Final Lab

Huong Vu, Aliyah Hsu, Omer Ronen

December 11, 2020

1 Introduction

The coronavirus (COVID-19) pandemic is an ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first case was identified in December 2019 and the World Health Organization (WHO) declared it as a pandemic. According to WHO, no country has suffered more cases or deaths than the US, which as of December 2020 has not been able to efficiently contain the virus or prevent it from spreading. When no measures of social distancing are taken, the number of COVID-19 cases grows exponentially in a given area, as a result hospitals may be ill-equipped to help those in need.

Accurate predictions of the number of deaths/cases in a given area can help decision makers to better distribute the equipment especially PPE for front-line workers and reduce the number of deaths caused by SARS-CoV-2. On this project, we will investigate the prediction methods proposed by the YU Group as well as propose two new methods for predicting the number of cases/deaths per county.

2 The dataset

To train our predictors, we used the data provided to us from Yu Group's covid19-severity-prediction repository. While many features were available, we decided to use the number of cumulative cases/deaths as our main features to predict the future values of the same variable. While this decision neglects many features that may be useful for improving accuracy, we prioritized getting a decent predictor, rather than risking a more complicated one that would be harder to implement. The earliest records of cases and deaths from the dataset are from January 27, 2020. For this project, we look at counties from Bay Area and South Central Valley. However, to train the shared model, we used data from all counties available rather than just the counties we were instructed to predict for, due to an improvement in accuracy.

3 Our Own Model Implementation

We choose to use time series model, Auto Regression Integrated Moving Average (ARIMA) and moving average to predict the cumulative Covid-19 cases and deaths. Since the purpose of this project is to help hospitals keep track of PPE supply, it is better to overestimate than underestimate. We test models with different time intervals, and we notice that the models perform better with smaller time gap. However, choosing too small time gap would make it hard on the logistic side of PPE supply i.e. the problems with ordering and shipping PPE between locations. Therefore, we decide to use 5 days as our time gap that would give enough time for shipping and maintain high performance of the models. In addition, we use the data since the beginning up until August 28 which is the minimum date that all counties have at least 1% of their population tested positive to Covid-19. We use the estimated population in 2018 to estimate the percentage. We choose this range to ensure all counties have enough data to have stable predictions. We will compare models' performance from September 2, 2020 (5 days after August 28, 2020) to November 17, 2020 to choose the best model

3.1 ARIMA model

Our first model is ARIMA. When implementing the model, we need at least two observations to make the prediction. Therefore, when a county only has one historical value, we use the average of all historical data of neighbor counties as our prediction. The case when a county does not have enough data usually happens at the beginning of the pandemic; hence, using all historical data of neighbor counties would capture both situations, beginning and the current situations of neighbor counties' data records. To implement ARIMA model, we use `auto.arima` function. We also include a condition that the prediction must be at least the last observation to ensure the monoticity of cumulative counts.

3.2 Moving Average

Our second model is a moving average model on the rate of change in cumulative counts. We first calculate the rate of change of cumulative counts over the last 7 days and use the average as the rate of change for the next 5 days. The prediction equals $y^c_{t-5}(1 + \frac{1}{7}\sum_{i=1}^{7}(r_{t-5-i}))^5$. We choose to calculate the average rate of change of 7 days because of the seasonality of the data over the week period. There are two special cases that we treat in the algorithm. The first one is when a county only has one observation, the treatment is the same as in ARIMA model. The second special case is when the cumulative counts of previous date is 0, we will impute 0.25 as the rate of change. The value 0.25 is chosen after trying different values and observing the predictions do not change much with value 0.25. For counties with the number of data points more than one but less than 7, we still use the average of previous days' rate of change to predict the future cumulative count.

3.3 Ensemble of ARIMA and MA models

We implemented the CLEP model in the paper with the linear and exponential models in the previous part, but we didn't tune the hyper-parameters c and μ. We just plugged in the same values as mentioned in the paper ($c = 1$ and $\mu = 0.5$). Now we'd like to look deeper into the choice of the hyper-parameters for our CLEP model built with the ARIMA and MA models, and we will look at it separately for bay area counties and the south valley counties.

As discussed in the paper, for μ, the authors tried with different values in the range of 0.5 to 0.9, while for c, they just used the default value 1. Inspired by the discussion in the paper, we tried with three different values separately for c and μ ($c = (1, 1.5, 2)$; $\mu = (0.5, 0.6, 0.7)$) and tested with the permutation of the values on the CLEP model to see which set of the hyper-parameters gave the best performance. Note the candidate lists of c and mu are set to be the same for both bay area counties and south central valley counties. To evaluate the performance of the CLEP models, we look at the following five evaluation metrics: coverage, normalized error, MAPE, rawMAE and sqrtMAPE. After comparing the evaluation results of the nine CLEP models for each set of counties, we found that the choice of $c = 2$ and $\mu = 0.5$ performed best in not only the bay area CLEP models but the south central valley CLEP models, and for both cases and deaths predictions. So we set the hyper-parameters to be $c = 2$ and $\mu = 0.5$ for our CLEP model of ARIMA and MA models.

