Predicting COVID-19 Confirmed Cases Growth Curves in U.S. Counties

By Davis Ulrich, Andrew Kaplan, & Victoria Austin

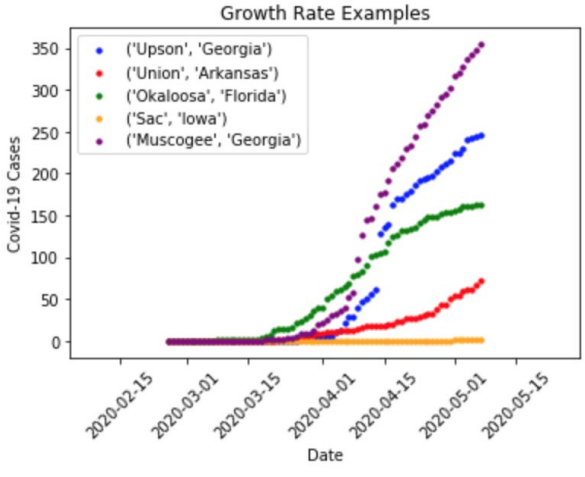
Abstract

The COVID-19 epidemic has abruptly and drastically affected everyone's lives on a global scale. With the uncertainty of not knowing when life will return to normal, our final project aims to find some tangible predictions of how COVID-19 will spread at the U.S. county level. Our research question asks, given a particular county's demographic data, can we generalize the growth curve of confirmed COVID-19 cases in similar counties to predict the growth curve of that particular county without information about its real growth? We used nonlinear least squares logistic growth to model the growth curve of the number of confirmed COVID-19 cases per county. Our results in general, are successful at predicting counties with small or average population sizes, with larger counties tending to be underpredicted. Our research suggests that counties with similar population size, geographic proximity, population density, along with other demographic features are good estimators for predicting the logistic growth for the number of confirmed cases for counties with no available COVID-19 data.

Introduction

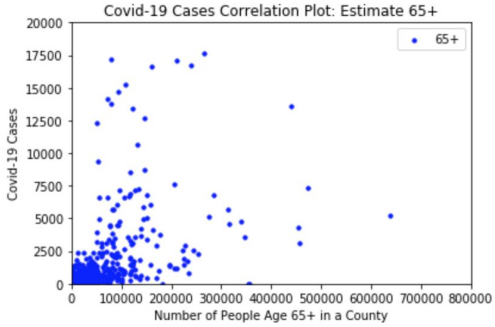
From the beginning of our exploratory data analysis, our team was set on predicting something more complex and applicable than the number of COVID-19 cases or deaths in a county. Albeit valuable, this information doesn't tell us anything about tomorrow, or the next week, or the next month of coronavirus cases that we can expect to see. This is what drove our research question, given a particular county's demographic data, can we generalize the growth curve of confirmed COVID-19 cases in similar counties to predict the growth curve of that particular county without information about its real growth? Being able to accurately estimate the logistic curve for a county's growth will give us an idea of how the virus will spread within that particular population over time. Through extensive exploratory data analysis,data cleaning, data transformation, model designing, and model fitting, we landed on a model that is able achieve our objective.

Here are some examples of different growth curves that we set out to predict:



