ANALYZING THE CHANGE OF THE INFLUENCE OF HUMAN SOCIAL MOBILITY ON COVID-19 SPREAD: A COUNTY LEVEL ANALYSIS

Yang Huang
EECS
UC Berkeley
Berkeley, CA
yang_huang@berkeley.edu

Kewei Chen
EECS
UC Berkeley
Berkeley, CA
kewei_chen@berkeley.edu

Runda Tian
CEE
UC Berkeley
Berkeley, CA
tian1998@berkeley.edu

December 14, 2021

ABSTRACT

This project is aimed at trying to answer the question relating to the spread of COVID-19 in California State. According to our EDA analysis, we make a hypothesis: the positive correlation between human social activities and the spread of COVID-19 is much weaker than before. In particular, we apply linear regression models to predict COVID-19 deaths for counties in California. Modeling and analysis have been carried out based on datasets detailing confirmed confirmed cases of COVID-19. This analysis comes at an appropriate time as several mutations of COVID-19 continue to spread globally. Creating data-driven models enables us to prove if lockdown policies should be extended. It also has potential to provide estimates on how human activities will be directly affected by this disease in the future.

1 PROBLEM INTRODUCTION

To date, the COVID-19 pandemic has claimed over 700,000 lives in the US alone. However, by conducting mask wearing enforcement and vaccination, we can see a great slowdown of the blast of COVID-19 cases and deaths. As a result, some protests against traditional prevention measures such as lockdown have erupted. During the outbreak period of COVID-19, mobility restriction played an important role in controlling the spread of the virus, but with the increasing rate of vaccination, it becomes hard to say if human activities should still be restricted. Understanding how the virus spreads, what factors contribute to the reduction of confirmed cases and deaths and how they contribute can give us evidence to answer this question. Here we analyze parameters at the county level with the goal of uncovering what human mobility attributes of a county make it more vulnerable to larger numbers of confirmed cases, as well as higher mortality rates. Our goal in this project is to identify the change of the importance of human mobility during this pandemic that could provide guidelines for policy making.

All in all, in this report our hypothesis is:

After getting COVID-19 vaccination, the positive correlation between human social activities and the spread of COVID-19 is much weaker than before.

To justify our hypothesis, we introduce a new dataset containing people mobility data from Google COVID-19 Community Mobility Report and use social mobility features to build multiple linear regression from January to August in 2021 to check the change of its correlation coefficient. The hypothesis will be confirmed if we observe a decreasing trend of the normalized correlation coefficient of human mobility in this model from Feb. to Aug. 2021, and it will be rejected if the positive correlation does not obviously get weaker during these eight months.

2 DATA AND ITS PREPROCESSING

We have used the pre-provided datasets, which contain data of population, mask usage, vaccination and Covid-19 cases. To test our hypothesis, we also introduced several new datasets, including time_series_covid19_deaths_US.csv" which lists the cumulative number of confirmed deaths of COVID-19 in the county every single day; 2020_US_Region_Mobility_Report.csv, provided by Google, which is based on the company's location-tracking capabilities and measures changes in mobility with respect to a baseline for Retail and Recreation, Grocery and Pharmacy, Workplaces, Transit Stations, Residential and Parks. The dataset provides data on visitor numbers (or duration for the residential category) to these categories of location each day with respect to the baseline; us_FIPS_county.csv, which has FIPS codes to identify each county; covid19vaccinesbycounty.csv, which includes information about county-level vaccination progress.

3 OPEN-ENDED EDA

In this Open-Ended EDA part, we mainly analyze two time-series datasets of California, Daily County-level Community Mobility Reports and Daily County-level Vaccinations and Cases. Based on our common sense and knowledge, we believe there should be correlation between vaccinations, social mobility and cumulative COVID-19 cases.

