Factors Affecting Infection Rate and Human Mobility Flow Analysis During the COVID-19 Pandemic

David Liau (A12867539)
Fengqi Yang(A99015722)
Gates Zeng (A13181719)
Department of Electrical and Computer Engineering,
University of California, San Diego
9500 Gilman Dr., La Jolla, CA 92093

Abstract—**The contagious nature of COVID-19 has allowed the virus to propagate quickly and globally. To mitigate the damage caused by the pandemic, it is vital to uncover the factors that most impact the spread of the disease so that policymakers can make informed public health decisions. In our project, we attempt to identify the features that most influences COVID-19 infection rates through statistical analysis. Additionally, we attempt to visualize the effectiveness of California's lockdown by utilizing human mobility flow data and combining it with county-level infection cases during the pandemic.**

Keywords: COVID-19, Human Mobility Flows, Generalized Linear Model (GLM), Principal component analysis (PCA), Random Forest, Census

I. Introduction

COVID-19 has become a global pandemic in a short period due to the highly infectious nature of the disease. It has become a challenge for governments to identify policies that would be maximally effective at reducing the virus spread while also being minimally invasive in the lives of their citizens. There is a need to identify the factors that will most impact the infection rates of the disease so that more targeted policies can be made. Sweeping policies such as travel restrictions and lockdown are undesirable but may be necessary to effectively control the spread of the disease. In order to address above problems, in our project, we have two parts implementing different methodologies:

Factors affect COVID-19 infection rates In the first part of our project, we select census variables that might be interesting to analyze and might be highly correlated with infection rates including ethnicity, age group, gender, etc. Utilizing Principal component analysis (PCA), correlation analysis, and statistical modeling methodologies, we analyze how different groups of individuals with different ethnicity, age group, gender, etc. are correlated with the pandemic spread.

Human mobility flow during COVID-19 In the second part of our project, we build graph-based methodologies to visualize human mobility flows across different counties within California combined with confirmed COVID-19 cases. Through network graphs, our objective is to visualize mobility flows changes combined with trends of COVID-19 confirmed cases.

II. Related Work

In this section we will discuss the method and/or results introduced in the report are compared with current related models and/or results in the literature.

As the COVID-19 pandemic spread quickly across different places over the world, many disciplines have joined the scientific community of COVID-19. Machine Learning and Data Science contributed with various works to support COVID-19 challenges such as modeling, simulation, predictions, social networks analytics, Geographic Information Systems (GIS) [1] for spatial segmentation and tracking, etc.

Previous work has attempted to use supervised machine learning techniques to identify the discriminatory factors that contribute to COVID-19 infections [2]. Methods such as randomized decision trees, multiple linear regression, and principal component analysis are applied on datasets captured across the 50 US states.

Studies of mobility flows have been considered powerful tools in planning and predicting pandemic spread as researchers demonstrate strong positive correlations between mobility flows and confirmed cases as well as deaths. It is widely accepted that restricting human mobility is an effective strategy used to control disease spread, and [3] developed a model that can quantify the potential effects of various intracity mobility restrictions on the spread of COVID-19.

In addition, Graph analytics are also been applied to the study of tracking COVID-19 spread across regions, the number of confirmed cases, prediction of deaths, etc. Data analysis through the graph-based method is now considered state-of-the-art in many applications of community detection. The combination between the graph's definition in mathematics and the graphs in computer science as an abstract data structure is the key behind the success of graph-based approaches in machine learning [4].

III. Dataset

In this section we will discuss the dataset used for our methodologies.

Our final, analysis-based dataset is consolidated from three separate datasets, which we explain in detail: **2020 Census** The 2020 census data has a large variety of datasets corresponding to multiple societal topics (e.g. income, location, employment, family) – and granular breakdowns for each category. Due to the complex computational need for full analysis, we select specific subcategories to integrate within our model. These subcategories include Income-Poverty Ratio, Age, Education Level, Population, and Population Density.

COVID-19 Infection Data COVID-19's effect on the country has been recorded consistently by the CDC, allowing for readily available data. This dataset contains daily cumulative infections and deaths by county.

County-level Adjacency Matrix To ensure a graphical component to our analysis, we obtained a dataset that represents, on a county level, neighboring counties to any of the 3000 points within the US.

