Characterizing the Spread of COVID-19 from Human Mobility Patterns and SocioDemographic Indicators

Avipsa Roy
avipsa.roy@asu.edu
Arizona State University
Tempe, Arizona, USA

Bandana Kar
karb@ornl.gov
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

ABSTRACT

Mobility is an indicator of human movement through space and time. With the increasing availability of geolocated data (from GPS, accelerometers, etc.), it is now possible to examine individual as well as group human mobility patterns. Human mobility is influenced by both intrinsic (i.e. personal motivations) and extrinsic (i.e., events like natural hazards or a pandemic like the COVID-19) factors. However, the intricate relationships between human mobility patterns and sociodemographic characteristics in the context of a pandemic are yet to be fully explored. Our goal is to overcome this gap by using human mobility data at the census block group level from mobile phones and combining those with social vulnerability indicators to examine the overall spread of COVID-19 at local spatial scales. We used 585,878 weekly visits to 37,871 points of interests (POIs) from Safegraph to quantify mobility indices and social distancing metrics in 2,820 census block groups in the city of Los Angeles (LA) - before and during lockdown as well as during the phase1 and phase 2 reopening. Finally, using supervised machine learning algorithms, we classified the census block groups in LA into High, Medium and Low categories that represented the vulnerability of these block groups based on the cumulative number of occurrences of COVID-19 cases till July 24, 2020. Our results indicate that the tree-based classifiers performed well in comparison to the Support Vector Machines and Multinomial Logit models. Gradient Boosting had the highest classification accuracy of 97.4% COVID-19 with an AUC score of 0.987. The block groups with high COVID-19 cases also had a high concentration of socially vulnerable populations, high human mobility index and a low social distancing index.

CCS CONCEPTS

• Information systems → Geographic information systems; **• Computing methodologies → Supervised learning**; **• Human-centered computing → Empirical studies in collaborative and social computing**.

ARIC'20, November 3–6, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8165-9/20/11…$15.00
https://doi.org/10.1145/3423455.3430303

KEYWORDS

Human mobility, COVID-19, Spatio-Temporal analysis, Supervised Learning, Social Vulnerability

ACM Reference Format:
Avipsa Roy and Bandana Kar. 2020. Characterizing the Spread of COVID-19 from Human Mobility Patterns and SocioDemographic Indicators. In *3rd ACM SIGSPATIAL Workshop on Advances in Resilient and Intelligent Cities (ARIC'20), November 3–6, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3423455.3430303

1 INTRODUCTION

The recent outbreak of the coronavirus disease (COVID-19) has significantly impacted millions of lives across the world. Human mobility has played a very important role in the spread of the pandemic [17, 23]. At the beginning of the outbreak, the Centers for Disease Control and Prevention (CDC) suggested the local governments across the U.S. to implement strict social distancing rules [31], which contributed to the reduction of the COVID-19 spread[19]. However, cities across the US are still experiencing a rise in cases as lockdowns have been lifted and people have resumed traveling to different locations. Specifically, the city of Los Angeles (LA) has been a global hotspot for COVID-19 for several reasons including increase in mobility and failure to follow strict social distancing guidelines owing to several underlying social and economic factors.

LA County followed a strict lockdown starting March 20,2020 which persisted until May 8,2020 [2]. On May 9th, 2020 [2], the county started a restricted phase 1 reopening with essential businesses reopening with limited capacity. Phase 2 reopening started on June 12th,2020 when more small and large businesses were reopened to mitigate the economic slowdown affecting thousands of businesses [2]. However, the failure to meet state benchmarks has put both the city and county of LA on the governor's watchlist [16, 34], which resulted in increased oversight and additional closures to combat the virus. Over the past few weeks (July 24, 2020 - August 10, 2020), the county reported 28,300 new cases that is way above the standard for disease transmission [5]. Nonetheless, the number of hospitalizations has been steady in the county with 1,896 patients with a confirmed or suspected case (as of August 10, 2020) of COVID-19. Less than 80% of ICU beds are occupied and at least 75% of ventilators are available [5] (as of August 10, 2020). The steady increase in the number of COVID-19 cases in the Los Angeles area makes it crucial to understand what factors have led to such an increase in cases such that effective actions can be taken to contain the disease spread and reduce economic slowdown.

Previous research conducted on mobility data indicates the average distance to activities and the area size of activity spaces –

are associated with neighborhood socioeconomic and spatial characteristics [27]. Other studies have highlighted the importance of accounting for varying income levels and human mobility to tackle the COVID-19 crises. Huang et al. [28] have shown that counties with higher income tend to react more aggressively in terms of reducing more mobility in response to the COVID19 pandemic. Some studies have also indicated that SARS-CoV and MERS-CoV spread rapidly due to travel that increases social contact [21], which is no different than what has been seen so far with COVID-19. Typically, crowdsourced fitness apps [41], GPS trajectories, and accelerometers [39] have been used to study human mobility. Mobility has been a key indicator to study infectious disease spread [36] including COVID-19 [32]. Hence, this study characterized the possible spatio-temporal spread of COVID-19 during the different time periods of the outbreak by integrating human mobility data and contextual information about the demographic, socioeconomic, and distribution of socially vulnerable populations.

The landscape of economic slowdown and the reduction in activities since the beginning of COVID-19 is similar to what is experienced following a large-scale extreme event [11]. Like other disasters, socio-economic characteristics as well as policies have contributed to the spread of COVID-19, which include a lack of strict social distancing measure[19], less use of masks in some neighborhoods, exposure in beaches, lack of equitable distribution of healthcare resources across neighborhoods [40] as well as differences in mobility patterns across neighborhoods. Among all these factors, mobility plays an important role in the disease spread than by other factors as seen in recent studies [17, 24, 32] as it captures the dynamics of daily human movement. Although teleworking has become the norm and has contributed to a reduction in mobility even after lockdown/shelter-in-place policies have been lifted, the service industry requires travel from home to work to meet daily demands. Hence, recent studies have indicated that epidemiological models should capture the effects of mobility pattern on COVID-19 so that mitigation strategies can be undertaken for effective reopening of the economy as well as reduction of a resurgence [42].

Social distancing is a long-established public health tool, which reduces opportunities for an infectious disease to spread as well as its overall transmission rate [15, 26]. Social distancing measures include maintaining distance among individuals in a public setting, limitations on gatherings and business operations, and shelter-in-place. The more infectious the disease, the more it is necessary to implement such social distancing measures at an early stage of the spread to control transmission at early stages [12, 30]. As was seen during COVID-19, social distancing has a detrimental impact on economic activities, which has contributed to increasing mobility.

