difficult to measure from recorded data alone. Discrepancies and irregularities in the number of tests performed and the number of cases confirmed can make the real spread of the virus difficult to track. While epidemiological models, such as the SIRD model, are not prone to these same problems, they are overly simplistic and only approach a true solution as the population size goes to infinity. Thus, the true parameters underlying the spread of the pandemic lie somewhere between the idealistic differential models and the noisy but discrete raw data.

Our objective in this project is to 1) find these true underlying parameters with the use of an SAA, and 2) find the features which are most important in determining these parameters by using an LME model with FFS. Populations were examined on a county/parish/borough level, with the goal of taking county demographic and geographic data and comparing it with the counties' estimated transmission rate and mortality rate. All codes were implemented in MATLAB using standard libraries.

The parameters which model the pandemic, transmission rate, and mortality rate are estimated for each US county by optimizing the SIRD model such that it closely resembles reported data. An SIRD model was implemented in MATLAB, and its predicted deaths for a given county over time were compared to the real number of deaths in that county. Measuring the Mean Absolute Error (MAE) in these predictions for the history of the pandemic quantifies how accurate or inaccurate the models' parameters are in predicting the observed deaths. Thus, by allowing for transmission rate and mortality rate to vary and holding other demographic data (e.g., initial total population) constant, it is possible to find which pair of parameters best describes the observed data with an SIRD model. In this way, the gap between the differential model and raw data may be bridged.

It is worth noting that time was normalized for this data so that counties could be compared on a similar epidemical calendar. Day zero on this calendar is the day that there were at least 10 recorded deaths due to COVID-19, and all deaths are measured sequentially from day zero. In this way, the SIRD model begins from a similar statistical point for each county. Additionally, it was decided to measure error using MAE rather than the Mean Squared Error (MSE). After initial trials, it was observed that the MSE over-prioritized outliers, causing weak parameter optimization. Using MAE allowed for the error to be evenly composed of data points, with the effects of far outliers being smoothed out by other strong matches.

The transmission rate/mortality rate pair that minimizes the MAE between predicted daily deaths from COVID-19 and observed daily deaths from COVID-19, for each county, was found using an SAA. Simulated Annealing was selected for this parameter optimization search because it allowed for an effective search of the MAE landscape without greedily pursuing local minima. If more parameters were being optimized, a Genetic Algorithm would have been considered. But because of the low dimensionality of the search space and thus the limited number of 'genes' to be selected, Simulated Annealing was determined to be a more appropriate algorithm. The transmission rate and mortality rate were found and recorded by iterating through each county. Counties that exhibited exceptionally poor statistics, having high MAE relative to the population, were omitted from the dataset. In the end, mortality rate and transmission rate data for 168 counties were passed into the LME model for feature selection.

The LME model was used to estimate the coefficients used to calculate the infection and mortality rates. Mixed models were developed to deal with complex data sets. This model is useful in estimating the desired output given different features. However, there are certain assumptions in the linear mixed-effects model that need to be understood. These assumptions are:

- The model assumes a linear relationship between the dependent and the independent variables.
- There is no collinearity in the data.
- There is no heteroskedasticity, meaning that it assumes that the variance within the data is the same.
- No influential data points are present.
- Data are completely independent. [9]

These assumptions need to be considered when using this model. It should be noted that no random effects were because they involve categorical variables that were not encountered in the current work. [10]

The features used in this work are summarized in table 1.

Table 1: Features used in this work.

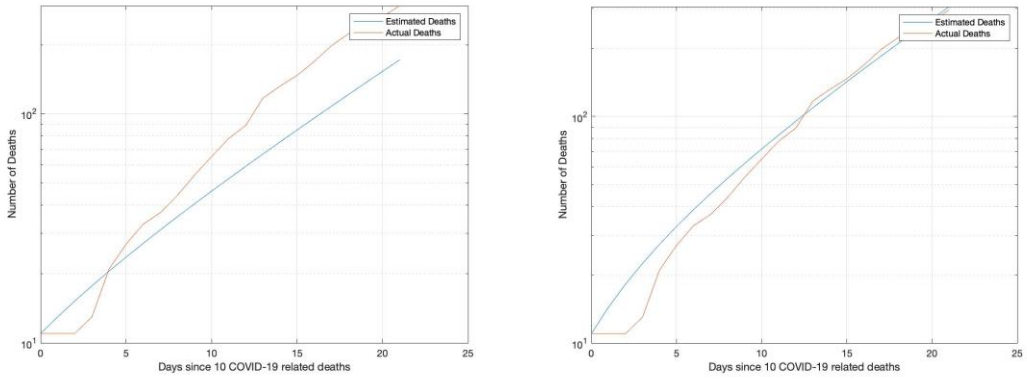
Feature	Reference
County Populations	https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
Urbanization Scale	https://www.kaggle.com/jieyingwu/covid19-us-countylevel-summaries
Poverty Level	https://www.kaggle.com/lithomas1/us-socioeconomic-indicators-by-county#countypoverty.csv
Median Household Income	https://www.kaggle.com/lithomas1/us-socioeconomic-indicators-by-county#countypoverty.csv
Number of ICU Beds	https://www.kaggle.com/jaimeblasco/icu-beds-by-county-in-the-us’
Housing Structures with 10 or More Units	https://www.kaggle.com/dannellyz/2018-cdcs-social-vulnerability-index-svi#2018_CDC_SVI.csv
Shelter in Place Orders	https://www.kaggle.com/hikmahealth/hikma-health-covid19-us-county-policies
Education Level	https://www.kaggle.com/jieyingwu/covid19-us-countylevel-summaries
County Temperature	https://www.kaggle.com/jieyingwu/covid19-us-countylevel-summaries
International and Domestic Migration	https://www.kaggle.com/lithomas1/us-socioeconomic-indicators-by-county#countypoverty.csv
Density Population	https://www.kaggle.com/mmcgurr/us-city-population-densities
Number of People in Each Age Group	https://www.kaggle.com/mchirico/population-by-age-and-race