4 Model Evaluation

4.1 Best Model

We plot the metric values for models: Linear, Shared Exponential, ARIMA, Moving Average, Linear-Shared Exponential CLEP and ARIMA-MA CLEP in Figure 1 to evaluate their performance on cumulative cases and deaths related to Covid-19 on the period from September 02, 2020 to November 17, 2020. On the radar plot, instead of using Coverage metric, we use Not Covered metric `(Not.Covered = 1 - Coverage)`. Hence, on Figure 1, we look for model that has the smallest circumstance as the best performed model. To calculate the values for these metrics, we use the medians of these metrics' measurements calculated from all the dates listed above across all counties (both Bay Area and South Central Valley). For cumulative cases, ARIMA-MA CLEP model performs the best in all metric aspects except for the Not Covered percentage. Linear model has the lowest non-coverage but comparing other metrics, linear model is not the best. ARIMA-MA CLEP model is also the best model in predicting cumulative deaths in all metrics. Since the objective of this project is to support the distribution of PPE supply, we choose ARIMA-MA CLEP model as our final model for having the best performance in predicting numbers of cumulative deaths. We make this decision because not everyone who is tested positive to Covid-19 would be admitted to the hospitals, only the severe cases. Therefore, the number of cumulative deaths would be more relevant in estimating the needed PPE supply in hospitals.

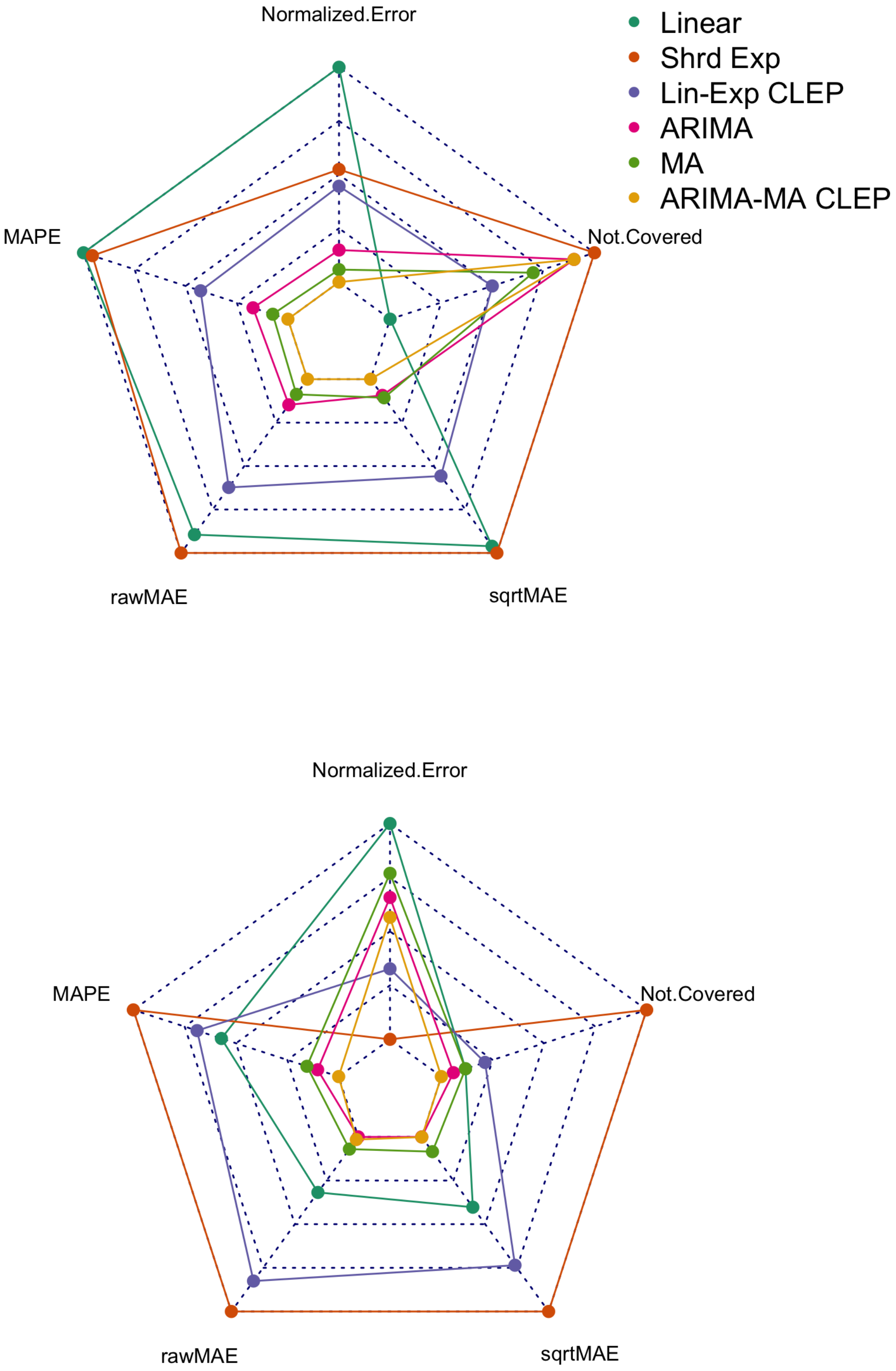

Figure 1: Radar plots for all models' performance on predicting cumulative cases (top) and deaths (bottom)

4.2 Ensemble Method

The CLEP method can be viewed as a past-performance-based ensemble method. The basic idea is to give more weight to the prediction of a model that has shown better results over the past week.