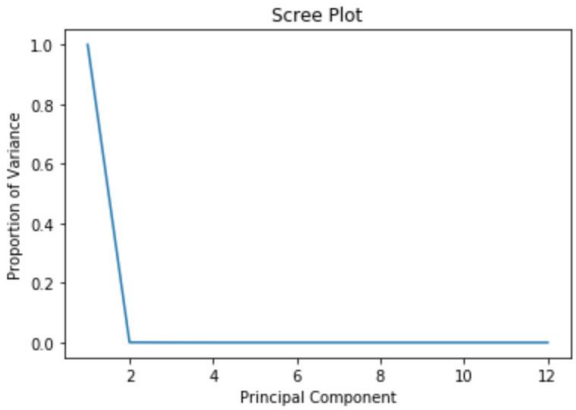
Exploratory Data Analysis

Our exploratory data analysis from a broad perspective consisted of understanding important features, extensive data cleaning, and principal component analysis. First, we updated our data to include all dates up until May 7th in order for our model to reflect the most accurate data as possible. Out of the five datasets provided, we only ended up using the county demographic data and the time series of confirmed cases. Our choice to predict COVID-19 cases instead of deaths was a consequence of cases having larger numbers that allow a more robust model. Between these two datasets, the FIPS code proved to be the most specific common feature and was therefore utilized as the primary key. To determine the value of each feature to our model, we initially constructed correlation plots for each feature. The most valuable features were distinguished by a graph similar to that on the left, and the others appeared to have no correlation with COVID-19 cases, such as on the right.



Next, in our data cleaning processes, we performed many crucial tasks such as imputing data into the columns we sought to keep, dropping the remaining null values / corrupting rows, and splitting New York City into its respective counties. We imputed data into columns that had null values but were still useful to our model, and dropped columns that did not add to our predictive accuracy. During the process of understanding relationships between columns, New York proved to be an extreme outlier and major influence on the prediction of many counties' growth curve. For this reason, cleaning the data collected about New York City was specifically important, so we transformed it to include further detail about each county than was provided.

Lastly, to determine the weighting scheme for features in our model, we performed principal component analysis and found that the population estimate for each county held the vast majority of the variance (~99%) and was therefore weighted heavier than other features. This weight was especially influential in finding similar counties that were used to estimate the new growth curve. The scree plot of the data is shown below, which showcases the variance contained in each principal component.



Methods

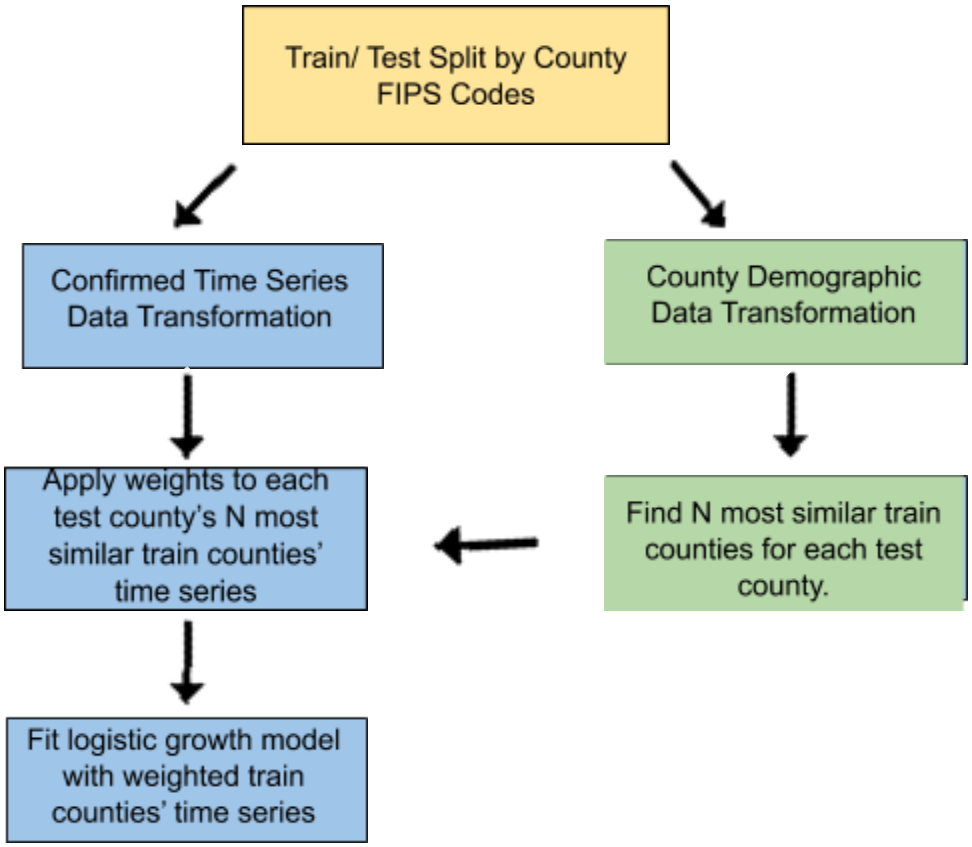
In order to model the growth curves of county's confirmed COVID-19 cases, we decided to use a logistic growth model since it characterizes increasing growth at earlier time periods, but decreasing growth at later periods, as the number of cases approaches its maximum. This is appropriate for modeling a county's COVID-19 confirmed cases, since the maximum possible number of confirmed cases is the total population size for that county. There has been significant research that shows that the beginning period of an epidemic shows exponential growth and the total period of an epidemic follows logistic growth. Since we are interested in being able to predict the total period of COVID-19 cases in a county, a logistic growth model is the logical choice.

