Spatiotemporal Analysis on Risk of COVID-19 in locations in and around the City of Los Angeles

by

Magdalene Mlynek, Pang-yu Liu, Haoxi Ma

University of Connecticut, 2020

Project Submitted in
2020 City of LA and RMDS COVID-19 Computational Challenge

Team name: Uconn STAT team
Mentor: Jun Yan

Summer 2020

Contents

INTRODUCTION

The purpose of this project is to outline public health and business measures useful for slowing the spread of COVID-19 in Los Angeles, and then provides better solutions of assessing risks real time with scores or ranks to be deployed.

In this paper, we try to utilize several Machine learning algorithms, such as Neural Network, Lasso to predict our critical response variable--Adjusted Death Rate, and then label risky locations as low risk, medium risk, and high risk. Using a second model, we will determine how the mobility of previous days across different businesses will affect the death rate, using the multivariate time series autoregressive method. In addition, with our precise prediction from methods mentioned above, we confidently provide suggestions for Los Angeles City and County officials to determine which businesses should remain closed in specific neighborhoods.

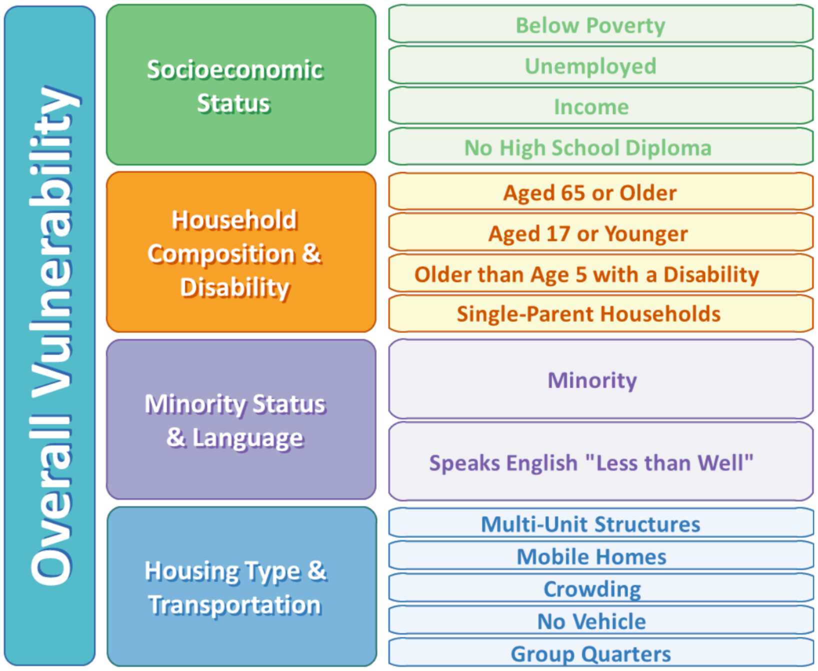
DATA DESCRIPTION

Dataset for Model1:

Variables	Description
Adjusted.case_rate	Number of cases / total population per 100,000, adjusted for age distribution of population
Adjusted.death_rate	Number of deaths / total population per 100,000, adjusted for age distribution of population
Adjusted.testing_rate	Number of people tested / total population, per 100,000, adjusted for age distribution
mean.Socioeconomic*	Mean of Socioeconomic Status within a neighborhood
mean.HouseholdComposition.and.Disability*	Mean of Household Composition & Disability within a neighborhood
mean.Minority.Status.and.Language*	Mean of Minority Status & Language within a neighborhood
mean.Housing.Type.and.Transportation*	Mean of Housing Type & Transportation
Diabetes_Percent	The proportion of population diagnosed diabetes within a neighborhood
Population	Total population
Latin	The proportion of Latinic population within a neighborhood
White	The proportion of White population within a neighborhood
Black	The proportion of Black population within a neighborhood
Native.American	The proportion of Native American population within a neighborhood
Asian	The proportion of Asian population within a neighborhood
Other.population	The proportion of all other population within a neighborhood
Area	Acreage of neighborhood(sqm)

∗ *These variables' definition shown below*

Table 2.1: Variables specification



∗ *Specific computation of these four variables can be found in Reference 1*

Figure 2.1: Definition of four variables mentioned above

Dataset for Model 2:

Variables	**Description**
day	Day number since March 10, 2020
date_dt	Date (mm/dd/yy)
new_case	Number of new daily cases
new_deaths	Number of new daily deaths
new_persons_tested	Number of new daily people tested
case.fatality.rate	# deaths / total # of cases at end of each day
retail_and_recreation_percent_change_from_baseline	Percent change from baseline* in mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.
grocery_and_pharmacy_percent_ch	Percent change from baseline* in mobility

ange_from_baseline	trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies
parks_percent_change_from_baseline	Percent change from baseline* in mobility trends for places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens
transit_stations_percent_change_from_baseline	Percent change from baseline* in mobility trends for places like public transport hubs such as subway, bus, and train stations
workplaces_percent_change_from_baseline	Percent change from baseline* in mobility trends for places of work
residential_percent_change_from_baseline	Percent change from baseline* in mobility trends for places of residence

*Baseline defined as the median value of the corresponding day of the week during the period January 3, 2020 - February 6, 2020.

Table 2.2: *Variables specification*

METHODOLOGY

3.1 Overview

In this project, we present two intuitive models that build machine learning techniques on top of a classic infectious disease model to make COVID-19 infections and deaths prediction then score risk in Los angeles(LA). Firstly, the major scope is the comparison of several machine learning algorithms -- Linear Regression, Ridge Regression, LASSO, K-Nearest Neighbors, and Neutral Network-- that will predict the death rates in the county of LA by each neighborhood. In this section, we describe the dataset we collected from neighborhoods of LA, the pre-processing steps, and the features that contain training data, as well as the machine learning models, and their evaluation metrics.

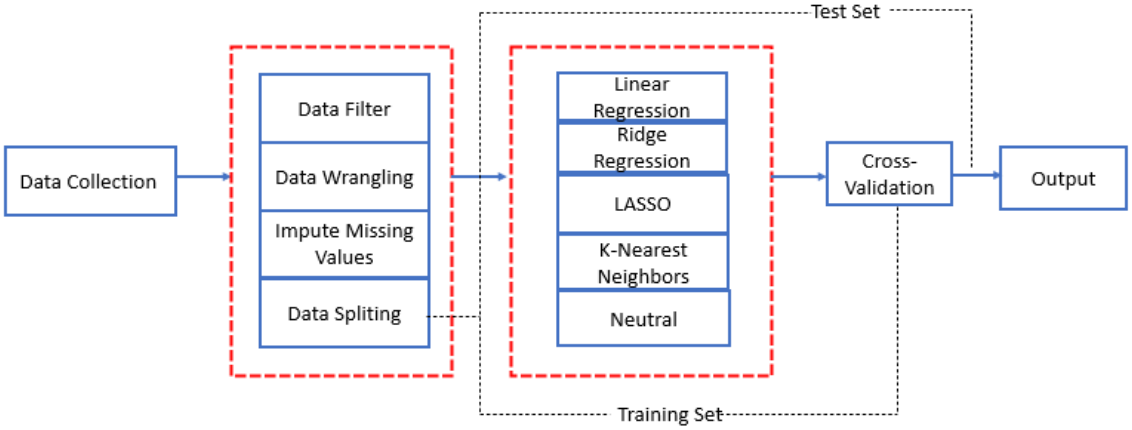


Figure 3.1.1: Methodology

In addition, in order to find a solution of assessing risk real time with rankings to be deployed, we firstly utilized Boxplot that is a method for graphically depicting groups of numerical data through their quartiles to define our risk range. In our Boxplot, the minimum is 0, the maximum is 169, the Median(50th Percentile) is 16, the First quartile(25th Percentile) is 8.5, as well as the Third quartile (75th Percentile) is 69. Base on our Boxplot, if the predicted adjusted death rate located from minimum to first quantile, those points were represented as low risk; if those points located from first quantile to second quantile, they were labeled as medium risk; the rest of points would be high risk if they were not seen as low risk and medium risk.