Given that the COVID-19 virus spreads from physical contact as well as through air, mobility patterns (a proxy for virus spread) along with COVID-19 cases and sociodemographic characteristics were used to (i) examine and quantify the interactions between mobility and social distancing measures, (ii) classify census block groups in LA city into high, medium and low risk of experiencing COVID-19 spread based on mobility, vulnerable populations and current cases of COVID-19. This information can be used by stakeholders for resource planning to reduce the spread of the disease while helping socially vulnerable groups that generally experience health disparity.

The remainder of the paper is organized into five sections. Sections 2 and 3 introduce to the study and site and provide a discussion of the datasets used in the study. The algorithm and analytics implemented are discussed in section 4. A discussion of results and concluding remarks are provided in sections 5,6 and 7.

2 STUDY AREA

This study focused on the city of LA, California in the USA. The LA metropolis is the second most populous metropolitan area in the United States and is coterminous with the County of LA. According to the U.S. Census [14], as of 2020, the city is home to about 10 million people in and more than 3.5 million households.

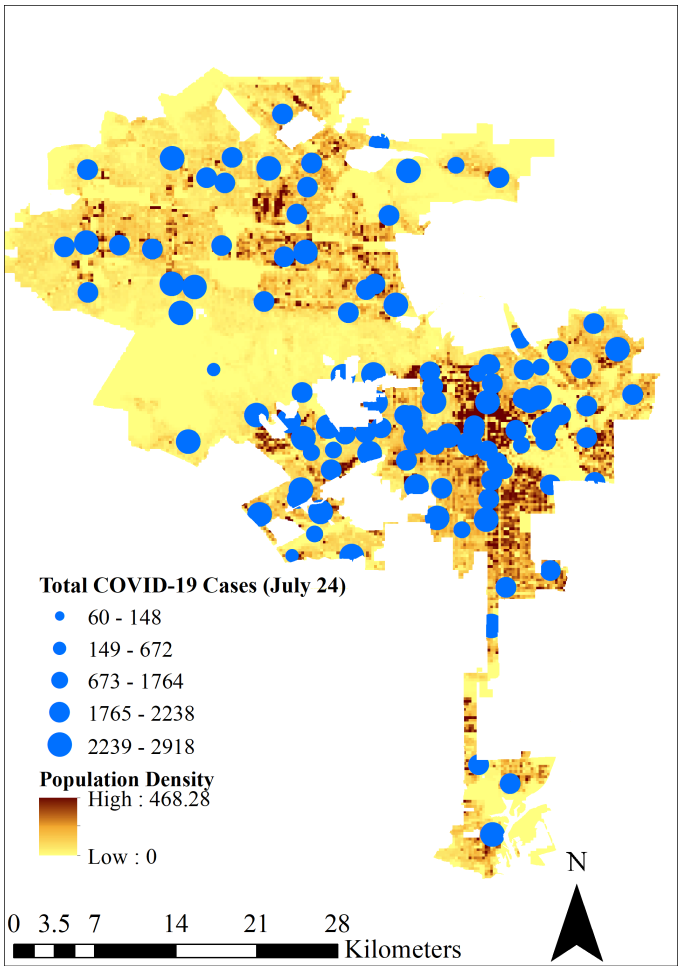


Figure 1: Map showing the population density in Los Angeles along with the total number of confirmed COVID-19 cases as of July 24, 2020.

In 2000, the city was occupied by a little more than one-third of the 2020 population (approx. 3.7 million people) and was the second-largest city in the nation. The demographic distribution within is quite significant (Figure 1) as a majority of the population resides in the southern and northwestern part of the city in and around downtown LA.

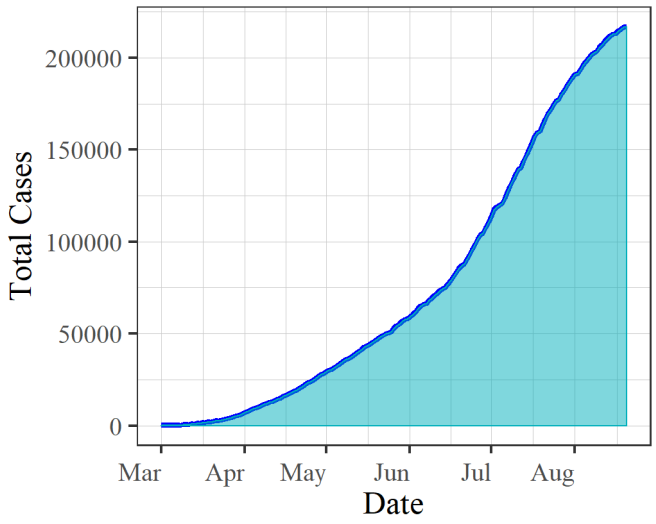


Figure 2: Cumulative number of daily cases in LA county based on recent data

The county of LA within which the city resides has become a hotspot for the COVID-19 cases in the US. Since Phase 2 reopening that started in (June 11, 2020), the county is experiencing a steep increase in the daily number of cases with a record high of 4500 cases in a single day in July as shown in Figure 2. As of July 24th, 2020, the LA Public Health department has identified 173,995 positive cases of COVID-19 and a total of 4,360 deaths across the county [3]. Figure 1 depicts the census block groups with the total number of confirmed COVID-19 cases along with the spatial distribution of population density (i.e the number of people per square kilometers at the census tract level) in LA.

3 DATA

Mobile phone data from SafeGraph [6] obtained between March 8th through July 24th, 2020 were used in the study. This mobility data were extracted from anonymized cell phone trajectories shared by SafeGraph [6]. SafeGraph is a commercial entity that compiles its dataset from several sources including mobile phone GPS data and governmental open data, to build a comprehensive listing of business establishments in the United States and Canada.

The mobility data captures weekly movements of thousands of people at the spatial resolution of census block groups (CBGs), which are geographical units that typically contain 600–3,000 people. The data consists of mobility pattern information to points of interest (POI), the number of visitors, visit counts, the median dwell time at each POI, the distance from a mobile device from its home location, and devices that were consistently found at home locations during lockdown periods. The mobile phone data for the 2820 CBGs in LA contained 585,878 unique visits to a total of 37,871 points of interest (POI). The top POIs with maximum visits included restaurants, grocery stores, religious establishments, fitness centers, and supermarkets. Figure 3 shows the average weekly visits to the different types of POIs before, during, and after the lockdown.