3.1 Vaccinations

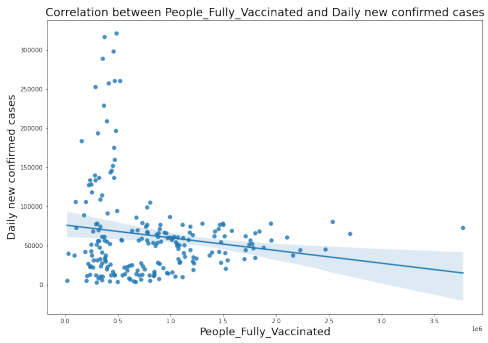

Figure 1: Correlation between vaccinations and daily confirmed cases

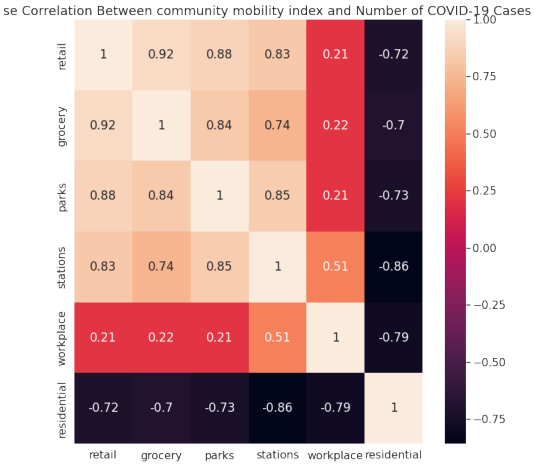

Figure 2: Heatmap of Community Mobility Reports

From Fig.1, we explore the correlation relationship of daily vaccination and daily confirmed cases in the US though linear regression plot, it's obvious that these two variables have negative correlation, which proves the effect of vaccination in preventing the spread of COVID.

3.2 Community Mobility

Community Mobility in this paper measures the movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. The data shows how visitors to (or time spent in) categorized places change compared to our baseline days. A baseline day represents a normal value for that day of the week, that baseline is the median value from the 5-week period Jan 3 – Feb 6, 2020. We only take use of the daily county-level community mobility changes in California while data exploration.

From the heat map of these six indexes2, we can find pretty obvious positive linear relationship between the percent changes from the baseline of retail and recreation, grocery and pharmacy, parks and transit stations. Among them, how visitors to grocery have the strongest correlation with how visitors to retail and recreation. In reality, these trips can be all classified as essential trips. Besides, the heatmap also reveals the inverse relationship between outdoor trips percent changes from baseline and the residential percent changes from baseline which is highly reasonable since when people stay at home they won't go out.

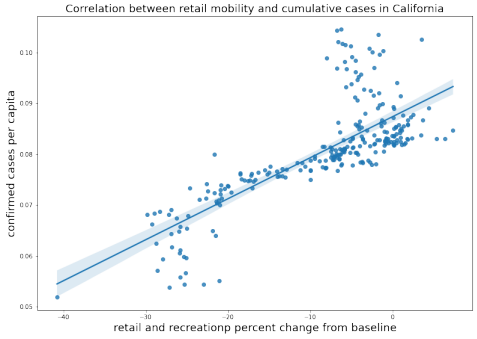

Figure 3: Correlation between retail mobility and daily confirmed cases

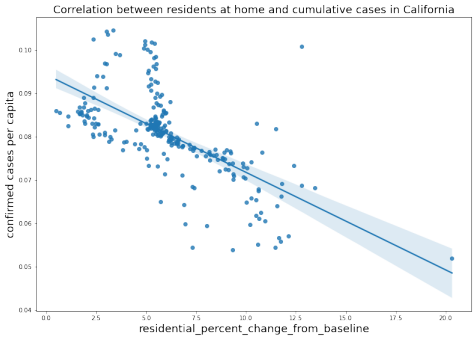

Figure 4: PCA analysis of vaccinations and cases

Further analysis discloses the correlation between these mobility indexes and cumulative cases. Fig.3 reveals the number of cumulative cases is positively associated with the foot traffic in retail and recreation and Fig.4 proves that cumulative cases have negative correlation with people's residential time at home. Using SVD, we decompose the standardized matrix X that contains the standardized current proportion of cumulative fully vaccinated and cumulative partially vaccinated, cases per capita, and six mobility percent changes from Community Mobility Report in county-level California from 01/01/2021 to 08/31/2021. Each row represents the county's figure during a certain day. From the PCA response in Fig.5, we can see the brighter the dots, the closer to the bottom. If a county one day has a large positive value of pc2, it is more likely to have a smaller value of cumulative cases per capita. Fig.6 reveals a clear linear relationship between these variables which guides us to use multiple linear regression in the modelling part.