Multi-scale Dynamic Human Mobility Flow Dataset [5] This dataset attempted to predict human mobility flow during the COVID-19 epidemic. First, the trajectories of anonymous cell phone users were obtained by SafeGraph[6] using GPS pings. Then, the user locations were spatially clustered to census administrative regions to protect the privacy and to allow for observation at different geographic scales. They then used this information to calculate the number of users (visitors) who traveled to a different administrative region. Since this data only was about a 10% sample of the entire population, American Community Survey (ACS) population data provided by the US Census was combined to infer the population level mobility flows. The formula the authors used is given in Equation 1. The human mobility flows were calculated across two-time intervals (daily and weekly) using three different scales (census tract, county, and state). In our paper, we focused on the county-to-county mobility flows.

$$population\_flows(o,d) = visitor\_flows(o,d) \times \frac{population(o)}{num\_devices(o)} \tag{1}$$

IV. Model and Methodology

In this section, we discuss proposed models used for the analysis of features from census data collected that affect COVID-19 infections, as well as the methodology used for visualization of mobility flow.

A. *COVID-19 Factors affecting Infections*

In the first part of our project, our primary goal was to identify, empirically, which features within our consolidated dataset most heavily affected county-wide COVID-19 infection rate, as well as the death rate. Through explorations and experiments, we found three models that are useful for our study:

Generalized Linear Model To start with, we utilized a Generalized Linear Model (GLM) from the statsmodels package[7] due to its ease of implementation and thorough summary. GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution[8]. In specific, we aimed to minimize the deviance score to extract feature coefficients, which would determine effectiveness at a roughly independent scale.

Random Forest Regressor We also sought to utilize scikit-learn's Random Forest Regressor (RFR) model, in large part due to its features_importance_ property, which displays and ranks certain features within the dataset that impact our selected target. Theoretically, features_importance_ property of Random Forest Regressor is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. So analyzing this property we can extract some useful information of key variables that influence the target variable we are interested in. In addition, Random Forest is usually more robust than a Decision Tree Model as it is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large[9].

Once we have extracted the feature coefficients from each model, we may look to combine the GLM and RFR concepts for future model choices.

Principal Component Analysis In our project, as both COVID-19 dataset and census datasets are relatively large and hard to navigate directly, so we conduct analysis using PCA to extract more useful insights. Principal Component Analysis is a technique for reducing the dimensionality of large datasets while also minimizing information loss. [3] This technique attempts to identify several principal components that maximize the variance of the data and captures the most important information in the data. We used PCA to confirm the results generated from our GLM as well as compare our results with that obtained in [2].

B. Mobility Flow Visualization

The methodology we use for mobility flow visualization is base on undirected Networkx graphs. By building network graphs where each county in California represent a node, we visualize both mobility flows and confirmed cases in one graph on a specific date during pandemic. The edges between nodes were represent if there is mobility flow between two nodes on a given date. In addition, the size of each node represents number of confirmed cases in the corresponding county.
The color of each node represents the magnitude of mobility inflow to that county where deeper blue means larger mobility inflow to that county on a given date and lighter blue means less mobility inflow from other counties. Given a specific date, our method will output a network graph according to above characteristics.

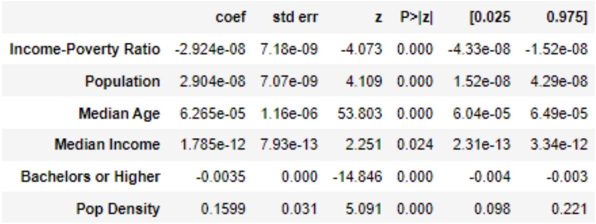
The reason for putting our focus on California only in this project is because visualizing mobility flow across counties which is a relatively large dataset requires both memory and processing power of CPU. We discussed all limitations and future works in the last section of our report.

V. Results

In this section, we demonstrate the results of our experiments using proposed methods.

A. COVID-19 Factors Affecting Infections

Generalized Linear Model Our Generalized Linear Model produced several coefficients that we lay out in the following image. Ultimately, population density seems to impact COVID-19 Death Rate per county the most, given its 0.1599 value. This, compared at scale to other values, seems to be the most impactful factor – a reasonable conclusion, given the high likelihood of COVID-19 infection spreading in a dense area. We also see, unsurprisingly, that a higher average level of education brings down the death rate by a minor amount. This assumption comes from previous work[2] which illuminate the education and mask-wearing correlation. The rest of the features all seem to impact the death rate a negligible amount.