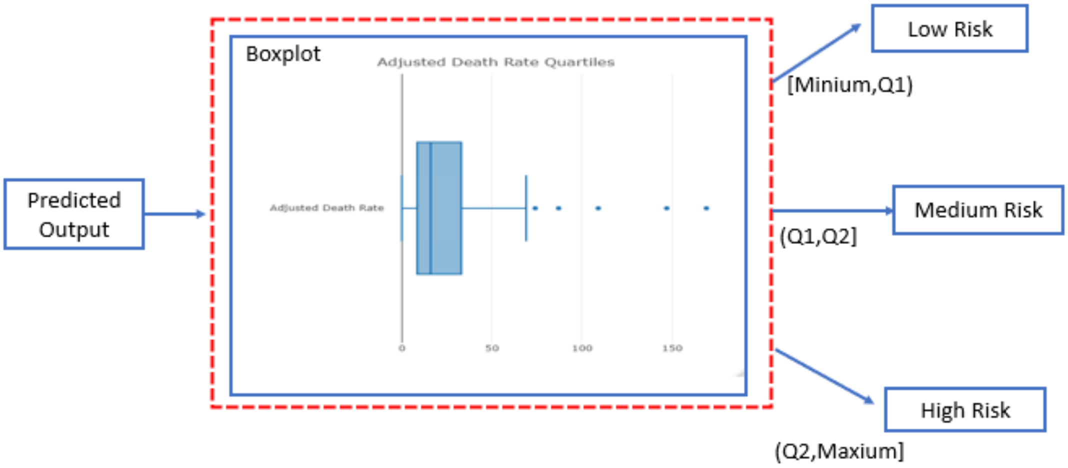


Figure 3.1.2: Methodology

To determine any further reasoning for the differences that exist between the risk levels across the neighborhoods in Los Angeles, we will take into consideration the mobility of the population over time and determine how stay-at-home orders are impacting this risk. More specifically, we constructed a Multivariate Time series model using Vector Auto-Regression using several lagged mobility variables to determine how the mobility of previous days affects and forecasts future daily deaths for day t. We did not standardize this outcome variable because the data is across the same population of Los Angeles County each day. This model will allow us to determine which categorized places should close or remain closed in high risk neighborhoods to prevent an increase in death rate. Additionally, this model will show which current restrictions or stay-at-home orders are significantly decreasing death rate. We also created ARIMA time series models for each of the mobility variables to determine how these mobility values change over time. Due to limitations on data access due to privacy restrictions on the mobility dataset, we constructed this time series model across the county level, not the neighborhood level. It should also be noted that due to data access restrictions, the outcome variable in this model is the number of new deaths for day t, and is not adjusted for the age distribution of the population and thus is slightly different from that of model 1.

3.2 Data Collection and Manipulation

The city of LA is the largest city in California. With an estimated population of nearly four million people which cover 469 square miles. It has more than one hundred neighborhoods.

We found several datasets about LA:

- “California_SVI.csv”[2]--contains all variables related to SVI(Social vulnerability refers to the resilience of communities when confronted by external stresses on human health, stresses such as natural or human-caused disasters, or disease outbreaks)
- “Census_population_race.csv”[3]--contains total population and race population by tract number.
- “diabetes.csv”[3]--shows proportion of people diagnosed diabetes by zip code
- “Name_zipcode.csv”[4]--shows all demographic changes in a neighborhood but we just use it to correspond neighborhoods to zip codes.
- “LA_CASE.csv”[5]--haves cases-related index for COVID-19. We feel interested in Adjusted Case Rate, Adjusted Death Rate and Adjusted Testing Rate.
- “google_mobility_LAcounty.csv”[6]-- shows percent change of mobility from baseline within Los Angeles county across a variety of categories (baseline is median value of corresponding day of the week over period January 3, 2020 - February 6, 2020)
- “date_table.csv”[7]-- shows number of total cases, new cases, total deaths, new deaths, total persons tested and new persons tested across Los Angeles county from March 10 to June 2, 2020

Here is procedures:

First, we select five variables we need from “California_SVI.csv” and get the data frame called *SVI.*

Then load “Census_population_race.csv” and merge it with “LA_spatial.csv” to match each Tract.Number with Neighborhood. Finally, get the data frame called *negi.*

After that, we can use FIPS to merge *SVI* and *negi*, group by Neighborhood and calculate all variables to get the data frame *Combine1*.

We are going to deal with some missing value and wrong value in *Combine1:* for missing value, we use the median value to replace it and for wrong value(Which looks extremely abnormal,like Population = -999), we just delete them. After doing these, we have a data frame called *Summary_data*.

Then, we load “LA_CASE.csv” and do some modifications on its neighborhoods’ names and merge it with *Summary_data* to get *Summary_new* data frame.

What's more, we need to add diabetes proportion information into the data frame but the dataset "diabetes.csv" only has diabetes information by zip code. Therefore, we also load "Name_zipcode.csv" which contains the zip code corresponding to each neighborhood. Then, we get the diabetes proportion from each Neighborhood by combining these two dataset. And add this variable in the data frame *Summary_new.*

At last, we select some variables we need from *Summary_new* and replace the missing value with mean value. Finally, we get our dataset "dataset_MODEL1.csv" for model 1.

To create the dataset for model 2, we will merge "google_mobility_LAcounty.csv" and "date_table.csv" by matching across dates, and removing any dates with missing case or mobility data.

3.3 Pre-Processing and Training

The collected data was then processed by using the dplyr package of the R Projects for data manipulation in RStudio. In order to access our model's performance lantern we divide the data set into two parts" a training set and a test set with setting seed(123). These models were trained with 75 percent of the dataset, which contained 71 observations. If any missing values of these measures existed in our training dataset, they were imputed by median of its column. Next, each of the Machine Learning techniques were trained, and acquired some performance metrics, such as Root Mean Square Error (RMSE) as well as R Square for us to find the best models.

3.4 Multiple Linear Regression

Linear regression is a parametric model and a supervised learning algorithm which uses a linear approach for a prediction problem. It tries to fit a line that explains the relationship between the independent variable and the dependent variable(s).

In our case, the independent variable is Adjusted.death_rate, and the dependent variables are Adjusted.case_rate, Adjusted.testing_rate, mean.Socioeconomic.x, mean.HouseholdComposition.and.Disability, mean.Minority.Status.and.Language, mean.Housing.Type.and.Transportation, Diabetes_Percent, Population, Latin, White, Black, Native.American, Asian, Other.population, Area. Although the model is simple, in some cases, it is shown to produce quite good results, as well as very fast predictions due to its simple form.

