City of Los Angeles & RMDS

2020 COVID Computational Challenge

ULV Analytics Team

Christopher Aquino, Tyler Feese, Honggang Wang, Xu Chen

June 8, 2020

Introduction

Summary

COVID- 19 as a new viral disease that now has become a real challenge to humans, especially in the US. Globally, there have been over 2 million confirmed cases and these numbers are increasing on a daily basis. As a result of this growing threat, the purpose of our project is to find determinants of exposure risk of COVID-19 and to minimize the impact on the economy across the all cities of Los Angeles. We created a model utilizing demographic, mobility, and the Center for Disease Control's Social Vulnerability Index. Using this model, we are able to create risk scores and understand the spread of this virus. The results reveal that there is a strong correlation between risk scores and socioeconomic factors.Other model variables demonstrate contribution towards increased case counts. With this understanding, we are able to make recommendations for preventative measures in containing the spread of this virus.

Data Handling Procedures

Through the data aggregation process, it was found that data was categorized by different granularities from different sources. COVID cases were provided under the neighborhood level, so this was the end aggregation goal. On the other hand, variables such as the Center for Disease Control's (CDC) Social Vulnerability Index (SVI) was provided by the risk by Census Tract, which is a much finer level of data. Additionally, Safegraph data was provided with the zip-codes of certain points of interest. To correct these differences, the tabulation of Census Tracts to neighborhoods was used from the City of Los Angeles' geohub and using the zip-code

to Census Tract tabulation from the government census website. However, there was an issue in the cross tabulation in which zip-codes could overlap across neighborhoods. Under the constraint of time, the decision was made to assign one census tract, and therefore one neighborhood, to each zip code for foot traffic data. With the freedom of more time, we would have liked to take a proportion of each zip code that lies in each neighborhood and more accurately distribute the traffic data to finer tune our model.

The CDC SVI inputted values of -999 in the case of missing values for their data. To avoid these outliers affecting the model's calculations drastically in an adverse way, the extreme values were replaced with the mean of the respective RPL ranking. In a similar way, certain neighborhoods had demographics inputted as 0. These values were adjusted to the mean of their respective column as well.

In the case of aggregating CDC SVI to the neighborhood level, the average for the census tracts in each neighborhood was taken to provide a total average for the neighborhood. In the case of foot traffic (visits/visitors), values were summed up to the neighborhood level.

Methodology

To create the model, tentative risk scores were given 10% of the neighborhoods. A score of 10 were given to highest risk neighborhoods and a score of 0 were given to the lowest risk neighborhoods. To determine which neighborhoods would be given a 10 or 0, the following equation based on case rates was used :

$$Mock\ Score\ =\ r1 * (r1 - r0)$$

Where r1 are the case rates 3 weeks after r0

This mock score has equal weights based on two factors: case rate and the difference between 3-week apart case rates.

The two factors are multiplied because whether a neighborhood showed low overall case rate or low increase in case rates should be classified as a “Low” risk

Based on the mock scores, the top 5% were given the 10 and bottom 0% were given the 0.

Utilizing JMP, a standard least regression was then applied on the neighborhoods with the tentative score with the tentative score as the target and the variables outlined in the introduction as the inputs.

The model was then applied to every neighborhood and the risk scores were calculated.

Based on the risk score distribution, the neighborhood is classified as “LOW, “LOW/MEDIUM”, “MEDIUM/HIGH”, or “HIGH” risk. The bottom 25% are “LOW”, 25%-50% are “LOW/MEDIUM” and so on.

Result

The model has a good fit with an RSquare = 0.93. The model summary it shown below

Standard Least Squares

Model Summary

Response	Risk Score Tentative
Distribution	Normal
Estimation Method	Standard Least Squares
Validation Method	None
Mean Model Link	Identity
Scale Model Link	Identity

Measure	
Number of rows	261
Sum of Frequencies	26
-LogLikelihood	43.564914
Number of Parameters	14
BIC	132.74318
AICc	153.31165
RSquare	0.93155
RSquare Adj	0.8683655
RASE	1.2925733

With p-values < 0.05, there is significant evidence that shows that the Rpl1 (Socioeconomic theme) has an effect on the risk score

