COVID - 19 United States County Level Prediction Project Statement

Jiahui Tang, Wenqi Chen, Xingyu Liu, Xiaohan Yang

November 18, 2020

1 Introduction

COVID-19 is at present one of the most important issues around world, as well as in the US. In order to mitigate the spread of COVID-19, various social interventions are taken place by different states in the US. Among those policies, the quarantine strategies and occupancy limit for restaurants bars lies at the center of the trade-off between economic development and epidemic control.

Therefore, a vital problem is to evaluate the effects of different strategies and predict the future infectious cases based on these strategies, offering a valid recommendation for the government. What's more, with the solid predictions and causal analysis, it can also encourage public to be actively involved and support the strategies.

2 Data Description

To address this question, we decide to conduct our analysis by deriving data from four different data categories, including Covid facts, mobility information, county profile, and relevant policies data. After obtaining and preprocessing the data, We then combine the datasets from different sources and try to explore the relationship between each other.

2.1 COVID Facts

For the COVID Facts, we intend to use the aggregated dataset from USAFacts as our primary source of Covid facts. In the dataset, there are daily numbers of newly confirmed cases of each county from 2020/1/1 to 2020/10/31. This is a time-series numerical data grouped by each county.

To explore the COVID Facts, we draw heatmaps of the average monthly confirmed cases in each county of US, observing whether there is a strong geographical relationship between the newly confirmed cases. Besides, we also draw the time-series change of the newly confirmed cases in each state across the 10 months. In this way, we can combine the policy data to see whether there is a significant influence of the policies.

2.2 Mobility Information

From the Bureau of Transportation Statistics, we got our mobility data, Trips by Distance. This dataset is originally provided by Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland.

This dataset is constructed from an anonymized national panel of mobile device data, and aggregated at national, state and county level. In particular, this dataset contains daily information about the number

of people stay at home each day and the number of trips occurred each day. Specifically speaking, the data provider defines a trip as a movement that stay at a different location (other than home) more than 10 minutes. Moreover, when counting the trips, the data provider also divide trips into different distance groups (e.g., less than 1 miles, 1-3miles and so on.)

The records in the Trips by Distance dataset ranges from Jan 1, 2019 to Oct 31, 2020. To see how people's behavior change in the COVID period, we plot the state-level stay-at-home ratio for these two years from Jan 1 to Oct 31. To be consistent, we removed the data points at Feb 29, 2020.

2.3 County Profile

Besides, we would also retrieve social, economics features of each county from the US census data source in 2018. This data includes households, immigration people ratio,marital status,employment and so on. Since the social-economics data is constant over the year, we don't think it would be a good predictor for predicting daily confirmed covid cases, but we think we can use the data to find some particular patterns between social-economics and covid cases later.

2.4 Policy Implementation

Last but not least, we would collect information about control policies from KFF and the Berkeley group. The data is on state level but not at a county level. We won't use this data at naive base model, but would consider to include this data source as a predictor in future refined models.

3 Exploratory Data Analysis

3.1 COVID data

The figure 1 shows the population distribution of different counties in US. Comparing 1 with 2, it can be seen that the population is strongly correlated to the newly confirmed cases in different counties.

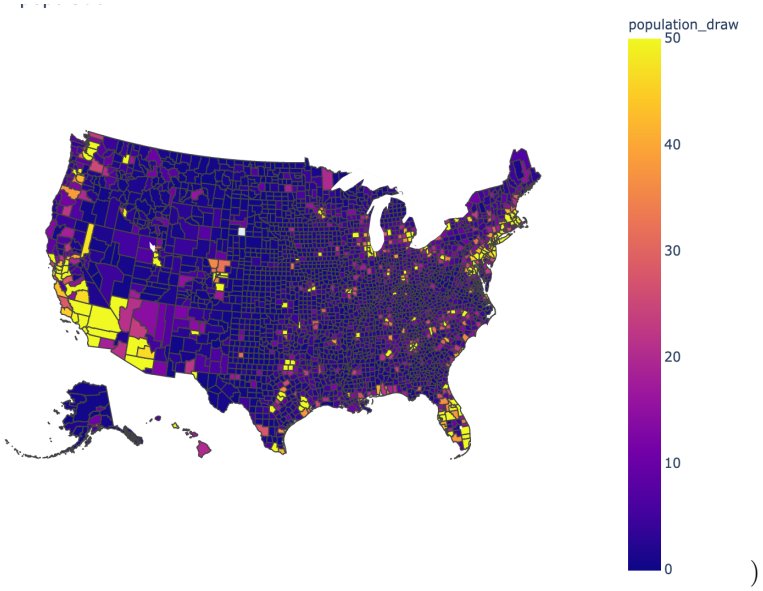


Figure 1: Heatmap of the population in different counties

The figure 2 shows the geographical relationship between the new cases in different areas. It can be seen

the cases mainly aroused from the states located in west and east coast and then the newly confirmed cases mainly came from the surrounding states.