3.5 Ridge Regression

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables). Because, unlike OLS regression done with lm(), ridge regression involves tuning a hyperparameter, lamada. In our situation, we can automatically find a value for lamada that is optimal by using cv.glment(), which uses cross-validation to work out. The optimal lamada here is 13.327.

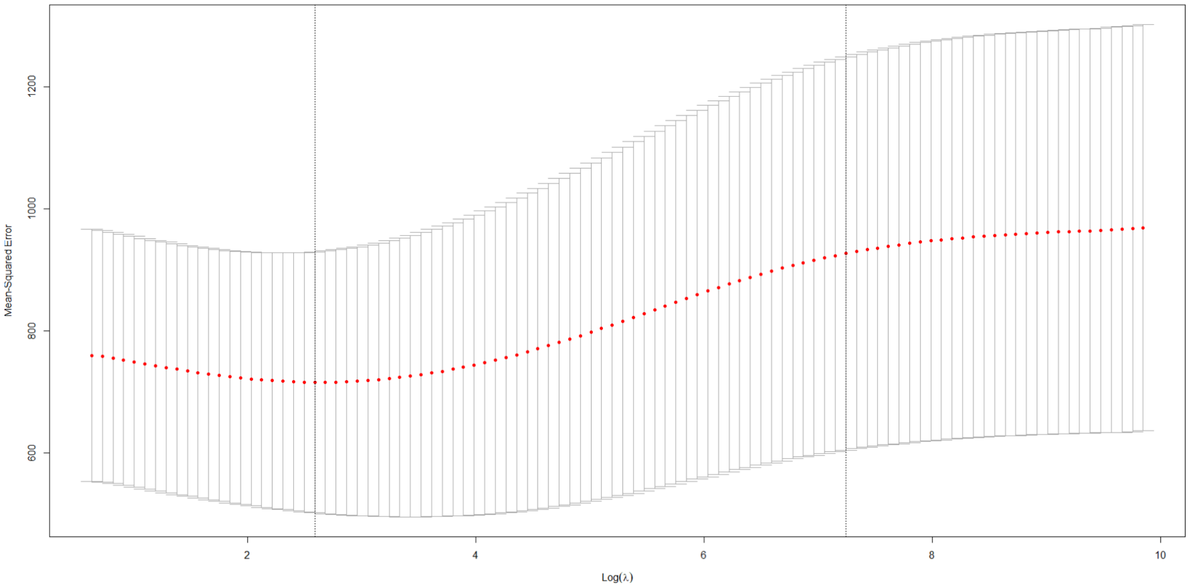


Figure 3.5.1: Plot for Log lamada vs MSE

3.6 LASSO

Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Similarly, we use the same method to get optimal lamda in our lasso model. The optimal lamada here is 2.291. In the lasso, one of the correlated predictors has a larger coefficient,such as the predictors below, while the rest are zeroed.

3.7 K-Nearest Neighbors

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

3.8 Neutral Network

Neutral Network is a popular and powerful tool to do supervised learning. To build a model between independent variables and dependent variables mentioned before, we choose to use one hidden layer and 15 hidden units Neural Network.

Besides, in order to overcome overfitting and choose a "best" model, we change the *threshold* option to specify a stopping criteria for the partial derivatives of the error function. Then calculate the R square for the training set and MSE for the test set.

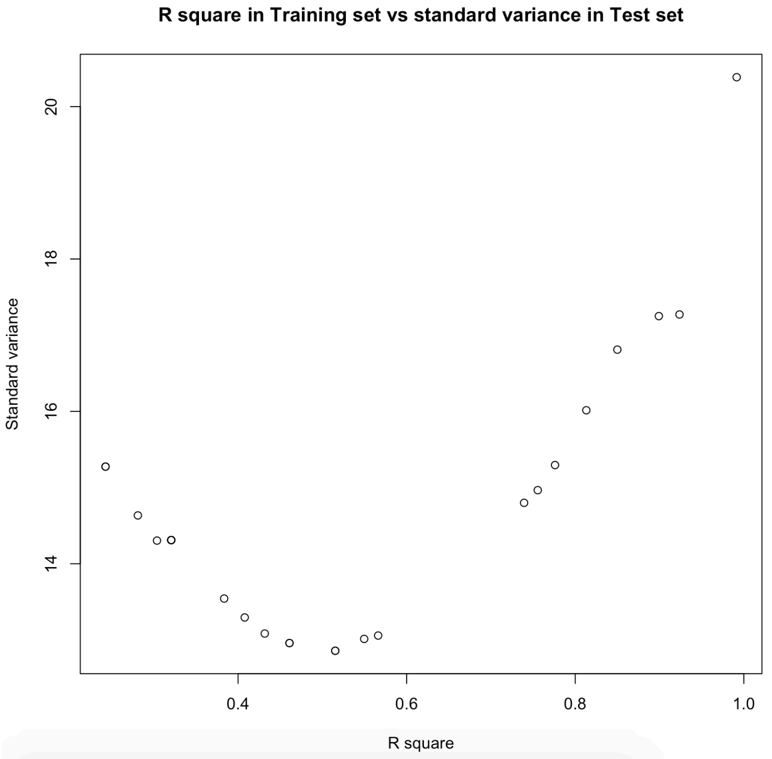


Figure 2: R square vs MSE for different stopping criteria

According to the figure above, we prefer the point where R square is 0.515 and standard variance is 12.858. This point is corresponding to *threshold=110*. Therefore, after choosing threshold=110, we can build a Neural Network model for prediction.

3.9 Time Series

To determine how time affects our outcome or risk variable, we used ARIMA (Auto-Regressive Integrated Moving Average) models to assess the possible trend variations or cyclical patterns in mobility data since March 10, 2020. The Auto-Regressive Integrated Moving Average (ARIMA) model is a linear time series forecasting equation in which the predictors are lags of the outcome variable or lags of forecast errors. In other words, to predict the value of day t, we will use the value of day t-1 or t-2 as predictors. In the ARIMA(p,d,q), p represents the number of autoregressive terms or lags, d is the number of nonseasonal differences needed for stationarity and q is the number of lagged forecast errors (Nau, 2019).

To incorporate the mobility variables in predicting the number of deaths due to COVID 19 in Los Angeles County, we will conduct a Multivariate Time Series Analysis that creates a model to forecast each of these variables according to lagged values. Due to the small number of data points, we only use the auto-regressive (VAR) model to prevent overfitting, however in future analysis where more data has been collected, we would recommend fitting a Vector Auto-Regressive Integrated Moving Average (VARIMA) model. The VARIMA model may perform better as it takes into consideration the ARIMA time series of the covariates.

3.10 Measuring Model Performance

To assess the quality of the adjusted death rate predictions, it is important to establish metrics that allow the comparison of the different methods. This performance evaluation must consist of a comparison between the adjusted death rate prediction result as well as actual adjusted death rate at the selected features, such as population. We used the following metrics:

- Root Mean Square Error (RMSE)—This metric corresponds to the square root of the mean of the squared difference between the observed and the predicted value.
- R-squared is a statistical measure that represents the goodness of fit of a regression model.

RESULT

4.1 Results from Model 1

	RMSE	R square
Multiple Linear Regression	14.517	0.410
Lasso	12.489	0.527
Ridge Regression	14.718	0.394
K-Nearest Neighbors	16.628	0.162
Neutral Network	12.858	0.515

Table 4.1.1: R square vs test MSE for different models

After comparing the results of all tools in Model1, we choose to use Neural Networks where R square is 0.515 and standard variance for prediction is 12.858. Here is an intuitive show of the Neural Networks model we got.