To better understand the time-series mobility report, we break the overall community mobility report into months and plot the relationship between movement trends at retail and recreation and its corresponding cumulative cases per capita over time from February to August in 2021. We replace those missing values with their average index within one week and remove outliers that were 3 σ from mean. To our surprise, there is an obvious change in the relationship between the mobility index and COVID-19 cases when the slope of the regression line over time gradually decreases and inversely increases. The result from February to May is reasonable since

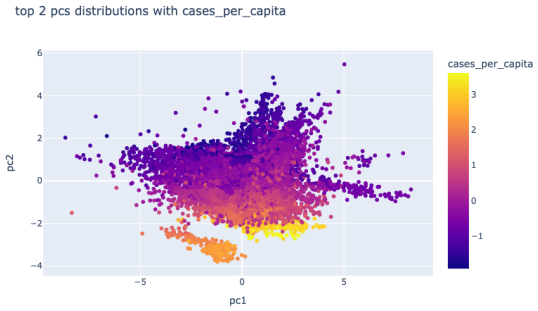

Figure 5: Top 2 PCs distributions with cumulative confirmed cases

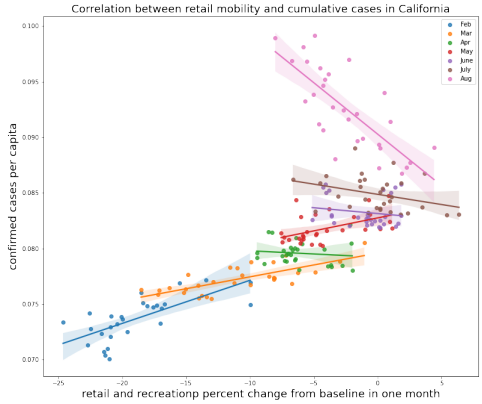

Figure 6: PCA analysis of vaccinations and cases

when people spend less time in public places with good social distance, they are less likely to get an infection. However, from June to August, it seems the even more people have active social mobility in public places like retail and recreation, the confirmed cases don't positively follow the trend which means positive correlation gradually disappears. Does social distancing still work? Why is there a change between mobility and COVID-19 transmission?

Combined with our analysis on vaccinations, we make a conjecture: in general, vaccinations and mobility reduction all have good effects on preventing new cases, but their influences change over time. At the start of COVID-19 with very few vaccinations, mobility have large influence in COVID-19 transmission and mobility reduction is almost the sole strategy in prevention but when fully vaccinated people keep accumulating, the influence of mobility gradually shrinks so that even we have active outdoor activities at public places, the confirmed cases won't increase dramatically. Based on this, we bring up our hypothesis.

4 MODELING

4.1 Baseline Model

We aim to build a model to predict the cumulative confirmed cases based on time-series vaccinations, time-series community mobility and other related variables. Then based on a effective model, we can measure the changes of the effects of human activities from Feb. to Aug. in 2021.

4.1.1 Model Choice

To confirm or reject our hypothesis, our model need to help us explore the strength of each input features on results. Though our final goal is not to build a model that can predict the most accurate index of COVID-19 cases, we still need an effective and discriminative model to act as a fair judge so that the model can measure whether the influence of different features on cases has changed. Our problem can be defined as a supervised learning problem, using current and past features to predict cases and our target value, confirmed cases per capita is a continuous quantity which means it can be predicted using regression.

In the basis of analysis in EDA part, we are also able to identify the clear linear relationship between both vaccinations and community mobility index with cumulative COVID-19 cases. Given these conditions, we decide to use linear regression to model our relationship. We first tried LASSO linear regression, however we found that it didn't fit out dataset. The R^2 score is negative when alpha is 1 and R^2 score with other alpha values are all no higher than the basic

linear regression model. Given a higher R^2 score and good interpretability, we decide to pursue a time-based linear regression model.

4.1.2 Features Selection

We designed our training and testing dataset on the basis of the joint data frame in EDA. We randomly hold 20 percent of the dataset as test set and 80 percent as training set.

At first, we brought in recent 2 days figures of confirmed cases per capita, total partial vaccinated data per capita, cumulative fully vaccinated data per capita, total deaths per capita, community mobility at retail, grocery, parks, transit station, workplaces and residential, and daily state-wide mask usage. To avoid noises, we identified data 4 std away from the mean value as outliers and removed them in the baseline model.