	coef	std err	z	P>|z|	[0.025	0.975]
Income-Poverty Ratio	-2.924e-08	7.18e-09	-4.073	0.000	-4.33e-08	-1.52e-08
Population	2.904e-08	7.07e-09	4.109	0.000	1.52e-08	4.29e-08
Median Age	6.265e-05	1.16e-06	53.803	0.000	6.04e-05	6.49e-05
Median Income	1.785e-12	7.93e-13	2.251	0.024	2.31e-13	3.34e-12
Bachelors or Higher	-0.0035	0.000	-14.846	0.000	-0.004	-0.003
Pop Density	0.1599	0.031	5.091	0.000	0.098	0.221

Fig. 1. Generalized Linear Model Regression Results

Random Forest Regressor In our experiment, the Random Forest Regressor revealed the top most important features in predicting the COVID-19 death rate. There was a surprisingly noticeable gap between one aspect and the rest – that is, a high education level significantly dwarfed every other feature in importance. Ranked similarly as important were the population density, median income, and income-poverty ratio, while population and median age were trivial. The findings here are interesting, given the assumption that population density would be a large portion of COVID-19 infection rates.

Principal Component Analysis In order to identify the feature importance using PCA, we normalized the data and then projected it across five principal components. Then we gathered the principal axis contribution of each feature as well the percentage of variance that was explained by each principal component, which can be seen in Table 1. From our table, it is seen that Income-Poverty Ratio and Population contribute significantly to the variance explained, followed by Median Income, Population Density, and Bachelors Degree or higher, respectively.

Overall Comparison Overall, each methodology used highlighted different features within our dataset: the GLM attributed population density with the highest coefficient, which means a higher correlation to the target variable. And the Random Forest Regressor marked education level as most important in explaining the target variable. In addition, the PCA method highlighted income-poverty and population variables as the most influencing factors that impact infections.

B. Mobility Flow Visualization

In this section we will show the visualization of the mobility flow across county in California and COVID-19 confirmed cases on a daily basis.

	Variance Explained	Income-Poverty Ratio	Population	Median Age	Median Income	Bachelors or Higher	Population Density
PC-1	51.38%	0.645	0.645	-0.017	0.030	0.105	0.394
PC-2	25.06%	0.015	0.015	0.006	-0.999	0.020	0.022
PC-3	19.84%	-0.283	-0.281	-0.000	0.012	0.027	0.917
PC-4	3.33%	0.062	0.059	0.010	-0.016	-0.994	0.067
PC-5	0.04%	0.018	0.002	1.00	0.007	0.011	0.006

