2020 COVID-19 Computational Challenge

Report

Team: LMU MSBA

Members: Eric Wu, Kayla Tanli, Rongxing Chen, Zuo Zuo

Mentor: Richard Zhen Tang

8 June 2020

Abstract

We define risk as the product of hazard and vulnerability, in which hazard is a universal event-specific negative impact, and vulnerability is location-specific socioeconomic features that determine to which extent the focal location will be influenced by the negative impact. Further, we classify the vulnerabilities into two categories—infection vulnerability and serious-condition vulnerability—to fit the nature of this Covid19 disease better. We build two SEIR models that have very strong predictive power on the total cases of both infection and deaths. The predictions then are used as the base of the hazard.

Our risk score system can capture the dynamics of the pandemic and relevant events (e.g., protests of George Floyd's death) through its hazard component and generate location-variant risk scores through its vulnerability component. A distinction between risks of infection and serious-condition gives a more nuanced risk evaluation.

Lastly, the main data that our risk score system uses is very simple and publicly available (from public Covid19 Cases dataset, US census, and American Community Survey), making our model and risk score system very accessible to general users.

Executive Summary

COVID-19, an infectious disease, has caused a world-class panic in 2020. As of June 2nd, there are about 6 million global confirmed cases and approximately 370,000 deaths around the world. Since no cure has been proved to work yet, we are relying on isolation methods such as city lockdown, shutting down businesses, and masks to cut off infections. In addition to the medical benefits from isolation, these necessary methods are actually causing sizable economic loss, and that brings us another question: when should we stop with those methods?

In this report, we are going to focus on the trade-off between shutting business and reopening business. In order to do so, we would first make a prediction on death and confirm cases using an epidemic model (SEIR). Then we would use the prediction and other data (in medical resources and demographic features) to come up with a measure of risk, quantifying each location's risk exposure of Covid19 in reopening. Based on the risk scores of locations, planners can more effectively target and support community-based efforts to mitigate and prepare for disaster events.

Table of Contents

Introduction

COVID-19 has been spreading throughout the world at an alarming rate. According to the World Health Organization, as of 2 June 2020, 216 countries have been affected worldwide. In addition to that, there have been about 6 million global confirmed cases, and approximately 370,000 deaths. As of now, there is no cure for the disease.

In Los Angeles County, there are 57,118 cases and out of those cases, 2,443 people passed away. According to NBC News, the first case detected in Los Angeles was in January of this year. When the first death occurred in California in early March, Governor Gavin Newsom urged the population to avoid events with large crowds. Since then, stay-at-home orders have been issued. It was not until recently that the California governor has relaxed the rules and standards. The county plans a full safe reopening as early as 4 July 2020. (KABC, “Coronavirus: Officials aim for 'safe reopening' of Los Angeles County as early as July 4”)

With the county slowly reopening California these past few weeks, there is a higher risk of people to be easily infected. The current protests on the death of George Floyd also have to be taken into account as the gathering of large crowds could lead to an increase in coronavirus cases.

Therefore, the objective of our study is to develop a location-based score system that could quantify the risk exposure to the COVID-19 in the process of reopening. In the current stage, we focus on locations in the city of LA.

Methodology

1. Definition of Risk

According to the formula that CDC uses to quantify community vulnerability to a disaster (Flanagan, et al. 2011), we define the risk exposure of one location to the Covid19 as:

$$(1)\ Risk = Hazard * (Vulnerability - Resources)$$

In the formula[1], the hazard is a condition posing the threat of harm, and vulnerability is the extent to which the focal entity will be impacted (Flanagan, et al. 2011). In addition to that, the hazard is event-based, and vulnerability is location-based. Therefore, our definition of risk is able to incorporate the risk exposure derived from the disastrous event and the focal locations under analysis, suitable to our objective in this study.

2. Quantifying Vulnerability

The location-based vulnerability can be quantified with the socioeconomic features that are available from the US Census of Population and American Community Survey[2]. In this study, we use an open dataset compiled from the two data sources by the City of LA[3]. Features such as poverty rate, traffic density, population density, senior population, asthma rates, and cardiovascular disease rates are arranged at the census tract level.

