Analysis of Performance Improvements and Bias Associated with the Use of Human Mobility Data in COVID-19 Case Prediction Models

SAAD MOHAMMAD ABRAR, NAMAN AWASTHI, DANIEL SMOLYAK, and
VANESSA FRIAS-MARTINEZ, University of Maryland, College Park, USA

The COVID-19 pandemic has mainstreamed human mobility data into the public domain, with research focused on understanding the impact of mobility reduction policies as well as on regional COVID-19 case prediction models. Nevertheless, current research on COVID-19 case prediction tends to focus on performance improvements, masking relevant insights about when mobility data does not help, and more importantly, why, so that it can adequately inform local decision making. In this article, we carry out a systematic analysis to reveal the conditions under which human mobility data provides (or not) an enhancement over individual regional COVID-19 case prediction models that do not use mobility as a source of information. Our analysis—focused on U.S. county-based COVID-19 case prediction models—shows that (1) at most, 60% of counties improve their performance after adding mobility data; (2) the performance improvements are modest, with median correlation improvements of approximately 0.13; (3) improvements were lower for counties with higher Black, Hispanic, and other non-White populations as well as low-income and rural populations, pointing to potential bias in the mobility data negatively impacting predictive performance; and (4) different mobility datasets, predictive models, and training approaches bring about diverse performance improvements.

CCS Concepts: • **Human-centered computing** → **Mobile devices**; **Ubiquitous and mobile devices**; • **Information systems** → **Location based services;**

Additional Key Words and Phrases: COVID-19 case prediction, mobility data, sampling bias, interpretable models

ACM Reference format:
Saad Mohammad Abrar, Naman Awasthi, Daniel Smolyak, and Vanessa Frias-Martinez. 2023. Analysis of Performance Improvements and Bias Associated with the Use of Human Mobility Data in COVID-19 Case Prediction Models. *ACM J. Comput. Sustain. Soc.* 1, 2, Article 16 (December 2023), 36 pages.
https://doi.org/10.1145/3616380

1 INTRODUCTION

The COVID-19 pandemic has mainstreamed human mobility data into the public domain and beyond academic networks. During the early stages of the pandemic, the importance of limiting mobility to contain the epidemic became evident, with cities, states, and countries taking various non-pharmacological interventions focused on mobility, such as national lockdowns or

S. Mohammad Abrar, N. Awasthi, and D. Smolyak contributed equally to this research.

Authors' address: S. Mohammad Abrar, N. Awasthi, D. Smolyak, and V. Frias-Martinez, University of Maryland, College Park, College Park, 8125 Paint Branch Drive, College Park MD 20740; emails: {sabrar, nawasthi, dsmolyak, vfrias}@umd.edu.

2834-5533/2023/12-ART16 $15.00
https://doi.org/10.1145/3616380

work-from-home approaches [10, 36]. To evaluate the effect of these interventions, public health experts, the CDC, city departments, and journalists explored the use of mobility data that, at the time, was made open and freely available. Companies like Apple, Google, SafeGraph, and Descartes shared different types of aggregated mobility datasets to characterize behaviors such as the volume of visits to specific places (e.g., schools, workplaces, or restaurants), the volume of trips between regions (e.g., trips between two counties), or the volume of trips by type of transportation (e.g., driving vs. public transit).

Beyond understanding the impact of mobility reduction policies, the increased access to mobility data sources has also supported research on regional COVID-19 case prediction models, with the assumption that how people moved within a region in the past could potentially provide additional information about how people become infected in the future. COVID-19 case prediction models focus on providing regional-level estimates for future cases in both the short and long term via lookahead analysis performance—that is, measuring region-level prediction performance for various temporal windows such as daily, weekly, or monthly [27]. For example, researchers have shown that SafeGraph data can help predict weekly COVID-19 cases at the county level in the United States, providing higher accuracy when compared to non-mobility baselines [35]. There exist a wide variety of models to predict regional COVID-19 cases, including epidemiological [6, 25, 52], machine learning [17, 33, 47], and statistical models [17, 27]. In this article, we focus on statistical models (linear regression and ARIMA) because we are interested in the deployment of models that are interpretable by decision makers rather than implementing black box predictive approaches that are harder to explain [42].

Nevertheless, there are several gaps in the current state of the art in regional COVID-19 case prediction using mobility data. First, performance results—measured as RMSE or correlation between actual and predicted regional COVID-19 cases—are reported as averages across regions, masking individual region-level performance, which is critical to inform local interventions and policies [27]. For example, past research has shown that mobility data enhances COVID-19 case predictions, on average, across counties in the United States; however, that average might be masking counties for which it did not work [9, 33]. Second, performance results are often not compared against non-mobility baselines, making it hard to measure the effectiveness of adding mobility data to the prediction model [9, 33, 47]. Third, prior work has shown that mobility data might suffer from sampling bias whereby certain demographic groups (e.g., Black, elder, and low-income individuals) can be under-represented in the data due to lower smartphone and cell phone ownership rates [7, 43]. Nevertheless, prior work focused on building COVID-19 case prediction models tends to ignore the bias present in the mobility data, which in turn might affect the performance of regional COVID-19 case prediction models depending on the population of that region [27, 33, 34]. Fourth, current approaches tend to provide narrow evaluations, focused on a few models, or on one or a few mobility datasets, with little research broadly looking into the impact of different prediction models, mobility datasets, and training approaches that use more or less data, on model performance. Given (1) the high cost of acquiring human mobility data for COVID-19 prediction purposes, now that it is no longer freely accessible, and (2) that COVID-19 case predictions are going to be used to assess non-pharmacological interventions such as mobility reduction, or vaccine distribution at the local level, we posit that it is critical to understand the conditions under which mobility data helps (or not) at the individual regional level so that it can adequately inform local decision making.