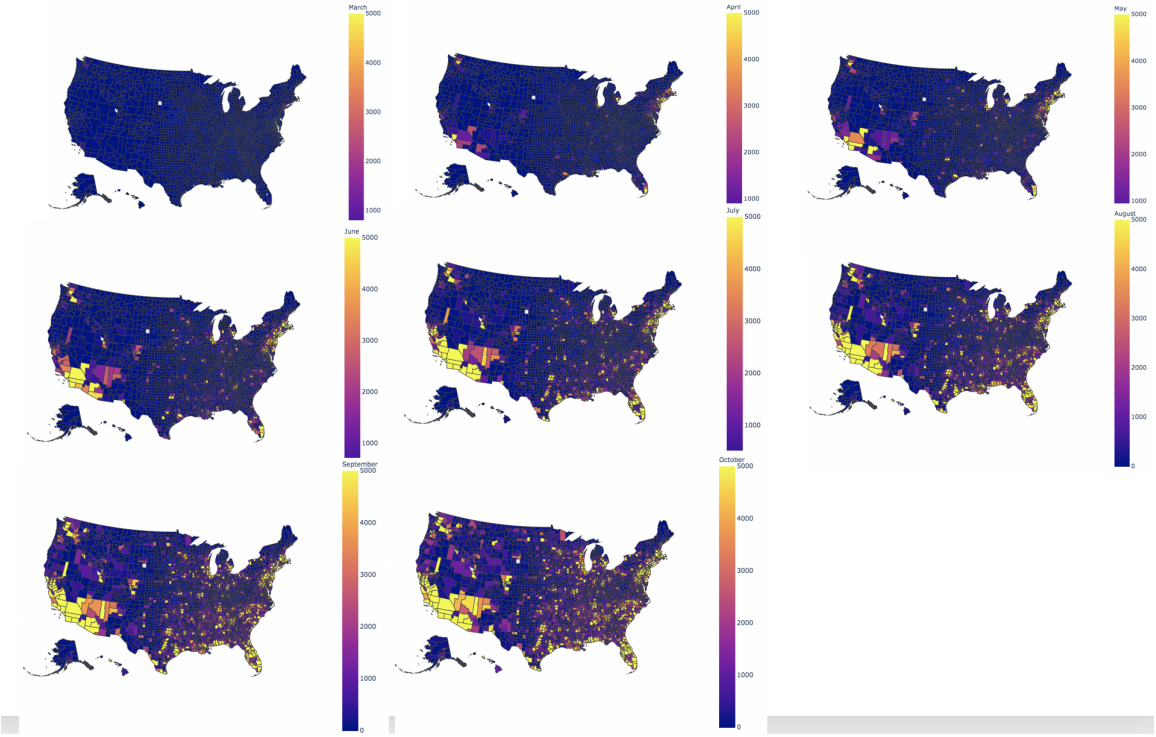


Figure 2: Heatmaps of monthly averaged cases from March to October

The figure 3 shows the new cases in different counties across the time from January to November. It can be seen that different counties have different increasing patterns, which we will take into account separately in our model.