2.1 Infection Vulnerability and Serious-condition Vulnerability

The innovation of our study is in classifying ‘vulnerability’ into two categories: one for getting infected, and one for getting into serious conditions. Theoretically, patients of Covid19

are in various conditions: while many patients can self-cure without any specific medical treatment, a significant amount of patents will become serious conditions. Mixing the risk of infection and the risk of serious-condition prevents us from having a comprehensive understanding of the disease. Empirically, features like poverty rate, traffic density, and population density are closely related to the vulnerability of infection, because the high density of people and traffic exposes people to many contacts, and poverty makes people unable to afford to stay at home, both increase the probability of infection. On the other hand, features like senior population, asthma rates, and cardiovascular disease rates are more relevant to the vulnerability of serious conditions rather than infection.

Thus, in our analysis, we first normalize the relevant features of each type of vulnerability as they are measured in different units and magnitudes. Then, we take the sum of the normalized scores. Lastly, we multiple the scores with the predictions of new death cases and infected cases to generate the risk scores for serious condition and infection, respectively.

We can observe that some minority groups experienced higher rates of infection and serious conditions; however, we do not want to put the race in our model to quantify the risk; we can avoid any unnecessary causal interpretation of our findings. We need to emphasize the higher rates in some minority groups is because of some preexisting unfairness in their employment and medical resources. We need to work together to erase the unfairness, and that’s all.

3. Quantifying Hazard

Hazard is the event-based negative impact to all locations. And again, the impacts can be more of a threat of infections or of serious conditions. Hence, we use the prediction of newly infected cases for the former and the prediction of the new deaths for the latter.

To make the prediction, we tried two approaches: (1) a Recurrent Neural Network (RNN) model with Long-Short-Term-Memory (LSTM) mechanism, (2) a compartmental model—SEIR. RNN with LSTM is purely a machine learning method; its predictive power relies on its ability to unveil unobvious patterns in data, which in turn requires a large amount of data. Given the fact that we only[4] have a few months' observations in LA County and even fewer for LA City, the RNN model does not perform well in our study. In contrast, the SEIR model is rooted in epidemic theory, and the embedded theory could mitigate the negative impact of the lack of data. Further, SEIR is much more interpretable than the RNN model. We then choose SEIR as our main prediction approach.

4. SEIR model

The SEIR Model[5] classifies population (N) into several compartments: susceptible (S) are those who are not infected; exposed (E) are those infected but not infectious yet; Infectious (I) are those infectious, Dead (D) are those unfortunately deceased, and Recovered (R) are those recovered (see Figure 1).

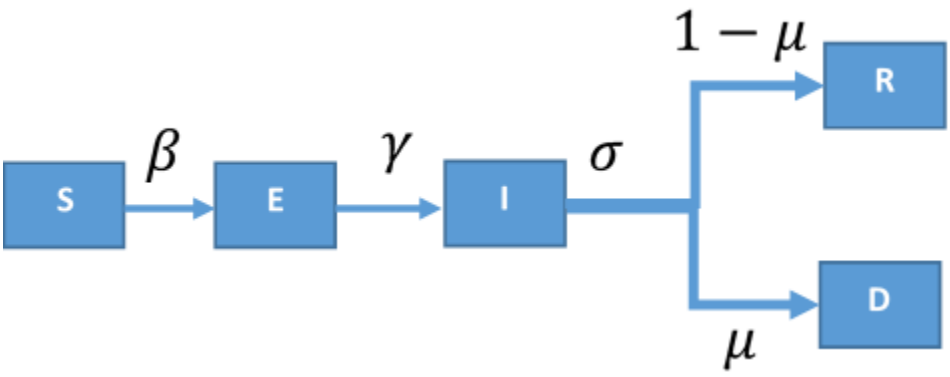


Figure 1. SEIR Model

Traditional SEIR models can be represented by the differential equation system below:

$$(2)\ \frac{dS}{dt} = -\beta \frac{S_t I_t}{N_t}$$

$$(3)\ \frac{dE}{dt} = \beta \frac{S_t I_t}{N_t} - \gamma E_t$$

$$(4)\ \frac{dI}{dt} = \gamma E_t - \sigma I_t$$

$$(5)\ \frac{dR}{dt} = (1-\mu)\sigma I_t$$

$$(6)\ r = \frac{\beta}{\sigma}$$

These models describe the flow of the population among the compartments at any given time. Solving the equation system with some starting value will give us a simulation of the status of the disease. Fine-tuning model parameters according to observed data will give us a model that could fit the reality.