In this article, we aim to analyze the conditions under which human mobility data provides an enhancement over individual regional COVID-19 case prediction models that do not use mobility as a source of information. Our main objective is to inform regional decision makers about the potential of region-level COVID-19 case prediction models that use mobility data, which we posit should be well understood given the high cost of human mobility data. The main contributions of this work are the following:

- Focusing on U.S. counties, we evaluate the number of counties that benefit from adding mobility data and quantify the improvements. Our analyses show that, at most, 60% of counties improve their performance over non-mobility baselines, and that those improvements are modest, happening mostly for longer-term predictions. Looking into the counties that benefit from adding mobility data, 50% of those counties show modest correlation improvements of at most 0.1 and 25% show correlation improvements of at most 0.3.
- We present and discuss an approach to assess whether mobility data bias—characterized by demographic and socio-economic characteristics of each county—might explain the differences in the performance of COVID-19 prediction models across counties. We show that correlation improvements were lower for counties with higher Black, Hispanic, and other non-White populations as well as low-income and rural populations, pointing to potential bias in the mobility data negatively impacting predictive performance.
- We analyze whether the differences in the performance of mobility-based models over non-mobility baselines vary depending on the mobility datasets, the predictive model, or the training approach. Our results reveal that the improvements brought about by mobility data are similar across mobility datasets, albeit with slightly better values for Apple and SafeGraph; linear regressions are associated with larger improvements; and the training approach might also affect the scale of the improvements.

2 RELATED WORK

2.1 Human Mobility Data and COVID-19 Case Predictions

Human mobility data has been used in the past to model and characterize human behaviors in the built environment [12, 21, 41, 46, 51], to support decision making for socio-economic development [11, 13, 14, 16, 22], for public safety [49, 50], as well as during epidemics and disasters [3, 19, 23, 24, 28, 48]. During the COVID-19 pandemic, human mobility has also played a central role in driving decision making, for example, with social distancing policies significantly reducing the spread of the virus [2].

Related work has shown that COVID-19 case prediction models can be enhanced using human mobility data when compared to non-mobility baselines [5, 27, 35]. For example, Ilin et al. [27] analyzed the use of mobility data to forecast COVID-19 cases using interpretable statistical models like regressions [27]. Working at various spatial scales (from county to state to country), the authors revealed that adding mobility data significantly helps decrease the mean percentage prediction error, and that the improvements were higher for longer forecasting lengths. Nevertheless, a significant amount of papers focused on COVID-19 case prediction using mobility data fail to compare model performance against non-mobility baselines [9, 33]. More importantly, several papers have revealed settings in which mobility data did not help. For example, Curtis et al. [8] and Venter et al. [45] found only a small correlation between COVID-19 cases and mobility in parks and natural areas (blue-green spaces), and Mehrab et al. [34] showed that the performance of mobility-based prediction differed considerably across 50 U.S. counties that correspond to land-grant universities.

Human mobility data is generally collected from smartphones and cell phones; however, due to the differences in access to that technology, not all individuals are equally represented in mobility datasets. In fact, prior work has shown that, for example, Black and elder individuals were under-represented in SafeGraph's dataset for the state of North Carolina [7]; wealthier individuals tend to be over-represented in cell phone data from several countries, including Sierra Leone or Iraq [43];

and the relationship between COVID-19 cases and mobility is stronger in urban areas than in rural areas [32]. Given this evidence, we posit that the sampling bias found in mobility data might also affect the performance of regional COVID-19 case prediction models.

In this article, we aim to provide a much needed systematic analysis to evaluate the conditions under which mobility data can enhance county COVID-19 case prediction models, and to quantify by how much when compared to non-mobility baselines. Thus, we will extensively evaluate performance across types of prediction models, temporal prediction windows (lookaheads), mobility datasets, training-testing approaches, and county-level demographic and socio-economic characteristics associated with potential mobility data bias that could affect performance.

2.2 COVID-19 Prediction Models

A wide variety of models exist to predict regional COVID-19 cases, including epidemiological, machine learning, and statistical models. Epidemiological models, such as SEIR and SIR models, have been used to predict infection rates, and in certain cases, related work has shown that the models can be improved with human mobility data characterizing how people (agents) travel and might infect others [6, 25, 52]. However, SEIR/SIR models have a number of pitfalls, such as the large number of parameters that need to be adjusted [54], or the complexities of adding mobility responses of the population as a function of time and space [40]. For example, Roda et al. [40] found simpler SIR models to be more effective in predicting COVID-19 cases than the more complex SEIR models.