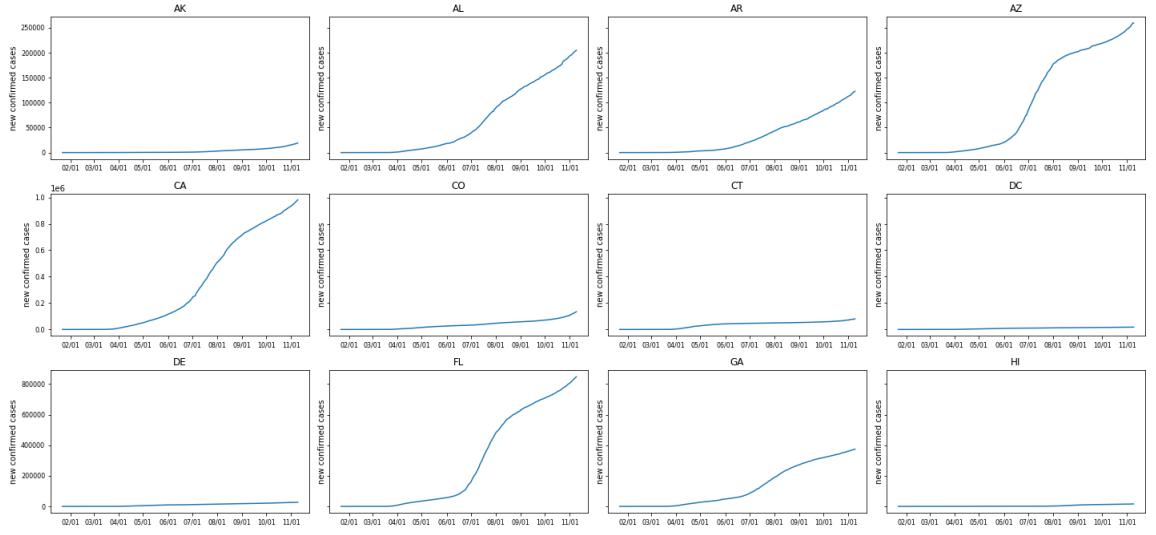


Figure 3: Figure of new cases in different counties across the time *See the full figure in Appendix A*

3.2 Mobility data

From Fig 4, we can tell the difference of people's mobility between the pandemic period and the normal time. Moreover, we also notice that the SAH (Stay-at-Home) ratio during March to June is the highest. The drastic increase at the end of March is probably related to close-down policies, while the drop at the end of May could be explained by Black-live-matters protest after George Floyd's death in May 26th. Moreover, this graph also shows the differences among each states.

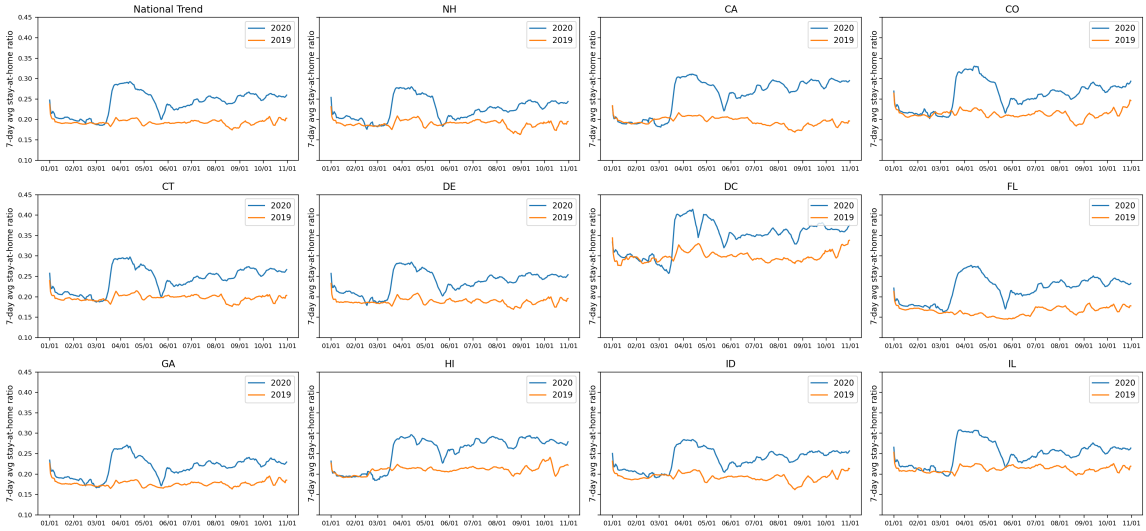


Figure 4: Figures of states' average stay-at-home ratio in 2019 and 2020 (see the full graph in Appendix A)

Furthermore, following heatmaps (Fig 5) demonstrate the county-level differences over this year. It can also be inferred that the counties with higher SAH ratio also tend to have fewer number of cases. In addtion, we could also observe the following patterns. First, most people still go out in the pandemic period, with only a few counties' SAH higher than 50%. Nevertheless, most people only go out several times a day (less than 5) and most trip distance is shorter than 10 miles.