However, to our surprise, the R^2 score for the model including past confirmed cases is around 0.997 and after visualizing the coefficient of each feature, we find confirmed cases is highly autoregressive with a near 1 weight. Compared to the unbalanced large coefficient of cases in past two days, the importance of other features are almost no higher than 0.0001, acting like random noises of our training data. Since our hypothesis mainly concentrates on exploring the dynamic changes of feature importance of human mobility, previous confirmed cases will push us away from the answer to hypothesis, so we exclude the time series confirmed cases. Besides this, the reason to exclude mask usage features is that the feature is based on state but other features in dataset are county-level, only in California.

4.2 Model Improvement

Before adapting new improvement method,we identified data 3 std away from the mean values as outliers and removed them in the baseline model.

4.2.1 Date Formatting

Since people's mobile pattern is different between weekdays and weekends, human mobility has a weekly periodical change. One problem is that our model is based on daily data, which will result in issues with predictions and poor performance if we compare day-to-day changes especially weekends with weekdays.

In order to differentiate the data attribute, we converted this feature into boolean value, whether the current day is a weekday. Besides from creating a new category of weekdays, we also standardize the datetime and add it as our new feature.

The result is shown in 5.1.1, the RMSE loss decreased and the R^2 score improved, which indicates that the solution will make a good difference to our model performance.

4.2.2 Calculating Autocorrelation Function

Another problem is that we are not sure how many former days should be included in our input, we computed the autocorrelation function (ACF). The ACF is a way to measure the linear relationship between an observation at time t and the observations at previous times. For example the ACF for a time series y_t is given by:

$$Corr(y_t, y_{t-k})$$

This value of k is the time gap being considered and is called the lag. A lag 1 autocorrelation (i.e., k = 1 in the above) is the correlation between values that are one time period apart. By calculating the correlation of the transformed time series we obtain the partial autocorrelation function (PACF), which can help us determine which lag would likely be feasible.

Fig.7 are the plots of PACF for six community places. We could see that the fluctuation of correlation coefficients is in a cycle of 7 days. So, we decided to bring 7 prior days time series data into our input features. The result is also shown in 5.1.1, the RMSE loss decreased and the R^2 score improved, which indicates that adding 7 prior days time-series data into our input will make a good difference to our model performance.