TABLE I
FEATURE CONTRIBUTION TO EACH PRINCIPAL COMPONENT

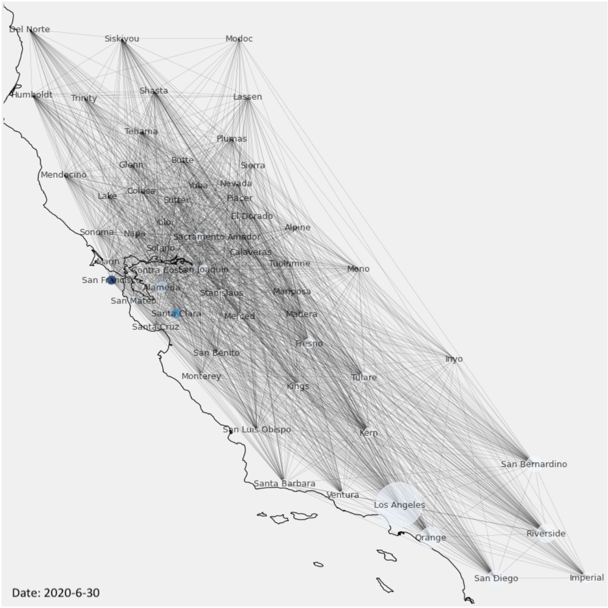


Fig. 2. COVID-19 confirmed cases and mobility flows on 2020-06-30

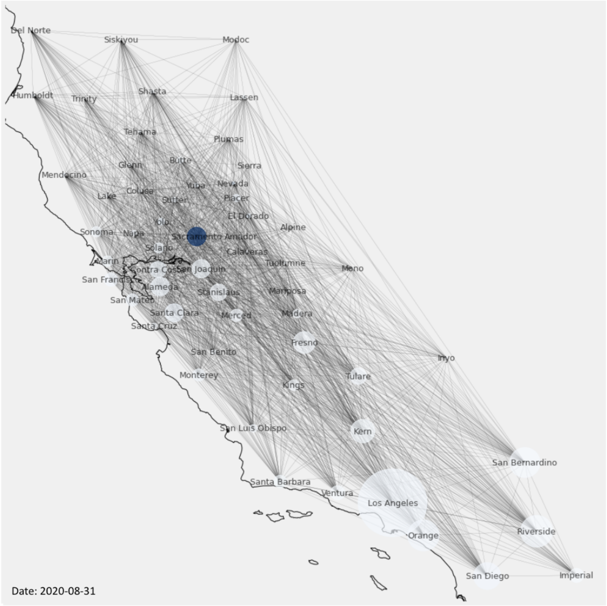


Fig. 3. COVID-19 confirmed cases and mobility flows on 2020-08-31

In Figure 1 - Figure 4, we provided the network graph on four specific date: 2020-06-30, 2020-08-31, 2020-10-31, 2021-03-01 which we believe represents different stages of pandemic spread and California lockdown policy: in the beginning of state lockdown, during lockdown and pandemic peak, after lockdown respectively.

Usefulness of the graphs From four figures, we can easily observe some trends and changes over time and directly visualize the state of COVID-19 confirmed cases as well as mobility flow for all counties in California. This is also our main objective of building this network graph-based methodology. By looking at the graph, policymakers or researchers can obtain direct messages of how large the mobility flows at different counties during the pandemic as mobility is a key factor that impacts infections. As previous studies show, human mobility restrictions in the city had a large effect on controlling the COVID-19 outbreak, especially when implemented in conjunction with efforts to reduce transmissibility,[10] we think visualizing the current mobility flow daily will help policymakers or other researchers to better control the pandemic. In addition, there are studies intend to assess the dynamics of respiratory

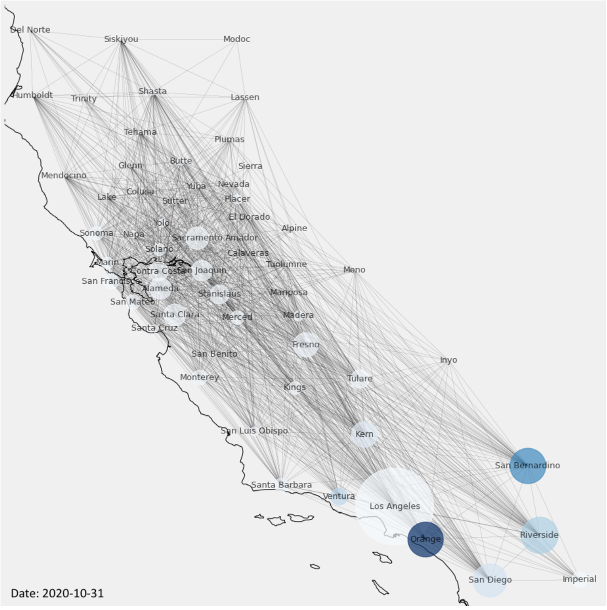


Fig. 4. COVID-19 confirmed cases and mobility flows on 2020-10-31

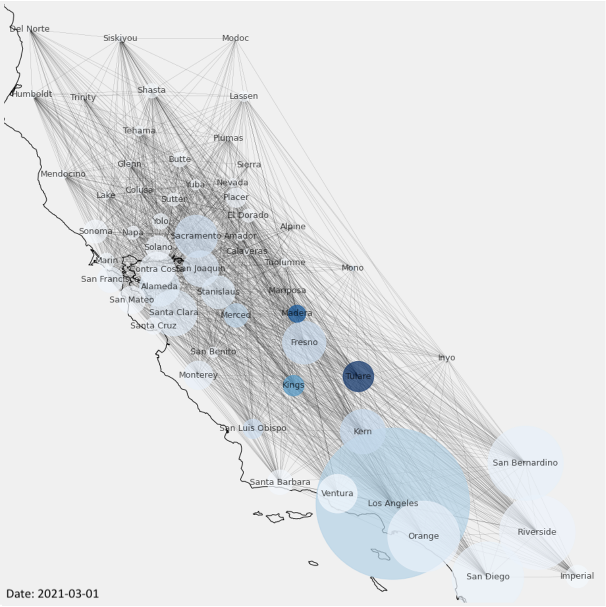


Fig. 5. COVID-19 confirmed cases and mobility flows on 2021-03-01

infectious diseases associated with human mobility. In Hou et al.'s study[11], researchers show that, by reducing the contact rate of latent individuals, interventions such as quarantine and isolation can effectively reduce the potential peak number of COVID-19 infections and delay the time of peak infection. So when combining the visualizations of our network graphs with those SEIR models, the policymakers and planners can quickly react to events like a sudden increase in mobility flows at certain counties to help manage COVID-19 outbreaks.

