# COVID-19 Risk Scoring in Los Angeles County

**UMICH-ZJU Team**
Litian Zhou, Wenxue Li, Zhangxing Bian, Yuxuan Cao, Li Xinyu, Weixiao Wang, Zixian Ma
Junhwan Kim, Zijin Chu, Yuxi Xie, Yueze Song, Chaoyi Wang, Ruopeng Wang, Linh Tran
Haozhu Wang (✉), L. Jay Guo (✉)
Email: guo@umich.edu, hzwang@umich.edu

## Abstract

Accurately estimating COVID-19 exposure risk is critical for implementing necessary mitigation and safely reopening the economy. In this work, we propose to assess the exposure risk for communities in the City of Los Angeles by forecasting new confirmed COVID-19 cases per 10,000 population. To this end, we integrate COVID-19 case data, mobility data, and social-economic data for training linear regression and LSTM models to forecast new cases. From the linear regression model, we draw insights on how mobility and social-economic status are associated with COVID-19 exposure risks. Due to the highly complex relationship between features and confirmed cases, the trained LSTM model achieves higher forecasting accuracy than the linear model. Thus, we build a web app for visualizing the estimated risk scores based on the LSTM model. Our risk scoring system could help the general public and policy makers implement timely COVID-19 risk mitigation.

## 1 Introduction

On June 8, 2020, there are more than 6.9 million confirmed COVID-19 cases in over 200 countries [1]. In the United States, over 1.9 million COVID-19 cases and 110,000 deaths have been reported [2]. Without available effective medication or vaccine, non-pharmaceutical interventions are widely implemented to slow down the spread of the disease. Interventions such as limiting public gathering, closing schools and restaurants, as well as shelter-in-place order, were used as the pandemic affects more and more individuals in the United States. These social distancing measures aim to change people's mobility patterns and consequently reduce the COVID-19 spread. Therefore, analyzing population mobility data enables us to assess the effectiveness of social distancing and to predict future infection risks.

In this technical report, we define the COVID-19 exposure risk as the new cases per 10,000 population. We aim to predict risk scores of regions in LA county using data available from the RMDS computational competition [3] by building a linear model and LSTM models. For training the models, we integrate heterogeneous data sources including COVID-19 case data, population mobility data, and social-economic data. For COVID-19 case count, we used data published by the Tracking Coronavirus in the Los Angeles County Project [4]. Daily and weekly risk scores are analyzed for ZIP codes in the LA county and for the LA county as a whole. Population mobility data is published by Google[5], Apple[6], and SafeGraph[7] that are listed on the RMDS website. Additionally, previous literature mentioned that socio-economic factors such as income level and profession profoundly determine whether one is able to conduct work from home to reduce their exposure risk to COVID-19[8]. Therefore, we incorporated the median house income of each zip-code region as a feature in our models.

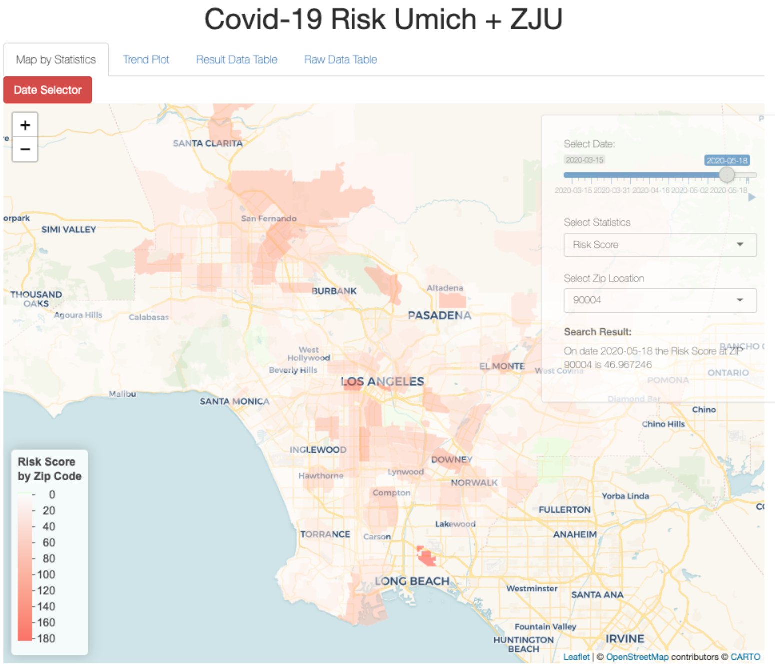
With the coefficients of the trained linear regression model, we identified features that are associated with high COVID-19 exposure risk. Our finding suggests that strict risk mitigation policies should be implemented for communities with high population density and/or low average income. Because the

2020 RMDS COVID-19 Challenge. Technical Report.

trained LSTM model achieves higher forecasting accuracy than the linear regression model, we build a website to visualize the forecasted risk scores based on the LSTM. We believe that our risk scoring system can facilitate efficient mitigation for COVID-19.