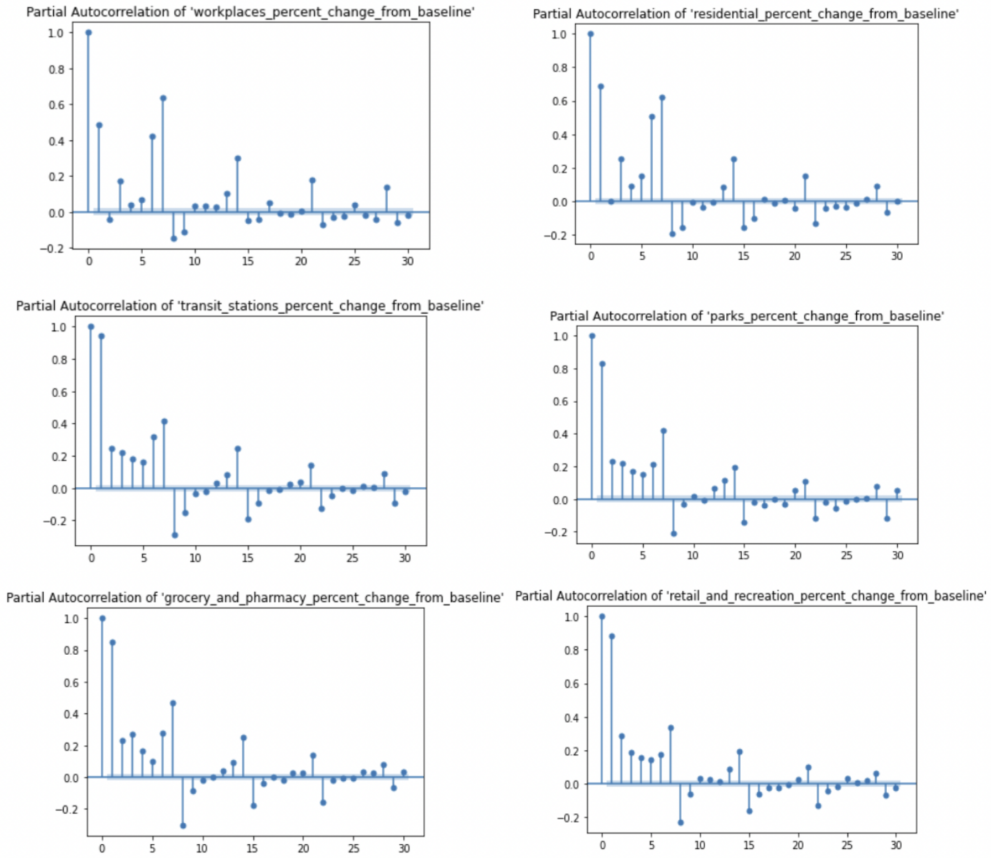

Figure 7: PACF for 6 community places

5 EVALUATION AND ANALYSIS

5.1 Model Evaluation

5.1.1 Result

- **Baseline model**
  We have an RMSE loss of 0.670 in our basic linear regression model. The standard deviation of predict values is 0.916. The R^2 score is 0.464 for our model.

- **Model improved by date formatting**
  The performance of the improved model is better than the baseline model so the improvement of formatting date is valid. We have an RMSE loss of 0.460 in our basic linear regression model which is lower than the baseline model. The standard deviation of predict values is 0.829. The R^2 score is 0.692 for our model which is higher than the baseline model.

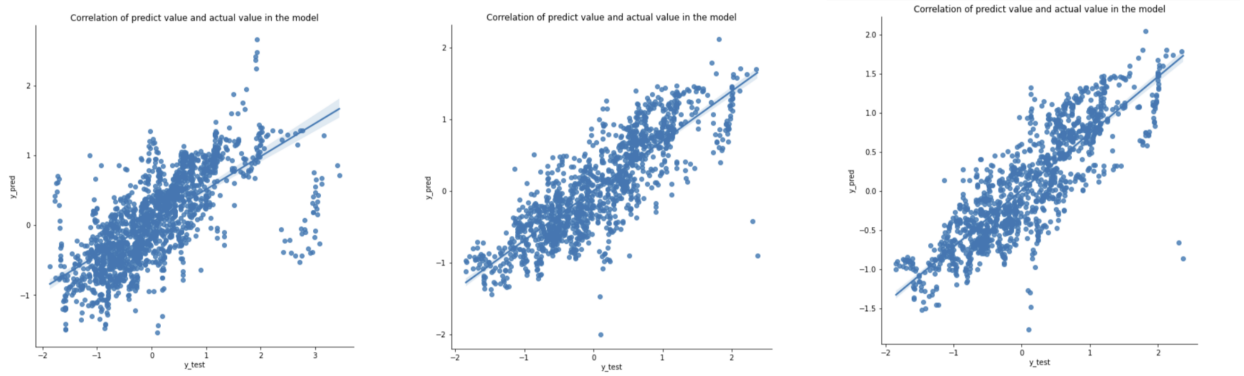

Figure 8: Correlation of Predicted Values and Actual Values in the Model (a)Baseline Model (b)Model improved by date formatting's prediction (c)Model improved by time series features' prediction

- **Model improved by time-series features**
  The performance of the improved model is better than the first improved so the improvement of adding time-series features is valid. We have an RMSE loss of 0.442 in our basic linear regression model which is lower than the model improved by date formatting. The standard deviation of predict values is 0.829. The R^2 score is 0.715 for our model which is higher than the model improved by date formatting.

5.1.2 Visualization

- **Correlation of Predicted Value and Actual Value in the Model**
  We draw the correlation graph of the predicted values and the actual value in our model. Fig.8.a is the baseline model prediction result. Fig.8.b is the result of model improved by date formatting. Fig.8.c is the result of model improved by time series features. The X-axis represents the actual confirmed cases in the test set while the y-axis represents the predicted confirmed cases. We could see that the area of 95% confidence interval for that regression is more narrow in the improved model than baseline model. We are also glad to see that the angle between the fitted line and the x-axis in the graphs is gradually close to 45 degree.