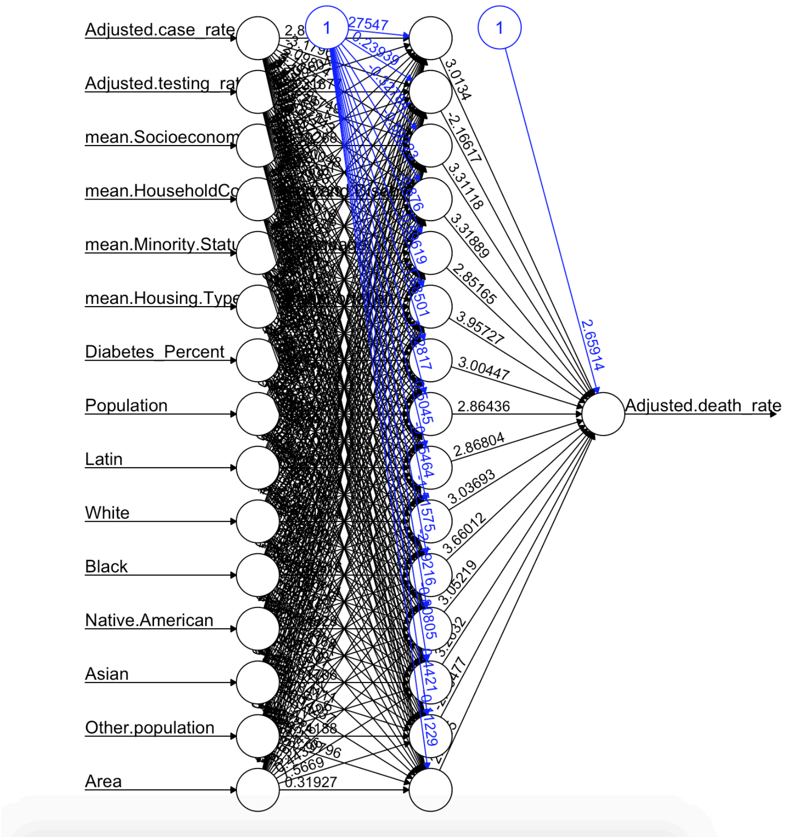


Figure 4.1.1: Neural Networks

Besides we also make a spatial data map to show the risk for each neighborhood.

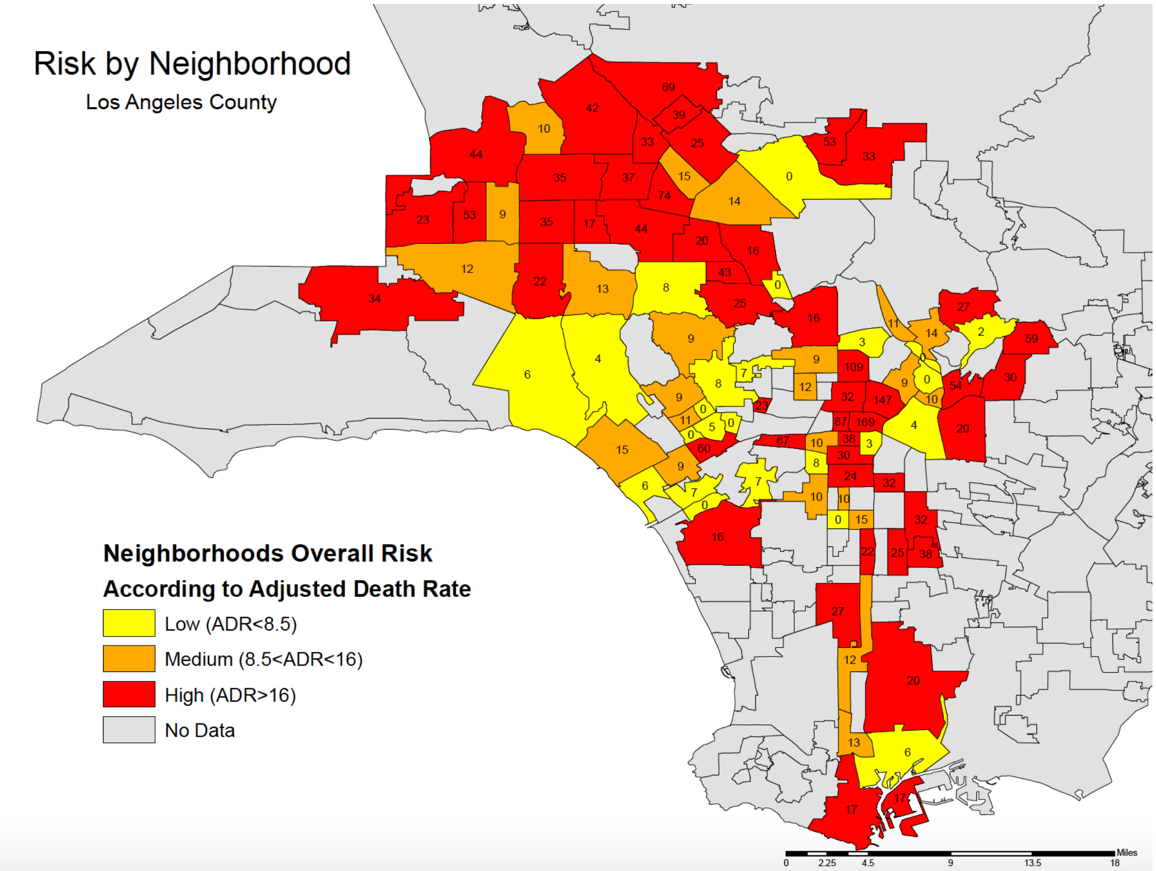


Figure 4.1.2: Risk plot for neighborhoods

4.2 Results from Model 2

From scatterplots of each of the variables of interest (i.e. mobility variables and new death variable) in Model 2, we observe a possible time series relationship and build an ARIMA model for each variable that will be included in our Vector Auto-Regression model.

a) The time series model of the outcome variable new deaths for day t, without any covariates, follows an ARIMA (2, 1, 2) model:

$$x_t = x_{t-1} + 1.07(x_{t-1} - x_{t-2}) - 0.171(x_{t-2} - x_{t-3}) – 1.807e_{t-1} + 0.887(e_{t-2})$$

where x_t is the value at time t and e_t is the error at time t. These parameter values are the same across all of the following models.