Details about the SEIR model setup and estimation are available in the attached R script.

4.1 Variable Base Reproduction Number and Lockdown Fatigue

Another innovation in our SEIR model is that we use the variable base reproduction number for different time periods. Reproduction number is the average number of people one infected case could infect. This is a core parameter for an SEIR model. In reality, we experienced a complete lockdown and now are gradually reopening. The reproduction number should be different for the two stages. Further, society could not be shut down for too long, as the shelter-at-home order lasts, people may not obey the order and thus increase the reproduction rate. Thus, in our model, we have a parameter, lockdown fatigue (*f*), to capture the dynamics in the reproduction number.

$$(7)\ \frac{dr}{dt} = f$$

4.2 Separate SEIR Models for Infected Case and Death

As this pandemic spreads the world quickly, many countries are not prepared with necessary resources like testing kits. We would argue that the reported infected cases are far from the true numbers, whereas the death numbers should be less noisy in this regard (patients who have died are more likely to get tested). Therefore, an SEIR model that fits the infected cases data well might not be able to predict the death cases well, and vice versa. Thus, we built two separate SEIR models for the prediction of infected cases and deaths, respectively.

4.3 Training the Model and Model Performance

The model is fine-tuned with grid-search with the prior values of parameters set according to the previous study[6]. In our case, grid-search is a better optimization algorithm than other algorithms that rely on gradient descending, because there are many local minimums in our objective function that traps the gradient-based algorithms.

Our models outperform the bash models in both types of the predictions (see the attached model fit summary for details).

Data

Even though the data we use in the study is from many different data sources (as we described below), we could run the model and generate the risk score with much less data too. To have our model and score system work, we just need to have Covid19 cases data for the SEIR model to make a prediction, and location-based features to generate the risk scores. Next, we introduce you to all the data we explored.

The data used in the analysis comes from many different sources that include:

1. The California Department of Public Health
    - COVID-19 Cases (https://data.chhs.ca.gov/dataset/6882c390-b2d7-4b9a-aefa-2068cee63e47/resource/6cd8d424-dfaa-4bdd-9410-a3d656e1176e/download/covid19data.csv)
2. Google
    - COVID-19 Mobility Reports (https://www.google.com/covid19/mobility/)
3. City of Los Angeles Open Datasets
    - City of Los Angeles COVID-19 Indicators (https://github.com/CityOfLosAngeles/covid19-indicators)