- **Model Parameters**
  The plot of the baseline and two improvement models' parameters are shown in Fig.9. We could find that the coefficients related to deaths decrease from baseline model to model improved by datetime, and from model improved by datetime to model improved by adding more time-series features. That means that our models depend less on deaths and more on vaccinations and mobility, which is exactly what we want. Since our purpose is to explore the correlation changes of mobility and cases over time.
  Also, we could confirm our guess in 4.2.2. Values of the date a week ago have a great autoregression relationship with the current date's values since the prior 6 day features have higher coefficients than other days.

5.1.3 Performance on Time Scale

- **Improved model performance on next k days**
  To discuss our improved model's performance on both short-term and long-term time scales, we use a metric of RMSE Loss to compare our model's predictions on k next days from present. Our predictions can be written as $f(x) = [x_{t+1}, x_{t+2}, ..., x_{t+m}]$ for some $m > 1$, using the features x that contain past and present quantities to predicts values for m days in the future.
  From Fig.10, the RMSE loss increases with m which means our model performs better on predicting short-term

Figure 9: Coefficients of Baseline and Improved Models

time scales than long-term time scales. However the large RMSE loss and small R_2 score represents it has weak ability in predicting future. After plotting Fig.11, we find it can capture the trend of the dynamic change of COVID-19 cases but it fails to capture the intercept of the linear function, requiring a good shift. We assume $y = g(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(\mu, \sigma)$, in this case,$\mu$ is not zero.

- **Comparison between Baseline Model and Improved Model on k days**
  Though the model in future prediction is pretty depressing, we still see an improvement between our improved

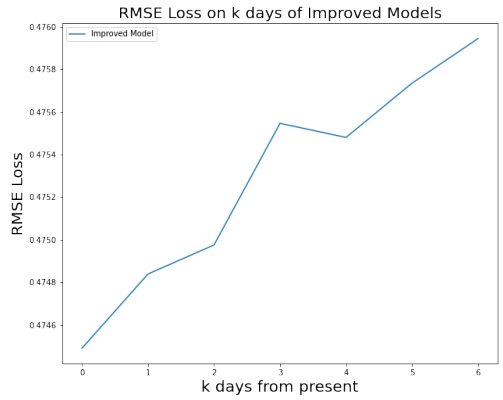

Figure 10: RMSE Loss on k days of Improved Models

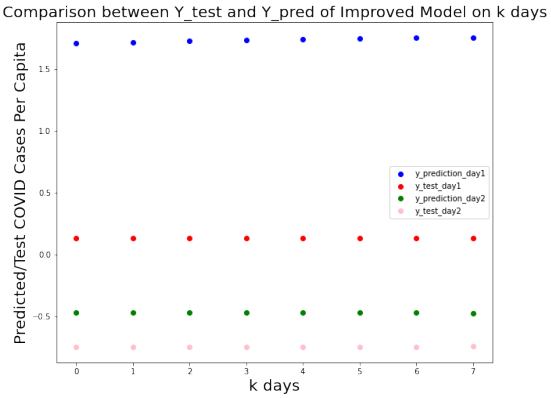

Figure 11: Comparison between Y test and Y pred of Improved Model on k days

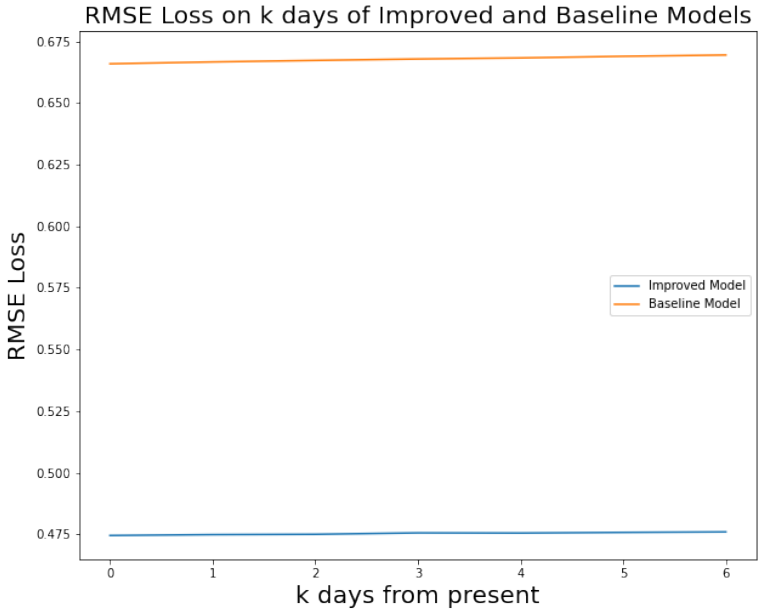

Figure 12: RMSE Loss on k days of Improved and Baseline Models

model and baseline model. From Fig.12, the RMSE loss of our improved model is around 0.475 but the baseline is much higher, around 0.670. That somehow signifies the importance of our model improvement.