We summarize our contributions:

1. We build a data modeling pipeline that integrates case count data, mobility data, and social-economic data for forecasting new cases.
2. We define the regional risk score as predicted new cases divided by the population.
3. We build a website[1] for visualizing the forecasted risk scores.
4. Our risk scoring system could be a useful tool for the general public and policy makers to practice timely risk mitigation.



Covid-19 Risk Umich + ZJU

Select Date: 2020-03-15 2020-05-18

Select Statistics: Risk Score

Select Zip Location: 90004

Search Result:

On date 2020-05-18 the Risk Score at ZIP 90004 is 46.967246

Figure 1: Screenshot of the risk score heat map from our website. Users can use the date slider to choose a specific date for the plot.

## 2 Data Prepossessing and Feature Extraction

### 2.1 Cumulative and new confirmed case data

The cumulative confirmed cases of neighborhoods in LA county and the coordinates of these neighborhoods are gathered from The Los Angeles Times [4]. These data are first cleaned up by removing the special characters such as * in their names, then we matched these names of places with their corresponding ZIP codes.

We used the python module geopy to transform the coordinates corresponding to various places into ZIP codes. We manually entered ZIP code coordinates that cannot be identified by geopy. Then, we calculated the cumulative confirmed cases for each area corresponding to a ZIP code. Also, the number of daily new cases by each ZIP code was then computed by taking the difference between cumulative confirmed case count of two consecutive days,

[1] https://biostat-umich.shinyapps.io/Covid-19_Risk_Umich_ZJU/

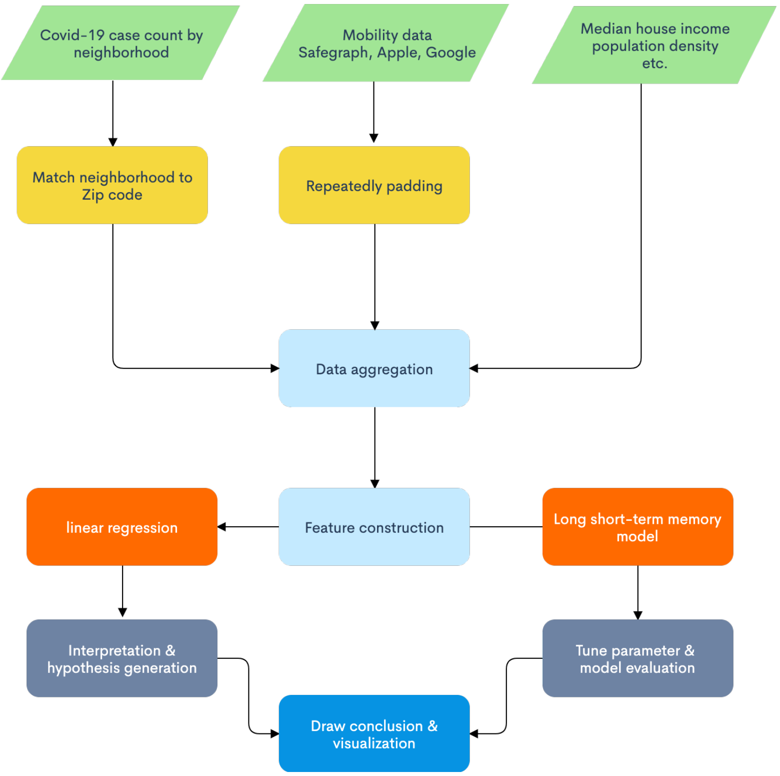

Covid-19 case count by neighborhood
Mobility data Safegraph, Apple, Google
Median house income population density etc.
Match neighborhood to Zip code
Repeatedly padding
Data aggregation
linear regression
Feature construction
Long short-term memory model
Interpretation & hypothesis generation
Tune parameter & model evaluation
Draw conclusion & visualization

Figure 2: The framework of this project. The color of nodes represent different stages of the modeling pipeline. Green: data collection; yellow: data cleaning; cyan: data preprocessing; orange: model training; grey: model evaluation; blue: conclusion and visualization.

