FORECAST-DRIVEN EVIDENCE OF HUMAN MOBILITY IMPACTING THE SPREAD OF COVID-19

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing the exact impact of various components of non-pharmaceutical interventions on a pandemic is difficult. We focus on studying whether human mobility has had a significant impact on the spread of COVID-19. We perform a forecast-driven analysis on two types of mobility patterns in Los Angeles: inter-region mobility among various census blocks in Los Angeles and mobility patterns based on business/purpose of visits. On both types of mobility patterns, we demonstrate that the mid-term forecast (up to 5 weeks ahead) does improve significantly by accounting for the mobility. Additionally, the proposed analysis and models can be used to extrapolate the effect of various mobility scenarios into the future to plan levels of lockdown.

1 INTRODUCTION

There has been a significant amount of debate on the levels of lock-downs necessary to contain the spread of COVID-19[1]. While there have been studies showing evidence of lockdowns impacting the epidemic spread (Atalan, 2020), it is non-trivial to extrapolate using such studies into measuring the effect of future interventions. We acknowledge that identifying the exact impact and causality relationships between human behavior and the epidemic is difficult. Instead, we take a forecast-driven approach, where we analyze if the foresight of human mobility patterns could improve the prediction of the spread of the epidemic. The advantages of such analysis are twofold: (i) If the foresight of mobility patterns improves the predictions, it would suggest that it has an impact on the epidemic spread. And, (ii) the learned model can enable the study of the impact of human mobility (as a proxy for various levels of lock-downs) on the epidemic in the future.

We perform our forecast-driven analysis on two types of mobility patterns in Los Angeles: inter-region mobility among various census blocks in Los Angeles and mobility patterns based on business/purpose of visits (POI: Points-of-Interest). For the latter, we further various aggregation strategies to generate mobility features: (i) aggregated mobility trends from Google, (ii) manual clustering of visits to POIs by their types, and (iii) automatic clustering of visits to POIs based on visit patterns. For all the mobility patterns, we demonstrate that the mid-term forecast (up to 5-week-ahead) does improve significantly by accounting for the mobility compared to ignoring mobility. We do so by measuring the errors of variations of SIkJα model (Srivastava et al., 2020) with and without mobility-related variables.

2 METHODOLOGY

2.1 THE SIKJα MODEL

We use variations of our SIkJα approach (Srivastava et al., 2020). This approach models the epidemic as a discrete-time process with temporally varying infection and death rates. The model considers many complexities such as unreported cases due to any reason (asymptomatic, mild symptoms, willingness to get tested), immunity (if any) or complete isolation, and reporting delay, and yet, it can be reduced to a system of two linear equations which can be fitted one after the other

[1] https://www.reuters.com/article/uk-factcheck-lockdowns/fact-check-studies-show-covid-19-lockdowns-have-saved-lives-idUSKBN2842WS

resulting in fast yet reliable forecasts. The forecasting methodology currently used with SIkJα to generate public forecasts does not account for recent mobility. The forecasts are performed by smoothing the data, learning the parameters by fitting the reduced model, and then simulating the model into the future using these learned parameters. The approach has the advantage of being extremely fast (approx. 5 mins for 20,000 locations on commodity hardware). Yet it has consistently performed among the top methods for various COVID-19 related tasks such as US state-level case and death forecasts (Srivastava et al., 2020; Srivastava, 2021), Germany/Poland case and death forecasts (Bracher et al.), and country-level death forecasts (Friedman et al., 2021). A preliminary version of SIkJα model was one of the winners used in the 2014-2015 CHIKV DARPA Grand Challenge (noa, 2015). In this paper, we will incorporate a foresight of future mobility into the model. While this is impractical for the purpose of real-time forecasting, retrospectively, this helps us assess relationships between mobility and the trajectory of the epidemic.

2.2 INTER-REGION MOBILITY

Inter-region mobility is incorporated using the same equation as the original SIkJα model Srivastava & Prasanna (2020b). We used the following approaches to generate variations of the model.