5.2 Hypothesis Evaluation (Answer to Hypothesis)

To answer our hypothesis, we need to analyze whether the effect of human activities on the spread of COVID-19 fades over time. Based on our quantitative analysis, we confirm our hypothesis. We have built a linear regression model to analyze the correlation between social mobility as the representative of human activities and COVID-19 cases. To examine the changes over time, we need to observe the changes of feature importance on social mobility, so we split our standardized data into months and feed the month-level features that have been discussed previously into our linear

regression model. Then, we derive the coefficients of each feature in these models since these coefficients can be used directly as a crude type of feature importance score. The higher the absolute value of the score, the more useful the feature is at predicting a target variable, our cumulative COVID-19 cases.

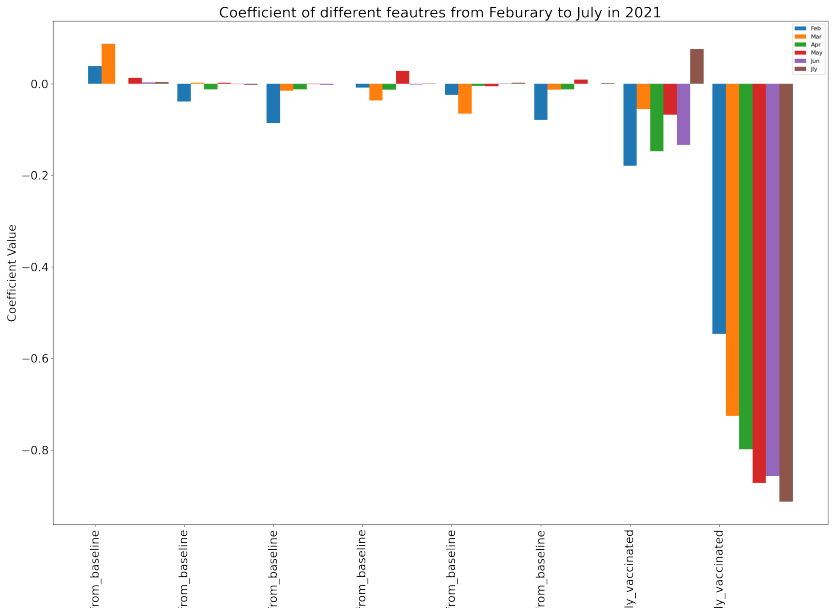

Figure 13: Coefficient of different feautres from Feburary to July in 2021

According to Fig.13, we plot the coefficients of six mobility indexes, fully-vaccinated cases and partially-vaccinated cases from February to July in 2021 to roughly observe the changes of the important scores over time. We can see the coefficients of mobility indexes get smaller and even approach zero from February to July but the coefficients of fully-vaccinated cases that always have the largest weight in these six months keep increasing by 35% which means the influence of mobility indexes on cumulative cases is fading while the effects of vaccinations go up. This observation confirms the validity of our hypothesis.

Figure 14: Normalized Coefficient of mobility of retail and recreation from February to July in 2021

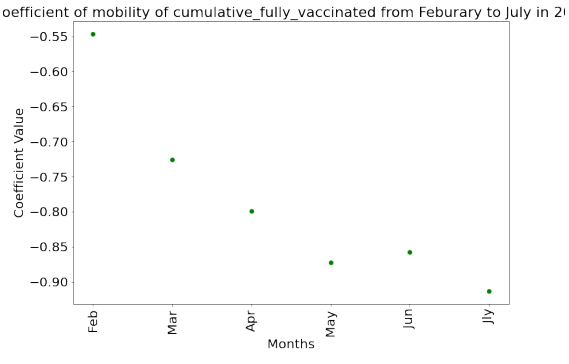

Figure 15: Normalized Coefficient of mobility of cumulative fully vaccinated from February to July in 2021

To be fairer, we normalize the coefficients using (2) in each model and plot the normalized importance score of mobility index of retail and recreation and vaccinations respectively in Fig.14 and Fig.15. These two trends further confirm a substantial change in the relationship between mobility and transmissibility, with a gradual decoupling or dampening of the relationship. As mobility gradually increased, the transmissibility still increased but more slowly than expected from the previously inferred relationship.