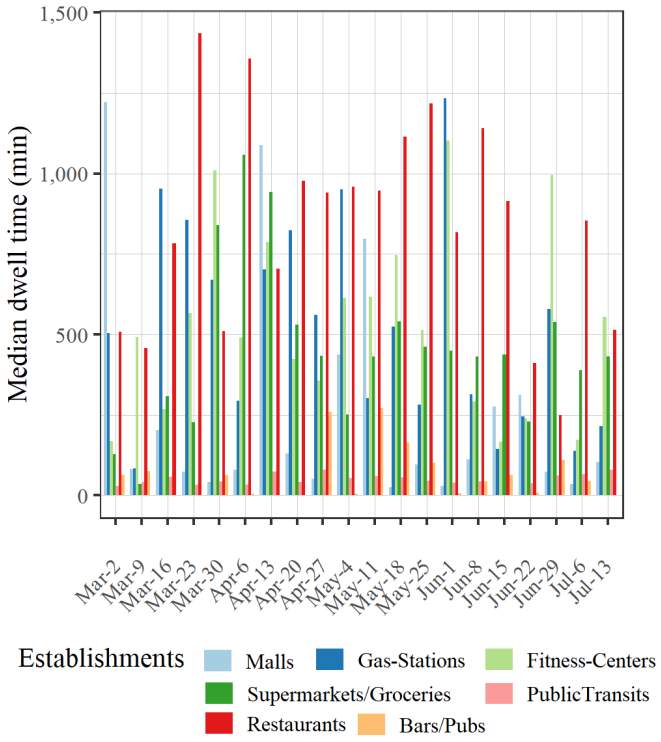


Figure 3: Distribution of median dwell times between March and July to different POIs in Los Angeles grouped by the type of establishment

From the SafeGraph COVID-19 Data, mobility data about service-oriented POIs such as retail shops, restaurants, movie theaters, and fitness centers were extracted. For each POI, besides its daily foot traffic, its geospatial location, and NAICS (North American Industry Classification System) codes [37] were obtained. Of those who died, the LA health department released the race and ethnicity information for 4,069 people. Hence, other datasets representing social vulnerability (i.e., social vulnerability index data from the CDC's social vulnerability indicators for 2018 [22]), unemployment rate from 2018 American Community Survey [14], 2020 population density from WorldPop [1] and Nitrogen Dioxide (NO2) from NOAA [10] were used to examine the relationship of COVID-19 mortality cases and spread with sociodemographic and environmental conditions as well as mobility patterns.

4 METHODS

Using disparate datasets, we extracted features from sociodemographic characteristics, mobility patterns (in this context, mobility index based on POI visit information was used) and COVID-19 cases in LA to fit supervised classification models (Random Forest, CART, Gradient Boosting along with Support Vector Machines and logistic regression) to predict high, medium, low number of COVID-19 cases based on mobility patterns by accounting for spatial-temporal changes in travel distance and stay-at-home as well as the effect of social distancing policies on human movement.

A three-step approach was implemented to classify the CBGs by the number of confirmed COVID-19 cases. First meaningful features from raw mobility, demographic and socioeconomic data were extracted. Second, tree-based classifiers (CART) [13], Support Vector Machines [18] and Multinomial Logistic Regression was fitted to the feature vector by splitting it into training and test sets. Finally, the accuracy of each classifier was examined and evaluated using confusion matrices and Area Under the Curve (AUC) [25] characteristics.The output provides information about potential hotspots within LA city at the census block group level. Both ArcGIS [20] and R 3.6.1 [38] statistical software were used to undertake the analytics. Each of these steps is explained further in the following sections.

4.1 Feature Extraction

The mobility data for LA city containing information for 37,871 POIs were obtained as a collection of flat text files. These files were parsed and filtered into four different time period bins representing - before lockdown (March 1 - March 19), during lockdown (March 20 - May 8), phase 1 reopening, and phase 2 reopening (June 12 - July 24), which were then joined to the CBGs (SafeGraph, 2020) for further processing. Table 1 lists the different features extracted along with the corresponding data sources .

Table 1: List of variables used to model COVID-19 spread

Features	Source
Mean weekly mobility index	SafeGraph[6]
Mean social distancing index	SafeGraph[6]
Density of POIs visited	SafeGraph[6]
Mean monthly NO_2 concentration	NOAA[9]
Social vulnerability index	CDC[22]
Population Density	US Census Bureau[14]
Unemployment Rate	LA County Open Data Portal[4]
Daily confirmed COVID-19 cases	LA Public Health Department[3]

Mobility Index: Weekly patterns data had three different mobility attributes - distance from home location of a mobile device(D_i), median dwell time at a POI location (T_i) and the total number of visits to the POI location during the week (N_i). Using this information, a mobility index (M_{index}) was computed by calculating a percentile rank of the sum of normalized scores for all three variables (Equations 1,2) and aggregated the values for each census block group (i). The normalized scores were computed as a ratio of the distance from home for each and the total distance for the same POI within the same census block group. The Mi_{norm} is the standardized mobility index computed as a range of values between 0 and 1 using a min-max equalizer.

$$M_i = Di_{norm} + Ti_{norm} + Ni_{norm} \tag{1}$$

$$M_{index} = 100 * \frac{M_i}{Mi_{norm}} \tag{2}$$

Social Distancing Index: The social distancing index was computed as a percentile rank (SD_{index}) of the normalized ratio (SDI_{norm}) of the number of devices completely at home (N_i) and the total device count (Tot_i) in each census block group (i).

$$SD_i = \frac{N_i}{Tot_i} \tag{3}$$

$$SD_{index} = 100 * \frac{SD_i}{SDI_{norm}} \tag{4}$$

POI density: A kernel density estimate [33] as per equation 5 was used to calculate the density of POIs visited in each census block group POI_i using the number of POIs (pop_i) within a standard distance d_i, from the mean center of each census block group i, and a search radius r. The density estimate was finally normalized using using a min-max equalizer to obtain the normalized POI density. The kernel density estimate at a location (x,y) is given by:

$$POI_i = \frac{1}{r^2} * \sum_{i=1}^{n} \left[\frac{3}{\pi} . pop_i \left(1 - \left(\frac{d_i}{r} \right)^2 \right)^2 \right]; where, d_i < r \tag{5}$$

Social Vulnerability Index: The Centers for Disease Control's SVI score is available for each census tract, which was resampled to get unique values for census block groups. The SVI indicates the relative vulnerability of every U.S. Census tract to a natural hazard, in this case the COVID-19 pandemic.

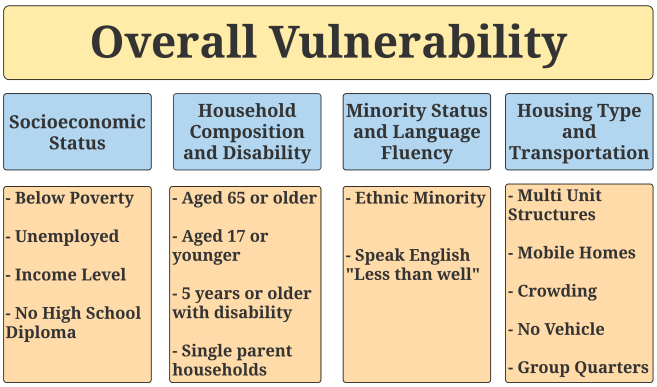


Figure 4: Variables used to compute the Social Vulnerability Index

The SVI was calculated as percentile ranks for each group of variables as per CDC SVI guidelines described by Flanagan et al. [22]. Each census variable was ranked from highest to lowest vulnerability across all census block groups in LA with a nonzero population. The variables were then grouped among four themes (Figure 4). A percentile rank was calculated for each theme for each census block group. Finally, an overall percentile rank for each census block group was calculated using the sum of all variable rankings.