Other features of interest were information about transportation and testing data. Unfortunately, no information was found on the county level for these two features. The features were then normalized and centered in order to fairly compare different features that have different scales. This was performed by subtracting the mean and dividing by the standard deviation of the results. Each feature had a mean of zero and a standard deviation of one.

In order to select which features are relevant in estimating the infection and mortality rates, the forward feature selection technique was implemented. Two categories were selected in choosing features, the coefficient, and the p-value. Features with the largest coefficient were picked if the p-value was within the acceptable range. Features were added until adding more features did not provide any significant p-values. At that point, the feature selection stopped, and the features used until that point were picked in estimating the mortality and infection rates.

Results and Discussions

The SAA parameter estimation process worked to match the SIRD model to raw data, minimizing the MAE between predicted deaths and actual deaths. An example of how this optimization process works is shown in Figure 1, where an SIRD model using the standard assumed transmission rate (2.5) and mortality rate (5%) is compared to an SIRD model with optimized parameters.



$r_0 = 2.50, m.r. = 5.00\%$

$r_0 = 2.53, m.r. = 8.42\%$

Figure 1: The SIRD model with standard parameters is compared with an SIRD model with parameters optimized by the SAA for Los Angeles county data.

Not all results from the SAA worked well in this optimization scheme. Some counties did not have deaths increasing in an evenly exponential manner and were quite disjointed. It is natural that the SIRD model could not fit these counties well. One county had an especially bad performance, and it was omitted from the analysis. The average MAE for each counties' SIRD model is 3.35 deaths on any given day, with a standard deviation of 7.38 deaths.

Three different results were obtained using the linear mixed-effects model in MATLAB. The first result corresponds to data obtained from the SAA, while the secondary results are from another parameter estimation method. The third results used raw county for deaths to correlate with the features from the FFS-LME.

First Results

The first result used the parameters from the simulated annealing algorithm (SAA) with the SIRD model. The goal was to predict the infection rate (R0) and the mortality rate from the collected county features. The features were selected using the forward feature selection algorithm (FFS) in the Appendix. Initially, a p-value of 0.05 was the maximum that was accepted, because this is the threshold that determines statistical significance in the data, anything above that is a sign of a weak correlation. The FFS and LME were run to predict the mortality rate from the features, but the algorithm could only find one feature alone that was of statistical significance, which happened to be the number of ICU beds in the county. The p-value for the ICU beds was calculated as 0.018 with a linear coefficient -0.18, indicating a negative correlation. It was thought that this negative correlation for the coefficient shows that as the number of ICU beds in a county increases, the mortality rate decreases. Rather than being interpreted this way, this feature could be a sign that the number of ICU beds is an underlying indication of the overall healthcare system in the county. That is, counties with more ICU beds simply have better health care systems, which in turn decreases the mortality rate for Covid-19.

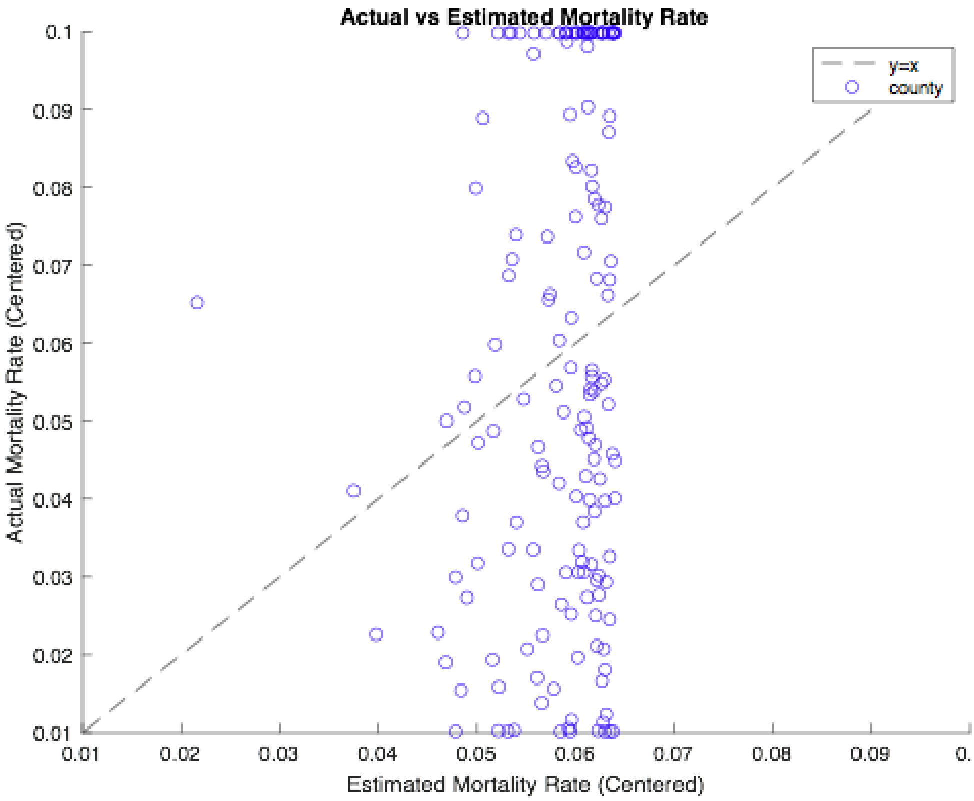


Figure 2: US County Mortality Rate, P < 0.05.