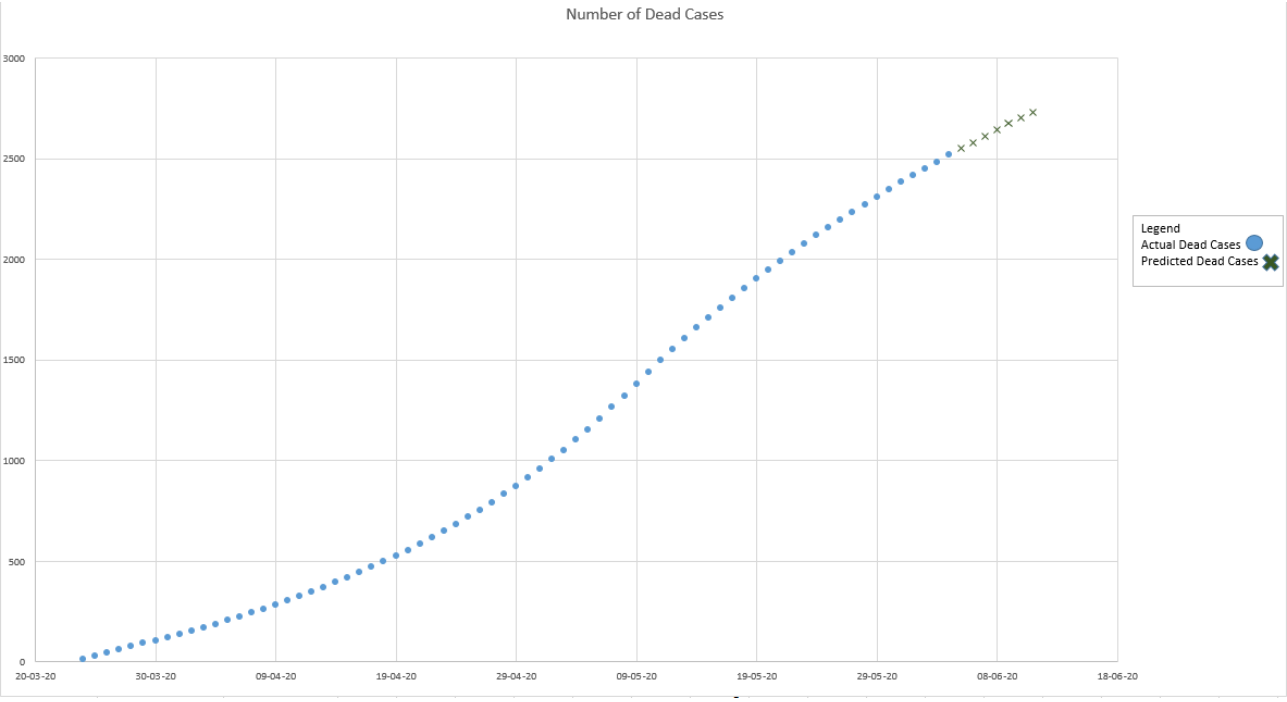
The final dataset to be used in the analysis. is a merged dataset based on the datasets mentioned above. The data was joined on the date, and on a county-level. It is also a time-series data. The dataset contains information on population mobility in different types of establishments, information on COVID-19 patients, the number of confirmed and death cases, and details on ICUs. The dataset contains the following columns (Note that we did not use all the data we list here. We list them so that interested readers can explore them more):

1. Date_new: the date that the data was recorded from April to May.

2. Retail_and_recreation_percent_change_from_baseline: The mobility trend for retail and recreation establishments. The baseline is the mobility rate for a five-week period from 3 January 2020 to 6 February 2020.
3. Grocery_and_pharmacy_percent_change_from_baseline: The mobility trend for grocery and pharmacy establishments compared to the baseline.
4. Parks_percent_change_from_baseline: The mobility trend for parks compared to the baseline.
5. Transit_stations_percent_change_from_baseline: The mobility trend for transportation services compared to the baseline.
6. Workplaces_percent_change_from_baseline: The mobility trend for workplaces compared to the baseline.
7. Total Count Confirmed: The cumulative total number of confirmed COVID-19 cases.
8. Total Count Deaths: The cumulative total number of deaths from COVID-19.
9. COVID-19 Positive Patients: The number of patients who are infected by COVID-19 is in hospitals. This number is not cumulative.
10. Suspected COVID-19 Positive Patients: The number of patients who show symptoms of COVID-19 and have tests that are pending confirmation.
11. ICU COVID-19 Positive Patients: The number of patients who are infected by COVID-19 and are in the ICU. This number is not cumulative.
12. ICU_available_count: The number of beds available in the ICU.
13. Performed: Number of COVID-19 tests performed.
14. Cumulative: Number of cumulative COVID-19 tests performed.

Results

Our model achieved good prediction performance without overfitting the data. More discussion on avoiding overfitting is in the attached R script. Here we just illustrate the prediction performance of our model compared with the benchmark model that uses the previous week's average as the prediction for the current week.