### 2.2 Mobility data

The mobility data used in this report are collected from three sources: Google, Apple and SafeGraph.

**Google Mobility Data.**[5] As shown in Fig.3, Google mobility data contain movement trends collected over time by apps like Google Maps. The dataset is created with aggregated, anonymized sets of data from users who have turned on the Location History setting. But no personally identifiable information, such as an individual's location, contacts or movement, is made available.

These mobility trends are aggregated into six different categories: retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Changes for each day are compared to a baseline value for that day of the week. The baseline is the median value, for the corresponding day of the week, during the 5-week period between Jan 3 and Feb 6, 2020.

Table 1: The details of Google Mobility Data.

| **Feature** | **Description** |
|---|---|
| Retail & recreation | Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters. |
| Grocery & pharmacy | Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies. |
| Parks | Mobility trends for places like national parks, public beaches, marinas, dog parks, plazas, and public gardens. |
| Transit stations | Mobility trends for places like public transport hubs such as subway, bus, and train stations. |
| Workplaces | Mobility trends for places of work. |
| Residential | Mobility trends for places of residence |

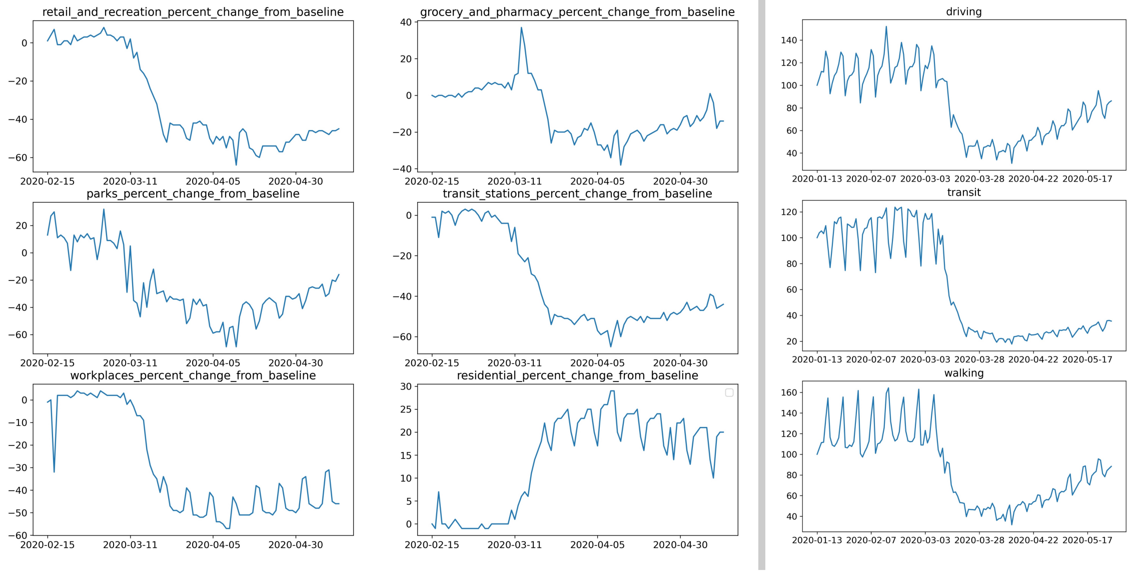


Figure 3: Google(left) and Apple(right) Mobility Data.

**Apple Mobility Data.**[6] As shown in Fig.3, similar to the Google mobility data, Apple mobility data is collected through Apple Maps. The dataset features daily changes in requests for directions by transportation types: walking, driving, transit. Data for May 11-12 are not available, so we used the data of May 10 and May 13 to linearly interpolate the missing values.