$$y(t) = \frac{c}{1 + a * e^{-bt}}$$

The formula for logistic growth is shown above. The aim for our model is to be able to predict coefficients a, b, and c that best represent the logistic growth for a county, with 'a' being the initial number of confirmed cases, 'b' being the initial growth rate, and 'c' being the maximum possible number of confirmed cases. Since it is not possible to rewrite a logistic function as a linear regression, we used Scipy's curve fit for nonlinear least squares estimation to find the a, b, and c coefficients that minimizes the least square error of our data's logistic growth curve. Scipy's curve fit function allows us to set the initial values of a, b, and c and their lower and upper bounds before attempting to fit our model. We set the initial value of 'a' to be the first entry of our confirmed cases, 'b' to be 0.057 according to previous COVID-19 studies, and 'c' to be the maximum population size out of our predicted counties, since our model has a tendency of underpredicting growth rates for counties with large populations. We set the

lower bounds for a, b, and c to be 0, because it is not possible for any of these metrics to be below zero. The upper bounds for ‘a’ is 100,000,000 , for ‘b’ is 5, and for ‘c’ is 1,000,000,000,000,000, in order to mitigate our model’s tendency to underpredict logistic growth rates. A general overview of our logistic growth model pipeline is shown below.

Logistic Growth Model Pipeline

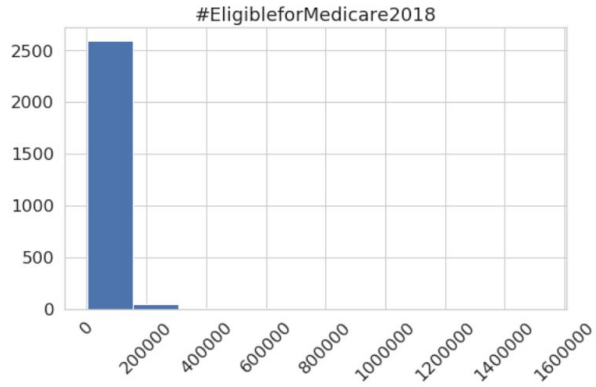


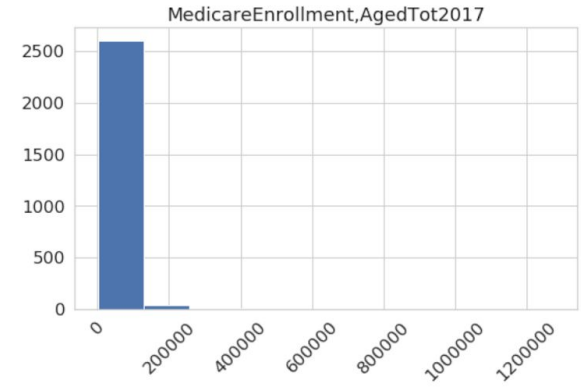
The first step in our logistic growth model pipeline is to split the confirmed time series data into train and test data frames by county FIPS codes. For our model, the testing data is the counties whose logistic growth curves we would like to predict and our training data is the counties we want to use for fitting and predicting our model. We used Sklearn’s model selection train test split function with a test size of 20% and a random state of 83. We made our test size to be 20% of our original data so that we could ensure our model is generalizable to any county in the U.S and that we aren’t overfitting to any specific set of counties.