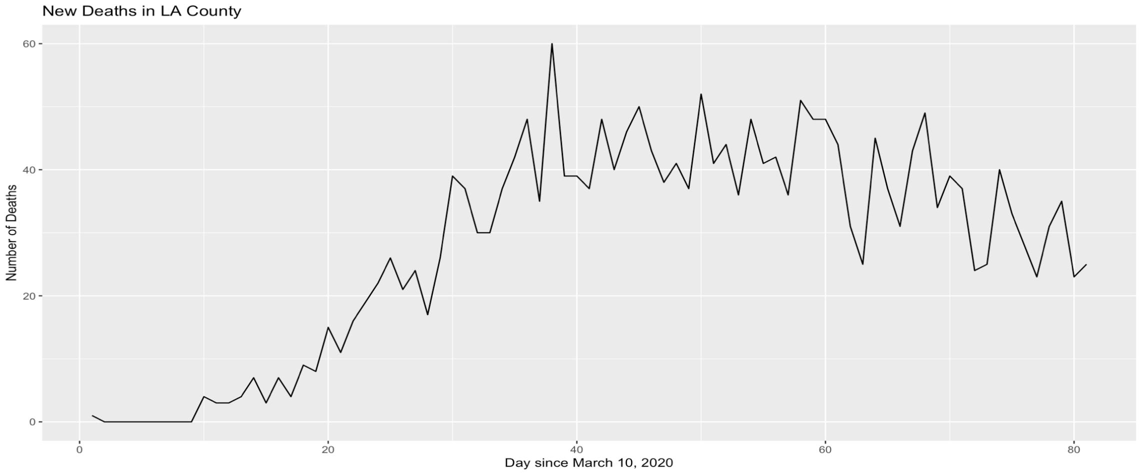


Figure 4.2.1： Time series plot of #new death

b) The change in the retail and recreation mobility data over time follows an ARIMA (0,2,1) model:

$$x_t = x_{t-1} + (x_{t-1} - x_{t-2}) - 0.954(e_{t-1})$$

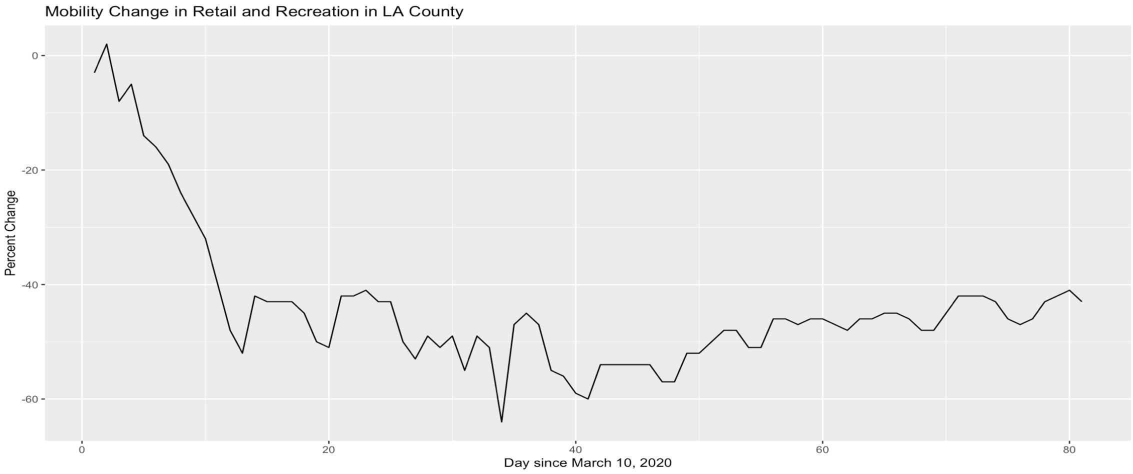


Figure 4.2.2: Time series plot of mobility change in retail and recreation

c) The change in the grocery and pharmacy mobility data over time follows an ARIMA (0,1,0) model:

$$x_t = x_{t-1}$$

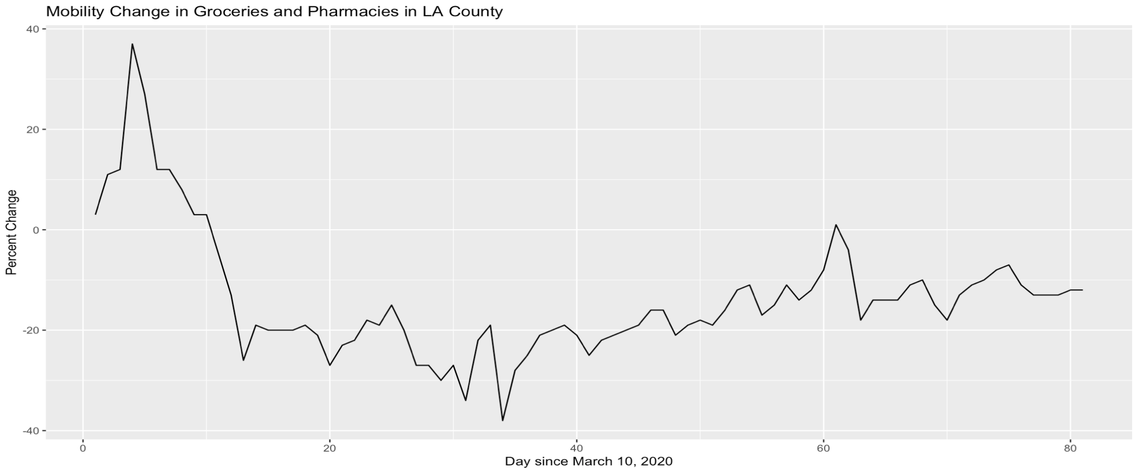


Figure 4.2.3: Time series plot of mobility change in groceries and pharmacies

d) The change in the parks mobility data over time follows an ARIMA (0,1,1) model:

$$x_t = x_{t-1} - 0.624(e_{t-1})$$

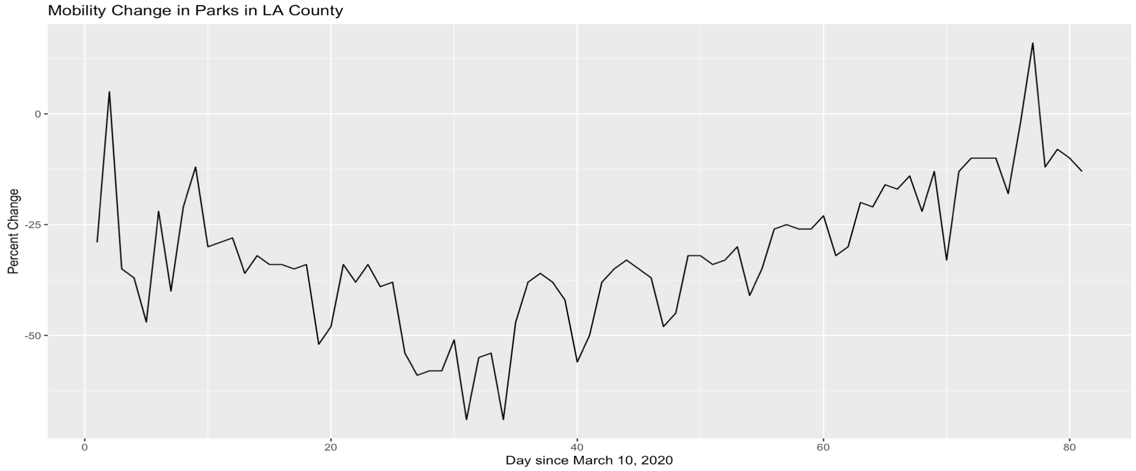


Figure 4.2.4: Time series plot of mobility change in parks

e) The change in the transit mobility data over time follows an ARIMA (2,2,1) model:

$$x_t = x_{t-1} + (x_{t-1} - x_{t-2}) - 0.434(x_{t-1} - x_{t-2}) - 0.389(x_{t-2} - x_{t-3}) - (0.8445)e_{t-1}$$