**SafeGraph Mobility Data.**[7] SafeGraph mobility data ("social distancing metrics data") is generated by a panel of GPS pings from anonymous mobile devices. Since devices are aggregated by home census block group and the metrics are further computed to the county level, data pre-processing is needed first to allocate the 12-digit block group number into LA ZIP codes according to largest residential address ratio. After removing unnecessary columns, we retain the following features: i) device_count and completely_home_device_count, ii) distance_traveled_from_home, iii) median_home_dwell_time and median_non_home_dwell_time. The following operation aims to generate daily data at ZIP code level. For columns like device_count and completely_home_device_count, we simply averaged them, and for remaining columns despite median_percentage_time_home , we used weighted averages with their device_count to get averaged results. The median_percentage_time_home data is computed by median_home_dwell_time and median_non_home_dwell_time. We applied similar processing methods as above to generate the weekly data and county-level data, and finally obtained four types of data: (i) ZIP code-level daily data (ii) ZIP code-level weekly data (iii) LA county-level daily data and (iv) LA county-level weekly data. Here, the weekly data starts with the day with first confirmed case in the ZIP code and last seven days. The last week with less than seven days are not considered.

### 2.3 Population and income data

ZIP code level population data and regional area data are gathered from world population review website [9] and an existing GitHub repository [10]. Only two ZIP codes (91355, 93590) don't have corresponding population data. Since these ZIP codes only each had one cumulative confirmed cases and since they are rural areas in Los Angeles county, we removed them from the dataset. Using all the aforementioned data, daily cumulative cases per 10,000 population, daily new cases per 10,000 population, and the population density were calculated for each ZIP code.

Estimated median household income for each ZIP code in LA county are obtained from the 2017 census data [11]. Using these data, these ZIP codes areas are categorized into one of three income levels: i) areas with median household income less than $45,000 are labeled as 'low income'. ii) areas with median household income between $45,000 and $140,000 are labeled as 'medium income' 2. iii) areas with median household income higher than $140,000 are labeled as 'high income'[12].

# 3 Modeling Methods

## 3.1 Problem formulation

Lagging effect exists for COVID-19 exposure and confirmation. The lagging effect is the time delay between the exposure of SARS-CoV-2 to the final confirmation of COVID-19 infection. After reviewing recent epidemiology and medicine journals[13, 14], we carefully chose this lagging effect as 7 to 10 days. In another word, if a person get exposed and infected by SARS-CoV-2, we believe it will take on average 7 to 10 days for the local disease surveillance system to receive and publish the infection confirmation information.

We consider the lagging time window to contain following 3 periods: the incubation period, the symptom development period, and the lab test and release period. The incubation period is the time from the moment of exposure to an infectious agent until signs and symptoms of the disease appear. The median incubation period of COVID-19 is 5.1 days (95CI: 4.5-5.8) [14]. The symptom development period is 1-2 day, and the lab test and release period is 2-4 days according to testing companies [15].

Given above considerations, we formulate the new case prediction task as a time-series forecasting problem. Specifically, our model takes the time series data for the past $n$ days $\bar{\mathbf{x}}_{t-n:t-1} = \{\mathbf{x}_{t-n}, \mathbf{x}_{t-(n-1)}, \dots, \mathbf{x}_{(t-1)}\}$ as the input to predict the average new cases $\bar{y}_{t:t+m}$ for the next $m$ days. Here, we set the input sequence length $n = 6$ and *prediction target* length $m = 4$ based on the lagging effect between 7 to 10 days. Finally, we assign the average new cases $\bar{y}_t$ as the risk score $r_t$, i.e.,

$$r_t \triangleq \bar{y}_{t:t+3} = \frac{1}{4}\sum_{i=0}^{3} y_{t+i}, \tag{1}$$

where $y_t$ is the confirmed new cases per 10,000 population on day $t$.

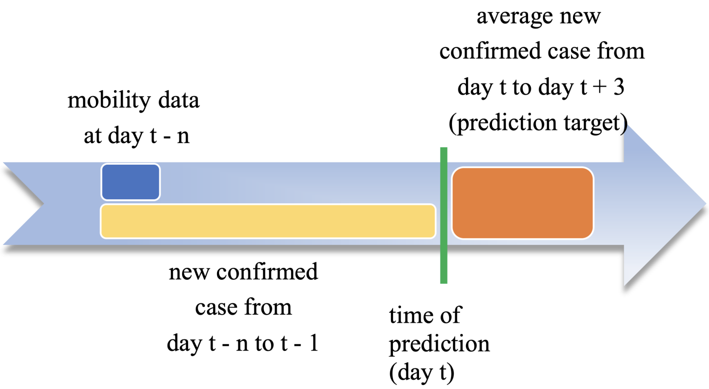


Figure 4: The temporal relationship between features and outcomes in one observation, the blue arrow represents the time direction. Day $t - n$ to day $t - 1$ is the input sequence, which is used to forecast the average cases in the prediction target window.