A number of machine learning techniques have also been widely applied for regional COVID-19 case prediction with mobility data, including tree-based and K-nearest neighbors models [33]. Furthermore, several works [17, 47] make use of a range of deep learning architectures, including sequential models like long short-term memory (LSTM) [35], gated recurrent units (GRUs), and recurrent neural networks (RNNs) [20], as well as spatio-temporal models like graph neural networks (GNNs) [31]. Some researchers have also incorporated static and dynamic mobility flows—characterizing average and daily mobility patterns between regions—as well as friendship networks to understand the spatio-temporal dependencies between regions that might affect infection rates and which could inform the prediction of regional COVID-19 cases [15, 47].

Statistical models have also been popular for predicting COVID-19 cases because they are transparent and simple to interpret, which is highly important for decision makers; these models can also easily incorporate mobility data along with other features. Statistical models such as autoregressive time series and linear regression have been used for COVID-19 case prediction showing that—early on during the pandemic—mobility data improved non-mobility baselines across countries, states, and other administrative levels [17, 27]. In this article, we focus on statistical models (linear regression and ARIMAX) because we are interested in the deployment of models that are interpretable by decision makers rather than implementing black box predictive approaches that are harder to explain [42].

3 DATASETS

To assess the effectiveness of mobility data on county-level COVID-19 case prediction models, we evaluate model performance across nine mobility datasets from four different companies: SafeGraph, Google, Descartes, and Apple. We describe each of the COVID-19 and mobility datasets in detail. All datasets used in this work were freely available during the onset of the pandemic. We focus on the period from March 18, 2020 to November 30, 2020 (258 days): the former marks the start of the consistent case data availability in the United States and the latter the date when vaccines were introduced (the Pfizer-BioNTech COVID-19 vaccine was made available on December 11, 2020). We focus on the pre-vaccine period to prevent immunity levels from acting as a confounder, since the relationship between mobility and infections post-vaccines is less clear in the literature [18].

COVID-19. We use the COVID-19 case data compiled by the New York Times[1] at the county level. To account for peaks in daily COVID-19 case counts due to delayed reporting, we use the 7-day daily rolling average of COVID-19 cases (computed as the average of its current value and the prior 6 days) instead of raw counts. We acknowledge that especially during the early stages of the pandemic, case numbers might not be reflective of the actual spread of COVID-19, in large part due to the lack of testing resources [39].

SafeGraph. Curated by tracking the movements of millions of *anonymized* users via mobile app SDKs, SafeGraph[2] open sourced the mobility patterns of app users at the onset of the pandemic. Based on the data available, we use two types of features from SafeGraph datasets: daily O-D flows [30] and daily visits to **Points of Interest (POIs)**. O-D (county-to-county) flows represent daily volumes of trips between pairs of counties across the United States, whereas visits to POIs represent the daily volume of visits to grocery stores, restaurants, religious organizations, and schools in a given county. For O-D flows, we retrieve from SafeGraph both Inflows (i.e., incoming flows to county D from county O) and outflows (i.e., outgoing flows from county O to county D). All mobility features are measured as changes in volumes with respect to a baseline of *normal* behavior computed by SafeGraph using mobility data from February 17, 2020 to March 7, 2020.

Google. Google[3] collects mobility data from users who have the location data collection option selected. During the pandemic, Google provided daily county-level mobility scores across different POI categories including parks, residential areas, and transit stations. Mobility scores are calculated as the ratio between the volume of visits on a given day during the pandemic and the volume of visits during a pre-pandemic baseline, with the baseline computed by Google as the median value for each day of week in the 5-week period from January 3, 2020 to February 6, 2020. Among all POI available, we selected workplace, the category with the greatest number of counties with daily data availability in our chosen time period.

Descartes Labs. This mobility data from Descartes Labs[4] is calculated using geolocation data from mobile devices and captures the median of the maximum distances traveled by individuals in each county each day. As with Google, this median is converted by Descartes to a ratio of pandemic mobility to baseline pre-pandemic mobility.

Apple. The mobility data of Apple[5] is collected from Apple Maps, and divides its categories by transportation method: driving, walking, and transit. We selected the *driving* category to measure the volume of individuals driving on a daily basis at the county level, because, similar to Google, this category had more consistent daily data availability than other categories.

Across all datasets, we only consider counties that have COVID-19 case data and mobility data available daily throughout the time period of study. Table 1 shows the number of counties that fit this criteria for each dataset; notably, Google has by far the fewest counties with 990 out of the total 3,143 U.S. counties being represented.

4 METHODOLOGY

In this article, we aim to provide a much needed systematic analysis to evaluate the conditions under which mobility data can enhance county COVID-19 case prediction models, and to quantify by how much when compared to non-mobility baselines. Our main objective is to inform regional

[1] https://github.com/nytimes/COVID-19-data

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

[3] https://www.google.com/covid19/mobility/

[4] https://github.com/descarteslabs/DL-COVID-19

[5] https://covid19.apple.com/mobility

Table 1. Number of Counties Considered for Each Mobility Dataset

Mobility Dataset	County Count
Apple	2,064
Descartes	2,551
Google	990
SafeGraph Inflows	3,116
SafeGraph Outflows	3,116
SafeGraph POI (Grocery Stores)	3,065
SafeGraph POI (Religious Organizations)	3,076
SafeGraph POI (Restaurants)	3,091
SafeGraph POI (Schools)	3,084

decision makers about the potential and pitfalls of COVID-19 case prediction models that use mobility data, given the high cost of acquiring such data. Next, we describe the prediction models we evaluate and their parameter adjustment, the training-testing approaches, and the overall evaluation approach.