Industrial Activity Recent studies have found that NO2 levels have a direct correlation with industrial activities which might lead to an increase in COVID-19 cases [7, 8], which indirectly could be used as a proxy for mobility owing to work trips or commutes. The average monthly NO2 concentration data from the TROPOspheric Monitoring Instrument (TROPOMI) instrument onboard the Sentinel-5 Precursor (Sentinel-5P) satellite developed by the European Space Agency's Copernicus Programme was obtained for LA from February through July, 2020. the TROPOMI instrument

monitors trace gases (O3, CH4, CO, NO2, SO2) as well as the index. The monthly NO2 concentration data used in this study is available at 3.5 km x 3.5 km spatial resolution.

4.2 Supervised Learning

A supervised classification approach using five different algorithms was used to categorize the amount of COVID-19 cases in each census block group. Each supervised classification was modeled using the number of COVID-19 cases as the response variable and the extracted features along with the time period (before lockdown / during lockdown/ phase1 reopening/ phase2 reopening) as independent variables. For the model implementation, the data were split into train and test sets with 70% data used for training and the remaining 30% for testing each model respectively.

Tree-based classification models, such as classification and regression tree (CART), Random Forest, and Gradient Boost have been used in previous studies [29, 43] and have been found to perform better in infectious disease spread modeling. Hence, these models were used for classifying the CBGs. In addition we also used Support Vector Machines (SVM), a non-probabilistic linear classifier and Multinomial Logit, a probabilistic logistic regression to compare their predictive accuracy to the tree-based classifiers.

CART, also known as a decision tree method, is a powerful and popular predictive machine learning technique that is used for both classification and regression. Decision tree models work by repeatedly partitioning the data into multiple sub-spaces so that the outcomes in each final sub-space is as homogeneous as possible. This approach is technically called recursive partitioning that allows the use of different possible splitting rules to effectively predict the category of COVID-19 cases (High/Low/Medium). Gradient Boosting (XGB) is a special case of a decision tree that builds trees one at a time such that each new tree helps to correct errors made by the previously trained tree. However, XGB tends to overfit. Random Forests (RF) train each tree independently using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data.

It was hypothesized that the census block groups with high number of vulnerable population and unemployment rate as well as a low social distancing index and high mobility index would have a higher probability to be classified as high or very high zones for COVID-19 cases and have the potential to emerge as hotspots.

4.3 Accuracy Assessment

Each of the five models was fitted with 70% training data using a repeated 10-fold cross validation with 3 repeats. Each model resulted in mean classification accuracy, class-wise sensitivity and specificity along with F1-score and balanced accuracy. The true negative (TN), true positive (TP), false negative (FN) and false positives (FP) were used to calculate the accuracy metrics for each class i given by Equations 6 - 10.

$$Sensitivity = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$Specificity = \frac{TN_i}{TN_i + FP_i} \tag{7}$$

$$BalancedAccuracy = \frac{Sensitivity_i + Specificity_i}{2} \tag{8}$$

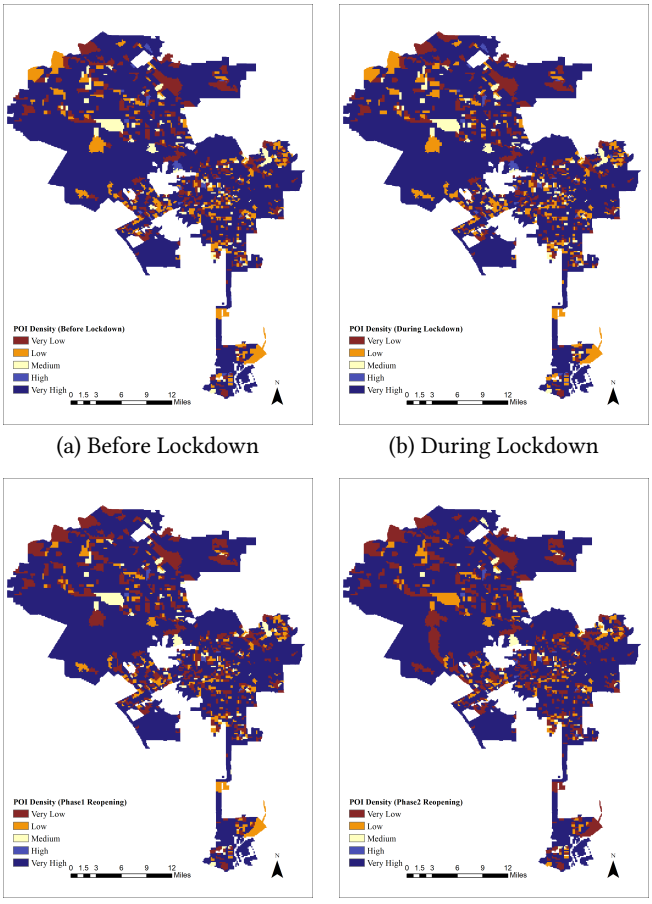
$$F1 - Score = \frac{2 * TP_i}{2 * TP_i + FP_i + FN_i} \tag{9}$$

$$OverallAccuracy = \frac{\sum_{i=1}^{k} \frac{\mathrm{tp}_i + \mathrm{tn}_i}{\mathrm{tp}_i + \mathrm{tn}_i + \mathrm{fp}_i + \mathrm{fn}_i}}{k}; \text{ where, k = no. of classes} \tag{10}$$

Cohen's Kappa statistic [35] which is used to measure the agreement of two raters (i.e., "judges", "observers") or methods rating on categorical scales were used to quantify how much both raters agree by chance. Equation 11 represents the Cohen's Kappa statistics.

$$Kappa(\kappa) = \frac{P_0 - P_e}{1 - P_e} \tag{11}$$

where P_0 is the proportion of observed agreement and P_e is the proportion of chance agreement. The values can range from 1 to +1, where 0 represents the amount of agreement that can be expected due to random chance, and 1 represents perfect agreement between the raters.