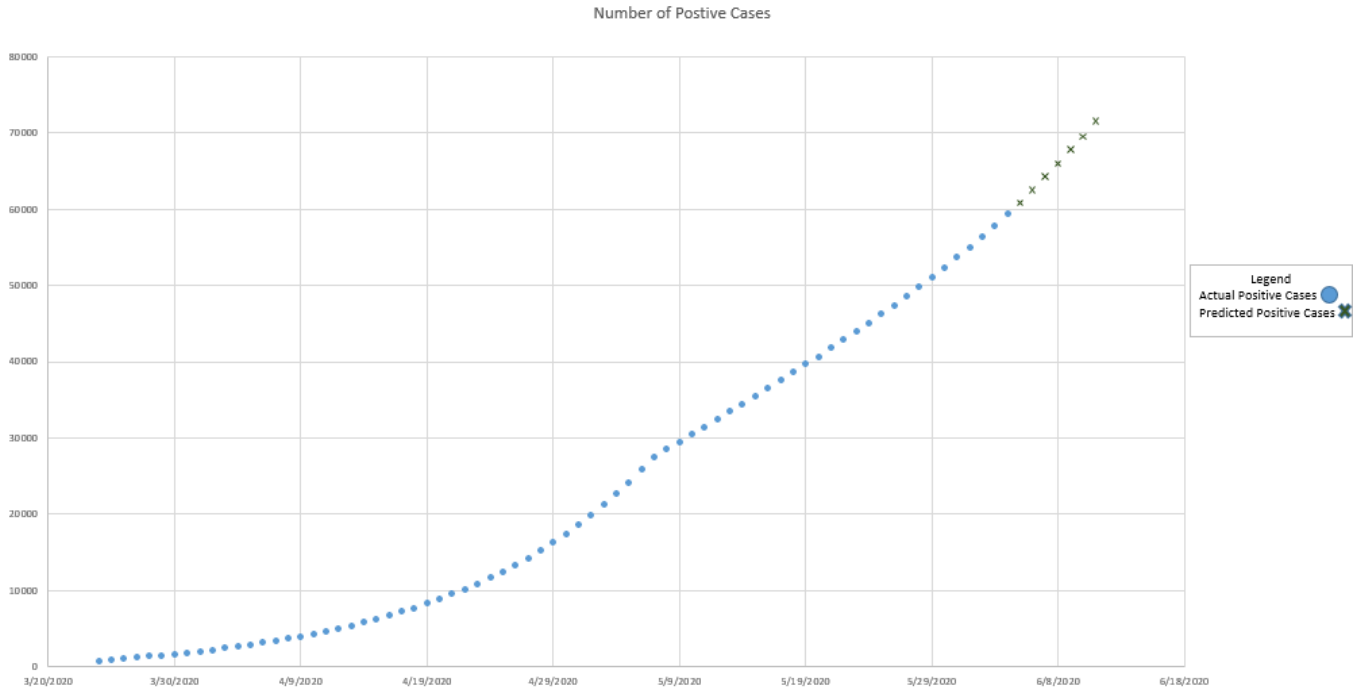


Figure 2. Predictions on death/positive cases

As shown in these charts, with our model we are able to make a prediction on the total dead/positive cases in the next period of time.