### 3.1.1 Linear regression model

We first fitted a linear regression model to predict new confirmed cases. There are three types of predictors from the data we collect, namely case count data of previous dates, mobility data and other social-economic data such as population density and median house income. The total number of features is 23.

To automatically select the predictors to balance the complexity and fitness, we performed step-wise selection method [16] based on root mean square error (RMSE) calculated by 10-fold cross-validation. Also, some predictors with low correlation with the outcome are dropped, which can increase the quality of our interpretation. Then, we plot the feature importance calculated by t-test statistics. See Figure 6.

### 3.1.2 Long short-term memory model

We also trained LSTMs to forecast new cases [17] on two datasets, one is daily dataset and another is weekly dataset. For daily dataset, it means that the units of daily dataset are features of one day. For weekly dataset, it means that the past history data is aggregated on the weekly level, then used as input to the LSTM model. Both of these datasets are of ZIP code level and contain 23 features including confirmed cases, social-economic data such as population and income level, and mobility data from Google, Apple Inc. and SafeGraph ("social distancing metrics data"), as mentioned above.

For the ZIP code level daily dataset, the types of predictors are similar to the linear model, which include case count data, social distancing data, google and apple mobility data and some social-economic data such as population and income level. Figure 4 shows the generic input and output of LSTM based on ZIP code level daily dataset.

To estimate the importance of data to the performance of LSTM, We experimented two different outputs. One was the new cases for day $t$ and another was the average new cases $\bar{y}_{t:t+3}$, both with the same input sequence $\bar{\mathbf{x}}_{t-6:t-1}$. We found that the network outputs $\bar{y}_{t:t+3}$ achieves better performance than the other one. We believe it's because the average new cases $\bar{y}_{t:t+3}$ has less variance than $y_t$.

The training details are described as follows. We first normalized all features and randomly divided the whole dataset into training set, validation set and test set, respectively. The proportions of the training set, validation set and test set is 70%,15%,15%. We use Adam optimizer to train the LSTM network. The initial learning rate is 0.0003, and we used an adaptive learning rate schedule. Specifically, when validation loss doesn't decrease for seven epochs, we reduce the learning rate. The number of training epoch is set to 1,500. The training process was early stopped when validation loss doesn't decrease for 24 epochs.

For the zip code level weekly dataset, the types of predictors are the same as the network using zip code level daily dataset. We split the training data and testing data temporally, i.e., for each Zip code, we partition the data into training/validation/test (70%//15%/15%). We input the data from the past two weeks to predict average new cases for the following week.

We also trained models to predict new cases in county level. However, the trained model has poor prediction accuracy because the training set is too small on the county level. Thus, we only report our results for zip-code level models in the next section.

## 4 Results

### 4.1 Interpretation of linear model

After step-wise selection, the model with best performance (RMSE = 3.357, R-square = 0.635) contains 17 variables. See Figure 5.

We can categorize the predictors into three groups: 1) case count, 2) social-economic data, and 3) mobility data.

The case count predictors have the largest coefficients and explained most of variance in our risk. The dates close to the time-of-prediction have higher coefficients than earlier dates. This aligns with our intuition that the COVID-19 cases grow in a changing rate, and the most recent daily new cases carry the most predictive weights for the following days.

From the social-economic data, the positive correlation between population, population density, and risk score shows that communities with higher population density in the LA county are associated with higher risk. Additionally, lower income levels are also associated with higher risk.

The mobility data (social distancing metric) from Safegraph has little correlation with the risk in our linear model. The reason could be spatial data has a more complex pattern which linear model cannot capture. However, the mobility data from Apple and Google can provide many insights. As far as the transportation is considered, driving, compared to walking and public transit, has a strong negative correlation with the risk. This indicates that driving during the pandemic has a significantly lower exposure risk to the virus compared to walking outside. Regarding to the venue of gathering, however, the message is less clear. The visits to grocery, pharmacy and parks also corresponds to low risk. The reason could be that at the end of May, when the infection has passed its peak, people