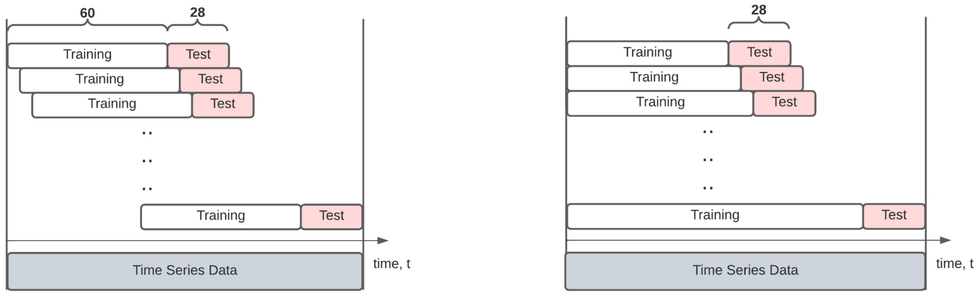
4.1 Prediction Models

As described in Section 2, we consider two types of predictive models commonly used by decision makers due to their interpretability: linear regressions and time series. In contrast to more complex epidemiological models that are hard to tune due to their parametric nature, and to deep learning models with black box architectures, linear models and time series are easier to interpret providing decision makers with the ability to clearly explain their policies [40, 42, 54].

To evaluate the effectiveness of mobility data in the prediction of county-level COVID-19 cases, we analyze the performance of linear regressions and time series using (1) only past county-level COVID-19 case data as an independent variable to predict future COVID-19 cases (these are the non-mobility baselines) and (2) both past county-level COVID-19 case data as well as county-level mobility data as independent variables, with the assumption that how people moved in the past could potentially provide additional information about how people get infected in the future. Both non-mobility baselines and mobility-based models will be evaluated across five different temporal prediction windows (a.k.a. lookaheads): 1-day, 7-days, 14-days, 21-days, and 28-days. Next, we provide further details for each predictive model, and in Section 4.4, we discuss details about the parameter tuning.

4.1.1 Linear Regression. We train one linear regression model per county and lookahead. For the non-mobility baselines, the number of county-level COVID-19 cases for a given lookahead is predicted using the COVID-19 cases from the previous lookahead value—for example, for lookahead 1, the COVID-19 cases in day x are used to predict cases for day x+1, whereas for lookahead 14, the COVID-19 cases in day x are used to predict cases in day x+14. For mobility-based models, the number of county-level COVID-19 cases for a given lookahead is predicted using (1) the COVID-19 cases from the previous lookahead value as well as (2) the 10-day lagged mobility features—that is, the mobility features from day x are used to predict COVID-19 cases for day x+10, since the infection gap (the incubation period from exposed to being able to spread the virus) has been associated with that lag by many reports in the literature [17, 27, 53]. Lagged mobility data is only included in the mobility-based models that will be compared against the non-mobility baselines.

4.1.2 ARIMAX. We train one ARIMA time series forecasting model per county and use it to predict county-level COVID-19 cases for all five lookaheads. Similarly to linear regression, we



(a) Short training window (STW) for ARIMAX and linear regression

(b) Long training window (LTW) for ARIMAX

Fig. 1. Training-testing approaches for LTW and STW. The LTW approach for linear regression is not represented because it is the standard 70/30 approach.

train non-mobility baselines exclusively with COVID-19 case time series, whereas mobility-based models incorporate 10-day lagged mobility data series. Mobility data is incorporated using ARIMAX models that allow to add exogenous covariates into time series forecasting models. ARIMAX forecasting models have three components that function as parameters: (1) p (autoregressive order, AR) indicates the number of lag observations that the dependent variable regresses with in the model, (2) d (integrated order, I) represents the number of times the time series needs to be differenced to achieve stationarity, and (3) q (moving average, MA) represents the size of the moving average window that models the relationship between the error terms of the moving average model and the lagged observations. We discuss the tuning of these parameters in Section 4.4.

4.2 Training-Testing Approaches

To explore the conditions under which mobility helps in improving predictive performance, we aim to test two different training-testing approaches that have been used in the literature: the **Long Training Window (LTW)** approach and the **Short Training Window (STW)**. The STW approach focuses on the use of small training datasets to test prediction accuracy for the next 28 days (lookaheads 1, 7, 14, 21, and 28), whereas the LTW approach uses much larger training datasets to predict values for the five lookaheads. Given the high cost of acquiring human mobility data—in fact, none of the companies described earlier offer their data for free anymore—the objective of evaluating these two training approaches is to understand the impact of cost on the performance of COVID-19 prediction models (i.e., buying more (LTW) versus less mobility data (STW)). Next, we explain the implementation of LTW and STW for linear regression and ARIMAX models (also depicted in Figure 1).

4.2.1 Linear Regression. The LTW approach creates one regression model per lookahead per county, trained on the first 70% of days in our dataset and tested on the remaining 30%. This splitting approach is consistent across lookaheads, easing interpretability, but also meaning that the split is not exactly 70/30 for each lookahead—that is, at higher lookaheads, the training set becomes smaller while the testing set remains the same size. The STW approach uses a 60-day sliding window for the training set and a sliding 28-day window for the testing set (smaller training sizes produced extremely low performing results and were not considered). Unlike the LTW approach, the testing set consists of one date for each lookahead (1 day after the end of the training set, 7 days after, etc.). Results are reported by averaging performance metrics across all testing datasets.