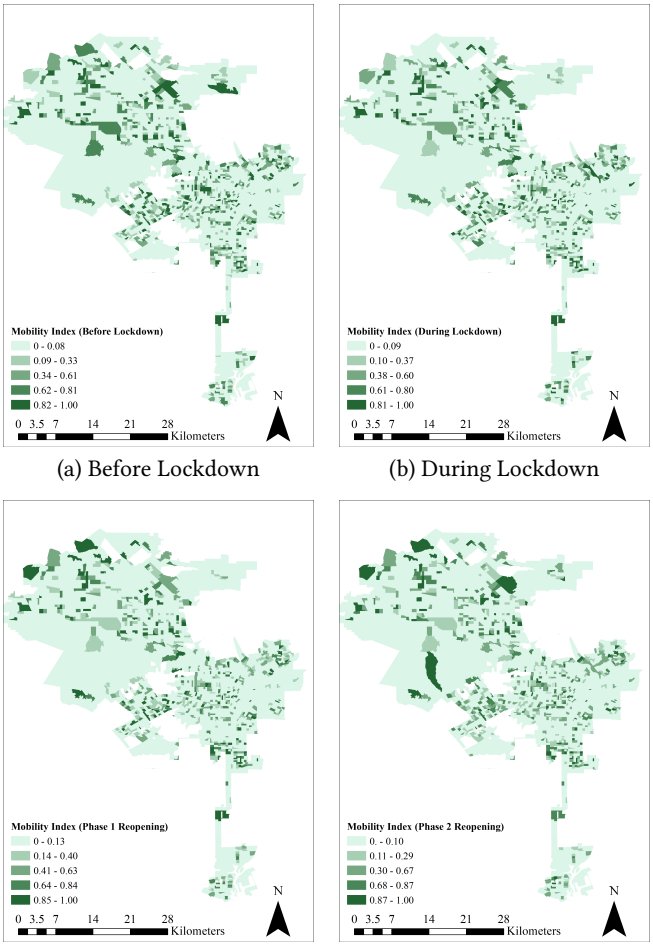
(a) Before Lockdown (b) During Lockdown

(c) Phase 1 Reopening (d) Phase 2 Reopening

Figure 5: Maps showing the kernel density of POIs visited before, during and after lockdown in Los Angeles

5 RESULTS

The features listed in Table 1 were calculated as part of the feature extraction step. The spatial distribution of the density of POIs visited at different periods of time - before, during and after lockdown regulations were imposed by the local authorities is shown in Figure 5. The dark red spots indicate areas where the POIs were more visited as opposed to blue areas that represent areas with a low number of POIs visited. POI density is also an indicator of how many business establishments were operational before, during, and after the lockdown and which communities were following social distancing guidelines more stringently. All the feature layers were converted to raster layers which were then used to fit the classification algorithms.

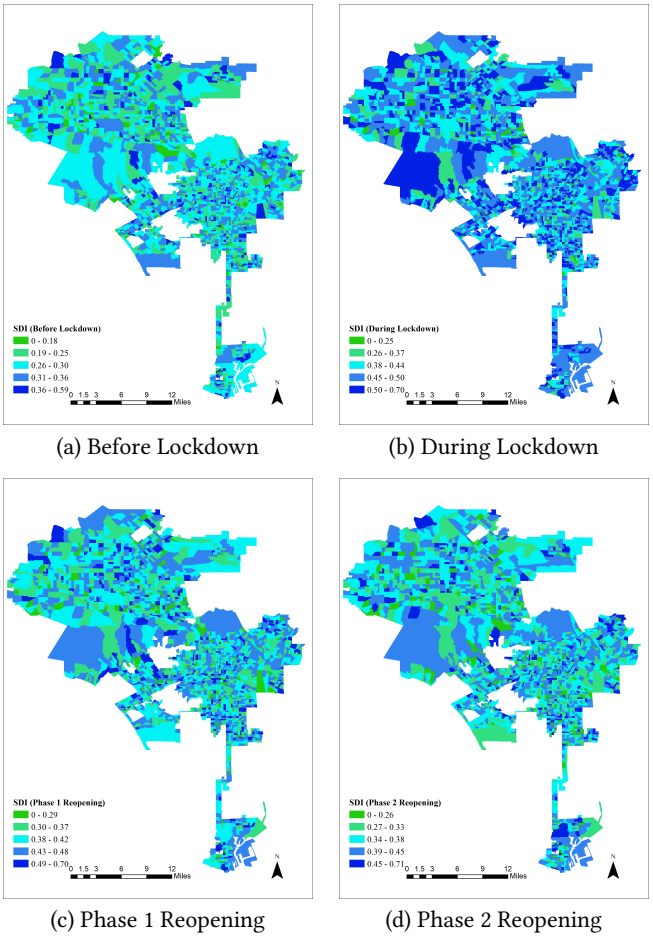


(a) Before Lockdown (b) During Lockdown

(c) Phase 1 Reopening (d) Phase 2 Reopening

Figure 6: Maps showing the spatial distribution of Mobility Index before, during and after lockdown in Los Angeles

Figure 6 visualizes the spatial distribution of the mobility index along with the number of confirmed COVID-19 cases during the periods of before lockdown (until March 19,2020), during lockdown (March 19 - May 8, 2020), phase 1 reopening (May 9 - June 11, 2020) and phase 2 reopening (June 12 - July 24, 2020) in LA. Evidently, the number of cases declined before and during lockdown when the mobility index declined and the social distancing index was high. It is clear that the restrictions imposed by the local authorities had a direct impact on the reduction in the number of daily cases across LA. Figure 7 shows how the social distancing measures changed over time spatially as more communities started to follow the distancing measures in comparison to certain communities where the mobility index was higher despite social distancing measures.



(a) Before Lockdown (b) During Lockdown

(c) Phase 1 Reopening (d) Phase 2 Reopening

Figure 7: Maps showing the spatial distribution of Social Distancing Index before, during and after lockdown in Los Angeles

The classification algorithms were used to classify the census block groups into three categories of high, medium, and low COVID-19 cases based on human mobility patterns, social distancing measures, industrial activity and social vulnerability indices. The hyperparameters for the random forest was set to 200 trees with a max depth of 5. A linear kernel was used for SVM, but the CART and Gradient Boost was trained with a maximum of 500 trees. Figure 8 shows the variability in the overall accuracy of all five models using the training data and Table 2 lists the Kappa Statistic along with AUC scores for each model to indicate the overall stability of the models.

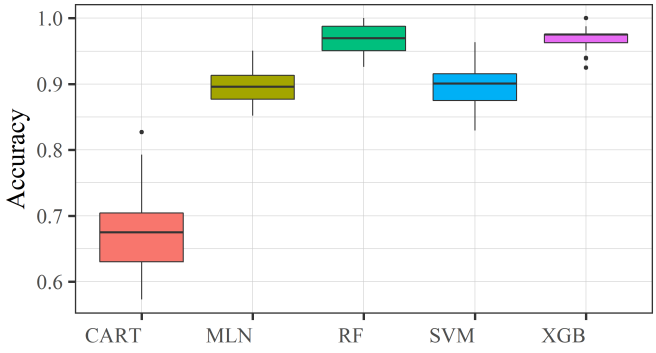


Figure 8: Model accuracy assessment across different supervised learning algorithms

From all the models, the Random Forests and Gradient Boost performed better overall compared to the other classification models. Gradient boost had a higher Kappa statistic and AUC score than Random Forests. Hence, Gradient Boost stood out to be the optimal choice with highest accuracy and lower chance of overfitting.