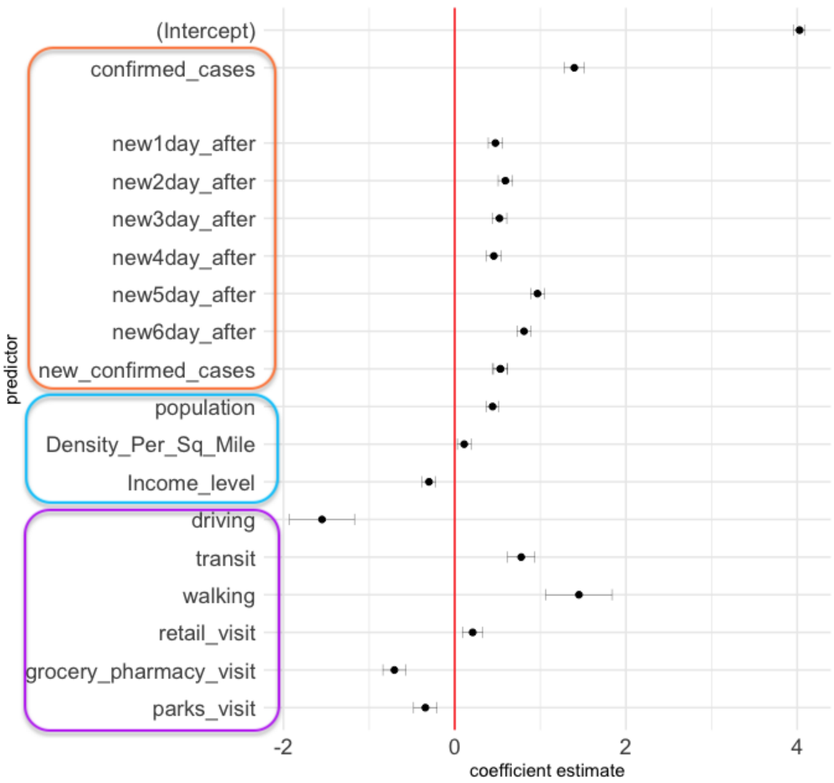


Figure 5: Coefficients of fitted linear regression model.

starts going out more. From the coefficients of the linear model, the impact of each predictor can be quantified in terms of average new case increment. For example, if driving percentage increases by one standard error, the average new confirmed cases will drop by 1.55 in the following risk score window. This is a considerable drop considering the average new confirmed cases across the LA county is only 4 per day.

### 4.2 Prediction accuracy of the LSTM model

We have conducted a range of testing to evaluate our proposed models in this section. In the following experiments, we use daily dataset in ZIP code level, and weekly dataset in ZIP code level to evaluate the trained LSTM models. The performance and comparison are as listed in Table 2.

Table 2: Test peformance of trained LSTM models.

| dataset | | RMSE | | dataset size | | |
|---|---|---|---|---|---|---|
| | | original | scaled | training set | validation set | test set |
| daily | predict 1 day | 4.5873 | 0.7946 | 5676 | 1216 | 1217 |
| | predict 4-day average | 2.2988 | 0.4473 | 5676 | 1216 | 1217 |
| weekly | temporal split | 20.7491 | 0.6190 | 827 | 170 | 170 |

Model outputs versus ground-truth targets on the test dataset are plotted in Fig.**??**.

For the daily dataset, we found that the LSTM network that predicts $\bar{y}_{t:t+3}$ achieves a better performance than the other one, because the average new cases $\bar{y}_{t:t+3}$ has less variance than a single day. We also found that the RMSE of LSTM using daily dataset to predict 4-day average new cases is smaller than the linear model's best performance. Thus, we will use the LSTM model to predict the final daily and weekly new cases results.