The second step in our model pipeline is to prepare the confirmed time series data and county demographic data to be used in our model. For the time series data, we collect the counties' FIPS codes and all the dates columns and take the cumulative sum across all the dates to get the total number of confirmed cases per day for each county. We are using the county demographic data to find the most

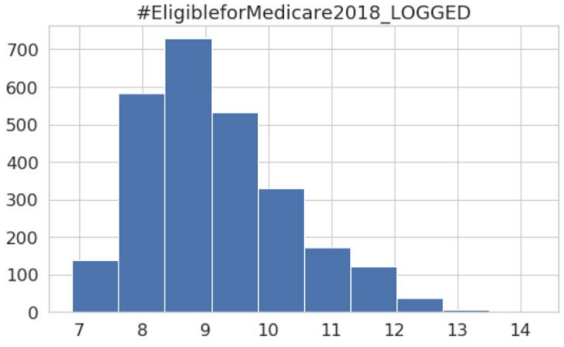
similar counties between our training data and our testing data. We use these similarity rankings to determine which counties should be used in our prediction, and how much weight should be applied to their time series data. To calculate similarity between counties, we used the numeric features in our county demographic data that are the most correlated to the number of confirmed cases' growth rates, as determined by our exploratory data analysis. Hence, we collect these features' columns along with their FIPS codes and we log the features that show significant skew in order to normalize the variance across observations. The effects of our log transformations are shown in the histograms below, with the x-axes being the number of people eligible for Medicare and the number of people enrolled in Medicare.

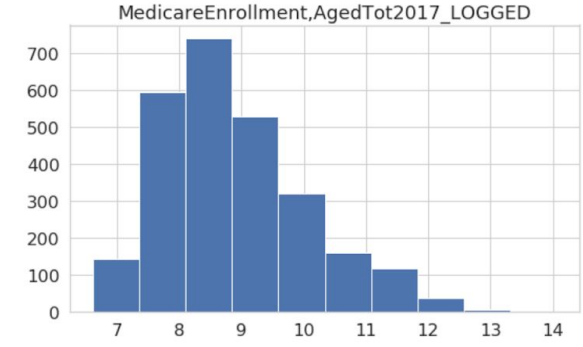
Columns Before Logging





Columns After Logging





As you can see, logging these columns helped to normalize these columns' distributions. Next, we take the z-score of each of our training data's county demographic features and use their mean and standard deviation to z-score our testing data's county demographics. Standardizing our data allows us to compare demographic features across our train and test counties with different normal distributions, leading to more accurate county similarity predictions.

The third step in our logistic growth pipeline is to calculate the similarity scores between each test county and train county's demographic features and rank the N most similar train counties for each test county. To calculate our similarity scores, we first find the haversine distance of the latitude and longitude between each test county and each train county to calculate the geographic distance between their locations. We then find the total euclidean distance between all the rest of our county demographic features between each train and test county. Our similarity scores then become the square of the euclidean distance plus the square root of our haversine distance between each train county and test county. We used haversine distance to calculate geographic distance to account for the spherical nature of Earth and used the euclidean distance for the rest of our features since this is a common and effective metric used to quantify the distance or similarity between two data points. The purpose of squaring the euclidean distance and square rooting the haversine distance is to allow our other features to have more weight for determining similarity than geographic proximity alone. To find the N most similar train counties for each test county, we simply take the N train counties with the lowest similarity scores and rank them in descending order, with the first county in the set being the most similar train county to the test county of interest, and so on.

Now that we have the N most similar train counties for each test county in rank order, the fourth step in our pipeline is to calculate the weights to apply to each of the N train county's time series data. To compute our weights, we split the total weight into a raw weight and a population weight component. We do this because we want the difference in population size between two counties to be weighted more heavily, since we found a strong correlation between population size and number of confirmed cases in our exploratory data analysis. For our raw weights, we make the rank #1 similar county to have a weight of 1, our rank # 2 similar county have a rank of ½, our rank # 3 county to have a weight of ⅓ , etc. This allows counties with stronger similarity scores to have more influence in our predictions. The population weight is the absolute difference between the population estimate of the test county and train county, according to 2018 census data. We then compute the final weight for each of the N most similar train counties for each test county to be their raw weight divided by their population weight, so that counties that are close in population size and have high similarity rank have the most weight in our logistic growth model.