We start by looking at Yu group's CLEP predictors:

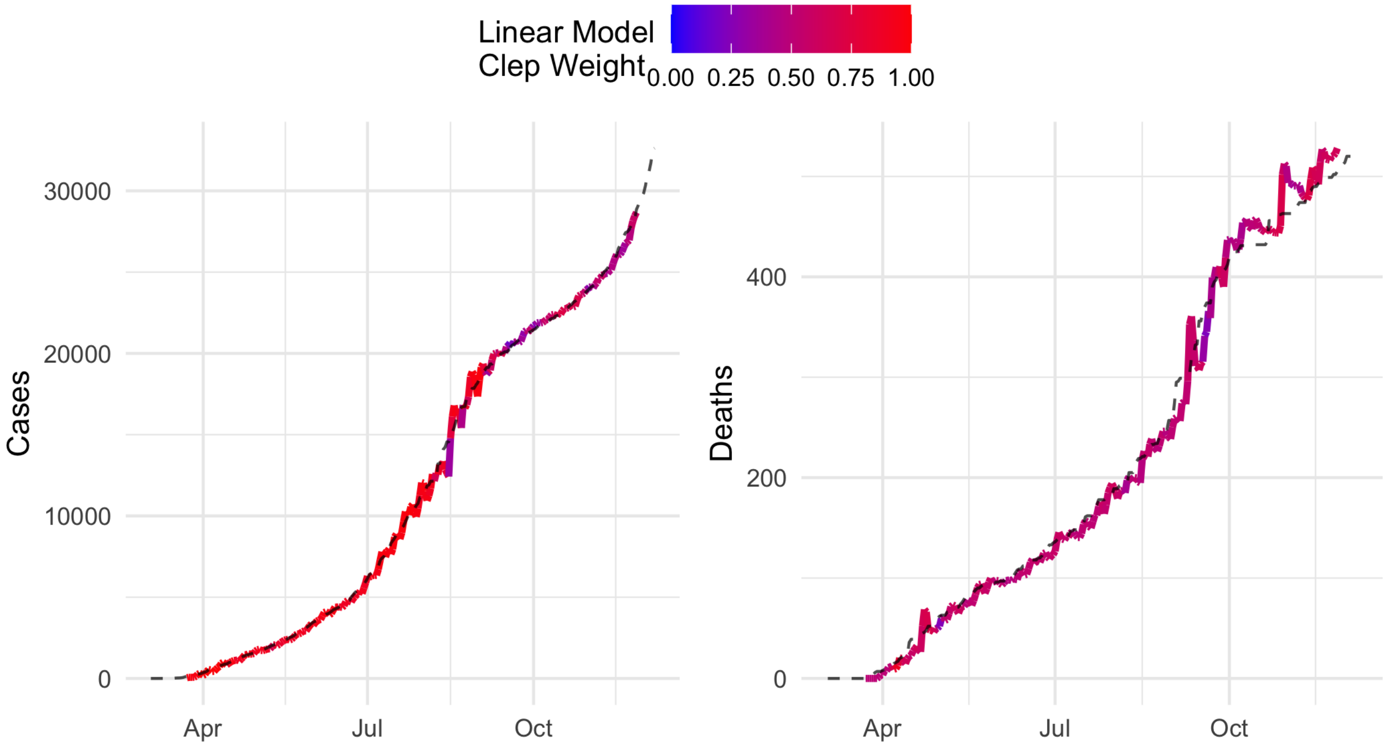

Figure 2: Clep prediction using the Yu group linear and shared exponential predictions. We show 3 days ahead prediction, the dashed line indicates the true values

Looking at the CLEP weights we see no particular trend, and that both methods obtain similar weighting. One particular interesting observation is that the shared model makes relatively poor predictions for the cases variable from April to August.