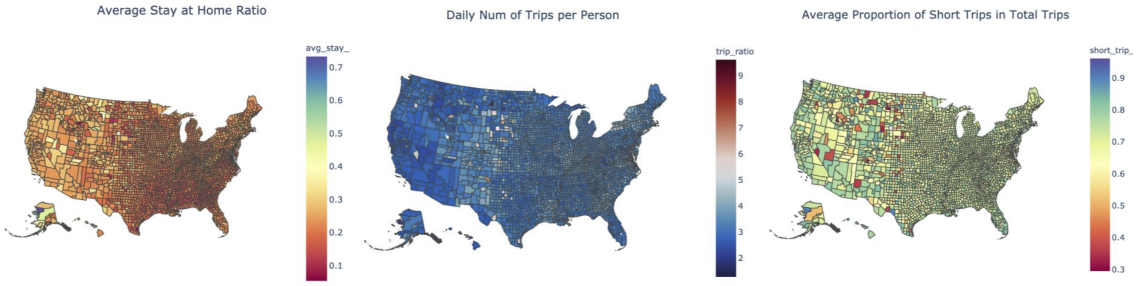


Figure 5: Figure of counties' daily average SAH ratio, trip ratio and short-trip ratio)

3.3 Social-economics data

The overall data has a size of $(3220, 1158)$, which means for each county, the data provides 1158 social-economics features. However, there are many features that are deemed not useful, e.g percentage of error for the estimated numeric feature like total household. Also, there are some NaN and string $'x'$ in the data. Therefore, for this milestone, we removed all the rows which contains missing value/NaN and removed all the columns which contains sting $'x'$. After that, the data shape is $(3142, 150)$, with all the features as numeric.

As we mentioned before, since the social-economics data is relatively constant over the year, and especially doesn't display significant variation in our investigation period, we don't think it would be an ideal predictor for predicting daily confirmed Covid cases, but we still could use the data to find some particular patterns between social-economics and Covid cases. To explore the patterns, we randomly select daily confirmed new cases in 2020/10/14 and build a ridge regression model based on the social-economics features. The absolute value of the model indicates the importance of the feature. Plotting the top 10 ranked features, we find that Veteran Status is the most important feature that explains the most variations in the Covid data, just as Figure 6 shows.

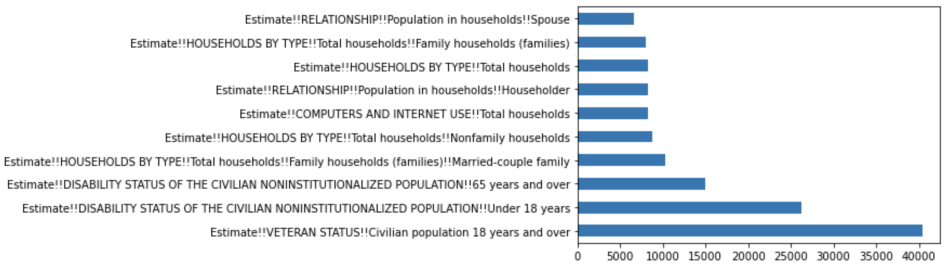


Figure 6: Social-economics feature importance

That's an interesting and surprising observation. Why does veteran status strongly correlated with Covid cases? We then took a further step to geographically visualize the distribution of veterans and confirmed Covid cases with the map below (Figure 7).

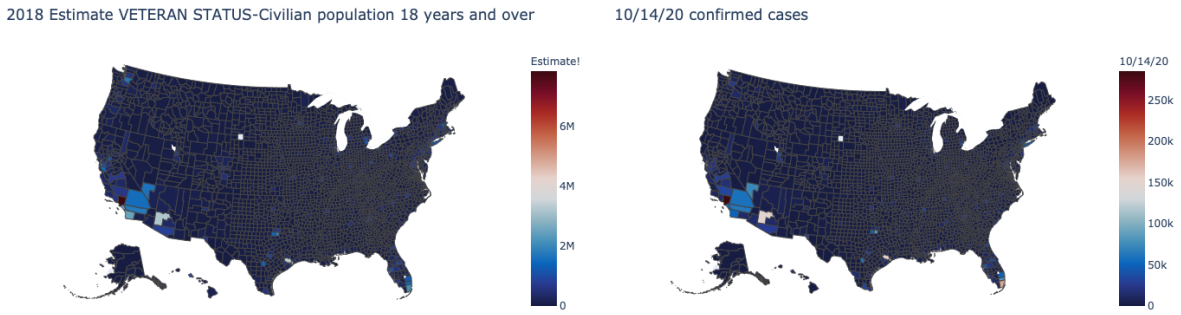


Figure 7: Comparison between veteran status and 10/14/2020 confirmed cases

Indeed, the geographic distribution of veteran status and Covid cases display a similar pattern: most of them are in California. We think it may due to the fact that some veterans would move to California to get better education and career.