If the model was capable of accurately predicting the mortality rate, then that data in Figure 2 should fall roughly along the dotted line, within statistical significance. However, there is no correlation evident in Figure 2. This implies meaningful correlations were not found by LME/FFS methodology and that further analysis is needed.

The infection rate (R0) was also modeled using the same technique behind the model in Figure 2, with a p-value threshold of 0.05. Only one feature was selected, which was the percent of adults completing some college or associate degree. The p-value was calculated as 0.0418 with a linear coefficient -0.16. This indicates that as education increases, the infection rate will decrease.

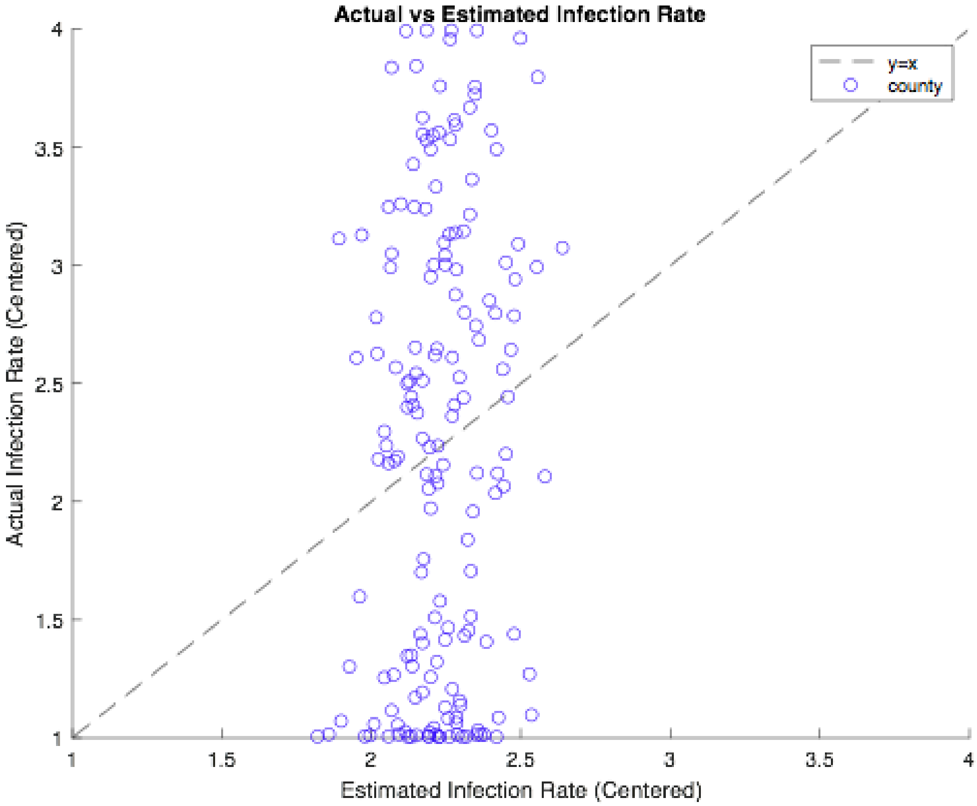


Figure 3: US County Infection Rate, P<0.05.

The same interpretation for Figure 2 is used for Figure 3. There is no correlation evident.

With the proper p-value threshold, only one feature was observed to affect the models. Therefore, in order to observe the importance of other features, the p-value threshold was increased to 0.1 as opposed to 0.05, which allows for the model to select more features in the FFS-LME. Doing this with the new p-value threshold of 0.1, the features modeling the infection rate go unchanged (same as Figure 3), which is the percent of adults completing some college or associate degree. This shows that the correlation of the features with the infection rate is unfortunately very poor. However, the features for the mortality rate change considerably with the new p-value threshold. The FFS algorithm was able to find five features for predictor variables, which are the ICU beds, Average April Temperature, Feb Temperature, Net Migration, Total Age over 85. The linear coefficients in order are, -0.15, -0.34, 0.44, 0.22, -0.34 and p-values are 0.052, 0.048, 0.012, 0.098, 0.013. The p-values are within the limits as specified by the algorithm, which is good. Looking at the coefficients, once again, ICU beds cause a decrease in the mortality rate. Temperature is an interesting variable here. Areas with higher April temperature average saw lower mortality rates, while areas with higher February temperatures saw increased mortality rates. Although this may simply be an issue of having poor relatively weak p-values, it may also show that counties with warmer temperatures in February, a time where prevention policies were not in action, people would simply tend to go outside more. However, a sufficient explanation for the negative correlation cannot be said. Neither can an explanation be made for the negative correlation with Total Age over 85, this may simply be bad data. The Net Migration coefficient of 0.22 being positive makes sense. An explanation is that a county in which people travel a lot causes the disease to spread more quickly. The results of the LME model from these five features is shown in Figure 4, no strong correlation is evident.