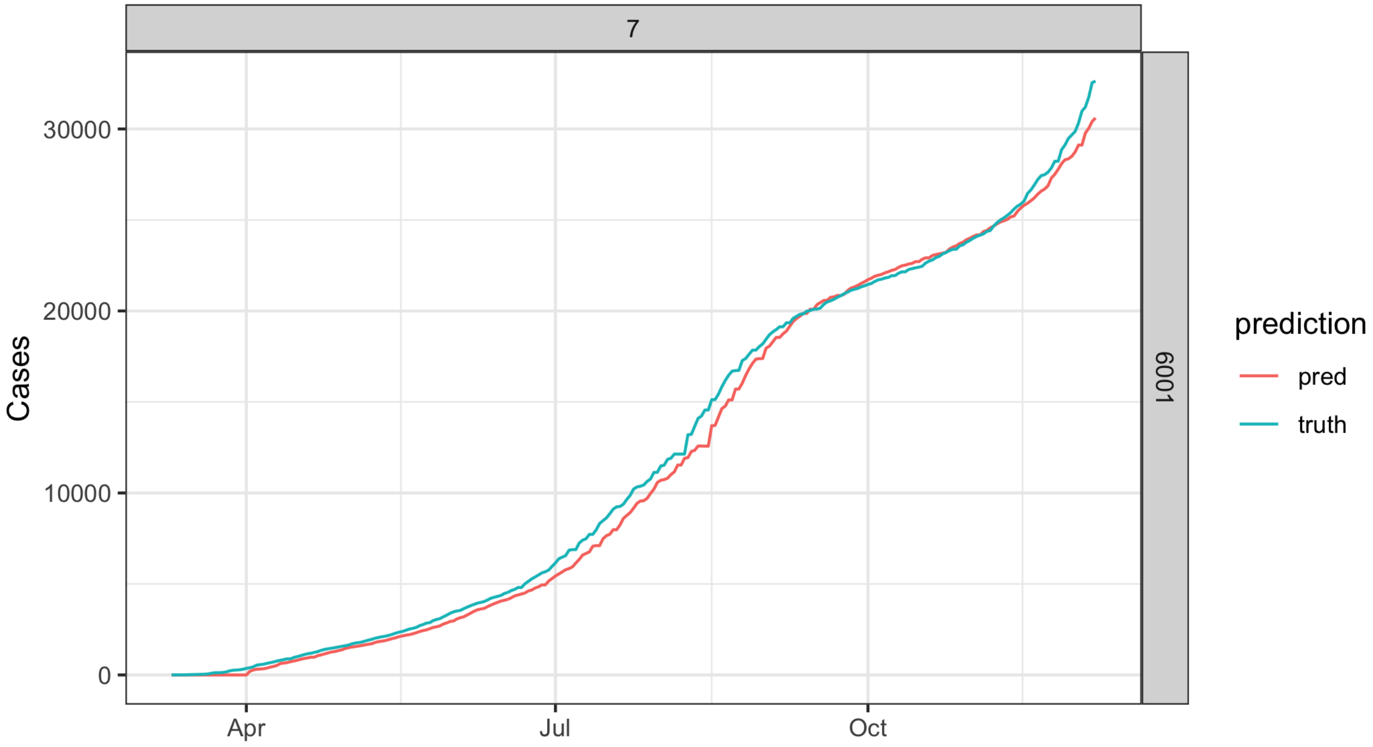

Figure 3: Shared exponential cases predicitons vs the true value

Observing the prediction, a clear flaw of this shared exponential is evident. It constantly underpredicts the number of cases for a long period.

We now perform a similar analysis to our ARIMA and MA CLEP method:

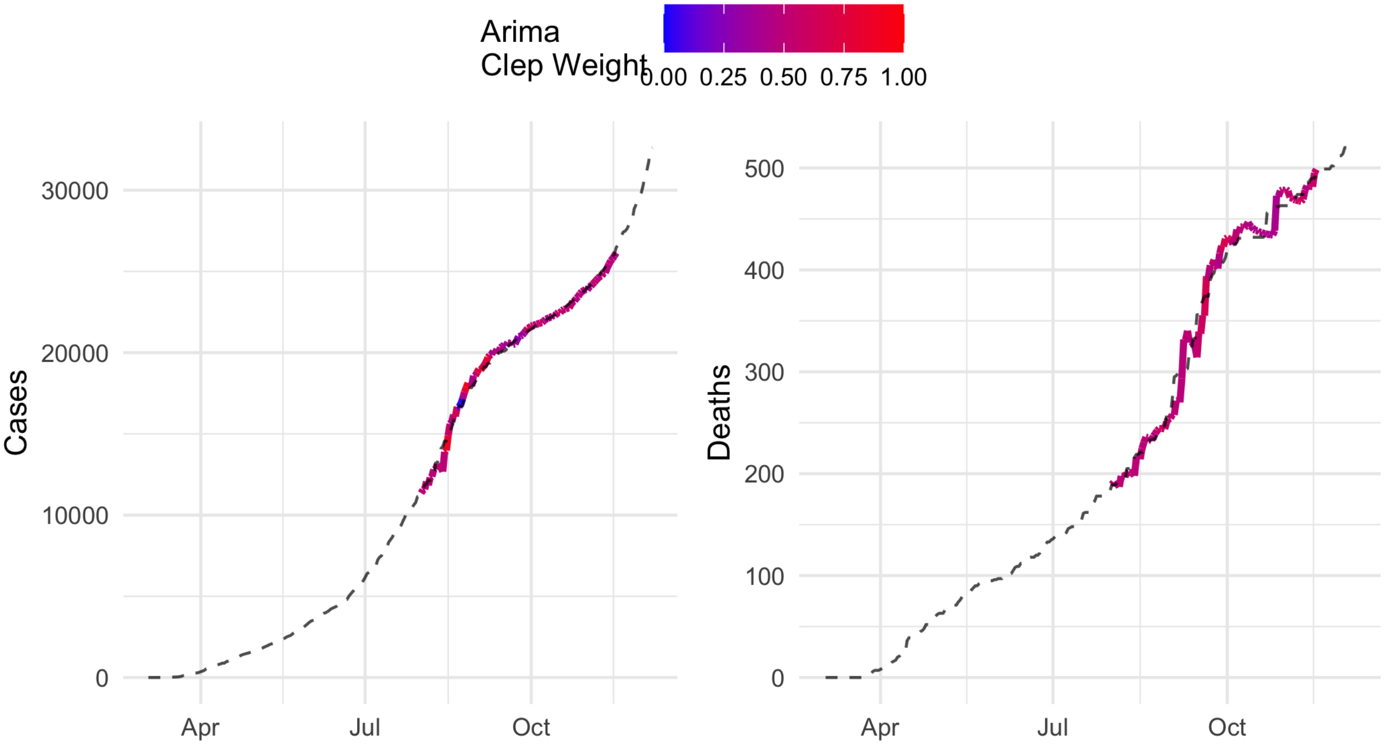

Figure 4: Clep prediction using our ARIMA and MA models. We show 7 days ahead prediction, the dashed line indicates the true values

We see CLEP chooses similar weights to the ARIMA and MA models, which makes sense as MA can be seen as a specific case of ARIMA model and we would expect those models to be similar in terms of performance.