Once we have the weights to apply to our N most similar counties in our confirmed time series data, it is time to go to the 5th stage in our pipeline and break our time series data into their x and y component to be used to fit our logistic growth model for each test county. We define x to be the number of days from 1 to the total number of days in our confirmed time series data set and our y to be the

cumulative sum of the number of confirmed cases per day for each of the N counties multiplied by their respective weights.

With our x and y ready, we move on to the 6th and final stage of our logistic model pipeline, which is to fit our model for each test county. Another feature we added to improve the accuracy of our model is to not just use the N most similar counties in our prediction, but to use the counties out of the N counties that will give us the lowest least square error. We then use the a, b, and c coefficients associated with the lowest least square error to predict the growth curves for each test county.

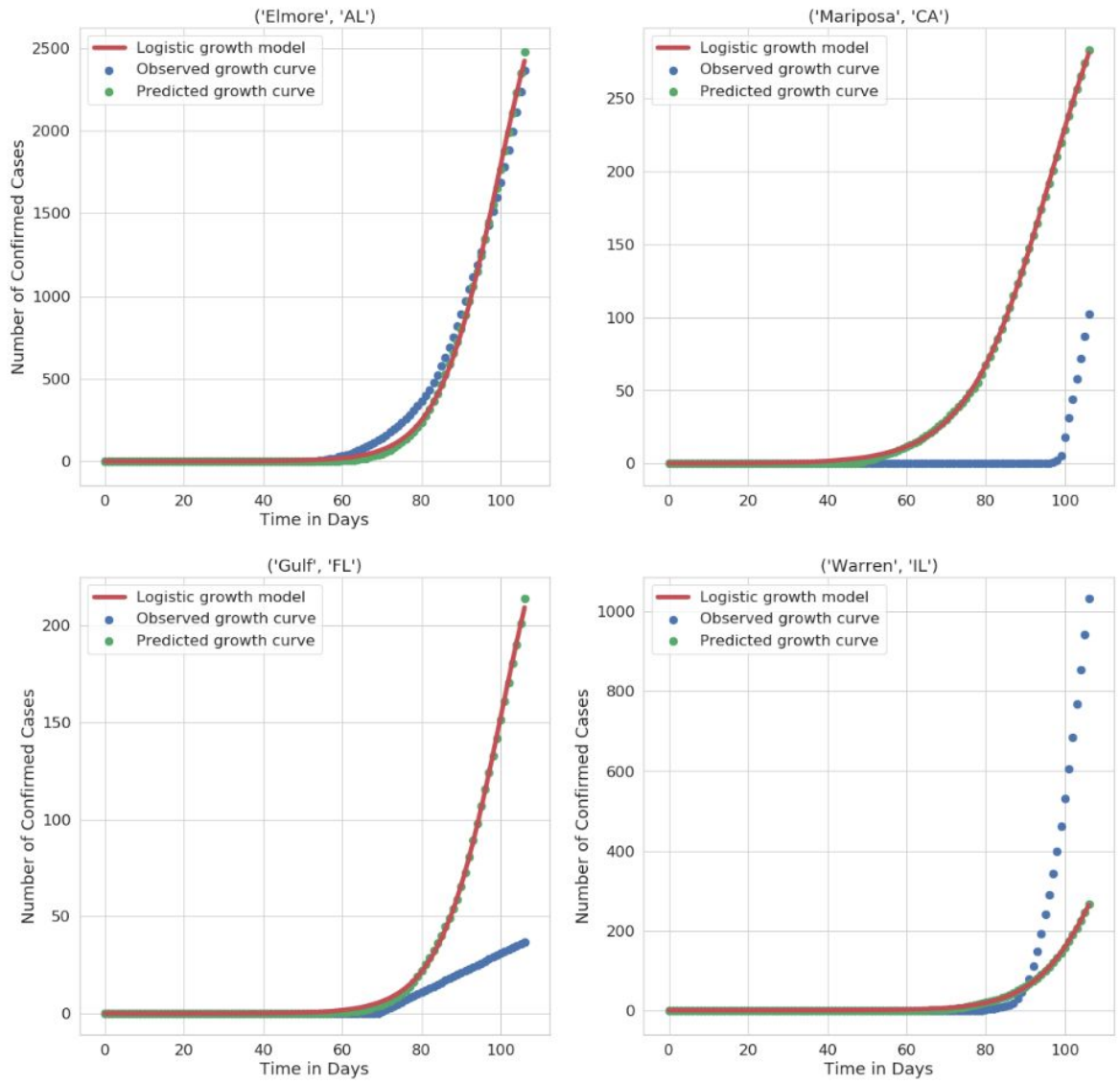
Results

The error metric of our model is the mean sum squared error between the cumulative number of confirmed cases over time, and the logistic growth curve modeling confirmed cases over time. We used a (.8, .2) train-test split, and randomly selected 40 counties from the test split to run the model. The results are as follows,

- **[Train]** mean sum of least-squared errors: 2,583,273.464
- **[Test]** mean sum of least-squared errors: 751,668,518.942

These values are clearly very far apart, but the story is better understood by breaking down the group into 3 categories, and finding whether we fit, overestimated, or underestimated the curve

- High confirmed counties (> 10,000 confirmed cases)
    - Fit: 4 Over: 1 Under: 3
- Medium confirmed counties (1,000 - 10,000 confirmed cases)
    - Fit: 3 Over: 1 Under: 6
- Low confirmed counties (< 1,000 confirmed cases)
    - Fit: 3 Over: 4 Under: 15



We believe the reason for the trend of underfitting small counties is due to the many of the similar small counties having 0 confirmed cases, impacting predictive performance

Discussion

I) What were two or three of the most interesting features you came across for your particular question?

The two most interesting features that improved the model's predictive performance were *'MedicareEnrollment,AgedTot2017'* and *'#EligibleforMedicare2018'*. Some small amount of data imputation needed to be done to make these columns usable, so we were interested in seeing how their inclusion impacts the model. When we saw that models using the log of the two metrics when calculating