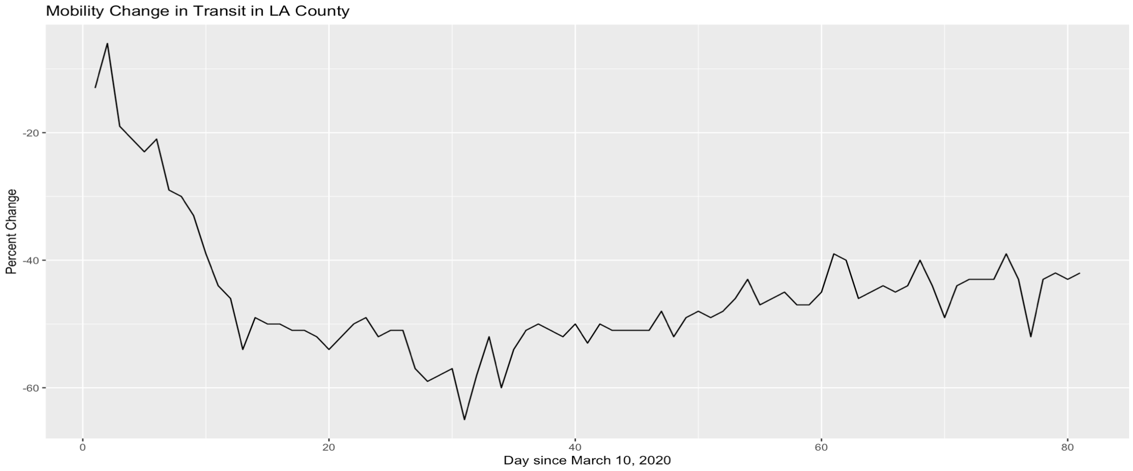


Figure 4.2.5: Time series plot of mobility change in transit

f) The change in the workplaces mobility data over time follows an ARIMA (0,1,2) model:

$$x_t = x_{t-1} - 0.233(e_{t-1}) - 0.289(e_{t-2})$$

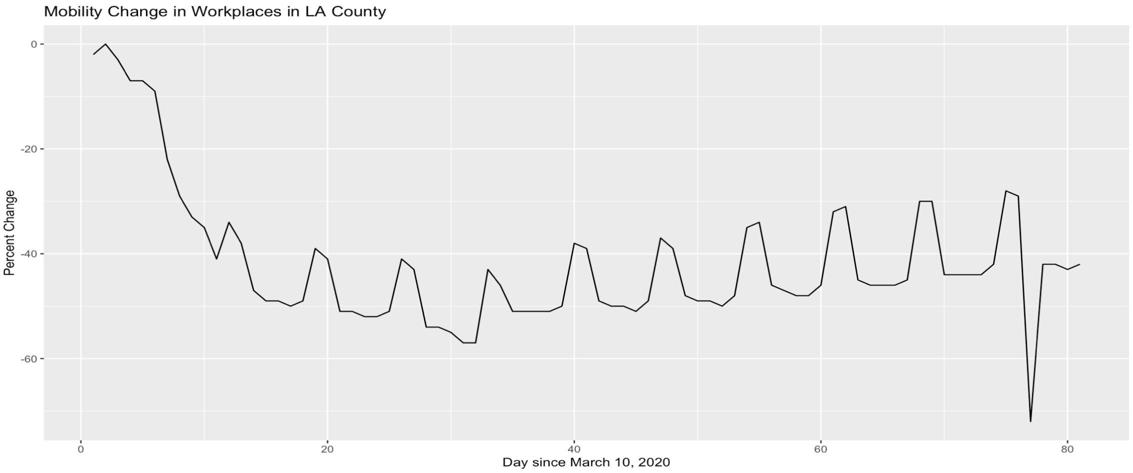


Figure 4.2.6: Time series plot of mobility change in workplaces

g) The change in the residential mobility data over time follows an ARIMA (1,0,2) model with drift:

$$x_t = 16.43 + 0.957(x_{t-1}) - 0.177(e_{t-1}) - 0.355(e_{t-2})$$

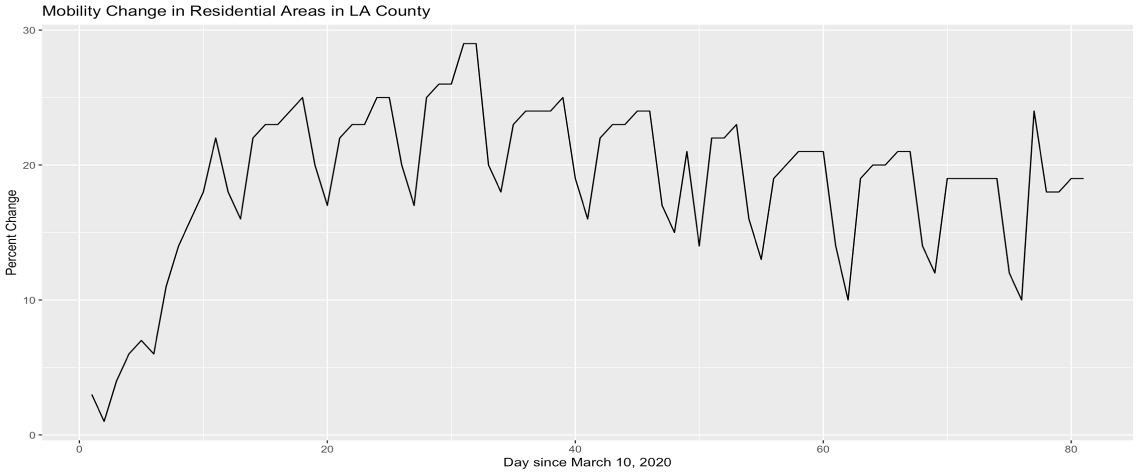
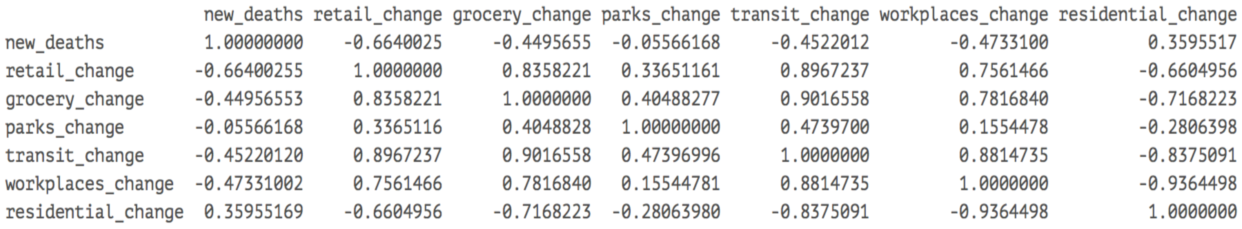


Figure 4.2.7: Time series plot of mobility change in residential area

Before finding the VAR model, we calculated the correlations between each of these variables to determine if collinearity will be an issue. From these correlations, we will remove the residential and transit variables, as they are highly correlated with the workplace variable.



```
                   new_deaths retail_change grocery_change parks_change transit_change workplaces_change residential_change
new_deaths         1.00000000    -0.6640025     -0.4495655  -0.05566168     -0.4522012        -0.4733100          0.3595517
retail_change     -0.66400255     1.0000000      0.8358221   0.33651161      0.8967237         0.7561466         -0.6604956
grocery_change    -0.44956553     0.8358221      1.0000000   0.40488277      0.9016558         0.7816840         -0.7168223
parks_change      -0.05566168     0.3365116      0.4048828   1.00000000      0.4739700         0.1554478         -0.2806398
transit_change    -0.45220120     0.8967237      0.9016558   0.47396996      1.0000000         0.8814735         -0.8375091
workplaces_change -0.47331002     0.7561466      0.7816840   0.15544781      0.8814735         1.0000000         -0.9364498
residential_change 0.35955169    -0.6604956     -0.7168223  -0.28063980     -0.8375091        -0.9364498          1.0000000
```