4.2.2 ARIMAX. The ARIMAX LTW approach uses an expanding window protocol [26, 38] that first trains ARIMAX with a data window that expands over 70% of the entire time series and is tested for the next 28 days to evaluate accuracy per lookahead. After that, the training data window is expanded by 1 day at a time, without dropping older data points, and tested for subsequent 28-day windows to assess accuracy at each lookahead. However, the STW adopts a sliding window approach [38] whereby the training window length remains fixed at each train-test step with a length of 60 days. Each training window is computed by shifting by one with respect to the prior window, effectively discarding older data points. Results are reported by averaging performance metrics across all testing datasets.

4.3 Model Performance

We measure individual county prediction performance as the correlation between the predicted volume of COVID-19 cases for that county and its actual case numbers retrieved from the COVID-19 official report dataset. To assess the effectiveness of enhancing COVID-19 case prediction models with mobility data, we will report the correlation improvement (ci) of the mobility-based models over non-mobility baselines. Given that the only difference between mobility-based models and non-mobility baselines is the mobility data, we posit that ci allows us to measure the effectiveness of using mobility data in COVID-19 predictive settings. Specifically, we compute ci as $ci = pcorr\_Mobility - pcorr\_Baseline$ with $pcorr\_Mobility$ representing model performance (correlation) when mobility data is used and $pcorr\_Baseline$ measuring the correlation when no mobility data is used in the COVID-19 case prediction model. Given that correlation values go between −1 and +1, ci will be within the $(-2, 2)$ range.

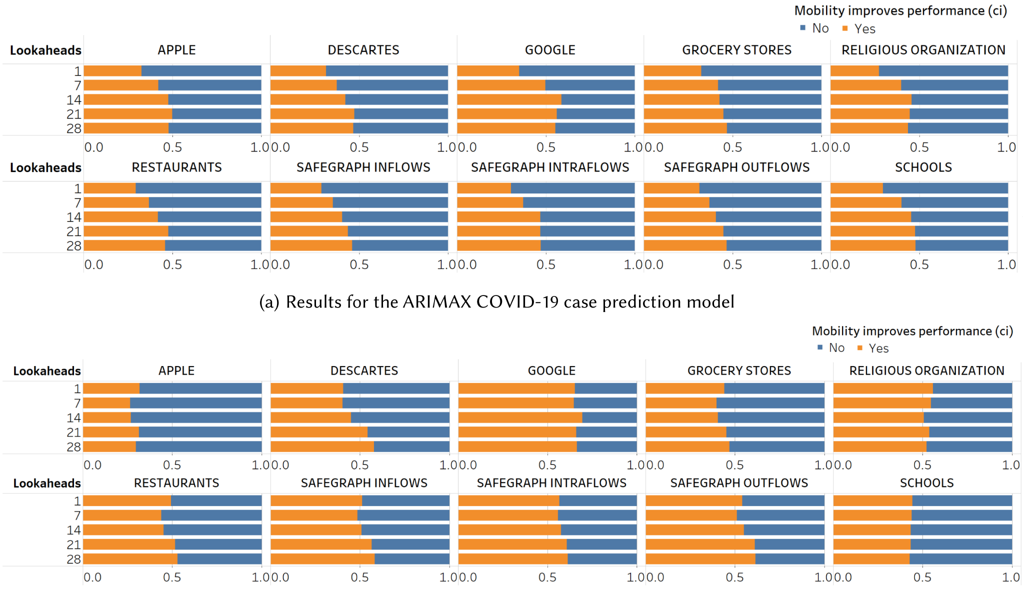
4.4 Adjusting Predictive Models

To identify the best linear regression implementation, we computed the average performance of Ridge, Lasso, ElasticNet, and OLS trained with and without mobility data for each lookahead and training approach (LTW and STW), and across all mobility datasets. We then selected the implementation with the majority of best performance values: Ridge for LTW and ElasticNet for STW. Appendix Table 7 shows the detailed numbers. However, the optimal p, d, and q values for the LTW and STW ARIMAX models were chosen based on a grid search and the minimum Akaike information criterion value. The p, d, and q values were selected across lookaheads, and the same values were used for mobility-based models and non-mobility baseline models for comparison analyses. Appendix Table 14 shows a summary of the p, d, and q values identified across all counties.

4.5 Evaluation Approach

To analyze the conditions under which mobility data provides (or not) an enhancement over county-level COVID-19 case prediction models that do not use mobility as a source of information, we propose to carry out the following analyses (results are presented in Section 5).

In Section 5.1, we analyze the number of counties for which mobility data improved the individual prediction performance of non-mobility baselines, and we quantify the improvements, with a focus on models trained with the LTW approach. In Section 5.2, we delve into the demographic and socio-economic characteristics of these counties to assess if mobility data bias—whereby certain demographic groups are over- or under-represented in the data—might explain the differences in the performance of COVID-19 prediction models across counties. We also analyze whether the differences in the performance of mobility-based models over non-mobility baselines across counties vary depending on (1) the training approach (STW versus LTW (in Section 5.3)), (2) the mobility datasets used (10 different data sources (in Section 5.4)), and (3) the predictive



(a) Results for the ARIMAX COVID-19 case prediction model

(b) Results for the Ridge Regression COVID-19 case prediction model

Fig. 2. Percentage of counties for which adding mobility data to the COVID-19 case prediction model improves the prediction performance ($ci \geq 0$). Results are for the LTW approach.