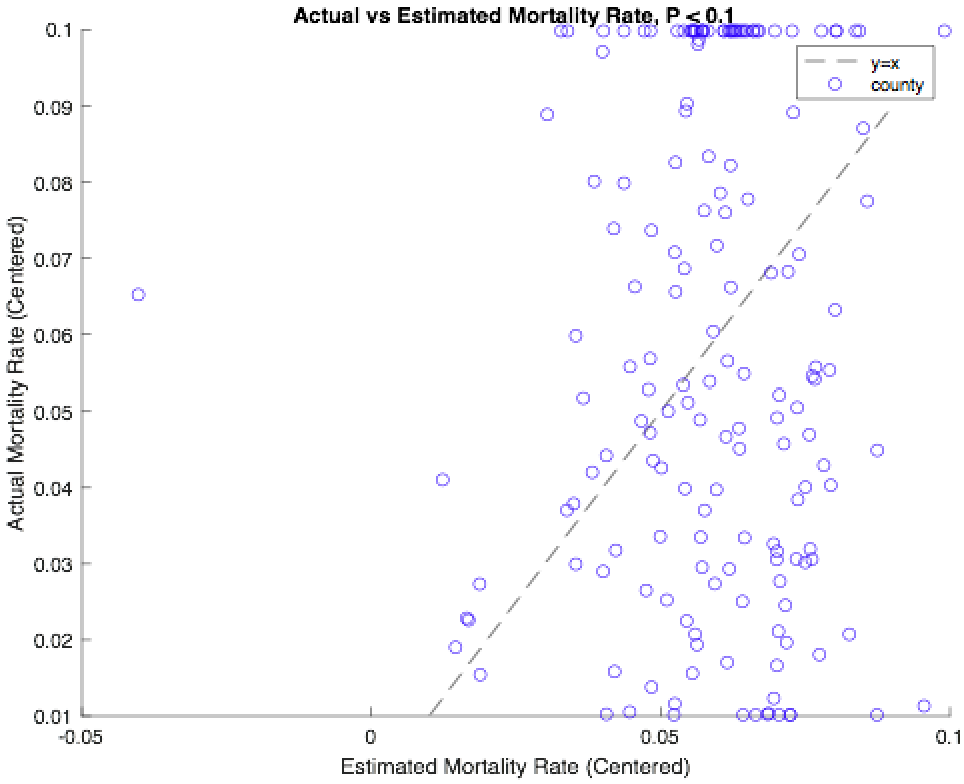


Figure 4: US County Mortality Rate, P < 0.10.

Second Results

Part of the reason the SAA is believed to have given poor results in the parameter estimation is that it was allowed to modify the mortality rates freely from a wide range, causing it to deviate from realistic mortality rate data. While data exists for the observed mortality rate in each county, the transmission rate is a true variable and can only be estimated. Thus, it was decided to use raw data for the mortality rate and take the transmission rate as the variable to be optimized.

As this is now a one-dimensional optimization rather than a two-dimensional optimization, significantly fewer data points are needed to effectively test the entire parameter space for a global minimum. An SAA was not used, but rather 50 evenly spaced samples of transmission rate were taken between 1.0 and 3.5. The SIRD model was evaluated for each county with each of these samples, and the one yielding the lowest MAE was selected. After extreme outliers were omitted, the transmission rates estimated for each county were saved to a file and passed on the LME/FFS process.

After using the iterative optimization to find the transmission rates, the LME/FFS method was utilized to find a relationship between the features and the transmission rate. The feature selection here did not perform as well as the results in Part 1; that is, the p-value threshold had to be increased to 0.2 to get any meaningful features to be selected. After running the algorithm with 0.2 as the maximum, it was seen that two features exhibit correlation within this range, which are Percent of Adults with Less than a Highschool Diploma, and Rural/Urban Continuum. The p-values are 0.16 and 0.01, and the linear coefficients -0.17 and 0.26. This does not make sense because the data is saying that as education gets worse, then the infection rate goes down, but in reality, it should be the opposite. It's likely that people with higher education levels are more likely to take the precaution of limiting exposure to the virus. Furthermore, the Rural/Urban continuum is also counter-intuitive. Rural/Urban continuum is a measure from 1-9 of how rural a county is, with 9 being the most rural. The data here is telling us that as the country becomes more rural, the infection rate increases when, in reality, it is more likely to be the opposite. The model is shown in Figure 5, once again showing no correlation.