the euclidean distance when finding similar counties were constantly more accurate, we were glad that our intuition was correct. Our thinking in including these features was that COVID-19 presents symptoms more seriously in elders, many of which are on Medicare. Because testing guidelines were so strict while the country was ramping up its testing capacity, symptomatic elders with access to quality healthcare were likely tested in high numbers, so many of the confirmed cases would overlap with the population on Medicare.

II) Describe one feature you thought would be useful, but turned out to be ineffective.

Although COVID-19 has spread throughout the country, the number of confirmed cases are not entirely dependent on demographic data, but also the geographical location. Counties that are close to one another tended to see confirmed cases rise in parallel. However, using the distance in miles between counties to calculate similar counties had a disastrous impact on the model before adjustment.

As described in methods, the similarity matrix is the sum of two components, the euclidean distance of the numeric categories, and the haversine distance in miles between geographical locations. With no operations performed on the prior to summing the euclidean matrix and haversine matrix, the haversine matrix had an enormously outsized influence when evaluating which counties were similar to the test counties. The counties returned were almost exclusively the closest counties, regardless of demographic differences, which led to major modeling errors. Only after adjusting the values of each matrix by taking the square of the euclidean matrix and the log of the haversine matrix were we able to find the correct balance between demographic similarity and geographical distance for finding similar counties, which in turn greatly improved the accuracy of our model's predictions.

III) What challenges did you find with your data? Where did you get stuck?

It wasn't until after completing our model that we realized all boroughs in New York have confirmed cases coded into New York county, New York and the boroughs themselves, had all zeros in their time series data. Therefore, any test counties that selected these boroughs as their most similar counties ended up getting significantly underpredicted. We spent days trying to figure out why our model was performing so poorly. In our EDA we were mostly concerned with finding NaN values, so it wasn't until we dug really deep into the data based on the results of our predictions that we were able to find the root cause of our problem. We ended up having to condense these boroughs into a single series to solve this issue, and this drastically improved the accuracy of our model.