- A1: Infection rate driven by a fixed parameter scaled up and down by changes in within region mobility (α in the $SIkJ\alpha$ model is set to 1 suggesting that the transmission rate does not change with time in the training data due to any factor other than mobility)
- A2: Infection rate driven by learning from the recent data without explicitly considering within region mobility (performing hyper-parameter search for α, which is the default learning approach)
- B1: Normalized mobility matrix, zero diagonal
- B2: mobility matrix with zero diagonal without normalization

With the above components we create the following models: M1 = A1, M2 = A2, M3 = (A1, B1), M4 = (A1, B2), M5 = (A2, B1), M6 = (A2, B2), and M7 = (A2, B2, with $\alpha < 1$ suggesting learning transmission rate as a product of intra-region mobility and a learned infection rate from recent data).

2.3 POI VISITS MOBILITY

We extended the SIkJα model to incorporate an arbitrary number of mobility features in the form of daily time series. Three hyperparameters need to be specified for each mobility feature: (i) **k:** number of time windows, (ii) **jp:** number of days in each time window, and (iii) **lag:** number of days to delay before first time window. For a particular mobility term, these three hyperparameters specify the time period the model should use as the mobility for a certain day d.

The modified model is given by

$$\Delta I_d = \sum_{i=0}^{M-1} \left(\sum_{j=0}^{k_i-1} \beta_{ij} \cdot \text{sus}\left(start_{d,i,j}\right) \cdot \Delta I\left(start_{d,i,j}, end_{d,i,j}\right) \cdot \text{mobavg}_i(start_{d,i,j}, end_{d,i,j}) \right) \tag{1}$$

Here, M is the total number of mobility features. sus(d) is the susceptible population on date d, $\Delta I(s,t)$ is the increase in cumulative infection from date s to date t, and mobavg$_i(s,t)$ is the average mobility score for the i^{th} mobility feature from date s to date t. "start$_{d,i,j}$" and "end$_{d,i,j}$" are the starting date and ending date of the j^{th} time window for the i^{th} mobility feature, calculated as $d - lag_i - (j+1) \cdot jp_i$ and $d - lag_i - j \cdot jp_i$ respectively, where k_i, j_i and lag_i are hyperparameters corresponding to the i^{th} mobility feature.

To generate mobility features from POI hourly visitor count data, we use the following method to estimate the level of visitor co-occupation. Let $\mathbb{S}$ be the set of POIs that are considered in a mobility feature, and let $\mathbf{c}_p$ be the hourly visitor count vector of a POI p. Then the mobility feature is generated as

$$\mathbf{m} = \sum_{p \in \mathbb{S}} {\mathbf{c}_p}^T (\mathbf{c}_p - \mathbf{1}) \tag{2}$$

We constructed four models using different mobility feature sets. *(i) Generic Model:* This model predicts future infections without any mobility feature. Its forecast is solely based on susceptible

population and active cases. *(ii) Google Mobility Model:* This model adds additional mobility features from Google Mobility Report. We chose dimensions "workplaces" and "retail" out of the six categories because this combination seems to yield the best prediction accuracy. *(iii) Agglomerative Hierarchical Clustering Model:* To construct this model, we first conducted agglomerative hierarchical clustering Pedregosa et al. (2011) on POI-level visitor pattern data. The clustering is done based on a place's hourly visitor count distribution vector over a week in February 2020. The number of clusters is set to 3. *(iv) Manual Clustering Model:* To construct this model, we manually clustered the POI-level visitor pattern data. The clustering is done by applying keyword filters and manually examine the nature of POIs. We selected features "recreation", "shopping", and "airport" out of all the clusters. We use linear regression to fit the above models.

3 EXPERIMENTS

3.1 DATA SOURCES

Inter-region Mobility To study the impact of inter-region mobility in Los Angeles, we considered census block groups as individual regions. The SafeGraph Mobility Data [2] are processed as mobility matrices elements representing the number of travelers between neighborhoods in Los Angeles. The neighborhoods are defined using the 2010 Census Block Group Data[3]. Confirmed COVID-19 cases for these neighborhoods are obtained from LA Times[4] to be used in our model.