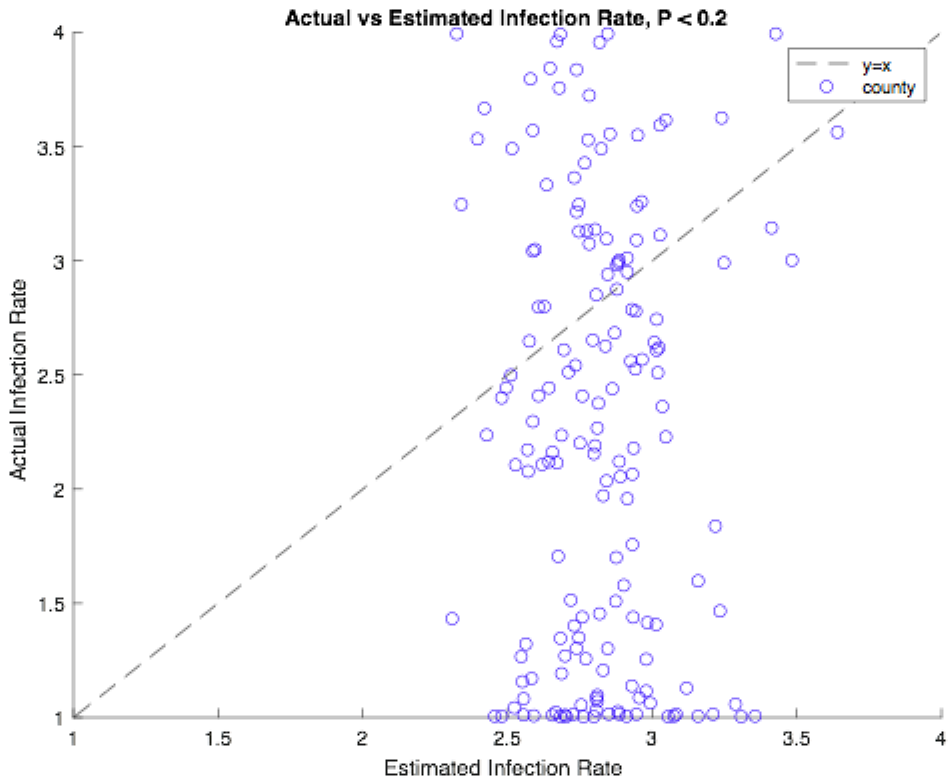


Figure 5: US County Infection Rate with Known Mortality Rate Known, P < 0.2.

Third Results

After realizing that there could be an error in the parameters themselves, an FFS-LME was run on the latest deaths and infection numbers (raw data) for each county in the US, so no parameter estimation is done in this case. For deaths, the p-values turned out to be much better than the previous results, with 11 features being selected with p-values less than 0.05. Although the p-values were good, the coefficients did not make sense, with a coefficient of -0.0310 for Total Age over 65, for example. It is known that elderly people are more affected by the virus, so the correlation should not be negative. Nonetheless, the results of the multicoefficient combined with the feature vectors are shown below (Figure 6); the results are not as they should be.

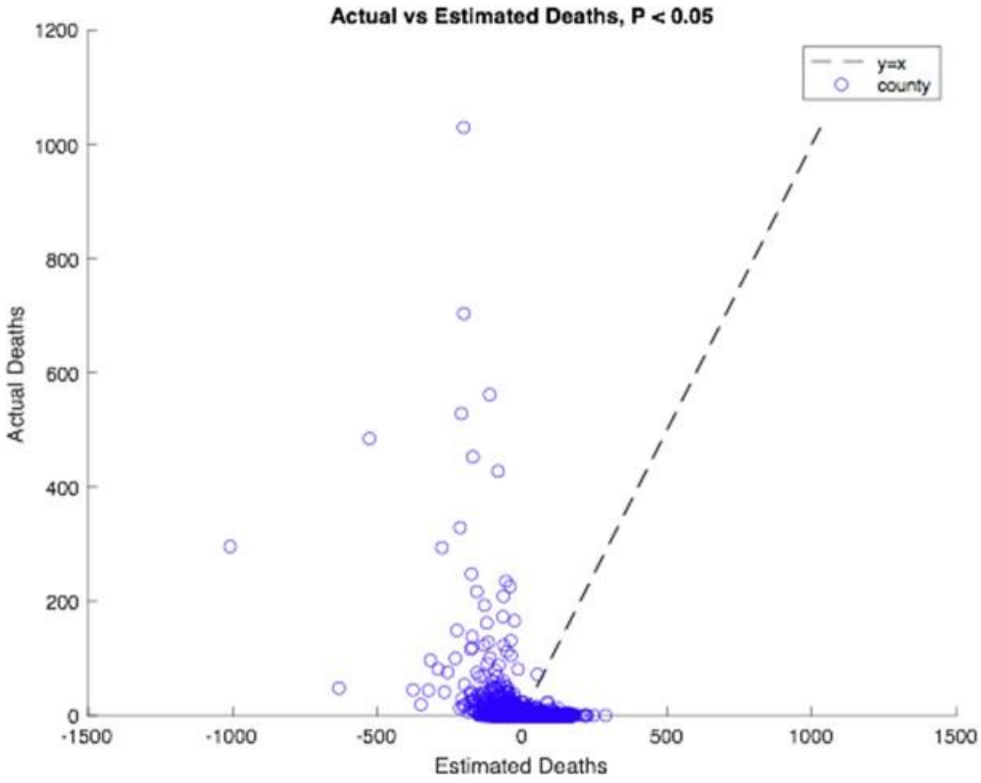


Figure 6: US County Deaths LME, P < 0.05.

The same process for the deaths was done for the number of cumulative cases. The model performed better in terms of finding features with good p-values, with a selection of 10 features with p-values less than 0.05 The coefficients unfortunately were once again counterintuitive, with a -0.0478 coefficient for Total Age over 65, which is opposite of what we're seeing in real life; old people are more susceptible to being infected. The LME model is shown in (Figure 7), once again, not ideal.