4.3 Model performance between clusters of counties and within each cluster

To see if there are any differences in the two clusters of counties, we plot the box plots of the results of the five evaluation metrics for each model, and for cases and deaths predictions separately. We first look at the evaluation results in cases predictions. We can see that coverage, rawMAE and sqrtMAE are generally higher in south central valley counties, while MAPE and normalized error are higher in bay area counties. Another thing worth noticing is that the variance of the normalized error of bay area counties are also generally larger than that of south central valley counties. Next, we take a look at the evaluation results in death predictions. It is obvious to see at first glance that the variance of the five evaluation results of south central valley counties in the shared exponential model is much larger than that of the others. Furthermore, the evaluation results in bay area counties are generally worse than that in south central valley counties. To be more precise, for deaths predictions, the coverage is lower, while the MAPE, normalized error, rawMAE and sqrtMAE are all higher in bay area counties than that in south central valley counties for all models except the shared exponential model.

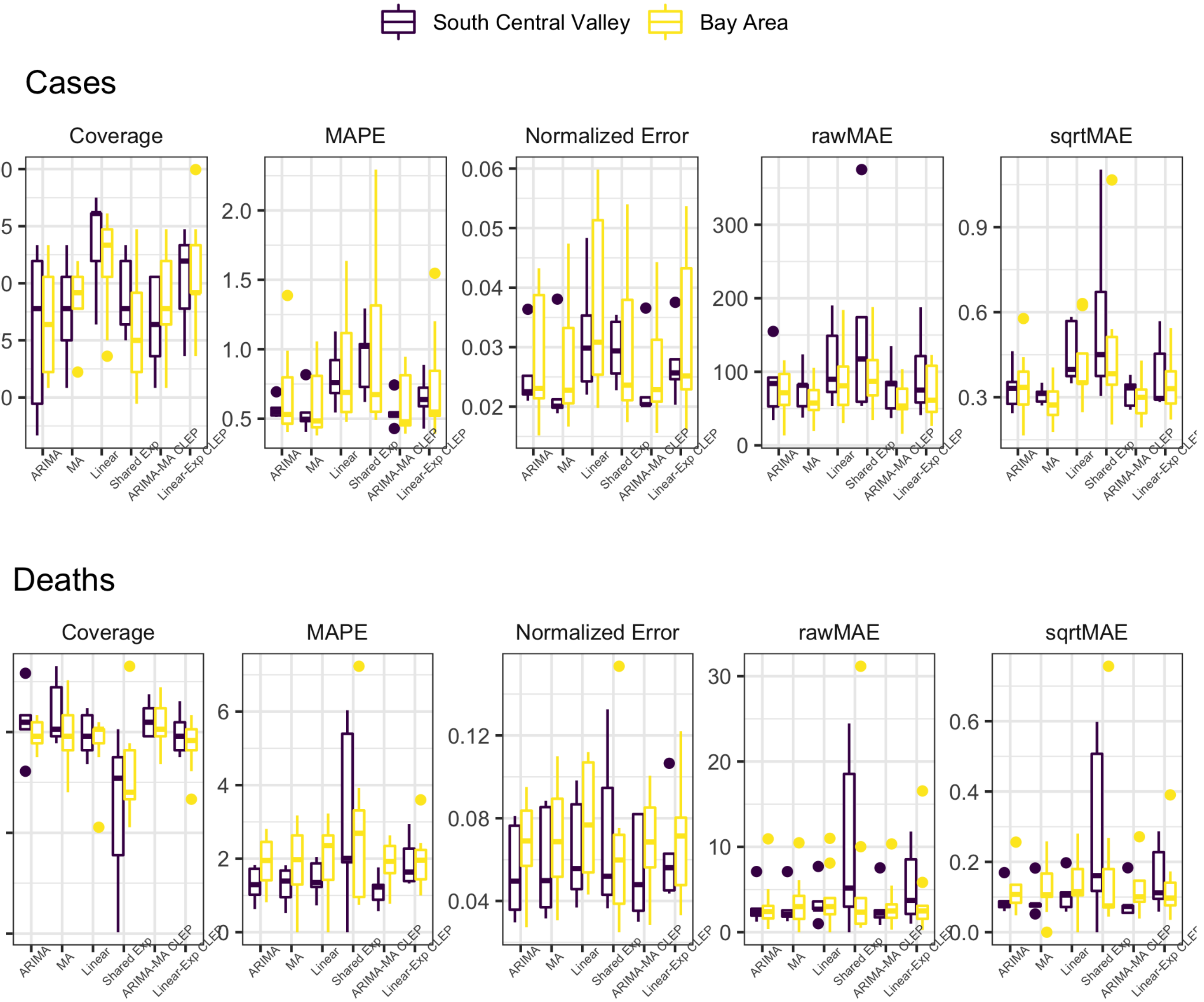

Figure 5: Models' performance between bay area counties and south central valley counties

Now we look at the performance of the models within each geographical area. We discuss the models' performance in cases prediction first. For the nine bay area counties, we can observe a similar trend of ranking of the model performances in most of the counties. For example, the ARIMA-MA clep model tends to perform best, and the linear and exponential models tend to perform worse in Alameda, Contra Costa, Napa, San Francisco, San Mateo, Solano and Sonoma counties. While the performance of the models behaves more differently in Marin and Santa Clara counties. In Marin, the ARIMA model seems to perform worst, and the performances of all the other models don't differ much. In Santa Clara, it is only clear that the linear model perform worst solely. In the five south central valley counties, the ARIMA-MA clep model performs the best, but the worst-performing models actually differ, and we can further divide the south counties into two groups based on them. For Madera, Merced and Tulare, it is the linear model that performs the worst, while for Fresno and Stanislaus, it is the exponential model that performs the worst. The radar plots for this analysis are excluded in the report due to the page limit, but they can be found at `other/bay_cases_model.pdf` and `other/south_cases_model.pdf`.