Figure 4.2.8: Time series summary

After splitting the time series into training (days 1 - 65) and holdout (days 66 - 81) sets, based on the AIC, HQ and FPE criterion, we will choose lag 5 for our VAR model, with no constant or trend. It is important to note that the VARIMA model may perform better than the VAR model over the long term because we previously saw that each of the variables have followed an ARIMA model. However, with only 81 data points, the VARIMA model would introduce too many covariates into the model, so we chose to remove the Integrated and Moving Average analysis from our multivariate time series model. The VAR model will forecast the values of deaths and each of the mobility variables according to lagged variables, but here we are only interested in forecasting the deaths. The model for new deaths for day t is as follows:

$$\begin{aligned}\text{Deaths}_t = {} & 0.24(\text{deaths}_{t-1}) + 0.733(\text{retail}_{t-1}) - 0.94(\text{grocery}_{t-1}) + 0.041(\text{parks}_{t-1}) + 0.104(\text{workplaces}_{t-1}) + \\ & 0.091(\text{deaths}_{t-2}) - 0.942(\text{retail}_{t-2}) - 0.833(\text{grocery}_{t-2}) + 0.168(\text{parks}_{t-2}) - 0.855(\text{workplaces}_{t-2}) + \\ & - 0.005(\text{deaths}_{t-3}) + 1.488(\text{retail}_{t-3}) - 0.832(\text{grocery}_{t-3}) - 0.256(\text{parks}_{t-3}) + 0.637(\text{workplaces}_{t-3}) + \\ & 0.473(\text{deaths}_{t-4}) - 2.013(\text{retail}_{t-4}) + 0.915(\text{grocery}_{t-4}) + 0.17(\text{parks}_{t-4}) - 0.323(\text{workplaces}_{t-4}) + \\ & 0.142(\text{deaths}_{t-5}) + 1.07(\text{retail}_{t-5}) - 0.715(\text{grocery}_{t-5}) - 0.214(\text{parks}_{t-5}) + 0.242(\text{workplaces}_{t-5})\end{aligned}$$

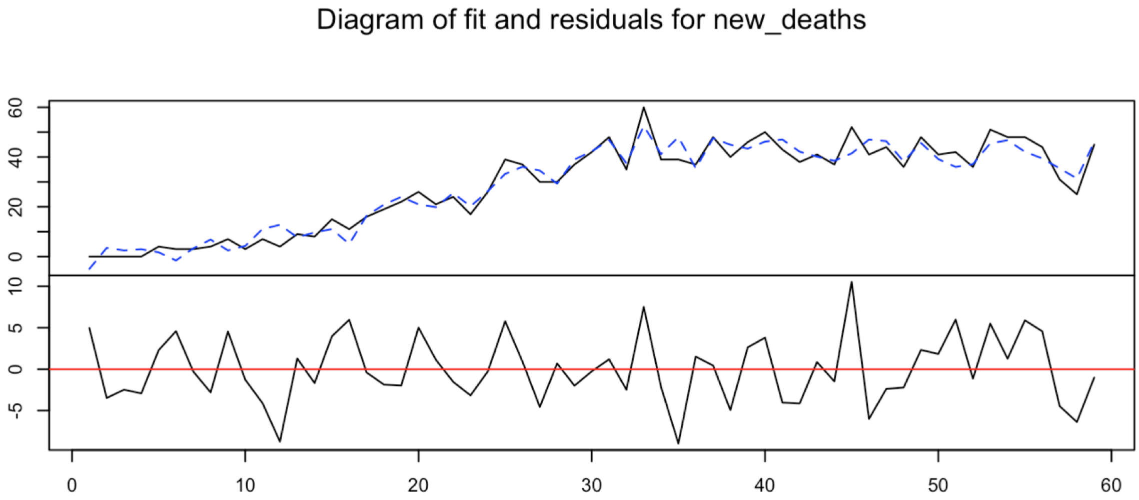


Figure 4.2.9: Diagram of fit and residuals for new deaths

This model performs very well on the dataset, with an adjusted R squared value of 0.9751 and p-value <0.0001. In addition, our model correctly predicted the holdout values 88% of the time (15 out of 17 holdout values were within the 95% CI of the values predicted by the model). The fanchart below shows the predicted new deaths for the next 30 days in Los Angeles County, according to our VAR model:

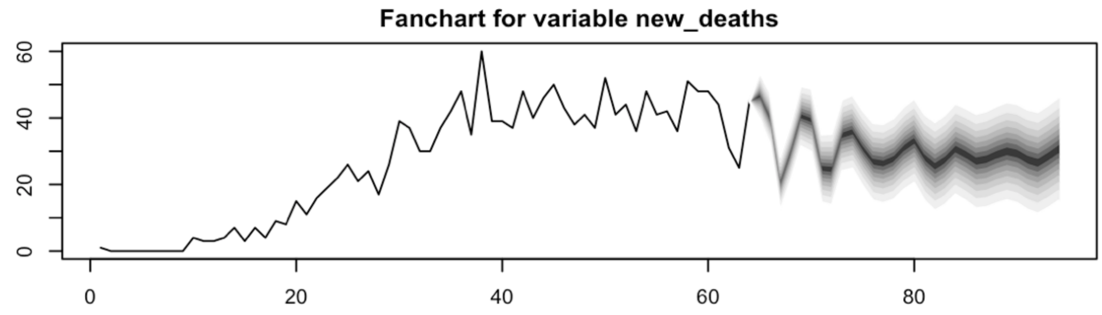


Figure 4.2.10: Fanchart for variable new deaths

More specifically, the table below shows a closer look at the true values of day t (May 29, 2020) and the predicted values of day t+1 (May 30, 2020) using the VAR model:

Variable	May 30	May 29	Change
Number of new deaths	25.46887	25	increase by 0.46887
Retail % change from baseline	-40.06994	-43	increase by 2.93006
Grocery and Pharmacy % change from baseline	-9.847839	-12	increase by 2.152161
Parks % change from baseline	-26.99329	-13	decrease by 13.99329
Workplaces % change from baseline	-38.87054	-42	increase by 3.12946

Table 4.2.1: Changes between two days

IMPLEMENTATION PROPOSAL

5.1 Advises on reuse model

To implement our model and get fairly great results, people should update the data:

1.Adjusted.case_rate, Adjusted.death_rate and Adjusted.testing_rate for MODEL 1

2.new_case, new_deaths, new_persons_tested, retail_and_recreation_percent_change_from_baseline, grocery_and_pharmacy_percent_change_from_baseline, parks_percent_change_from_baseline, transit_stations_percent_change_from_baseline, workplaces_percent_change_from_baseline, residential_percent_change_from_baseline for MODEL 2

In addition, while our VAR multivariate time series model did well in forecasting our outcome variable, we would suggest running a VARIMA model in future analysis, as the mobility variables followed an ARIMA model. As the sample size increases and we collect more data, it is likely that the VARIMA model will perform better than the VAR model.