IV) What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?

The main limitation to our analysis is that we cannot adjust our logistic growth model based on confirmed case information from the past. In its current setup, the model has no way to take a subset of the time series (the first week after the first confirmed case, for example), and use it to predict future outcomes. This functionality would increase the use cases of our predictions, and likely increase their accuracy. Because the model makes use of only numeric data, we are implicitly assuming that data about when shelter-in-place orders were issued do not impact the curve, which may prove incorrect.

V) What ethical dilemmas did you face with this data?

Throughout this project we faced two major ethical dilemmas. The first being any time we were faced with having to drop a county from our data sets. Since dropping a county means we have less data to use for our predictions and limits us from being able to predict as many counties' curves as possible, we dropped counties with hesitation and caution. We only dropped counties when we knew it would benefit the entire model as a whole, and we ended up trying to impute as many counties with missing data as possible. The other ethical problem we faced was our model has a tendency to underpredict logistic growth curves, rather than overpredict. Everyone in our group agreed that this was an ethical issue since for situations as sensitive as a global pandemic, it is best for our predictions to lean on the conservative side in order to encourage the proper decisions to be made to prevent the virus from spreading.

VI) What additional data, if available, would strengthen your analysis, or allow you to test some other hypotheses?

Even now, we do not understand the true scope of COVID-19 confirmed cases. Because the virus can be carried asymptomatically, many people who have been infected have not been tested to add to an accurate confirmed case count. Time series data tracking the number of tests performed each day would be very useful understanding how the number confirmed cases would be growing if testing initiatives were more comprehensive from the outset of the outbreak. Data regarding public transit usage and percentage of essential workers may also help increase the accuracy of county comparisons to find the best counties to model test cases with.

VII) What ethical concerns might you encounter in studying this problem? How might you address those concerns?

When working with living data, it is always important to consider the ethical implications that your models could have if utilized in the decision making process of the distribution of scarce resources. With COVID-19 data in particular, projections from our model could potentially be used to estimate the confirmed case growth curve for counties who have kept partial or unreliable data regarding the severity of the outbreak. This could be used in determining how much aid a county receives for its recovery. An underestimate would mean too little resources are distributed, while an overestimate could harm other counties which may be hurt by the shrinking pool of total aid resources for allocation. The best way to address this problem would be to greatly improve the accuracy of the model, and continue thorough testing to make sure no systematic bias exists in the predictions. In its current state, we would recommend against our model being used for any decision making in this time of crisis.

Evaluation and Limitations

The logistic growth curve model our group developed has a level of high complexity relative to other models we have designed and used through the semester, which made iterative testing time consuming and limited the number of counties we were able to model simultaneously. In the future, more efforts could be made to increase the efficiency of model predictions so that a larger number of county growth curves can be modeled simultaneously. The model can also clearly be improved upon, and effort towards eliminating bias towards underestimating the number of confirmed cases is small confirmed case counties would be made in work continues into the future.

Our model is limited by design to take in no information about the confirmed cases of the test county, but at an early stage in the design process, we considered being able to feed the model time series information up to a certain point, at which the model would use the known time series data to construct the rest of the confirmed case growth curve. Our model is also retrospective. It does not make proactive predictions about the future number of confirmed cases that could allow decision-makers to use the model to preemptively distribute supplies to areas that will be hit hard in the future, and consolidate supplies from areas that have already seen the worst of the outbreak.

Future Work

Future work on the project would first involve further testing to continue improving the prediction accuracy of the model. After the model had been tweaked, and systemic bias towards overfitting or underfitting certain types of counties had been completed, we would begin addressing all of the limitations described above. The ultimate goal would be an adaptable model that can take any subset of time series data, accurately predict the rest of the known time series, and give valuable predictions

about the future growth curve to help decision makers best plan how to allocate limited medical personnel, personal protective equipment, ventilators, and other medical supplies.