Of course, during EDA process, just simply choosing one day (2020/10/14) data and build such a simple ridge regression model would yield large randomness and uncertainty. We will keep exploring the relationship between social-economics and Covid confirmed cases deeper in the future.

4 Problem Statement

From EDA, we verified some of our assumptions about the influential factors on predicting future number of cases as well as gained more insights about some exciting questions we are about to pose. In general, our project question is to continuously focus on investigating the impact of different policies under different circumstances as well as predicting the future confirmed Covid-19 cases in different counties with various sources of influential factors. In particular, we are interested in how extending or lifting the policies, especially the different levels of quarantine policies, will change the number of COVID cases in different counties.

More specifically, we intend to investigate or consider the following questions through our model.

1. **How do the cases of surrounding counties influence the target county?**

   From the heatmaps(figure 2), it can be seen that geographical relationship is a vital factor in the spread of COVID. This insight motivates us to use SIR model to take the geographical relationship into consideration when constructing prediction model.

2. **How do the different levels of policies affect the spread of COVID?**

   From the figures of Stay-at-Home ratio in different states(figure 4) and the implementation of different policies, we found that the policies can have a significant influence on the ratios. Besides, from the heatmaps(figure 5) and time-series figures(figure 3), we can see the people's behavior is significantly affected and has unique pattern during pandemic, which can be a strong predictor of the COVID spread. Therefore, it is reasonable to involve the SAT ratio and other social distancing factors as predictors of the infectious rate as the representative of influence of policies. Then the next step is to explore the detailed relationship between the social distancing factors and the policies.

3. **How does the profile of the different counties affect the spread of COVID? What may be the important factors?**

   From the naive feature importance(figure 6) of the profile factors of different counties, it can be seen that some factors, like the retired veteran rate, are far more important than others. It is reasonable to believe that some of the profile factors can represent the particularity of the counties, contributing to increasing the prediction accuracy and excluding other influencing factors when investigating causal influence of the policies.

4. **How will the number of cases change in the future in different counties?**

   As is shown in the figure of time-series cases change in different counties, it can be seen that different counties have different patterns, due to the influential features of different counties. The final goal of our project is to reasonably predict the future change of cases with the factors, which will be verified as powerful with the above research questions.

5 Baseline Model

The baseline model is a fitted Ordinary Linear Regression Model focus on daily Covid confirmed cases prediction on each county level.

We use the data between Jan and Oct as the training data, and the data after Oct is used as our testing data.

Training features includes the a sliding window of previous 7 days' daily confirmed cases number and previous 7 days' daily social distance features. The response y is evaluated day's confirmed cases number. Thus, we have nearly $1,000k$ training points (i.e $3k$ counties * 300 in a rolling basis of 7 days period) and $10k$ testing points.

To get October testing results, we use two methods:

Approach 1. Every time we have a previous 7 days($T-7$ to $T-1$) period of true data to predict the target day's (T) response.

Approach 2. First, we use previous 7 days true data to predict the current day's (T) response. Then for the following predictions, we will re-utilize this prediction. For example, for the next day ($T+1$) use previous 6 days' data ($T-6$ to $T-1$), plus yesterday's prediction (T) to predict the target day's ($T+1$) confirmed case, and so on.

Our metrics is the average percentage of error over Oct. We got each county's predicted daily confirmed cases[1] in Oct and calculate the error per county of the two methods[2]. From the results file we found that the average percentage of error for approach 1 is **0.29**, which is smaller than the approach 2 on average, indicating that approach 1 is obviously better than approach 2.

Taking San Bernardino County as an example, we visualized and compare the prediction in two approaches with testing data's real label.

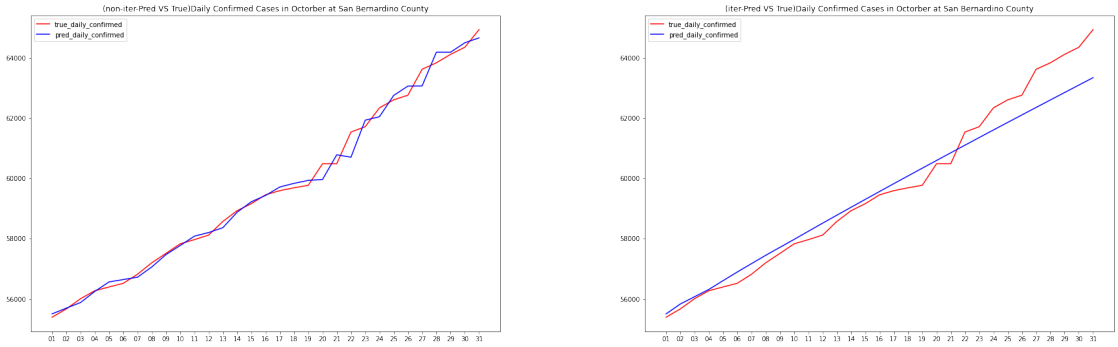


Figure 8: (non-iter-Pred VS True) and (iter-Pred VS True) Daily Confirmed Cases in Octorber at San Bernardino County