Moreover, the graphs we produced can also provide an easy way to see the number of confirmed COVID-19 cases at each county on a specific date besides mobility flow, which further helps policymakers and planners to get a general understanding of the current state of epidemic infections at each county. For example, if policymakers and planners use our graph on 2020-10-31, Orange county has large mobility inflows represented by node's color and the number of confirmed cases was relatively higher than most north California counties. This may be an indicator of a jump in confirmed cases in next coming days as there is an interval period between infected and confirmed. So policymakers and planners potentially can make arrangements or actions to prevent further outbreaks on 2020-10-31.

Last but not the least, our network graph can be expanded to visualizations of more regions other than California by incorporating COVID-19 infection data and mobility flow across counties data. Also, output a visualization on a specific date can be achieved by running our Python script and changing the date filter. These graphs can also be applied for pandemic or epidemic infections other than COVID-19.

Observations and Findings First, Look at different timestamps, both the magnitude of mobility flow and the number of confirmed cases is increasing over time which is as expected as counties reopening gradually after lockdown and the number of confirmed cases is cumulative. So the size of nodes becomes larger and the amount of deeper color nodes is increasing. Secondly, it's obvious that at the beginning of lockdown and during lockdown the mobility across the county at different time snapshot did not change much in northern California which is represented by the color of the nodes. However, in southern California, the county mobility inflow was larger, and confirmed cases were more than in north California especially in October 2020 and March 2021. Last but not the least, from the graphs, we can observe that northern California and southern California are forming two communities over time in which have similar trends of both mobility flows and confirmed cases. For example, most counties in northern California are small and light color nodes. However, in southern California, most nodes are larger and deeper color nodes. We think the main reason for this community formulation is because of geographic location. During a nationwide lockdown, most non-essential travel was restricted especially those by air. Most essential commute or travel occurred within neighbor counties that can use a car, train, etc. as transportation tools. So neighborhood counties are more likely to have similar trends of mobility flows as well as the spread of the pandemic.

VI. Comparison with Existing Works

A. *COVID-19 Factors Affecting Infections*

As referenced earlier, several attempts have been made to explain COVID-19 infection sources and reasons[2]. Such works have applied supervised machine learning approaches in identifying key factors that influence infections. Previous study carry out regression analysis to pinpoint the key pre-lockdown factors that affect post-lockdown infection and mortality, informing future lockdown-related policy making. While our methodology utilize supervised machine learning techniques, as well as easy to understand statistical regression models as well as feature decomposition approaches to understand key factors across the nation instead of across different timestamp of the lockdown. While different attempts utilized a varying amount of features, population density stood out as a common factor in increased COVID-19 cases.

In future works, we look to engineer further data points together, selecting and combining several features between each work.

B. Mobility Flow Visualization

Mobility and travel restriction together with lockdown are important public health tools and has been used in the US as well as globally to control the COVID-19 pandemic. In Zhou et al. study[10] of effects of human mobility restrictions on the spread of COVID-19, the authors show the effects of various types and magnitudes of mobility restrictions on controlling COVID-19 outbreaks at the city level in Shenzhen, China using a modified susceptible–exposed–infectious–recovered (SEIR) compartmental transmission model. The authors state that the model could help policymakers to establish the optimal combinations of mobility restrictions during the COVID-19 pandemic, especially to assess the potential positive effects of mobility restriction on public health because of the potential negative economic and societal effects. In our project, instead of using the SEIR model to simulate or predict the effects, we take the approach of visualizing the effect of mobility by building a network graph. As visualization is a supplementary task [4], and the main objective of this part in our project is to provide a direct and easy-to-understand graph showing mobility flow, as well as COVID-19, confirmed cases on a specific date rather than simulation. With the help of data visualization, when illustrating the current state of mobility flows and confirmed COVID-19 cases, we believe visual graphs can describe the output of our analysis in concise and direct ways than the mathematical results or any textual description for this part in our project.