Table 2: Accuracy metrics of different models with 10-fold cross validation and 3 repeats

	Model	Accuracy	Kappa	AUC
1	SVM	90.5 %	0.853	0.955
2	MLN	91.3 %	0.867	0.960
3	RF	96.8 %	0.951	0.984
4	CART	65.9 %	0.457	0.806
5	XGB	97.4 %	0.960	0.987

Figure 9 shows the intensity of the number of COVID-19 cases that are categorized into Low, Medium, and High classes. The census block groups classified as High typically have more populations with a social vulnerability index between 0.5 and 0.75 indicating low socioeconomic status (SVI Theme 1), more disabled and elderly population (SVI Theme 3), and higher unemployment rates (SVI Theme 4).The language fluency of residents (SVI Theme 2) within a census block group did not seem to affect the occurrence of COVID-19 cases as the SVI were below 0.5 for all census block groups.

Additionally, the effect of household composition and disability is more pronounced in Phase 2 reopening with High cases being reported more frequently in CBGs with a more high concentration of disabled population and single-parent households (Figure 10). Social distancing was also found to have direct influence on the number of confirmed COVID-19 cases. Most census block groups classified as High had an overall social distancing index of less than 0.5 (Figure 11). This indicates that CBGs that followed strict social distancing guidelines had a lower number of reported COVID-19 cases; therefore, they have a lesser risk of spreading the disease.

The misclassification rate of each model using confusion matrices of the trained data is presented in Table 3. Figure 12 shows the confusion matrix of the Gradient Boost model as it had the highest accuracy with the actual labels and predicted labels based on the testing data.

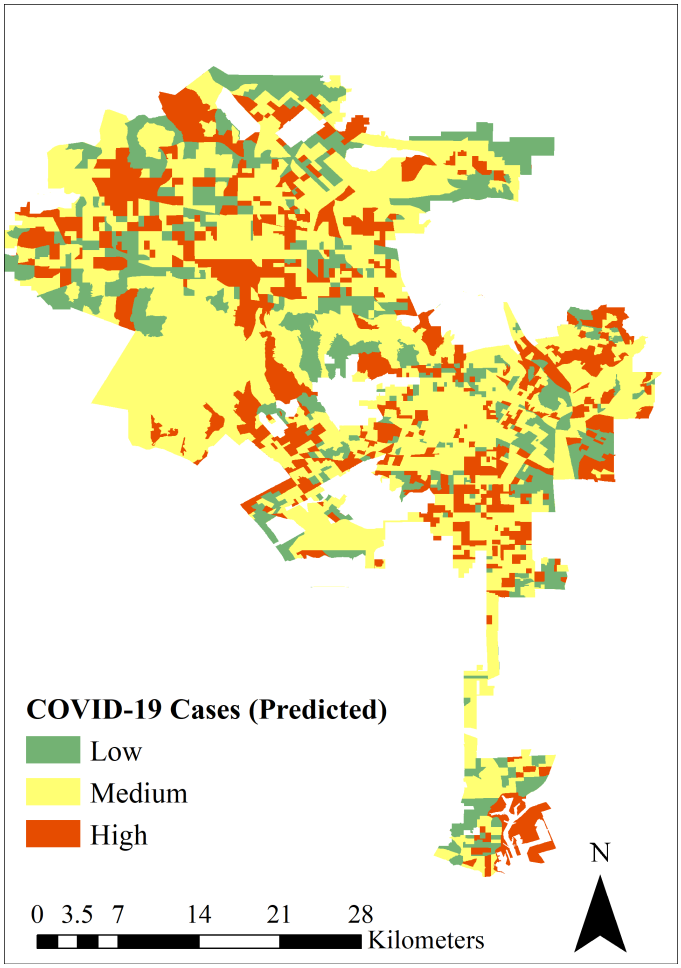


Figure 9: Map showing the predicted category of COVID-19 cases in each census block group of the city of Los Angeles

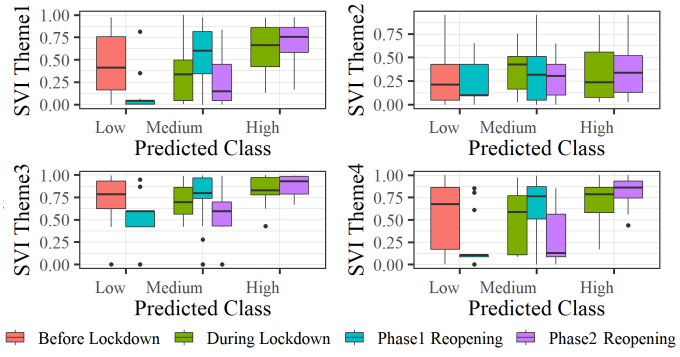


Figure 10: Variability of Social Vulnerability Indices by different themes based on predicted classes using test data

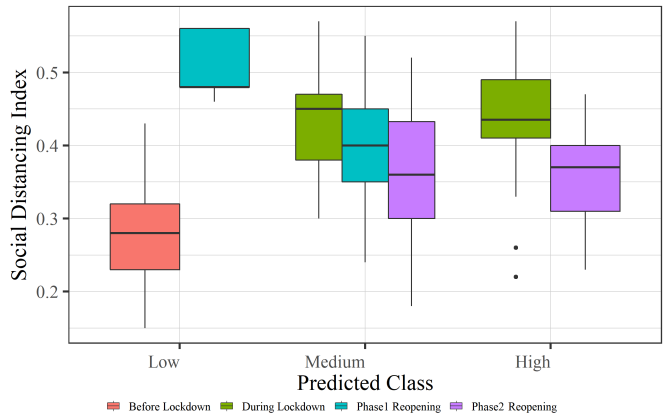


Figure 11: Variability of Social Distancing Index by predicted classes using test data

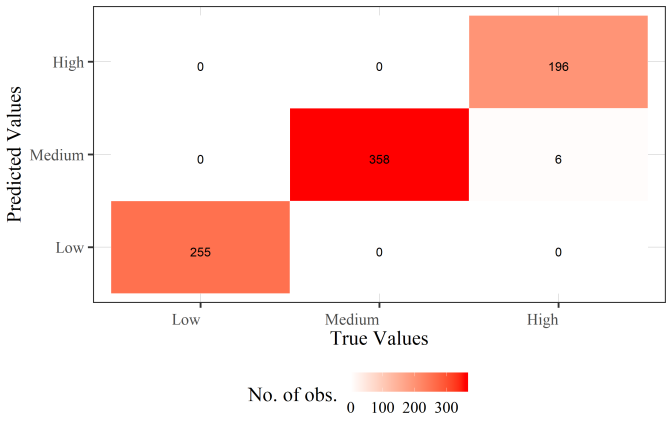


Figure 12: Confusion Matrix of predicted COVID-19 case in Los Angeles using Gradient Boost