From the above two figures 8, we could clearly observe that that simple linear model is not good enough for iteration prediction because its error would become increasingly large as time goes by. Besides, an iterative approach has worse prediction error and performance comparing to the non-iterative approach.

6 Discussion and Next Steps

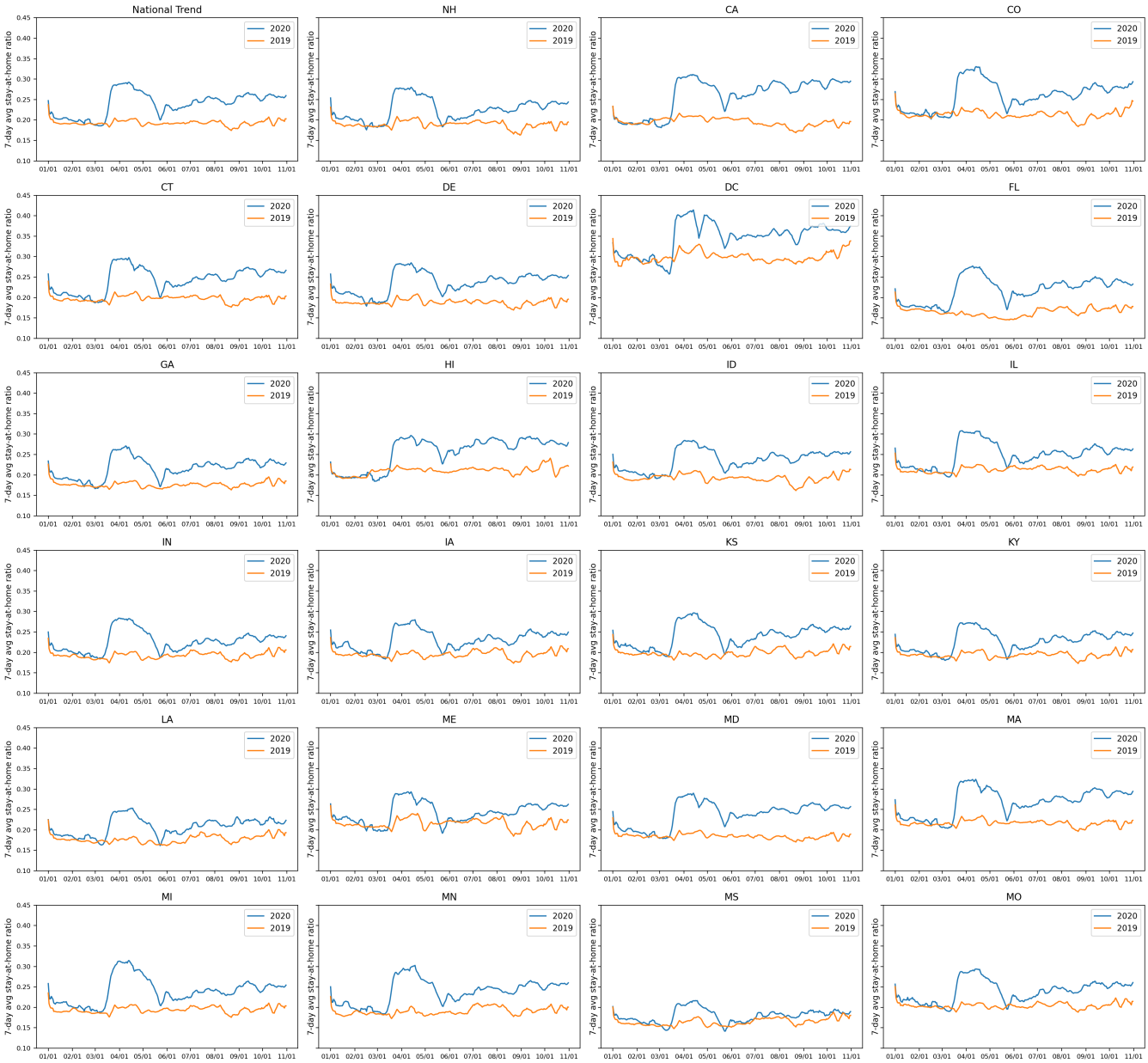
For the project plan in the coming weeks, as we observed from our base model, the performance doesn't display a satisfying result. We would explore our future direction in the following ways:

- Further refine our Linear Regression
  - Training Regularization terms
  - Working on PCAs to extract most important features instead of using all features
  - Combine more predictors or useful variables such as policy information
- Working on other Models for predictions
  - Decision Tree, Random Forest, Boosting Model

[1]Result File 1: Daily confirmed cases prediction per day per county file
[2]Result File 2: Prediction error per method per county file

- Incorporate SIR Model from epidemiology and decease spread. Estimating coefficients and ratio in all stages of epidemiology models
- Working on Neural Networks if neccessary, for example, non linearity patterns that are hard to capture
- Ensemble the above mentioned steps to produce a refined model

Appendix A: States' average stay-at-home ratio in 2019 and 2020



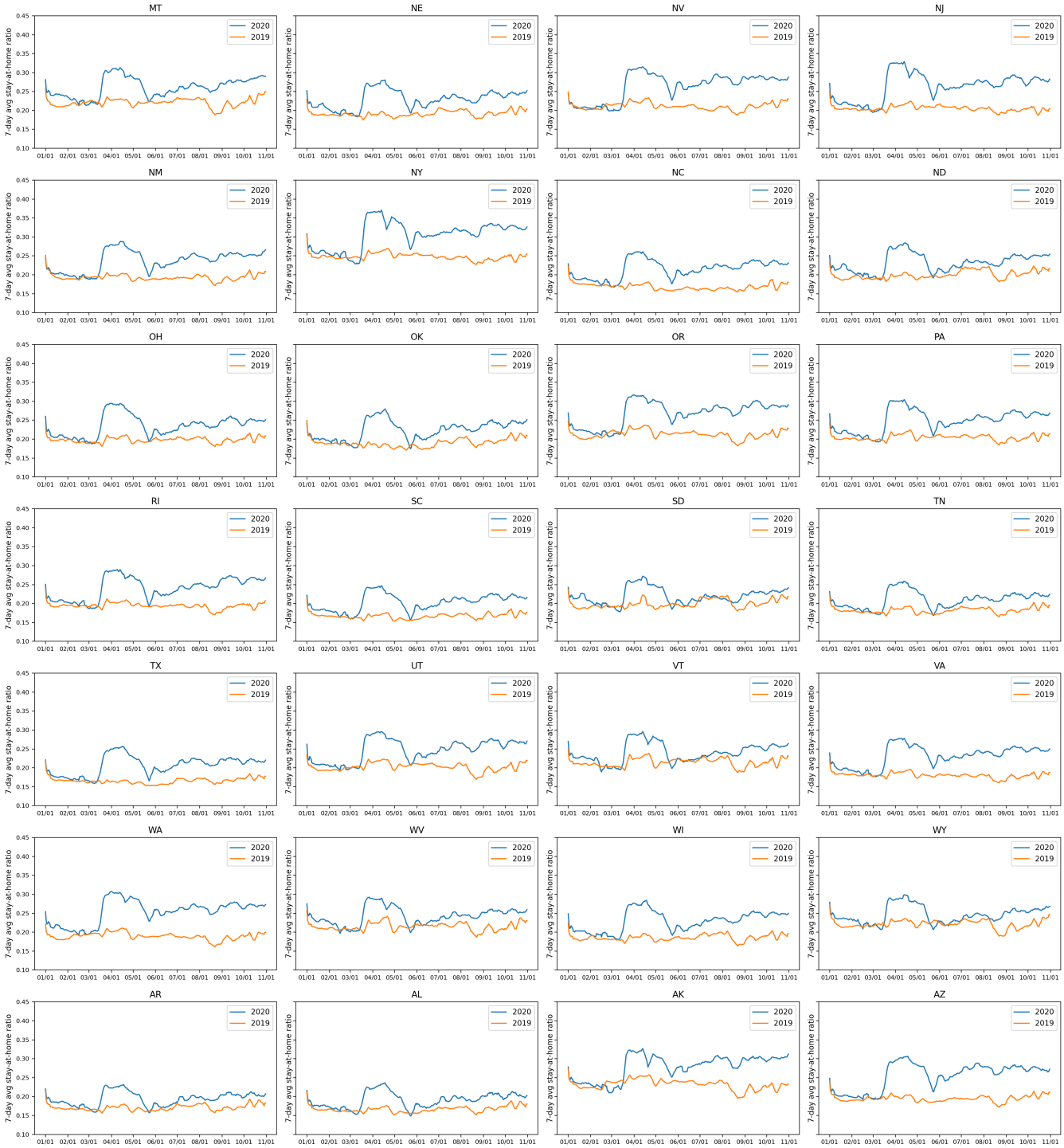


Figure 9: States' average stay-at-home ratio in 2019 and 2020 (full)