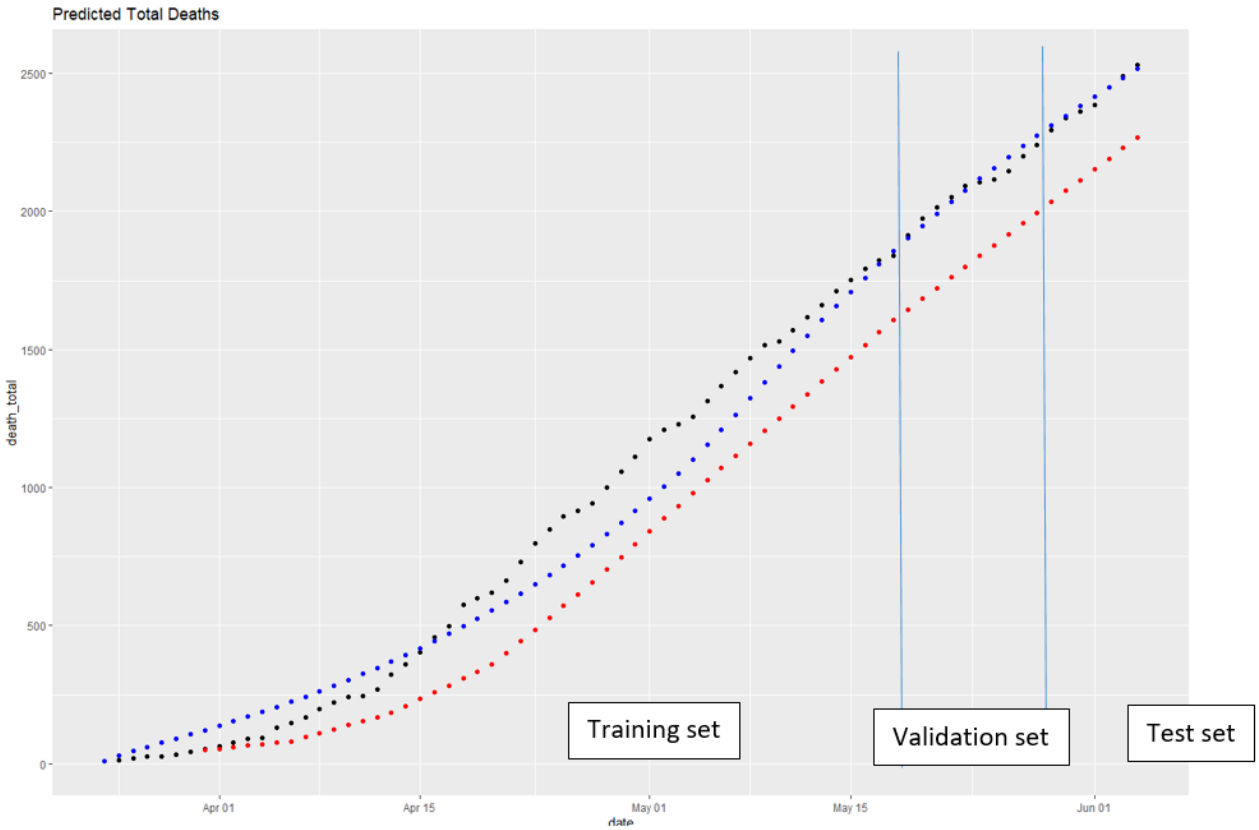


Figure 3. Model comparisons on death cases

For total deaths prediction: black is ground truth; blue is our model prediction; red is the base model prediction (using previous week's average as prediction). Our model outperforms the base mode, and achieves very good performance, especially in testing data.