Table 3: Summary of misclassification rate for predicted classes

Metrics	CART	MLN	RF	SVM	XGB
Sensitivity	0.69	0.91	0.92	0.89	0.99
Specificity	0.81	0.95	0.97	0.95	0.99
Precision	0.62	0.92	0.99	0.91	0.98
Recall	0.69	0.91	0.98	0.89	0.99
F1 Score	0.63	0.91	0.99	0.90	0.98
Balanced Accuracy	0.75	0.93	0.98	0.92	0.99

6 DISCUSSION

The gradient boost stands out as the best model in terms of accuracy and AUC score. Random Forest appears to have a high accuracy as well, however, there is a higher variability around the mean accuracy as shown in Figure 8 which could be due to model overfitting. Since gradient boosting does not use the individual trees, but rather averages all the trees together, therefore, for a particular data point (or group of points) the trees that over fit those points in the model will be an average of the under fitted trees and the combined average is adjusted accordingly by subsampling the features randomly.The confusion matrix for XGB (Figure 12) indicates that the model classified the low and high categories compared to the high cases. The misclassification rate was highest for Multinomial Logit and SVMs. The inaccuracies in classification mostly occur as a result of missing mobility data. Since, the SafeGraph data is only a sample of the actual mobility patterns across the study area, places where we had missing information on mobility index and social distancing are more prone to misclassification. With more data, such inaccuracies in the model can be gradually overcome.

Despite their overall accuracy, it is clear from all the models that the effect of mobility is more pronounced in census block groups with high socially vulnerable populations and those maintaining low social distancing measures ash shown in figures 10 and 11. Based on the themes of the social vulnerability indices, it was found that after the reopening phases, areas with a social vulnerability index of 0.5 to 0.75 based on socioeconomic status (SVI Theme 1), housing type and transportation (SVI Theme 4) shown in Figure 10 experienced high COVID-19 cases (classified as 'High'). The social distancing index varies from 0.4 to 0.5 for these census block groups and from 0.25 to 0.45 for CBGs classified as 'Low'. However, areas with mobility index between 0.5 and 0.75 falls into 'High' categories of COVID-19 cases compared to areas with mobility index between 0.25 and 0.5 which are classified as 'Low' (Figure 13).There are some census block groups where the mobility is high but also follow a strict social distancing measure, report a lower number of COVID-19 cases as a joint effect of both factors is more pronounced in such cases. Census block groups with a Medium number of COVID-19 cases have a mix of the socially vulnerable population residing in those areas along with high mobility and less pronounced social distancing measures (Figure 9). There a more census block groups near downtown LA which has a high number of COVID-19 cases as shown in Figure 9, which are caused not just due to high mobility but also due to a lack of access to healthcare facilities [40].

7 CONCLUSION

The findings are based on the SafeGraph data, which represents weekly human mobility patterns in LA. Hence, this data can be considered as a good approximation of the actual human movement and social distancing practices being followed over time in Los Angeles. From these data and reported COVID-19 cases, it is clear that social distancing is a useful tool to reduce COVID-19 spread, which in itself is not surprising. What is surprising and of value is that the block groups reporting a high number of cases also have high socially vulnerable populations including high unemployment rate and disabled populations. This is a cause for concern because (i) based on the race and ethnic profile of diseased COVID-19 patients reported by the LA Health Department, these block groups are occupied by the high-risk population groups who if infected have a high probability of mortality, and (ii) these CBGs are occupied by unemployed and disabled populations who will be infected if anyone of their family or community has a positive case and may

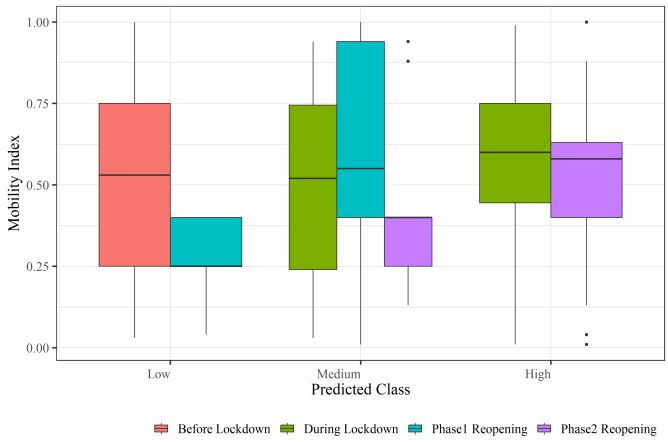


Figure 13: Variability of Mobility Index by predicted classes using test data

not be able to receive medical support due to disparity in access to health facilities [40].

Although lockdown is a good solution to maintain social distancing and spread of the disease, it is not an effective solution from economic perspective. Because social distancing appears to be an effective measure to reduce exposure to the virus, mitigation strategies like wearing masks and maintaining a certain distance in public spaces should be continued to ensure the virus spread is contained, especially, within the socially vulnerable population groups. Given that the block groups with high number of cases have high mobility index and high percentage of vulnerable population groups, it can be concluded that these block groups are probably occupied by people working in the service sectors. While it is not possible to stop these groups from traveling between home and work, another strategy to contain the spread would be to have targeted testing sites in the disadvantaged neighborhoods.

A major contribution of this study is that the framework described in this study can be useful for practitioners to understand heterogeneity across POIs, demographic groups, and neighborhoods for future mitigation strategies. Although the COVID-19 disease has a global footprint, its impacts are felt at a local level with a disparity in the number of cases and rates of spread. Hence, more fine-grained assessments of the effects of reopening policies need to be carefully addressed by local authorities to control further spread of the pandemic.

ACKNOWLEDGMENTS

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05- 00OR22725 with the U.S. Department of Energy. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. This manuscript was also supported in part by an appointment with the National Science Foundation (NSF) Mathematical Sciences Graduate Internship (MSGI) Program sponsored by the NSF Division of Mathematical Sciences. This program is administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and NSF. ORISE is managed for DOE by ORAU. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of NSF, ORAU/ORISE, or DOE.