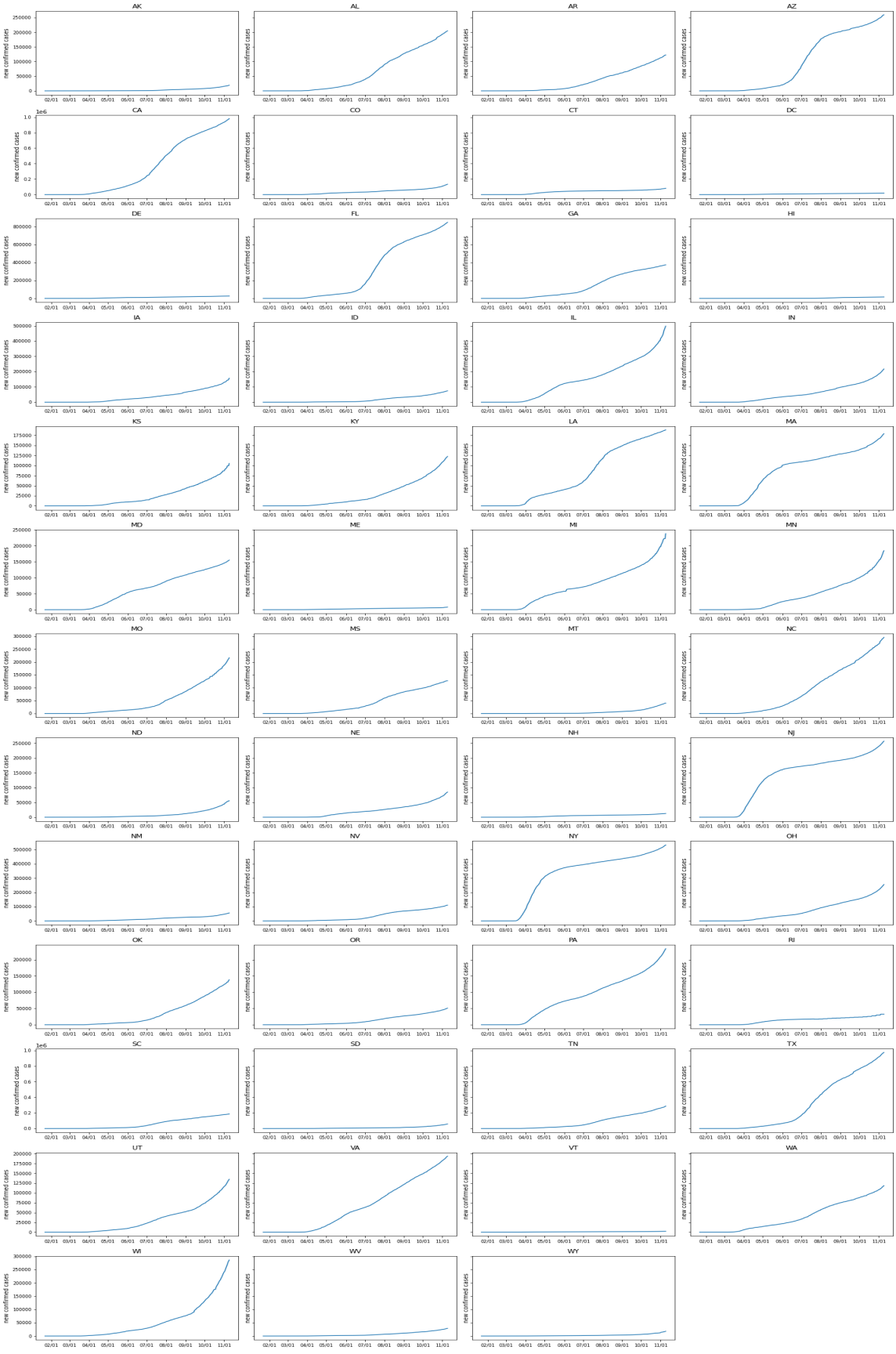


Figure 10: Figure of new cases in different states across the time