VII. Conclusion

In Summary, we proposed two methodologies for analyzing and visualizing COVID-19 infection spreading. The first methodology required extracting and building a master dataset of features, which we fed into three models: a Generalized Linear Model, a Random Forest Regressor, and a Principal Component Analysis table. While each model produced different results, we do not believe this undermines our analysis – rather, it demonstrates the complexity of COVID-19 infection sources and the combination of features that affect them.

Our second methodology proposed is to visualize the mobility flow across California on a given date through building network graphs. In our experiments, we build four graphs on different dates, where each representing a different point of COVID-19 county-wide infections within the state. We identify useful insights into mobility flows within different regions, and the relationship between flows and confirmed COVID-19 cases at the certain county. And we suggest that these graphs may potentially help policymakers mitigate increased infection amounts in addition to controlling the pandemic.

We believe that the two established methods can and will help with quelling the COVID-19 infections that have ravaged the world since 2019 – as well as prevent future pandemics from occurring at such a globalized pace. We hope that this paper, along with other research, can inform and educate others within the human race so that future situations can be resolved at a faster, and more optimized, rate.

VIII. Limitation and Future Work

Due to the limitation of time and resources available, we did not conduct experiments on all state-of-arts machine learning models to identify correlated features. So in future works, we plan to explore more possibilities of different models especially Gradient Boosting Tree or Neural Networks as these models usually achieve higher performance in many real work problems. Additionally, we look to improve the granularity of our dataset, perhaps across cities or sorted by urbanicity as the available census data contains many other variables that may be helpful for our study.

Moreover, in our graph-based methodology of analyzing mobility flow, it may be helpful to create an interactive dashboard that gives the end-user the flexibility to navigate through different dates. However, due to the time limit of this project, our visualizations can only be produced by running the python script we built. Another opportunity for improvement is to include more regions other than California in our graphs so end users can visualize nationwide mobility flow together with infection numbers.

References

[1] C. Zhou, F. Su, T. Pei, A. Zhang, Y. Du, B. Luo, Z. Cao, J. Wang, W. Yuan, Y. Zhu, and et al., *Covid-19: Challenges to gis with big data*, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666683920300092.

[2] S. Roy and P. Ghosh, *Factors affecting covid-19 infected and death rates inform lockdown-related policymaking*. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371\%2Fjournal.pone.0241165.

[3] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, Apr. 2016. DOI: 10.1098/rsta.2015.0202.

[4] Z. A. E. Mouden, R. M. Taj, A. Jakimi, and M. Hajar, *Towards using graph analytics for tracking covid-19*, Nov. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920322961.

[5] Y. Kang, S. Gao, Y. Liang, M. Li, and J. Kruse, "Multiscale dynamic human mobility flow dataset in the u.s. during the covid-19 epidemic," *Scientific Data*, pp. 1–13, 390 2020.

[6] *Safegraph: The source of truth for physical places*. [Online]. Available: https://www.safegraph.com/about.

[7] *Statsmodels: Generalized linear models*. [Online]. Available: https://www.statsmodels.org/stable/glm.html.

[8] *Generalized linear model*, en, Page Version ID: 1018579650, Apr. 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Generalized_linear_model&oldid=1018579650 (visited on 06/09/2021).

[9] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," en, in *Ensemble Machine Learning*, C. Zhang and Y. Ma, Eds., Boston, MA: Springer US, 2012, pp. 157–175, ISBN: 9781441993250 9781441993267. DOI: 10.1007/978-1-4419-9326-7_5. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-9326-7_5 (visited on 06/09/2021).

[10] Y. Zhou, R. Xu, D. Hu, Y. Yue, Q. Li, and J. Xia, "Effects of human mobility restrictions on the spread of covid-19 in shenzhen, china: A modelling study using mobile phone data," *The Lancet. Digital Health*, vol. 2, no. 8, e417–e424, Aug. 2020, ISSN: 2589-7500. DOI: 10.1016/S2589-7500(20)30165-5.

[11] C. Hou, J. Chen, Y. Zhou, L. Hua, J. Yuan, S. He, Y. Guo, S. Zhang, Q. Jia, C. Zhao, and et al., "The effectiveness of quarantine of wuhan city against the corona virus disease 2019 (covid-19): A well-mixed seir model analysis," *Journal of Medical Virology*, vol. 92, no. 7, 841–848, Jul. 2020, ISSN: 1096-9071. DOI: 10.1002/jmv.25827.