REFERENCES

[1] 2018. WorldPop. https://dx.doi.org/10.5258/SOTON/WP00674
[2] 2020. California reopening: Tracking progress across counties. https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/reopening-across-counties/
[3] 2020. LA County Department of Public Health. http://publichealth.lacounty.gov/
[4] 2020. LAC Open Data: LAC Open Data. https://data.lacounty.gov/
[5] 2020. Report on LA County COVID-19 Data Disaggregated by Race/ Ethnicity and Socioeconomic Status. http://publichealth.lacounty.gov/docs/RacialEthnicSocioeconomicDataCOVID19.pdf
[6] 2020. "SafeGraph", a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places. To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an establishment in a month from a given census block group. https://docs.safegraph.com/docs/weekly-patterns
[7] José M Baldasano. 2020. COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain). *Science of the Total Environment* (2020), 140353.
[8] Jesse D Berman and Keita Ebisu. 2020. Changes in US air pollution during the COVID-19 pandemic. *Science of the Total Environment* 739 (2020), 139864.
[9] KF Boersma, HJ Eskes, JP Veefkind, EJ Brinksma, RJ Van Der A, M Sneep, GHJ Van Den Oord, PF Levelt, P Stammes, JF Gleason, et al. 2007. Near-real time retrieval of tropospheric NO 2 from OMI. (2007).
[10] K. F Boersma, J. P. Eskes, and E. J. Veefkind. 2013. Near-real time retrieval of tropospheric NO2 from OMI. https://sos.noaa.gov/datasets/nitrogen-dioxide/
[11] Giovanni Bonaccorsi, Francesco Pierri, Matteo Cinelli, Andrea Flori, Alessandro Galeazzi, Francesco Porcelli, Ana Lucia Schmidt, Carlo Michele Valensise, Antonio Scala, Walter Quattrociocchi, et al. 2020. Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15530–15535.
[12] Martin CJ Bootsma and Neil M Ferguson. 2007. The effect of public health measures on the 1918 influenza pandemic in US cities. *Proceedings of the National Academy of Sciences* 104, 18 (2007), 7588–7593.
[13] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.
[14] US Census Bureau. 2020. American Community Survey Data Releases. https://www.census.gov/programs-surveys/acs/news/data-releases.html
[15] Peter Caley, David J Philp, and Kevin McCracken. 2008. Quantifying social distancing arising from pandemic influenza. *Journal of the Royal Society Interface* 5, 23 (2008), 631–639.
[16] State of California. [n.d.]. County variance info. https://covid19.ca.gov/roadmap-counties/
[17] Serina Y Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2020. Mobility network modeling explains higher SARS-CoV-2 infection rates among disadvantaged groups and informs reopening strategies. *medRxiv* (2020).
[18] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
[19] Daniel Duque, David P Morton, Bismark Singh, Zhanwei Du, Remy Pasco, and Lauren Ancel Meyers. 2020. Timing social distancing to avert unmanageable COVID-19 hospital surges. *Proceedings of the National Academy of Sciences* 117, 33 (2020), 19873–19878.
[20] ESRI. 2011. ArcGIS Desktop: Release 10.6.
[21] Aidan Findlater and Isaac I Bogoch. 2018. Human mobility and the global spread of infectious diseases: a focus on air travel. *Trends in parasitology* 34, 9 (2018), 772–783.
[22] Barry E Flanagan, Elaine J Hallisey, Erica Adams, and Amy Lavery. 2018. Measuring community vulnerability to natural and anthropogenic hazards: the Centers for Disease Control and Prevention's Social Vulnerability Index. *Journal of environmental health* 80, 10 (2018), 34.
[23] Edmilson D Freitas, Sergio A Ibarra-Espinosa, Mario E Gavidia-Calderón, Amanda Rehbein, Sameh A Abou Rafee, Jorge A Martins, Leila D Martins, Ubiratan P Santos, Mariangeli F Ning, Maria F Andrade, et al. 2020. Mobility Restrictions and Air Quality under COVID-19 Pandemic in São Paulo, Brazil. (2020).

[24] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special* 12, 1 (2020), 16–26.

[25] David J Hand and Robert J Till. 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning* 45, 2 (2001), 171–186.

[26] Richard J Hatchett, Carter E Mecher, and Marc Lipsitch. 2007. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences* 104, 18 (2007), 7582–7587.

[27] Lingqian Hu, Zhenlong Li, and Xinyue Ye. 2020. Delineating and modeling activity space using geotagged social media data. *Cartography and Geographic Information Science* 47, 3 (2020), 277–288.

[28] Xiao Huang, Zhenlong Li, Yuqin Jiang, Xinyue Ye, Chengbin Deng, Jiajia Zhang, and Xiaoming Li. 2020. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the US during the COVID-19 pandemic. *medRxiv* (2020).

[29] Jayson S Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia, and Nicholas A Christakis. 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* (2020), 1–5.

[30] Joel K Kelso, George J Milne, and Heath Kelly. 2009. Simulation suggests that rapid activation of social distancing can arrest epidemic development due to a novel strain of influenza. *BMC public health* 9, 1 (2009), 117.

[31] Stephen M Kissler, Christine Tedijanto, Marc Lipsitch, and Yonatan Grad. 2020. Social distancing strategies for curbing the COVID-19 epidemic. *medRxiv* (2020).

[32] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* 368, 6490 (2020), 493–497.

[33] H Läuter. 1988. Silverman, BW: Density Estimation for Statistics and Data Analysis. Chapman & Hall, London–New York 1986, 175 pp.,£ 12.—. *Biometrical Journal* 30, 7 (1988), 876–877.

[34] Alix Martichoux. 2020. Coronavirus watch list: 35 California counties where COVID-19 is getting worse. https://abc7news.com/california-county-watch-list-counties-on-covid-watchlist-ca/6265270/

[35] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[36] Robert Moss, Elham Naghizade, Martin Tomko, and Nicholas Geard. 2019. What can urban mobility data reveal about the spatial distribution of infection in a single city? *BMC public health* 19, 1 (2019), 1–16.

[37] ESMD naics@census.gov. 2019. North American Industry Classification System (NAICS) Main Page. https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf

[38] R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[39] Avipsa Roy, Daniel Fuller, Kevin Stanley, and Trisalyn Nelson. 2021. Classifying Transportation Mode from Global Positioning Systems and Accelerometer Data: A Machine Learning Approach. *Transport Findings*, (Sep 2021). https://doi.org/10.32866/001c.14520

[40] Avipsa Roy and Bandana Kar. 2021. A Multicriteria Decision Analysis Framework to Measure Accessibility to Healthcare Facilities in the Wake of COVID-19 (Under Review). In *TRB 2021 Annual Meeting*.

[41] Avipsa Roy, Trisalyn A Nelson, A Stewart Fotheringham, and Meghan Winters. 2019. Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. *Urban Science* 3, 2 (2019), 62.

[42] Efrat Shadmi, Yingyao Chen, Inês Dourado, Inbal Faran-Perach, John Furler, Peter Hangoma, Piya Hanvoravongchai, Claudia Obando, Varduhi Petrosyan, Krishna D Rao, et al. 2020. Health equity and COVID-19: global perspectives. *International journal for equity in health* 19, 1 (2020), 1–16.

[43] Lauren A White, James D Forester, and Meggan E Craft. 2018. Disease outbreak thresholds emerge from interactions between movement behavior, landscape structure, and epidemiology. *Proceedings of the National Academy of Sciences* 115, 28 (2018), 7374–7379.