We next look at the models' performance on death predictions for each geographical area. In the nine bay area counties, we observe approximately two trends of rankings of models' performance. In Contra Costa, Marin, San Mateo, Solano and Sonoma, the linear and MA models seem to perform worse, and the exponential model tend to perform best. While in Alameda, Napa, San Francisco and Santa Clara, the exponential model is the worst-performing model, and the MA performs the best. For the five south central valley counties, in Fresno, Stanislaus and Tulare, the exponential model is the worst-performing model, and the ARIMA-MA is the best model. However, the ranking of the models' performance is not so clear in Madera and Merced. To be more precise, there is not a distinct winner model or a loser model in Merced. In addition, it is only clear that for Madera the exponential model is the best-performing, while the rest perform comparably worse. The radar plots for this analysis are excluded in the report due to the page limit, but they can be found at `other/bay_deaths_model.pdf` and `other/south_deaths_model.pdf`.

5 Best Model Investigation

5.1 Evaluate the fit of the model

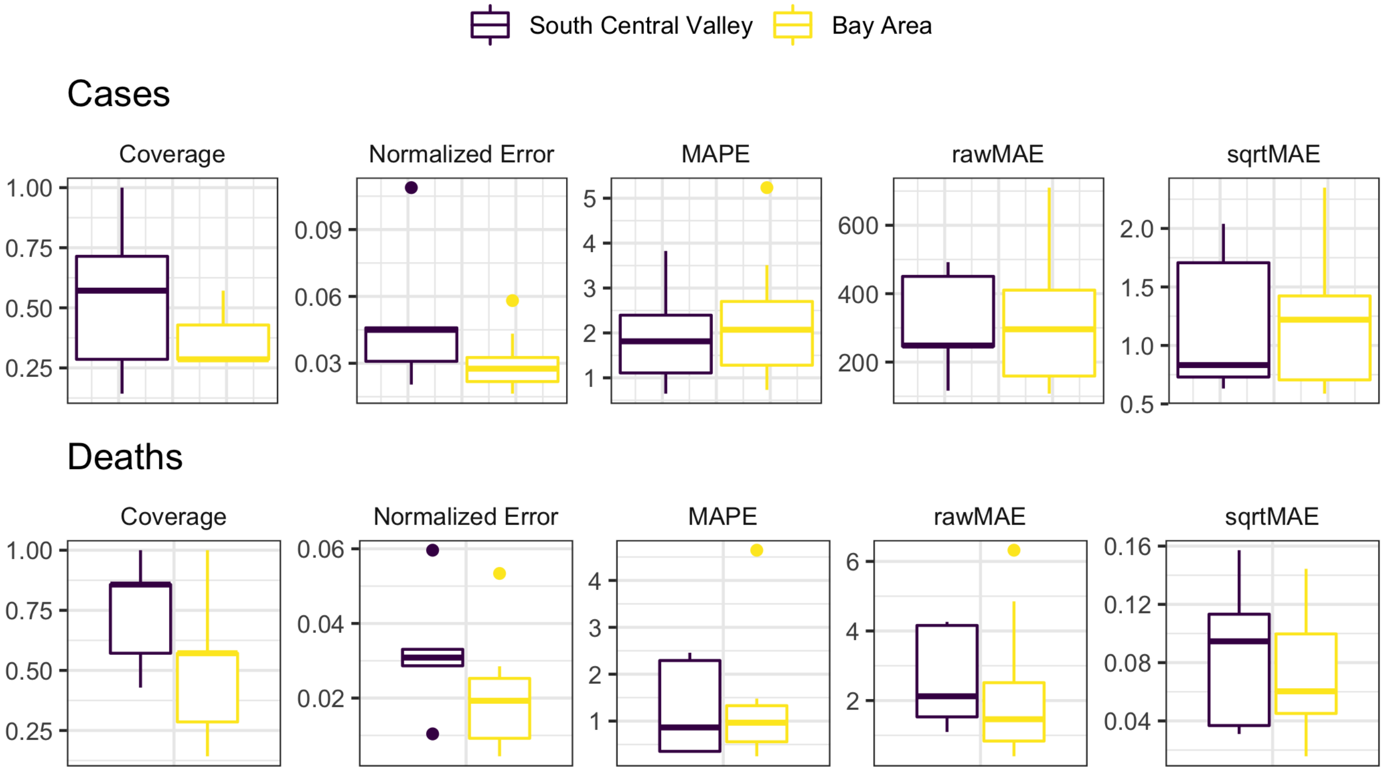

Figure 6: ARIMA-MA CLEP model's performance on November 28, 2020 to December 04, 2020 on predicting cumulative cases and deaths for Bay Area and South Central Valley counties

Comparing between Bay Area and South Central Valley counties, our model makes smaller normalized error among Bay Area counties with smaller difference within the group in number of cases and larger within group difference in number of deaths compared to South Central Valley. For MAPE, Bay Area counties slightly have higher value but their within group differences on both number of cases and number of deaths are smaller than South Central Valley counties even though we have more counties in the Bay Area.

We mainly have outliers in normalized error and MAPE from Figure 6. In predicting the cumulative cases, Stanislaus county from South Central Valley and Napa county from Bay Area are the upper outliers from the their county group in normalized error metrics while Marin county (Bay Area) is the upper outlier in MAPE. For cumulative deaths, Madera county and Tulare county (South Central Valley) are upper and lower outliers in normalized error metrics. Bay Area does not have outliers in normalized metric. For MAPE, South Central Valley does not have outliers but Bay Area has Napa county as the upper outlier. Madera and Napa counties have very low population density compared to other counties in the same clusters and they are reported to have outbreaks in number of cases and hospitalizations which are traced back to the transmission either at work settings (among farmers and business in and out of counties) or family gatherings, especially in low income and Latinx community.