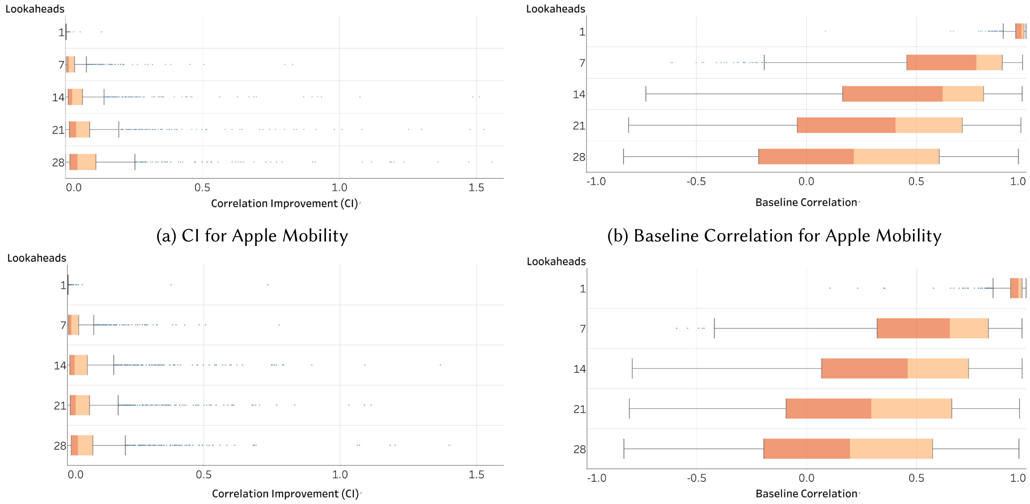
model (linear regression vs. ARIMAX (in Section 5.5)). As stated earlier, our main objective is to bring light into the use of mobility data in county-level COVID-19 case prediction models, to understand when it works (or not), and why, so as to inform decision makers assessing effectiveness-cost tradeoffs given that mobility data is not freely accessible.

5 RESULTS AND ANALYSIS

5.1 Does Mobility Data Help, and by How Much?

Figure 2(a) and (b) show the percentage of counties for which the prediction performance of mobility-based models improves over its corresponding non-mobility baselines for ARIMAX and Ridge regression models, respectively, using the LTW approach (a comparison with STW is discussed in Section 5.3). In other words, these plots show the percentage of counties for the correlation improvement $ci > 0$. These plots show several important insights: (1) incorporating mobility data into county-based COVID-19 case prediction models helps in, at most, 60% of the counties analyzed, leaving at least 40% or more of the other counties with prediction performances lower than their non-mobility counterparts (i.e., mobility data is frequently hurting prediction performance), and (2) mobility data appears to help more in longer-term predictions (lookaheads 14, 21, and 28) than in shorter-term predictions—that is, in the short term, COVID-19 statistics are generally informative enough and provide predictions that are more accurate than those when mobility data is added to the model, whereas for longer lookahead predictions, adding mobility data to predictive COVID-19 case models provides additional information that frequently improves the predictive accuracy of the non-mobility baseline models. This trend appears clear in the ARIMAX model, whereas for Ridge, it is more apparent for the Descartes dataset and for SafeGraph inflow and outflow datasets.

We have shown that adding mobility data to COVID-19 case prediction models improves their performance for, at most, 60% of the counties, across datasets, models (ARIMAX and Ridge) and



(a) CI for Apple Mobility

(b) Baseline Correlation for Apple Mobility

(c) CI for SafeGraph Inflows Mobility

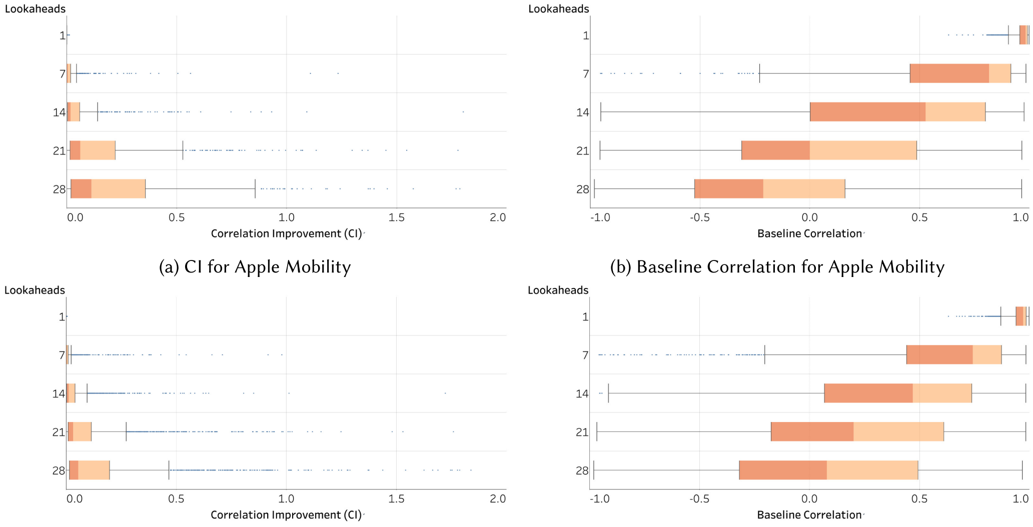
(d) Baseline Correlation for SafeGraph Inflows Mobility

Fig. 3. Correlation improvements (*ci*) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with the LTW approach using Apple mobility data (a) and SafeGraph Inflow mobility data (b).

lookaheads for the LTW approach. Next, we aim to quantify the performance improvement—that is, does mobility data produce small or large prediction improvements when compared to non-mobility baselines?