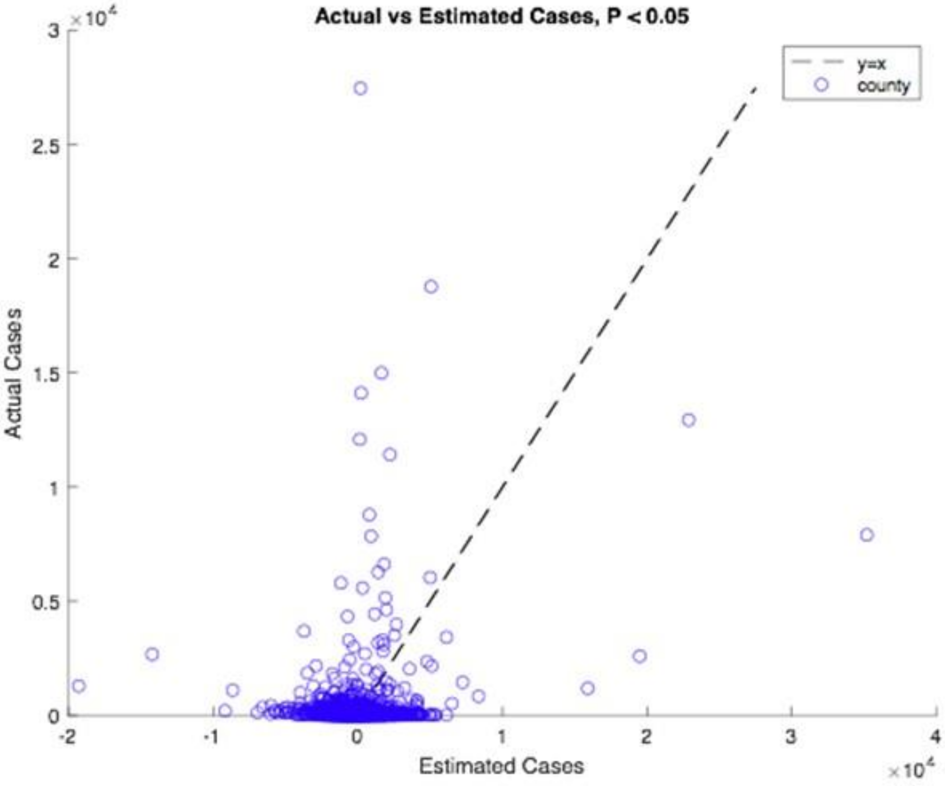


Figure 7: US County Cases LME, P < 0.05.

After going through third results, it can be seen that the issue is not necessarily with the parameter estimation; it could be in other areas such as underlying factors in the LME model or the FFS selection. Perhaps the combination of features here are not enough to determine a proper estimation of any parameters (infection rate, mortality rate, number of infections, number of deaths); the problem could simply be too complex.

Conclusions and Recommendations

The results obtained from this approach do not provide any immediately meaningful conclusions. The LME was not able to accurately predict mortality rate or transmission rate from the best-selected features, and the p-values associated with features indicate that the features are insignificant in determining the quantities of interest. This is due in part to the results gathered from the SAA and iterative parameter estimation. Many counties' predicted mortality or transmission rates tend to cluster around the upper and lower bounds of the parameter's range, indicating poor behavior from the parameter estimation process.

The driving factor in this process was the MAE between the SIRD model and the death rate data for each county. It is possible that the poor results communicate the inability of the SIRD model to predict the actual transmission of the virus. And this is entirely likely. As discussed previously, the SIRD has many assumptions that do not correspond to physical reality. For example, the model gains accuracy as the size of the population increases to infinity. The virus is assumed to spread continuously over time, rather than in discrete random transmissive events. And a person who is infected is assumed to be contagious for a fixed amount of time and is just as likely to be killed by the virus as someone who is elderly or in poor health. These and other assumptions cause the model to deviate from a real representation of epidemic spread. Thus, the SIRD model introduces harsh dissimilarities with observed death rates that cannot be reconciled by an SAA or iterative parameter estimation process.

Errors also may have arisen in the LME/FFS process. Although 24 features were included in this model, only seven showed any statistical significance in predicting mortality rate and transmission rate. Some important demographics could not be found at the county level. The use of public transportation, for example, should theoretically be a significant indicator of transmission rates. While this information was available for some cities, it was not available for counties and thus was not included in the analysis. The model would benefit from an expanded set of features and a more robust feature selection process.

This project is an important step in discovering the challenges and limitations associated with epidemiology. Future efforts to identify key demographic and geographic features should use epidemic models more complex than the SIRD to evaluate the accuracy of parameter estimation. When the mortality rates and transmission rates can be reliably estimated with an effective epidemic model and optimization algorithm, a more robust mixture model should be investigated to map the feature space to the parameters. Then, after expanding the feature set, interesting results may be obtained, which will shed important light on which communities are most prone to epidemic spread. With this information available, it will be possible for governments and civil authorities to make highly informed decisions in response to future epidemics and similar outbreaks, resulting in improvements to public safety, quality of life, and economic stability.

References

[1] Johns Hopkins Coronavirus Resource Center. 2020. *COVID-19 Map*. [online] Available at: <https://coronavirus.jhu.edu/map.html> [Accessed 3 May 2020].