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	15.140504	20.376804	0.5520883	0.4575	-24.7973	55.078306
Female perc	9.0358079	17.740877	0.2594083	0.6105	-25.73567	43.807288
Log(Rpl1)	-2.925769	1.4282547	4.1963201	0.0405*	-5.725096	-0.126441
Log(pop dens)	-2.133265	1.2997295	2.6939142	0.1007	-4.680688	0.4141577
Males perc	-13.11767	20.550126	0.4074598	0.5233	-53.39518	27.159833
Log[r1+1]	0.1626199	0.3663557	0.197034	0.6571	-0.555424	0.8806638
Rpl2	2.6520396	2.3092036	1.318972	0.2508	-1.873916	7.1779954
Rpl3	2.6007776	2.6107424	0.9923809	0.3192	-2.516183	7.7177386
Rpl4	-0.713118	3.6856575	0.0374363	0.8466	-7.936874	6.5106379
Rpl Overall	0.3946694	1.474502	0.0716434	0.7890	-2.495301	3.2846401
Ln(Visitors)	5.7447377	9.1678406	0.3926506	0.5309	-12.2239	23.713375
Ln(Visits)	-5.886342	9.5395186	0.3807484	0.5372	-24.58346	12.810771
Log[r1-r0]	0.498415	0.2980812	2.7958457	0.0945	-0.085813	1.0826434

Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	1.8279747	0.506989	13	0.0003*	0.8342946	2.8216549

The parameter estimates from the model were used to calculate a risk score for all neighborhoods and the output is the “Risk_Score+Category” csv file.

The model can be improved with more variables such as social distancing metrics or better aggregated variables as with the case of summing visitors for neighborhoods.

Implementation Proposal

According to our findings, we suggest implementing a two month proposal to minimize COVID-19 exposure risks. With the neighborhoods divided into 4 categories, LA county should lessen restrictions and open businesses in the “Low” risk neighborhoods first. After a two-week period, the new risk scores need to calculate and be assessed whether the next quarter of neighborhoods

should be allowed to lessen restrictions. If there is an insignificant amount of increase in case rates then the process is repeated for “Low/Medium” risk neighborhoods and so on.

However, if the case rates significantly increase, the process should be postponed for at least a week.

If a faster time frame for re-opening is wanted, then a 1-month plan can be used lessening restrictions and opening businesses for “Low” and “Low/Medium” risk neighborhoods first.

Risk Mitigation Recommendations

The risk of exposure to COVID-19 can be alleviated in a number of ways. Based on our findings that people who live in the densely populated areas of LA are more likely to be affected. Hence, it is advisable to keep social distance to make less contact. Additionally, we suggest that the City of Los Angeles should be aware of the different risk areas and appropriately address each of those areas. For those areas at Low/Medium risk, the city should slowly lighten the restrictions, allowing businesses in these areas to re-open. After a two week period, re-assessment of cases should be conducted to decide whether the businesses in the areas should remain open. All areas across LA should be assessed with constant due diligence over the period of the time regardless of the risk levels. Some actionable plans also can be taken into consideration. For instance, conducting COVID-19 tests more often in addition to carrying out the regulation of facial coverings.

Acknowledgement

We thank RMDS, CDC, the City of Los Angeles, Safe Graph, UCLA Computational Medicine, ESRI and Gartner for their training and hosting this competition. We also thank Dr. Honggang Wong for his encouragement and guidance as we worked through this project.

References

Safegraph Weekly Pattern data provided as part of the Safegraph Data Consortium. Data definitions from Safegraph can be found at https://docs.safegraph.com/docs/weekly-patterns.

The CDC's SVI 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. The Geospatial Research, Analysis, and Services Program (GRASP) created and maintains the CDC's Social Vulnerability Index. More info and datasheet available at https://svi.cdc.gov/.

Demographic data taken from Los Angeles City geohub at http://geohub.lacity.org/datasets/demographics-of-neighborhood-councils?geometry=-119.703%2C33.622%2C-117.118%2C34.419.

Case rates data taken from the GRMDS github at https://github.com/GRMDS/2020Covid19_challenge

Government Census Tract to Zip code taken from

https://www2.census.gov/geo/docs/maps-data/data/rel/zcta_tract_rel_10.txt

Appendices

Variable Definitions:

Neighborhood is the Los Angeles neighborhood of focus.

Social Vulnerability variables are taken from the Center for Disease Control (CDC) Social Vulnerability Index (SVI).The CDC SVI was created to help public health officials and emergency response planners identify and map the communities that will most likely need support, before, during and after a hazardous event. The SVI uses raw data estimates and percentages from each variable, and the variables are defined as follows (from SVI2018Documentation) :

Rpl1 (RPL_THEME1 in SVI documentation) is the Sum socioeconomic factors, including "Percentile Percentage of persons below poverty estimate," "Percentile of civilian (age 16+) unemployed estimate," "Percentile per capita income estimate," and "Percentile Percentage of persons with no high school diploma (25+) estimate." These values are then taken as a percentile ranking.