Quantifying Improvements. For this analysis, we look into two metrics per dataset and model: (1) the distribution of the county correlation coefficients *ci* for each lookahead, to explore the correlation improvements brought about by adding mobility data to the county prediction models (since we are assessing the improvement over non-mobility baselines, only counties with correlation improvements are considered), and (2) to contextualize these numeric improvements, we also look into the distribution of the correlations (between predicted and actual cases) for the non-mobility baseline models, and only for the counties that showed a correlation improvement so as to match the distributions in the first metric. By analyzing both the improvements brought about by adding mobility data, and by comparing these improvements with the actual baseline correlations, we are able to provide nuanced insights about when mobility data aids COVID-19 case prediction models. In this section, we focus on the LTW approach. Discussions about differences across training approaches, datasets, and models will be covered in Sections 5.3, 5.4, and 5.5. Figures 3 and 4 show an example of the distribution of the county correlation improvement (*ci*) and the distribution of the non-mobility baseline correlation values for Ridge and ARIMAX, respectively, across lookaheads and using the LTW approach. For clarity purposes, both figures only show distributions for Apple (a) and SafeGraph Inflow (b) datasets. Plots for the remaining datasets can be found in the appendix (see Appendix Figures 6 and 7).

The correlation improvements (*ci*) across datasets show that median *ci* values are between 0.0 and 0.1—that is, for counties where adding mobility data improves the prediction accuracy, it does so by a maximum of 0.1 for 50% of the counties, across models (Ridge/ARIMAX), datasets, and lookaheads (see Figures 3 and 4 as a sample trend, and Appendix Figures 6 and 7 for the full spectrum trend)—with the largest correlation improvements associated with higher lookaheads. These are modest correlation improvements that might not change the strength of the non-mobility baseline correlation—for example, a baseline moderate correlation of 0.5 will still be moderate after a



(a) CI for Apple Mobility

(b) Baseline Correlation for Apple Mobility

(c) CI for SafeGraph Inflows Mobility

(d) Baseline Correlation for SafeGraph Inflows Mobility

Fig. 4. Correlation improvements (*ci*) and baseline correlation distributions (non-mobility) across lookaheads for Ridge regression with the LTW approach using Apple Mobility data (a, b) and SafeGraph Inflow Mobility data (c, d).

0.1 improvement [1]. Looking at upper quartile values (Q3), we observe slightly better correlation improvements in the 0.0 to 0.4 range, with the majority of Q3 values under 0.3, across models, datasets, and lookaheads, revealing *ci* values that could improve correlation strengths. Maximum values (Q4) are in the 0 to 0.9 range across models, datasets, and lookaheads, with the majority of maximum values under 0.5, and outliers can reach correlation improvement values of up to 1.9. These less frequent, and extreme, correlation improvements point to situations were negative baseline correlations—from the non-mobility models—are being changed to positive correlations, revealing counties where adding mobility helps reverse bad-performing models to good ones, although these large improvements happen only for a handful of counties.

The correlation improvement plots also show that for higher lookaheads, the median value and the right skewness of the correlation improvement also increases across datasets and models. This finding shows that adding mobility data to COVID-19 prediction models produces better better-performing models over non-mobility baselines for longer-term predictions, whereas short-term prediction models (next day) barely benefit from adding mobility data, with correlation improvement values close to zero. We posit that this might be due to the fact that non-mobility baselines achieve high correlations for lower lookaheads, making it very hard to improve the baselines when adding mobility data. The baseline correlations plots in Figures 3 and 4 (as well as Appendix Figures 6 and 7) show that correlations for non-mobility baselines and lookahead 1 (next-day prediction) have median values of at least 0.9, whereas for higher lookaheads (21 and 28), we observe that non-mobility baselines have much lower correlations, with median values in the range $(-0.21, 0.21)$.

In summary, these findings show that mobility data helps, at most, in 60% of the counties in the datasets analyzed, and that correlation improvements range from minimal for next-day predictions to small for higher lookaheads, with 50% of the counties showing modest correlation improvements of at most 0.1 and with 25% of the counties showing correlation improvements of at most 0.3. Larger outlier improvements of up to 1.9 in higher lookaheads are found for a handful of counties.