[2] Gouglas, D., Thanh Le, T., Henderson, K., Kaloudis, A., Danielsen, T., Hammersland, N., Robinson, J., Heaton, P. and Røttingen, J., 2018. Estimating the cost of vaccine development against epidemic infectious diseases: a cost minimisation study. *The Lancet Global Health*, 6(12), pp.e1386-e1396.

[3] T. L. Wiemken and R. R. Kelley, "Machine Learning in Epidemiology and Health Outcomes Research," *Annual Review of Public Health*, vol. 41, no. 1, pp. 21-36, 2020/04/02 2020, doi: 10.1146/annurev-publhealth-040119-094437.

[4] B. Hu and J. Gong, "Support vector machine based classification analysis of SARS spatial distribution," in *2010 Sixth International Conference on Natural Computation*, 10-12 Aug. 2010 2010, vol. 2, pp. 924-927, doi: 10.1109/ICNC.2010.5583921.

[5] H. W. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review*, vol. 42, no. 4, pp. 599-653, 2000/01/01 2000, doi: 10.1137/S0036144500371907.

[6] D. P. Word, J. K. Young, D. Cummings, and C. D. Laird, "Estimation of seasonal transmission parameters in childhood infectious disease using a stochastic continuous time model," in *Computer Aided Chemical Engineering*, vol. 28, S. Pierucci and G. B. Ferraris Eds.: Elsevier, 2010, pp. 229-234.

[7] B. F. Finkenstädt and B. T. Grenfell, "Time series modelling of childhood diseases: a dynamical systems approach," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 49, no. 2, pp. 187-205, 2000.

[8] S. Zhou, S. Bin, and G. Sun, "Modeling and Simulation of Infectious Propagation Based on Cellular Automata," in *2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 31 May-3 June 2019 2019, pp. 28-30, doi: 10.1109/ECBIOS.2019.8807868.

[9] B. Winter, "Linear models and linear mixed effects models in R: tutorial 11," *arXiv preprint arXiv:1308.5499*, 2013.

[10] G. K. Hajduk. "Introduction to Linear Models." https://ourcodingclub.github.io/tutorials/mixed-models/ (accessed 05/03/2020, 2020).

Appendix

Table of Contents

```
% CSCE 666
% COVID-19 Project
% Iterative parameter estimation
%
% The SIRD model estimates deaths from an epidemic using an ODE.
% The solution to the ODE is parameterized by:
% r0 : reproduction number
%
% This script finds the transmission rate which best fits the SIRD
 model
% to real data for 210 counties in the USA using an iterative
 approach.
%
% The resulting r0 for each county is saved in the file 'r0_list.csv'
% for further analysis and feature correlation. Counties with poor
% statistics are omitted.

clear; clc;
```

Load Files, Prepare Data

```
population   = csvread('us-county-population_2.csv');
num_counties = length(population);
FIPS = population(1,:);

cases_data = readtable('us-county-cases_2.csv');
mortality_rates = csvread('Mortality_Rate_Summary.csv');

% Specify the SAA options for in-training monitoring
options = optimoptions('simulannealbnd','PlotFcns',...
          {@saplotbestx,@saplotbestf,@saplotx,@saplotf});
options.MaxStallIterations = 500;
```

Evaluate Transmission Rates, Find Optimum for Each County

Create a variable to store parameter values for each county

```
params = zeros(4, num_counties);
params(1,:) = FIPS;
```

```
r0_list = linspace(1.0, 3.5, 50);
trials = zeros(length(r0_list), num_counties);
for county = 1:num_counties-1
    % Population of county
    d = mortality_rates(2, county);
    pop = population(2, county);
    init_cases = cases_data.(county);
    Iinit = init_cases(2);

    fprintf('\nCounty: %g\nFIPS:   %g\nM.R.:   %f\n', [county,
 FIPS(county), d])
    % If ONLY optimizing transmission rate:
    for i = 1:length(r0_list)
        x = [r0_list(i) d county pop Iinit];
        trials(i, county) = mse_sir(x);
    end
    [~,ind] = min(trials(:, county));
    fprintf('r0:     %f\n', r0_list(ind))
end
```

Export Results

```
r0_eval = zeros(3, num_counties);
r0_eval(1, :) = FIPS;
for county = 1:num_counties
   [mser,ind] = min(trials(:, county));
   r0_eval(2, county) = r0_list(ind);
   r0_eval(3, county) = mser;
end

tbd = [];
for county = 1:num_counties-1
    if r0_eval(2, county) == 1
        tbd = [tbd county];
    elseif r0_eval(2, county) == 2
        tbd = [tbd county];
    end
end

r0_eval(:,tbd) = [];
csvwrite('r0_list.csv', r0_eval)
```

Published with MATLAB® R2018a

Table of Contents

```
% CSCE 666
% COVID-19 Project
% SAA parameter estimation
%
% The SIRD model estimates deaths from an epidemic using an ODE.
% The solution to the ODE is parameterized by:
% r0 : transmission rate
% d  : mortality rate
%
% This script finds the paramter values which best fit the SIRD model
 to
% real data for 210 counties in the USA using a Simulated Annealing
 Algorithm.
%
% The resulting parameters for each county are saved in the file []
% for further analysis and feature correlation.

clear; clc;
```