To compare the ARIMA-MA CLEP model's performance on this one-week period with that on a longer period, from September 02, 2020 to November 17, 2020, we need to compare Figure 6 with Figure 5. For cases predictions, to predict on a longer period generally yielded better performance for the ARIMA model. Not only that coverage is higher, but MAPE, normalized error, rawMAE and sqrtMAE are all lower, for both the bay area counties and south central valley counties, when predicting on the longer period in Figure 5. However, for the deaths predictions, this is not necessarily the case. For deaths predictions, predicting on a longer period only gave a higher coverage, but actually also yielded higher MAPE, Normalized error, rawMAE and sqrtMAE than what we have when predicting on the one-week period in Figure 6, for both the Bay Area counties and South Central Valley counties.

5.2 Fitted Parameters Variance

ARIMA-MA CLEP predictions are based on ARIMA and Moving Average models' predictions. We used the same tuned parameter c and μ when making predictions over the most recent one week.

ARIMA model is in the form `ARIMA(p, d, q)` where `p` is the order (number of time lags) of the autoregressive model, `d` is the degree of differencing, and `q` is the order of the moving-average model. Corresponding to `p` and `q`, we would have p `ar` parameters for autoregressive models and q `ma` parameters for moving-average model. For each county, on different dates, the model might have different values for `p`, `d` and `q` and corresponding different parameter values. Merced, Marin and Tulare counties generally have a more complicated model (4 to 5 parameters) compared to other counties. The number of parameters in the model for a county tends to not change over the day, but the values of each parameter vary day by day. However, overall, the variation is not significant.

With Moving Average model, we use the 7 days of historical rate of change. Between counties, the rate of change ranges between 0% to 1.9%. Marin county has 0% rate of change in number of cases continuously over several days and counties with high rate of change are Santa Clara, Napa and Stainislaus. we have the rates of change vary within in less than 0.1% for each county over different days. The range of rate of change in number deaths is smaller, between 0% to 0.8% with the variation rate bewteen days less than 0.1%. Napa and San Mateo counties have the lowest rate of changes while Madera and Sonoma counties have the highest values.

5.3 Model Feature Importance

Since ARIMA and MA model only uses historical cumulative counts in cases and deaths, the most important features are cumulative cases (in predicting cases) and cumulative deaths (in predicting deaths).

5.4 Model Stability In The Future

The question of assessing the model's stability can be addressed in many ways, we propose to look at the CLEP component and ask the question of how dependent are our predictions on accurate reporting of the deaths over the last week.
The motivation for this question is by the assumption that the reported number of deaths, may not be completely accurate due to human errors in the reporting process for example.
We look at deaths predictions and sample the noise from a Poisson distribution ($\lambda = 3$). We add/subtract the noise with equal probability.

To gain a basic understanding, we will add noise to our labels we use to perform CLEP and check visually what our predictions look like.