POI Visits Mobility Experiments in this section rely on three batches of data: LA county's cumulative infection time series, POI-level visitor patterns, and macro-level mobility patterns. *Cumulative Infection:* The time series is obtained from LA County Public Health[5] (under the website's Table: Tests By Date). The time series of infection data is smoothed using a Kolmogorov–Zurbenko filter. *POI-level Visitor Patterns:* The data is obtained from "Weekly Places Patterns v2" published by SafeGraph Inc. After the LA County data is extracted from the global data-set, the column "visits_by_each_hour" is constructed as time series for each point of Interest. There are 45481 POI locations in LA County that have complete hourly visitor count data during this date range. *Macro-level Mobility Patterns:* The data is obtained from Google Mobility Report. After extracting LA county's data from the data-set, the mobility time series vector $\mathbf{m}$ in each macro-level place category is then calculated as $\mathbf{m} = (\mathbf{s} + 100)^{\gamma}$, where $\mathbf{s}$ is the mobility score time series evaluated by Google and γ is a hyperparameter that represents the intensity of human interactions.

3.2 RESULTS

We evaluate all the models for a given setting (Inter-region or POI visits) on the task of 1-6 week ahead incident case forecasts. Each forecast is performed from data up to a Sunday, and a week is defined from Sunday-Saturday in accordance with the ongoing CDC forecasting efforts (Reich-Lab, 2020). We use absolute error as the metric defined by the absolute difference between predicted and the observed data. Note that the evaluation for inter-region mobility is based on mean absolute error calculated separately over case data of the neighborhoods. On the

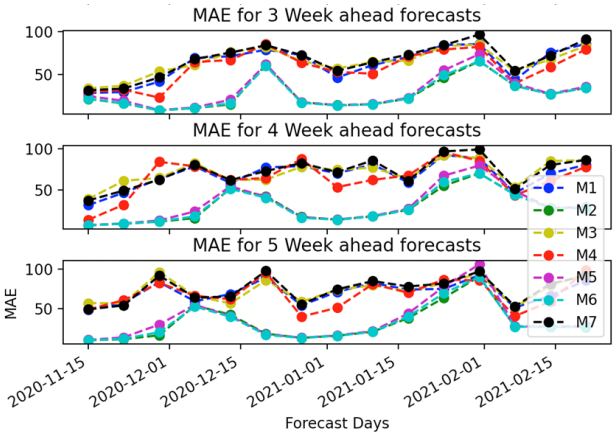


Figure 1: MAE over all Census Block Groups

[2] https://www.safegraph.com/

[3] https://geohub.lacity.org/

[4] https://github.com/datadesk/california-coronavirus-data

[5] http://dashboard.publichealth.lacounty.gov/covid19_surveillance_dashboard/

other hand, evaluation on POI visits mobility is based on the absolute error of cases data of Los Angeles as a whole.

3.2.1 INTER-REGION MOBILITY

Figure 1 shows the MAE obtained over all the neighborhoods in Los Angeles. We observed that M5 and M6 consistently perform the best suggesting that mobility plays an important role in the trajectory of the epidemic. On the other hand, including mobility explicitly in the transmission term as done in M3, M4, and M7 does not produce significant improvements over no-mobility. This indicates that mobility is not the only factor dictating transmission, and explicitly scaling transmission with mobility may not reflect the true process. Normalizing the mobility matrix results in a slight reduction in the MAE (M6 vs M5), which helps resolve the distortion of ranges in the mobility data we collected from SafeGraph. The code for replicating these results is available on Github[6].

3.2.2 POI VISITS MOBILITY

Figure 2 shows the results of all the models based on POI visits in Los Angeles. We observe that inclusion of mobility features consistently produce lower errors. Further, models that utilize POI-level features are better than Google Mobility model that uses macro-level feature. We believe this is due to the fact that the POI-level feature additionally contain hourly visitor counts that help us estimate local interactions given by Equation 2. Finally, both manual clustering and automatic agglomerative clustering result in good forecasts. The good performance of manual clustering is expected as different location types will have different levels of impact on the epidemic. We believe, the good performance of automatic clustering suggests that the underlying pattern need not be explicitly identified by location types. The code for replicating these results is available on Github[7].