6 FUTURE WORK

6.1 Limitations

This project predicts the epidemic trends of COVID-19 in California by constructing a multiple linear regression model and considering human social mobility factors. The aim here is to provide a reference for clarifying the change of the correlation between social mobility and COVID-19 cases and deaths, scientifically formulating an epidemic prevention and control plan, and for promoting the resumption of work and production in an orderly manner. However, this study has the following limitations. First of all, due to the large fluctuations or partial absence in the county-level data of confirmed cases and deaths, there will inevitably be a big deviation between the predicted results and the actual data. Second, due to the restrictions of access to larger datasets, the data we used to build models only ranges from Jan.2021 to Aug.2021 when the epidemic has spread for a year. We can expect to see a more clear change of the effectiveness of human activities restriction in the epidemic control if larger datasets can be introduced. Third, due to the limitations of objective factors such as virus incubation period, the detection capability, and medical and health facilities, the current official epidemic data may be lower than the actual infection data, which will also lead to uncertainty in the final evaluation results. Same inaccuracy also exists in the Google human mobility database. Fourth, from Jan.2021, The emergence of many new mutations of the COVID-19 that are more infectious or virulent, some of which are even resistant to the vaccine, may lead to an increase in the number of cases or deaths at a certain time and place and affect the influence of the vaccine and population mobility in the spread of the COVID-19. Finally, this project only predicts the epidemic trend at the county level in California, and does not put forward effective prevention and control measures, thus restricting the research depth and application value of this paper.

6.2 Future Work

First, using autoregression, the current features and models could capture the regression problem on COVID-19 cases and deaths to a certain extent, but not a hundred percent. Some other features such as the economic level and hospitality could also influence the spread of COVID-19. In the future, we could conduct field research and gather more information to better fit the model and make more accurate predictions.

Second, other granularity in terms of time can be used to build models. In this project, every row in the final dataframe is a daily summary. Although it can mostly retain the preliminary information, it can cause overfitting when we train the models to a certain extent. We could try weekly or monthly to generalize the main trend of COVID-19 cases and at the same time keep the errors on an acceptable level.

Last but not least, we can take 2020's data into our models. We only focus on the 2021 data in this project, which is not very sufficient to train a robust model to justify our hypothesis. The lockdown or social mobility restrictions were enforced mainly in 2020. Although the vaccination rate back then was not high, grabbing the confirmed cases and social mobility data at the end of 2020 can provide a clearer picture of how the weight of mobility restrictions changed over longer periods of time.

7 ACKNOWLEDGEMENT

The authors would like to thank Prof. Fernando Perez, Prof. Alvin Wan, and all GSI's and TA's for their help.

References

[1] Fajnzylber, J. et al. Sars-cov-2 viral load is associated with increased disease severity and mortality. Nat. Commun. 11(1), 1–9 (2020).

[2] Lafzi, A., Boodaghi, M., Zamani, S. et al. Analysis of the effectiveness of face-coverings on the death ratio of COVID-19 using machine learning. Sci Rep 11, 21675 (2021). https://doi.org/10.1038/s41598-021-01005-y

[3] Lalmuanawma, S., Hussain, J. Chhakchhuak, L. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. Chaos Solitons Fract.1, 110059 (2020).

[4] Koh WC, Naing L, Wong J. Estimating the impact of physical distancing measures in containing COVID-19: an empirical analysis. Int J Infect Dis. (2020) 100:42–9. doi: 10.1016/j.ijid.2020.08.026

[5] Sun J, Kwek K, Li M and Shen H (2021) Effects of Social Mobility and Stringency Measures on the COVID-19 Outcomes: Evidence From the United States. Front. Public Health 9:779501. doi: 10.3389/fpubh.2021.779501

[6] https://www.google.com/covid19/mobility/

[7] https://online.stat.psu.edu/stat501/lesson/14/14.1