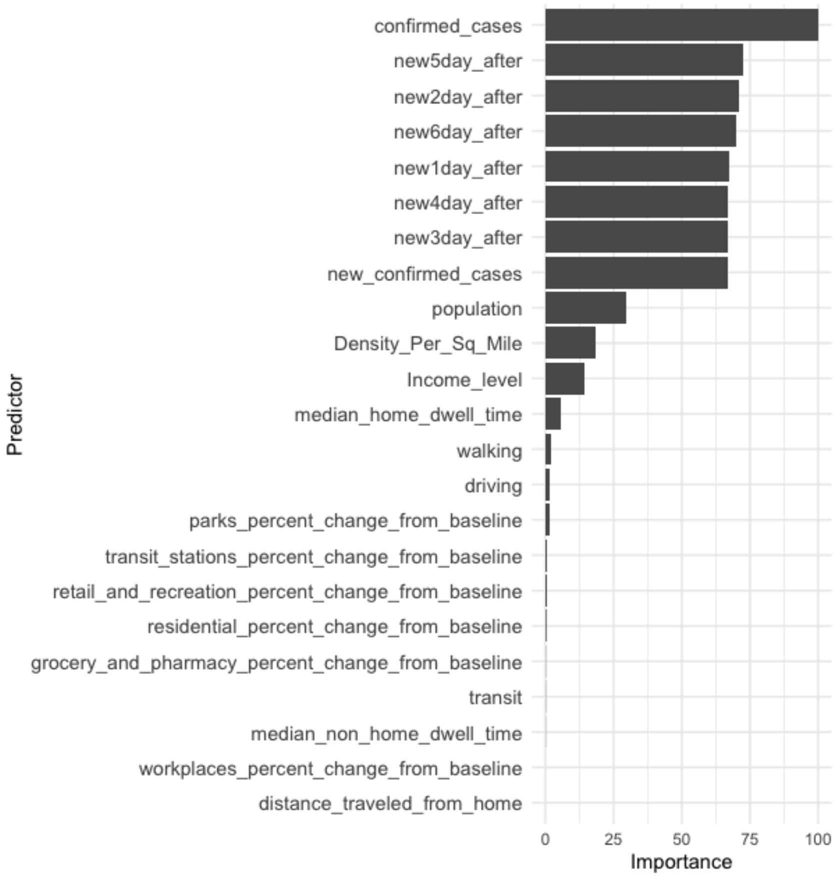


Figure 6: Feature importance of the linear regression model (count = 17). The importance is measured by the absolute value of the t-statistic for each model parameter is used. Here the most important variable: confirmed_cases is set as 100.

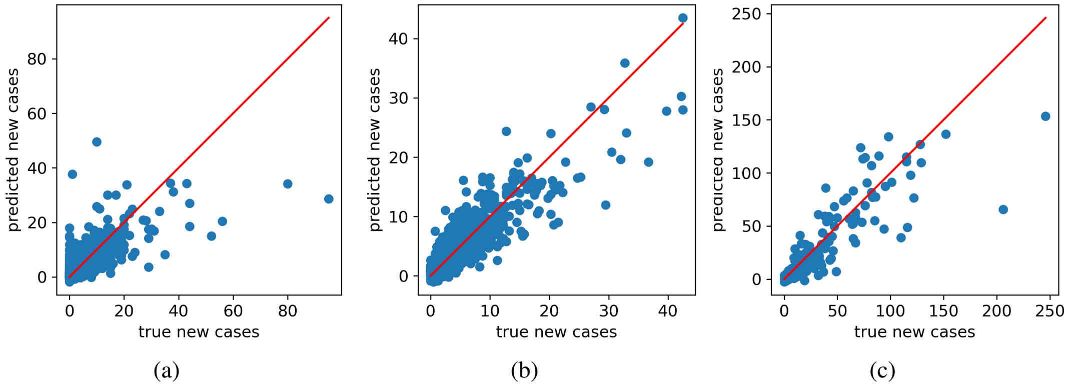


Figure 7: Scatter plots for the prediction on daily dataset (predict 1 day (a), predict 4-day average (b) ) and weekly dataset with a temporal split (c).

### 4.3 Web app

We built a website using Rshiny to visualize the risk score and other statistic. The risk score, mobility and infection rate of each ZIP code are shown on the map with the specified date. The trend plot shows these statistic across the timeline including a historical and a predicative part. Also we provide a dynamic data table of the raw data used in our model for users who like to download and reproduce our findings.

## 5 Conclusion

Both of our models are based on an assumed lagging effect of 7 to 10 days between COVID-19 exposure and confirmation. The limitation of the definition of lagging effect is that the actual delay may change throughout our study period due to many factors, which may affect the forecasting accuracy of our trained models. For example, the public awareness of the COVID-19 pandemic and lab efficiency dramatically increased in May, which potentially shortened the lagging time. Lagging time at the beginning of the spread could be longer than the one we used due to the lack of disease information and testing unavailability.