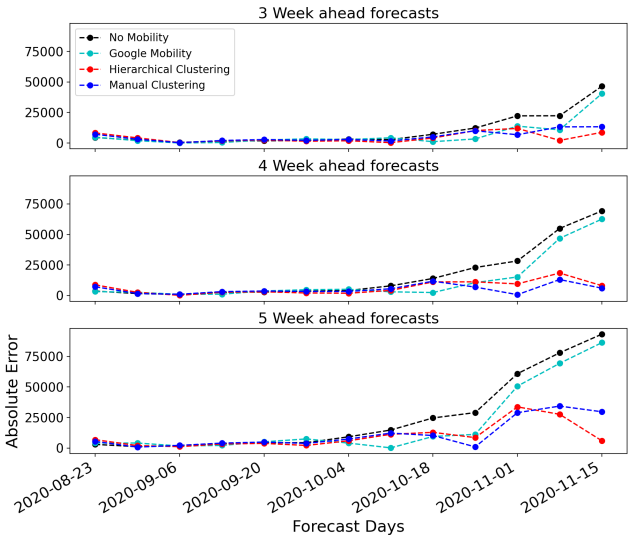


Figure 2: Absolute error of all the POI mobility models.

4 DISCUSSIONS

Other modeling strategies that incorporate mobility exist and can be used as presented in this paper. Our main goal was to demonstrate that the foresight of mobility consistently improves forecasts of cases, and thus, it is likely to be a crucial and measurable factor in the spread of the epidemic. A detailed comparison with other mobility-based models is left for future work. Further, the learned parameters can be reused to generate projections for mobility scenarios in the future that may assist in designing policies. We have not tested our approach on longer-term forecasts (beyond 5 weeks) to avoid the effect of true prevalence/immunity/under-reporting. This factor, to the best of our knowledge, has no reliable estimation algorithm (Srivastava & Prasanna, 2020a), but it has a significant impact on long-term forecasts.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation Award No. 2027007 (RAPID).

[6] https://github.com/HaiwenC/dslab
[7] https://github.com/JonnodT/mobility_model_COVID19

REFERENCES

CHIKV Challenge Announces Winners, Progress toward Forecasting the Spread of Infectious Diseases, 2015. URL https://www.darpa.mil/news-events/2015-05-27.

Abdulkadir Atalan. Is the lockdown important to prevent the covid-19 pandemic? effects on psychology, environment and economy-perspective. *Annals of medicine and surgery*, 56:38–42, 2020.

Johannes Bracher, Jannik Deuschel, Tilmann Gneiting, Konstantin Görgen, and Melanie Schienle. *Assembling forecasts of COVID19 cases and deaths in Germany and Poland*. URL https://jobrac.shinyapps.io/app_forecasts_de/.

Joseph R. Friedman, Patrick Y. Liu Liu, and Samir Akre. How have the models performed?, 2021. URL https://covidcompare.io/model_performance.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Reich-Lab. The COVID-19 Forecast Hub, 2020. URL https://covid19forecasthub.org/.

Ajitesh Srivastava. COVID-19 Forecast Evaluations, 2021. URL https://scc-usc.github.io/ReCOVER-COVID-19/#/leaderboard.

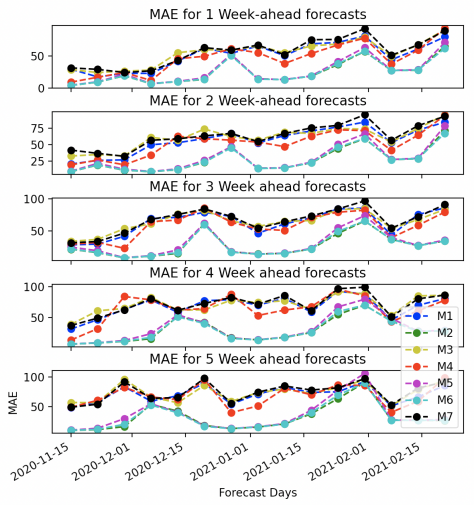
Ajitesh Srivastava and Viktor K Prasanna. Data-driven Identification of Number of Unreported Cases for COVID-19: Bounds and Limitations. *ACM SIGKDD international conference on Knowledge discovery and data mining (Health Day), arXiv preprint arXiv:2006.02127*, 2020a.