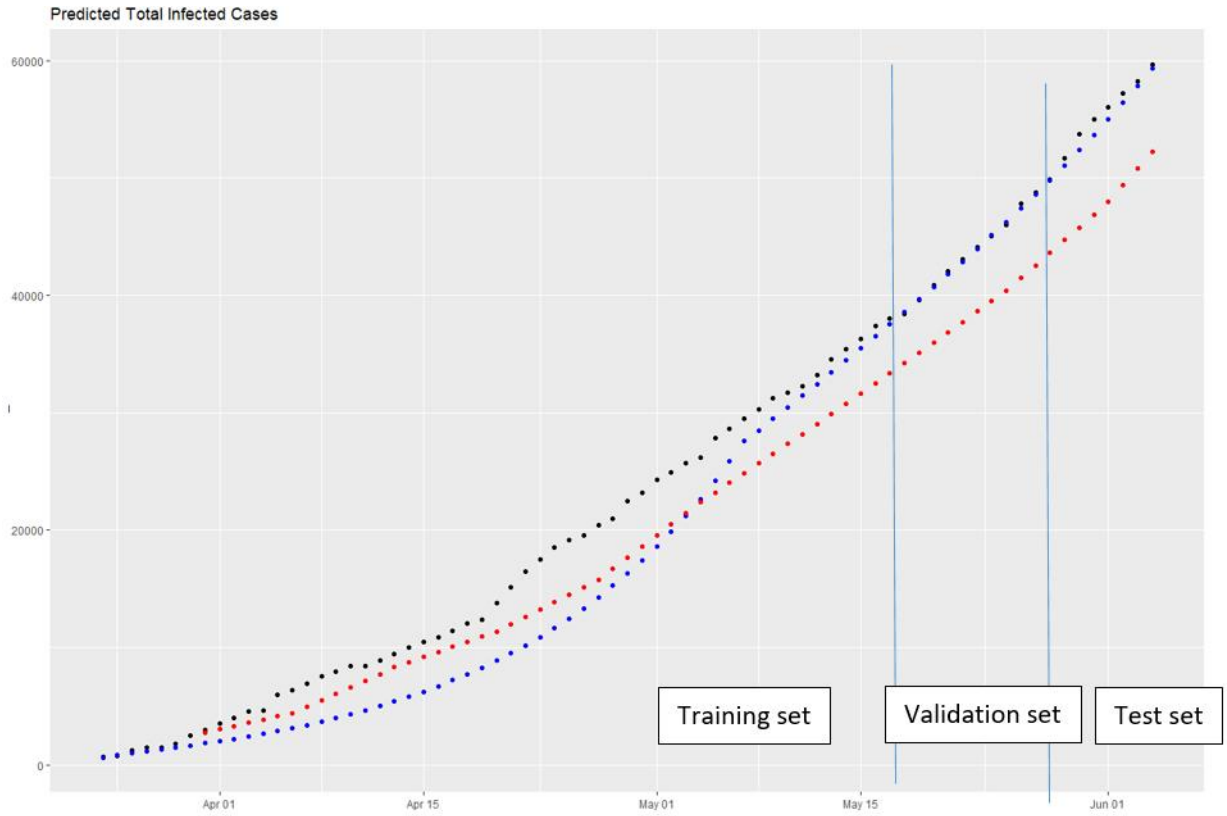
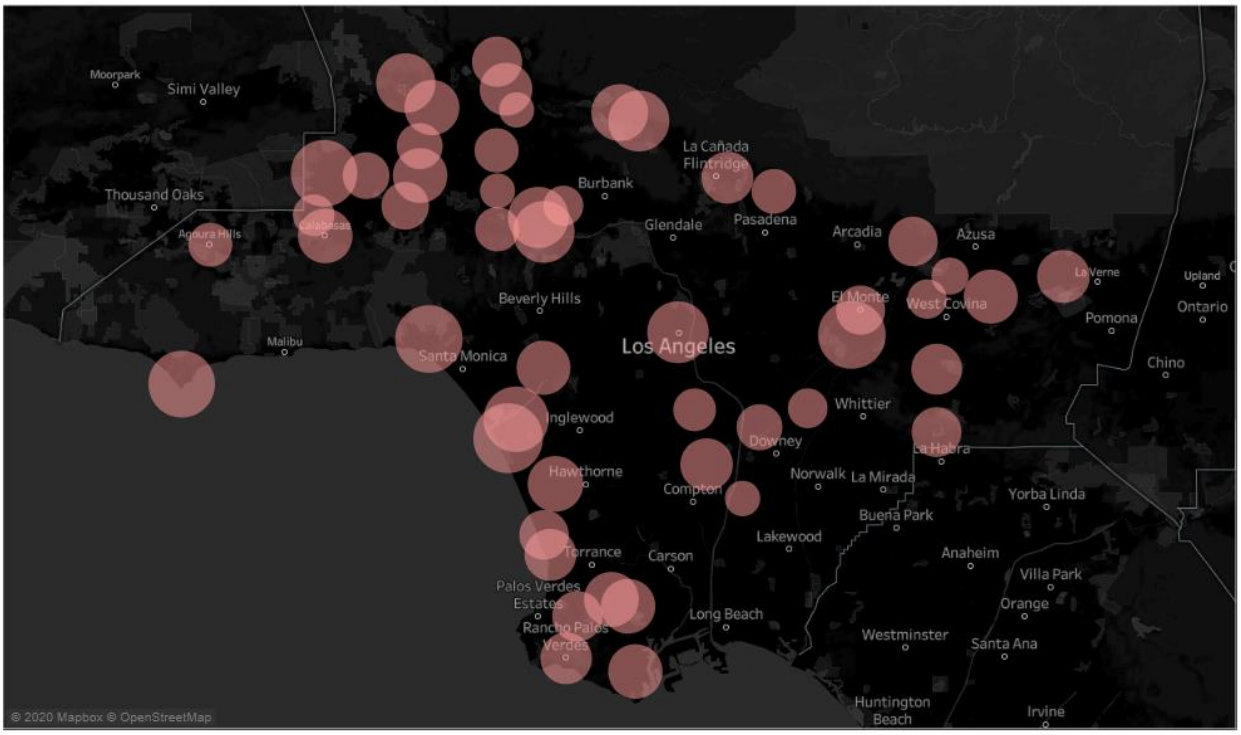


Figure 4. Model comparisons on infected/positive cases

For total infected cases, prediction: black is ground truth; blue is our model prediction; red is the base model prediction (using previous week's average as prediction). Our model outperforms the base model and achieves very good performance, especially in testing data.