Our current models are trained on case data, mobility data, and socio-economic data. The trained LSTM achieved a reasonable performance, outperforming the linear regression model. However, the forecasting accuracy is far from perfect. In the future, we plan to incorporate three more factors when training our models to further boost the forecasting accuracy. The first type of data is transaction of debit cards and payroll, which can be used to measure the impact of COVID-19 across different business sections. The second factor is the recovery/death data. The risk score should decrease as more and more patients recovery and gain immunity. However, since the health agencies release the recovery/death data irregularly, we cannot decide the proportion of immunity population at a certain time. Also, infection is too low for each ZIP code area to obtain herd immunity. Thus, We will estimate the immune population by the positive test rate and asymptomatic infection rate from other data sources. At last, since different ZIP code does not have real boundaries, the exposure risk of one community is dependent on the exposure risks of its neighboring communities. To give a more accurate risk score and provide suggestions on reopening the LA city, we will use the proximity between ZIP codes and aggregate “internal risk” and “external risk” into a more informative score.

In conclusion, we define the risk score on day $t$ as the average new confirmed COVID-19 cases per 10,000 population for the following four days. Linear regression and long short-term memory network are trained for to forecast new cases. From linear regression, we obtained insights from three categories of predictor: the confirmed case count, social-economical and mobility data. The LSTM model achieved better than the linear regression model. Although the linear regression model can be trained faster than LSTM and is more interpretable, it is difficult to fit the nonlinear relationship in the time-series data. Thus, we deployed the LSTM model for forecasting future risk scores and visualize the results with a website. At the same time, due to time and computation resource constraints, our current LSTM model is relatively simple, and there is still much room for improvement. In the future, we plan to train graph neural networks [18] to better model the spatial-temporal structure of the data.

### 5.1 Risk Mitigation Recommendations

To mitigate the future COVID-19 exposure risk, policy makers should consider maintaining strict risk mitigation measures for the high-risk communities. When travel is necessary, government should encourage people to use personal vehicles and avoid public transportation. Thus, drive-thru service is an effective way to limit the spread of the virus. The social distancing rules should be strictly observed during outdoors activities especially in places involving walking such as parks. The area with high population density and low income level needs more effort to flatten the curve. The business in such area should consider extending their lock-down period or gradually reopening when the forecasted exposure risks start to drop.

## References

[1] WHO. Coronavirus disease (covid-19) pandemic. `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`, 2020.

[2] CDC. Cases in the u.s. `https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html`, 2020.

[3] RMDS. 2020 covid-19 computational challenge. `https://grmds.org/2020challenge`, 2020.

[4] The Los Angeles Times. The tracking coronavirus in los angeles county project. `https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/los-angeles-county/`, 2020.

[5] Google LLC. Google Mobility Data. `https://www.google.com/covid19/mobility/`, 2020.

[6] Apple Inc. Apple Mobility Data. `https://www.apple.com/covid19/mobility`, 2020.

[7] SafeGraph Inc. SafeGraph Social Distancing Metrics. `https://docs.safegraph.com/docs/social-distancing-metrics`, 2020.

[8] Abigail Adams-Prassl, Teodora Boneva, Marta Golin, and Christopher Rauh. The large and unequal impact of covid-19 on workers. *VoxEU. org. Library Catalog: VoxEU. url: https://voxeu.org/article/large-andunequal-impact-covid-19-workers*, 2020.

[9] worldpopulationreview.com. worldpopulationreview. `https://worldpopulationreview.com`, 2020.

[10] github reposity. population by zipcode. `https://github.com/sharding1023/Springboard-Capstone-Project`, 2020.

[11] Los Angeles Almanac. Median Household Income by Zip Code Los Angeles County, 2017. `http://www.laalmanac.com/employment/em12c.php`, 2020 (accessed June 1, 2020).

[12] U.S. Census Bureau. Hinc-01. selected characteristics of households by total money income: All races. `https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.html`, 2020.

[13] Charles Courtemanche, Joseph Garuccio, Anh Le, Joshua Pinkston, and Aaron Yelowitz. Strong social distancing measures in the united states reduced the covid-19 growth rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed covid-19 cases across the united states. *Health Affairs*, pages 10–1377, 2020.

[14] Bi Q et al. Lauer SA, Grantz KH. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Ann Intern Med*, 2020.

[15] LabCorp. Labcorp's testing for covid-19. `https://www.labcorp.com/assets-media/2330`, 2020.

[16] Zhongheng. Zhang. Variable selection with stepwise and best subset approaches. *Annals of translational medicine vol. 4,7 (2016): 136.*, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.