Ajitesh Srivastava and Viktor K Prasanna. Learning to Forecast and Forecasting to Learn from the COVID-19 Pandemic. *arXiv preprint arXiv:2004.11372*, 2020b.

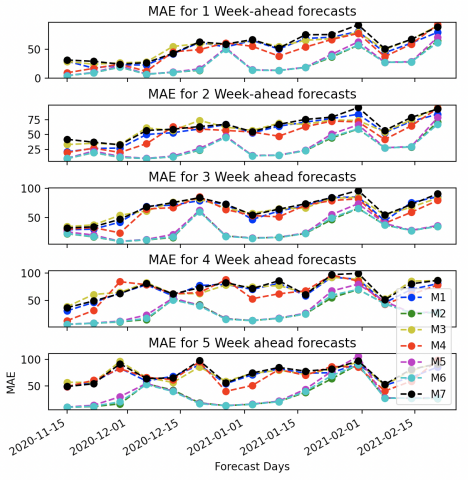
Ajitesh Srivastava, Tianjian Xu, and Viktor K Prasanna. Fast and Accurate Forecasting of COVID-19 Deaths Using the SIkJ alpha Model. *arXiv preprint arXiv:2007.05180*, 2020.

A APPENDIX

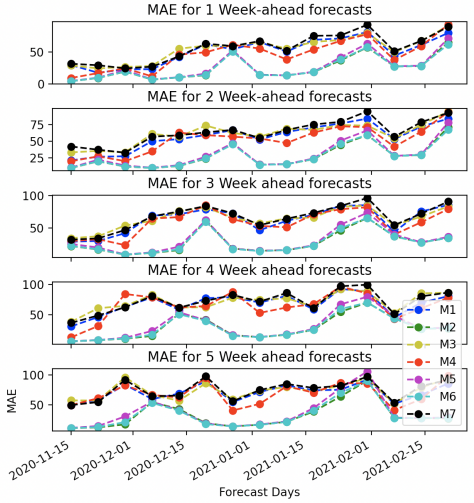
A.1 ADDITIONAL PLOTS



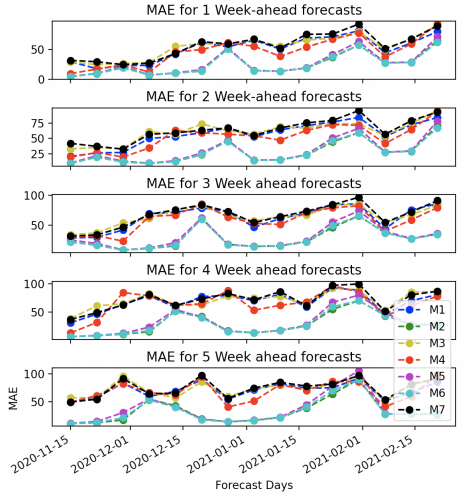
(a) All Neighborhoods



(b) Population size ≥ 5000

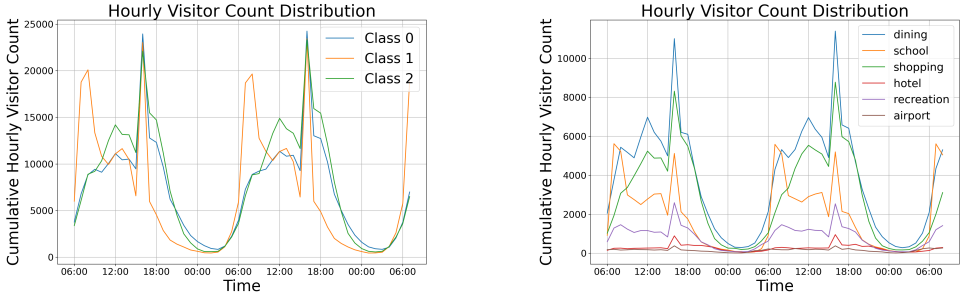


(c) Population size ≥ 10000



(d) Top third population size

Figure 3: MAE for Neighborhoods in Los Angeles



(a) Agglomerative Clusters

(b) Manual Clusters

Figure 4: Snapshots of hourly visitor count distributions aggregated over clusters from Feb. 11 to Feb. 13, 2020