Weekly Confirmed Risk



Weekly Death Risk

Figure 5. Risk scores in different regions

Based on the elderly, asthma, and cardiovascular level in each city, we are able to come up with a risk score for each city on the death risk. It turns out Los Angeles city has the highest death risk for the COVID-19. Similarly, we could calculate the infection risk score for each city based on traffic and population. Again, Los Angeles city has the highest infection risk for the COVID-19.

In Figure 5, the upper panel is the risk score of infection based on the infected cases prediction over the next seven days. The bottom panel is the risk score of serious condition based on the death cases prediction over the next seven days.

Implementation Proposal and Risk Mitigation Recommendations

In recent months, we have observed large crowds gathering for various reasons. For example, there were protests in various states to ask for businesses to open up. As Los Angeles starts opening up public places, swarms of people have entered these locations. According to ABC7 news, there was an uptick in cases when beaches and businesses started to open (https://abc7.com/coronavirus-covid-19-los-angeles-la-county/6238256/). In addition to that, the recent protests nationwide on the death of George Floyd has drawn large crowds. All of these events have the ability to increase the transmission rate of the virus. If the proper measures are not taken in order to control the virus, Los Angeles, and even the United States will face a second wave. This is alarming especially since we have no vaccine for the virus.

Our risk score system through its hazard component could properly incorporate the dynamics in the environment into the scoring, helping the government to grasp the development of the pandemic.

From our findings, we believe that in re-opening the county, strict social-distancing measures have to be taken. As mentioned previously, Los Angeles city has the highest death risk based on factors that can affect the severity of the disease in a patient (ie. old age, asthma, and cardiovascular disease), In addition to that, LA city has the highest infection risk. Hence, while steps are being taken to re-open the county in a safe manner, the county has to be strict on social distancing. Social distancing influences the reproduction number. With the parameters in our SEIR model, we can calculate the number and understand its development. And thus, inform the government on when to strengthen or release the social distancing requirements.

Test-trace-isolate is the golden rule for controlling infectious disease, our SEIR model could predict the increase of infected cases and thus inform the government on how many test kits it needs to prepare.

“Flatten the curve so that the medical system won’t be overwhelmed” is another golden rule for saving lives. Our model could predict the increase of death cases; the prediction can be used as a leading indicator for the medical system to prepare for the potential spike in cases in the reopening process.

Lastly, our location-based score system distinguishes locations with their underlying risk of infection and serious-condition; therefore, medical resources can be distributed accordingly.

Acknowledgment

We appreciate RMDS, the City of Los Angeles, and the County of Los Angeles for organizing this event and assembling all the resources for us.

We also appreciate the sponsor and partner organizations of this event: Safegraph, Snowflake, Esri, UCLA Computational Medicine, Gartner, and so forth. We learned so much from the training seminars.

We are grateful to many mentors who volunteer their time to help us in this challenge. Their insights helped us a lot.

Lastly, we want to make our own contribution to the battle with the Covid19 through this learning and research experience. With this regard, we thank ourselves.

Notes

[1] The CDC's formula is, risk = hazard * (vulnerability - resources). Due to the lack of time in our study, we do not put resources measures in our analysis yet. This is also one of our priorities in improving the work.

[2] The vulnerability features compiled by CDC: https://svi.cdc.gov/data-and-tools-download.html

[3] Available here: https://lahub.maps.arcgis.com/home/item.html?id=8659eeee6bf94eabb93398773aa25416&view=list#overview

[4] We use "only" purely from the perspective of data availability; we wish we do not need to have the data in the first place.

[5] More about the SEIR model: https://www.idmod.org/docs/hiv/model-seir.html#:~:text=The%20SEIR%20model%20assumes%20people,return%20to%20a%20susceptible%20state.

[6] https://gabgoh.github.io/COVID/index.html

Reference

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., & Lewis, B. (2011). A social vulnerability index for disaster management. *Journal of homeland security and emergency management, 8*(1).

“ Coronavirus: Officials aim for 'safe reopening' of Los Angeles County as early as July 4”, KABC News. Web. 20 May 2020