5.2 Advises on business re-open

In an effort to advise the county of Los Angeles on which businesses should remain closed or are safe to reopen, we will use the results of models 1 and 2. Model 2 will show us which businesses or categories will result in an increase in deaths due to COVID if they are opened. Then we will apply these conclusions to determine where these businesses should close or open. More specifically, in high risk neighborhoods we will advise that these businesses be closed and in low risk neighborhoods we will advise that they open.

In the VAR model, the lagged variables that were significant in predicting the number of deaths of day t, were (1) change of mobility in groceries and pharmacies on day t-1, (2) change of mobility in workplaces on day t-2, (3) change of mobility in retail on day t-3, (4) change of mobility in workplaces on day t-3, (5) # new deaths on day t-4, (6) change of mobility in retail on day t-4, (7) mobility in groceries and pharmacies on day t-4, (8) mobility in retail on day t-5, (9) mobility in groceries and pharmacies on day t-5. Based on the signs coefficients of these significant moderators, if the mobility in grocery stores decreased by 10% on

1) day t, the expected number of deaths would decrease 9.15 on day t+4
2) day t, the expected number of deaths would in increase 9.4 on day t+1

3) day t, the expected number of deaths would increase by 7.15 on day t+5

If the mobility in workplaces decreased by 10% on

1) day t, the expected number of deaths would increase by 8.55 on day t+2
2) day t, the expected number of deaths would decrease by 6.37 on day t+3

If the mobility in retail decreased by 10% on

1) day t, the expected number of deaths would decrease by 14.88 on day t+3
2) day t, the expected number of deaths would increase by 20.13 on day t+4
3) day t, the expected number of deaths would decrease by 10.7 on day t+5

From these results, we can see that the change in the number of predicted deaths is different across different lags of the same variable, and further research should be done to determine specific reasoning behind these differences. From the overall effect that the mobility of grocery stores and workplaces would have on the number of deaths, we can conclude that decreasing mobility in these areas and closing these businesses would not create a significant impact in decreasing the number of deaths due to COVID. This finding may indicate that grocery stores and workplaces are taking appropriate precautions to minimize contact between people. However, from the overall model according to all significant lagged variables of the mobility in retail stores, if the mobility in grocery stores is 10% less than baseline for each of the previous 5 days, we can expect an overall decrease in 5.5 deaths on day t. Therefore, we can conclude that only retail stores in high risk neighborhoods should remain closed to prevent deaths.

RISK MITIGATION RECOMMENDATIONS

6.1 Features related to risks

According to the risk definition in 3.1, we label each neighborhood using the latest data we got. Aiming to analyze the relationships between risk and all variables, we group all variables by risk (risk = 1 means low risk, risk = 2 means medium risk and risk = 3 means high risk) and calculate their mean with each group.

Variables	Low risk	Medium risk	High risk
Adjusted case rate	354.720	406.409	758.333
Adjusted test rate	8107.08	6762.045	6780.922
Socioeconomic	0.374	0.487	0.642
Household Composition and Disability	0.351	0.402	0.510
Minority and Language	0.408	0.548	0.684
Housing Type and Transport	0.43	0.552	0.606
Diabetes Population proportion	7.806	9.217	9.363
Population	22801.68	32050.545	41474.745
Latin Population proportion	0.282	0.360	0.513
White Population proportion	0.463	0.349	0.259
Black Population proportion	0.101	0.139	0.099
Native American Population proportion	0.00072	0.0003	0.0006
Asian Population proportion	0.137	0.137	0.116
Area	4.036	3.893	4.280

Table 1: Summary of variables by risk group

From the table above, we have some inferences:

1. Adjusted case rate has a positive correlation with risk. The more cases a neighborhood has, the more dangerous it is.
2. Adjusted testing rate has a negative correlation with risk. This makes sense that if people infected by COVID-19 do not accept a COVID test, they won't know they are virus-carriers and will touch more people. Besides, in the table, the high risk neighborhoods have higher adjusted case rates which make things worse.
3. Socioeconomic, Household Composition and Disability, Minority and Language as well as Housing Type and Transport represent different aspects of the vulnerability of a neighborhood. They all have positive correlations with risk.
4. Diabetes Population proportion is positively correlated to risk which means people diagnosed diabetes is more dangerous in the disease outbreak.
5. Populations have a positive correlation with risk. High population may lead to high mobility responsible for high risk.
6. For all race proportion variables, we can notice Latinic and White population proportions play an important rule in infecting risk compared to the others. The impact of Covid-19 on the health of racial minority groups is still obvious. Based on our finding, Latino are facing higher risk than other races, such as White or Asian in Los angeles.
7. For area acreage variables, this may relate to population mobility in the area. Again, the higher population mobility is, the higher risk it makes.

6.2 Actionable steps for mitigating risk

In this part, we will give some actionable steps to mitigate risk corresponding to the inferences mentioned above.

1. To lower the case rate, governments need to raise awareness of the dangers of the virus to make people pay more attention to self-protection and reduce their exposure to the virus.
2. In order to higher testing rate, the government needs to increase the supply of kits to reduce the cost of testing, which will encourage more and more people to be tested for the virus.
3. For neighborhoods having high vulnerability, society should better allocate resources to help these regions face the epidemic and improve social conditions in the long run.

4. Because people who have diabetes disease are more vulnerable in this disease outbreak, the government should give them more help to tide them over
5. Population and area may relate to population mobility, so some restrictions on people's travel can effectively reduce the risk.
6. To reduce the burden of COVID-19 on the Latino population, we can implement COVID-19 education programs and conduct further analyses on these high risk areas.

ACKNOWLEDGEMENT

We would like to acknowledge professor Jun Yan. This project would not have been possible without his concern and help. We are highly indebted to him.

We would also like to extend our gratitude to our family, friends and peers. Additionally, we would also like to thank the graduate and undergraduate department faculty, other professors whose classes we have taken and learned a lot from them and staff for making our time at University of Connecticut a wonderful and an enriching experience.

Besides, we also appreciate Los Angeles GeoHub which gives us lots of useful dataset.

Reference

1. CDC. (2018, September 12). *CDC's Social Vulnerability Index (SVI)* https://svi.cdc.gov/index.html
2. Los Angeles GeoHub.(2016, December 13)*Los Angeles Index of Neighborhood Change* http://geohub.lacity.org/datasets/los-angeles-index-of-neighborhood-change/data
3. Los Angeles GeoHub.(2016, March 16). *Census Blocks 2010 Population* http://geohub.lacity.org/datasets/census-blocks-2010-population/data?geometry=-119.442%2C33.621%2C-117.382%2C34.418
4. Los Angeles GeoHub.(2018, May 3). *Prevalence of Adult Diabetes, 2013-2014* http://geohub.lacity.org/datasets/census-blocks-2010-population/data?geometry=-119.442%2C33.621%2C-117.382%2C34.418
5. County of Los Angeles Public Health. COVID-19 Surveillance Dashboard http://dashboard.publichealth.lacounty.gov/covid19_surveillance_dashboard/
6. Google. COVID-19 Community Mobility Reports https://www.google.com/covid19/mobility/
7. County of Los Angeles Public Health. COVID-19 Surveillance Dashboard http://dashboard.publichealth.lacounty.gov/covid19_surveillance_dashboard/