Rpl2 (RPL_THEME2 in SVI documentation) is the sum of Household composition factors and represents the vulnerability associated with different household types. The contributing factors

include “Percentile percentage of persons aged 65 and older estimate,” “Percentile percentage of persons aged 17 and younger estimate,” “Percentile percentage of civilian noninstitutionalized population with a disability estimate,” and “Percentile percentage of single parent households with children under 18 estimate.” These values are then taken as a percentile ranking.

Rpl3 (RPL_THEME3 in SVI documentation) is the Sum of minority status/language theme and represents diversity. The factors contributing to this theme are “Percentile percentage minority (all persons except white, non-hispanic) estimate” and “Percentile percentage of persons (age 5+) who speak English ‘less than well’ estimate.” These values are then taken as a percentile ranking.

Rpl4 (RPL_THEME4 in SVI documentation) is the vulnerability associated with the type of housing and mobility of an area. Contributing factors include “Percentile percentage housing in structures with 10 or more units estimate,” “Percentile percentage mobile homes estimate,” “Percentile percentages households with more people than rooms estimate,” “Percentile percentage with no vehicle available estimate,” “Percentile percentage of persons in institutionalized group quarters estimate.” These values are then taken as a percentile ranking.

Rpl_overall (RPL_THEMES in SVI documentation) is the overall percentile ranking based on the preceding RPL variables.

Raw_visitor_counts/ln(raw_visitors) is an aggregation of point-of-interest (POI) foot traffic from SafeGraph mobility data. An individual visitor is a specific person visiting the POI during the timeframe (April). Safegraph defines POI's as a specific physical location which someone may find interesting. Restaurants, retail stores, and grocery stores are examples of POI's. Safe graph organizes this data by the POI and its location. For this challenge, POI's are summed up to the zip-code and then to the neighborhood level to aggregate at a level that would fit with COVID cases numbers for modeling.

Raw_visit_counts/ln(raw_visits) is an aggregation of point-of-interest (POI) foot traffic from SafeGraph mobility data. An individual visit is one person visiting a POI during the timeframe (April), for example if one person were to make a repeat visit in the same day, it would be recorded as two raw visits and one raw visitor. Safegraph defines POI's as a specific physical location which someone may find interesting. Restaurants, retail stores, and grocery stores are examples of POI's. Safe graph organizes this data by the POI and its location. For this challenge, POI's are summed up to the zip-code and then to the neighborhood level to aggregate at a level that would fit with COVID cases numbers for modeling.

Demographic Data:

White_pop/White_pop_percent reflect the estimated population and percentage of the respective neighborhood that are of white ethnicity. This value is coming from the City of Los

Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Afri_Amer_pop/Afri_Amer_pop_percent reflect the estimated population and percentage of the respective neighborhood that are of African American ethnicity. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Asian_pop/Asian_pop_percent reflect the estimated population and percentage of the respective neighborhood that are of Asian ethnicity. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Pacific_Isl_pop/Pacific_Isl_pop_percent reflect the estimated population and percentage of the respective neighborhood that are of Pacfici Islander ethnicity. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey

(ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Other_pop/Other_pop_percent reflect the estimated population and percentage of the respective neighborhood that are of ethnicity that would not be categorized into the other groupings. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Two_or_more_pop/Two_or_more_pop_percent reflect the estimated population and percentage of the respective neighborhood that are of mixed (two or more) ethnicity. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Tot_pop reflects the estimated total population for a certain neighborhood, and not a sum of each of the individual ethnic populations. This value is coming from the City of Los Angeles'

geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Males/Males_percent columns reflect the estimated population and percentage of the respective neighborhood that are of Male gender. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Females/Females_percent columns reflect the estimated population and percentage of the respective neighborhood that are of female gender. This value is coming from the City of Los Angeles' geohub, which is derived from the 2016 5-year American Community Survey (ACS) estimates used for demographic and population data. Errors in population count may be due to the weighting process used to reconcile census block data at the neighborhood council district level. *NOTE: all totals and counts are estimates, as detailed by the ACS.*

Case Rates:

R0 column shows case rates for the neighborhoods for 5/12/2020. Case rates are calculated by the number of patients who tested positive for COVID-19 per 100,000 people