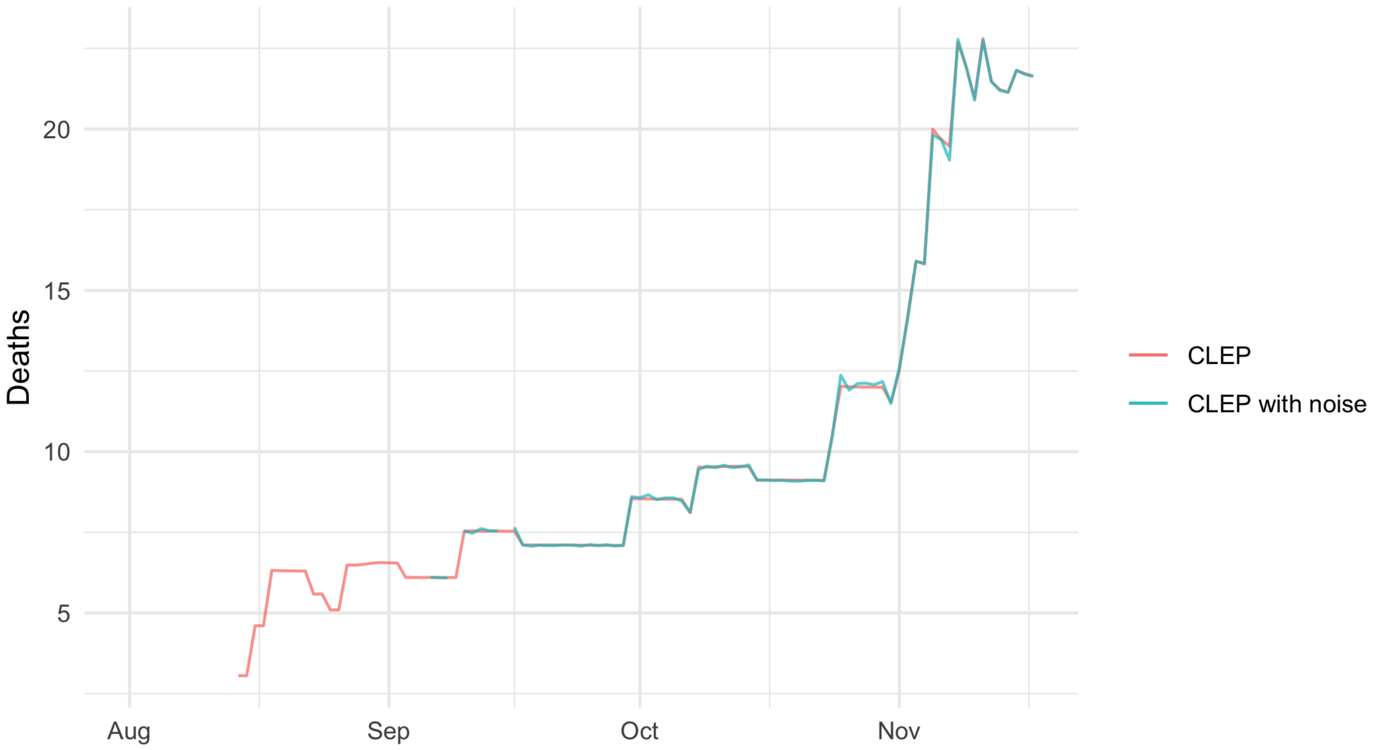

Figure 7: Comparing our method's prediction using true labels and noisy labels for the CLEP operation

We can see that since both models make very similar predictions the CLEP operation is not sensitive to noise.

Another aspect of stability we look at is the variability of the counties between rural and urban areas. We use our model on both Bay Area (urban) and South Central Valley (rural) counties. Even though the model's performance is not the same on both groups of counties, but their individual performance is acceptable. Therefore, we believe that our model would work on different set of data coming from different demographic areas. In addition, our model is trained on different period of the pandemic (the beginning, before and after big holidays). Since our model is time series based, the model cannot pick up sudden change in cumulative cases and cumulative deaths caused by external factors. However, given enough time observed, the model will start picking up those new changes in trends or seasonality (yearly data for holiday trend) and will perform better.

6 Predictions on December 10 to December 16, 2020

To produce the prediction, we use 80% confidence interval from `auto.arima` for ARIMA predictions and 80th quantile for MA model to produce the lower and upper predictions. To ensemble the predictions from ARIMA and MA models, we use the latest weight for each county from the training result. We save the prediction in `other/7-day-ahead.csv`.

7 Discussion

1. Discuss your predictions of cases vs. the predictions of deaths. Was predicting one more challenging than the other? do you trust your methods and predictions? Why and why not?

We feel that our model is a fairly good predictor for both cases and deaths, with the main strength of this predictor is the fact that it is trained every day which is how a model with so few parameters can show great flexibility in picking up changing trends. The deaths prediction presented us with numerical challenges in the cases where no new deaths were reported for a few days. We feel that as long as the dynamic of the pandemic does not change dramatically, by introduction of vaccine or unprecedented lockdown our model should make accurate predictions. Otherwise we are looking at making predictions on out of distribution data, a setting under which the model's behavior can be unexpected.

2. Do you think your methods will work for other times during the pandemic? In the past? What about the future? Explain?

Our model is best suited for the pandemic and was trained and evaluated on pandemic-time data. As long as there's no drastic change in the dynamics of the pandemic our model should utilize the daily re-trains to produce fairly accurate predictions. Once, for example, people would start getting vaccines we suspect that our model may have a harder time to pick up the new trend.

3. Discuss your initial collaboration plan. Did you end up following that plan or did things change as the project progressed? What worked and what did not? What parts of this collaboration did you find most challenging? Is there anything you would have done differently?

We have our `final_plan.xlsx` in \other folder which includes the details of our execution plan. We had a general outline of what tasks needed to be done, but did not assign members to specific tasks at the beginning. We went through the list of tasks needed to be done in each meeting, brainstormed sub-tasks and divided the work. Since we had a list of specific tasks to do, it was easy for us to assign the work and keep track of the workflow. During the project, we did not meet the deadlines on some tasks and had to postpone those tasks and sometimes we switched task assignments between members for the time constraints and difficulty of the tasks. The most challenging part of this project was implementing methods from the paper and evaluate the results. If we get to start the project again, we would spend more time discussing the format and structures of functions we implemented in this project, so it would be smoother when we combine everyone's work.

8 Bonus section

I initially called 1 (800) 782-4264 (option 6) and was told to call (909) 387 4859 since the phone number was for Covid-19 community services information. I tried to call (909) 387 4859, but no one answered. Then, I called (909) 387-9146, obtained from http://www.sbcounty.gov/main/Pages/departments.aspx, and the representative transferred me to Covid-19 hotline but no one picked up so I left the message. The link http://www.sbcity.org/about/covid19/default.asp listed (909) 387-3977 as San Bernardino County COVID-19 Public Information Line but no one picked up the phone. On the website https://sbcovid19.com/, which cannot be accessed sometimes, (909) 387-3911 is listed as contact. From this phone number, I was able to reach someone who had some information related to the data reporting.

1. Where does the data come from? There is a team called Surveillance Team and they are responsible for collecting the data and managing data upload to sbcovid19.com. The team the person I talked to is only responsible to help give the information related Covid-19 to those who don't have access to the internet.

2. How do you determine whether or not to attribute a death to COVID-19? The death is attributed to COVID-19 is based on the death certificate signed by doctors.

3. How do you decide what day to report the death? They report data everyday and it will be updated to sbcovid19.com website to the public.

4. Can you share the data with me? The access to the data source is under control of a higher level team. She only has access to information posted on sbcovid19.com under a tab called Dashboard. The website actually has a lot of detailed data related to Covid-19 information in San Bernardino county. In case sbcovid19.com cannot be accessed, this link might take you to data link. (https://sbcph.maps.arcgis.com/apps/opsdashboard/index.html#/44bb35c804c44c8281da6d82ee602dff)