Load files, prepare data

```
population   = csvread('us-county-population_2.csv');
num_counties = length(population);
FIPS = population(1,:);

cases_data = readtable('us-county-cases_2.csv');
mortality_rates = csvread('Mortality_Rate_Summary.csv');

% Specify the SAA options for in-training monitoring
options = optimoptions('simulannealbnd','PlotFcns',...
          {@saplotbestx,@saplotbestf,@saplotx,@saplotf});
options.MaxStallIterations = 500;
```

Prepare and Run Simulated Annealing Algorithm

Create a variable to store parameter values for each county

```
params = zeros(4, num_counties);
params(1,:) = FIPS;
r0_list = linspace(1.0, 3.5, 50);
```

```
trials = zeros(length(r0_list), num_counties);
for county = 1:num_counties
    % Population of county
    pop = population(2, county);
    init_cases = cases_data.(county);
    Iinit = init_cases(2);

    fprintf('\nCounty: %g\nFIPS:   %g\n', [county, FIPS(county)])
    % Initial guess for [r0 d] for a given county with population
 'pop'
    x0 = [2.5 0.05 county pop Iinit];
    % bounds defined as: [r0 d]
    % note that 'county' and 'pop' are not allowed to change
    lb = [1.0 0.01 county pop Iinit];
    ub = [4.0 0.10 county pop Iinit];
    [x, mserror] = simulannealbnd(@mse_sir, x0, lb, ub, options);
    params(2, county) = x(1);
    params(3, county) = x(2);
    params(4, county) = mserror;

    fprintf('r0  = %f \nd   = %f\nPop = %e\nMSE = %f\n', [x(1), x(2),
 pop, mserror])
end
```

Export Results

```
csvwrite('params.csv', params)
```

Published with MATLAB® R2018a

Table of Contents

THIS IS THE IMPLEMENTATION OF THE MIXTURE MODEL WITH FORWARD FETAURE

SELECTION

```
clc, clear all, close all
% READ THE DATA INTO MATLAB AND STORE THE FEATURE NAMES IN A SEPRATE
% VARIABLE
B=readtable('Excel_Sheet_Covid19-filtered_normalized.csv');
var_names=B.Properties.VariableNames;
% INITIALIZE A BOOK OF USED FEATURES AND OTHER VARIABLES FOR THE WHILE
 LOOP
usedfeatures=[];
p=[.01;.01];
counter=0;
coeffinal=0;
index=0;
pval=0;
```

FFS ALGORITHM and Linear Mixed Effects Model

```
while max(p(2:end,:))<0.2 & max(p)>0
    coeffinalfinal=coeffinal;
    pvalfinal=pval;
    ii=index;
    indices=[];
    counter=counter+1;
    clear coef
    clear pval
    for i=7:31
        if ismember(i,usedfeatures)==false
            st1=var_names(i);
            st1=char(st1);
                if counter==1;
                    st="R0 ~ 1 +" + st1;
                else
                    st="R0 ~ 1 +" + newstring + st1;
                end
            lme = fitlme(B,st);
            x1=lme.Coefficients(:,2);
            coef(:,i-6)=double(x1);
            x2=lme.Coefficients(:,6);
            pval(:,i-6)=double(x2);
        else
            continue
        end
    end

    coeffinal=coef;
    coef=abs(coef);
    maxvalue=max(coef(end,:));
    index=find(coef(end,:)==maxvalue);
    p=pval(:,index);
    [aa,indices]=sort(coef(end,:),'descend');
    counter2=1;

    while max(p(2:end,:))>0.2 & max(p)>0
        counter2=counter2+1;
        maxvalue=aa(1,counter2);
        index=indices(1,counter2);
        p=pval(:,index);
    end
    index=index+6;
    usedfeatures=[usedfeatures;index];
    if counter==1
       newstring=char(var_names(index));
       newstring=newstring+"+";
    else
       usedstring=char(var_names(index));
       newstring=newstring + usedstring + "+";
    end
end
```

Published with MATLAB® R2020a

```
function  e = mse_sir(in)
    % This function evaluates the Mean Squared Error in an SIRD Model
    % given a variable transmission rate 'r0' and mortality rate 'd'.
    % The MSE is measured between predicted deaths and recorded deaths
 due
    % to COVID-19 on each day past Day 0. Day 0 = first day at which
 there
    % were ar least 10 deaths due to COVID-19.

    % Load and format data
    r0 = in(1);
    d  = in(2);
    N  = floor(in(4));
    county = floor(in(3));

    county_data = readtable('us-county-death_2.csv');
    death_data  = county_data(2:end,:);

    deaths = death_data.(county);
    deaths(isnan(deaths)) = []; % eleminate NAN values
    days   = length(deaths);
    tspan  = linspace(0, days-1, days);

    k = 1./14;
    b = r0*(1+d)*k;

    % Set up ODE
    f = @(t,y) [-b*y(1).*y(2)./N; (b.*y(1).*y(2)./N)-(k.*(1+d).*y(2));
 k.*y(2); k.*d.*y(2)];

    Iinit = in(5);
    Rinit = 0;
    Dinit = deaths(1);
    xinit = [(N-Iinit-Rinit-Dinit) Iinit Rinit Dinit];
    [T,Y] = ode45(f, tspan, xinit);

    % Calculate MSE
    e = immse(Y(:,4), deaths);
end
```

Published with MATLAB® R2018a