5.2 What Types of Counties Benefit from Mobility Data? A Mobility Data Bias Analysis

As stated in Section 1, we hypothesize that a major reason adding mobility data is not helping to improve the accuracy of COVID-19 case prediction models in many more counties might be the bias present in the mobility data—that is, counties with large percentages of racial minority, elder, or low-income populations might be under-represented in the mobility datasets, thus negatively impacting the prediction performance [7, 43]. To assess that hypothesis, we carry out regression coefficient analysis for each mobility dataset and predictive model using demographic and socio-economic county characteristics as independent variables and the positive correlation improvements across lookaheads as the dependent variable. Our analysis will reveal the statistically significant associations of demographic and socio-economic features with correlation improvements, and the *direction* of that association (positive or negative).

Based on prior work on mobility data bias, we consider the following demographic and socio-economic county variables from the 2019 Census data: *age 65+* (percentage of county residents age 65 or older), *income* (median household income for that county), *Black* (percentage of county residents that identify as Black), *Hispanic* (percentage of county residents that identify as Hispanic), *race-Other* (percentage of county population that identifies as not White, Black, or Hispanic), and *urban-rural* (National Center for Health Statistics (NCHS) Urban-Rural Classification from 1 to 6, where 1 is *Large Central Metro* and 6 is *Non-core* (extreme rural)). We also consider—as independent variables in the regression model—all possible paired interactions between the variables described by multiplying the values of each pair of variables and by standardizing them. For interaction interpretability, we also ensure that higher values for each feature match with our hypothesis for lower performance. Thus, we change the directionality of the income by negating its values to match our hypothesis that lower income might be associated with worse coverage of mobility data and thus worse performance (all other features stay the same).

Finally, we would like to clarify that the COVID-19 case data itself might also suffer from different types of bias due to inaccurate data collection processes [43]. Nevertheless, since that bias affects both mobility-based and non-mobility baseline models, and since we are looking at the correlation improvement differences between the two, we can claim that any performance differences observed can be attributed to the mobility data, which is the only difference between the two models. Next, we discuss the demographic and socio-economic variables that were found to be statistically significantly related to correlation improvements for the Ridge and ARIMAX models with the LTW training approach (results for the STW approach will be discussed in Section 5.3). A detailed presentation of the coefficients and their significance can be checked in Appendix Tables 8 and 10.

Findings. For the Ridge model, we observe that the percentage of Black and Hispanic population, rurality, and age above 65 years are significantly and negatively related to correlation improvements across most datasets—that is, counties with higher percentages of minority race and ethnicity, rural, or elder populations benefit less from the addition of mobility data, with increases in these populations related to lower performance improvements over the non-mobility baselines. We posit that this is probably due to potential sampling bias in the mobility datasets, with under-representation of race, ethnicity, old age, and rurality in mobility datasets pointing to worse overall predictive performance. Lower income, however, was significantly positively associated with higher correlation improvements when mobility was added to individual county-level COVID-19 case prediction models. In principle, this result was counter-intuitive since we were expecting that lower incomes would be associated with lower access to smart phones. Nevertheless, looking at the interaction terms between low income and age 65+ as well as low income and race/ethnicity (Black and Hispanic), we observe significant negative coefficients, pointing to the fact that counties with higher percentages of low-income Black and Hispanic groups, as well as a low-income elder population, are associated with lower correlation improvements possibly due to these groups not being well represented in the datasets.

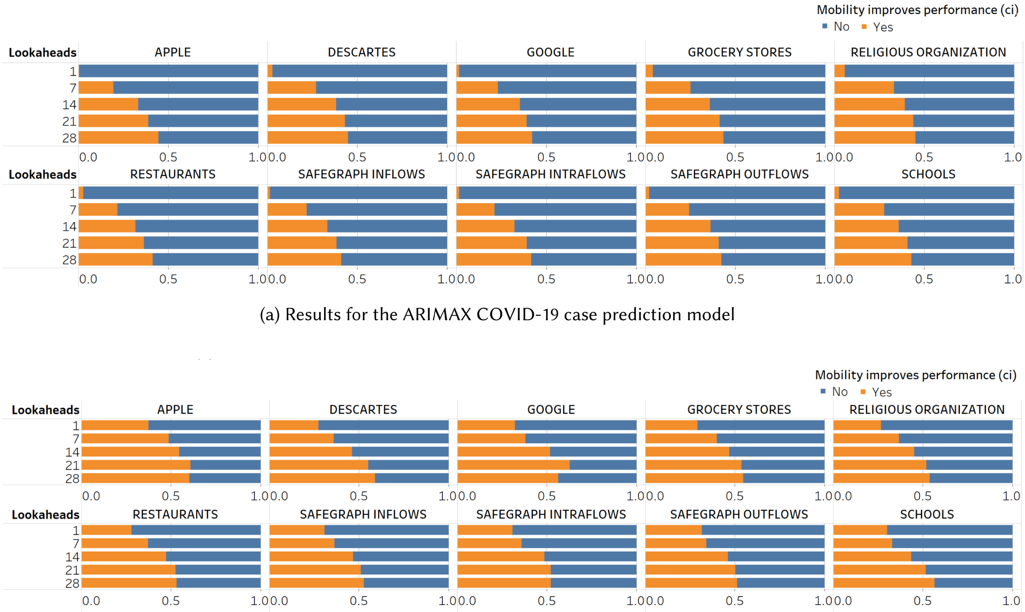
For the ARIMAX model, the trends were not pervasive across mobility datasets, with specific coefficients observed for different datasets. Age 65+ and rurality were negatively related to correlation improvements for the Apple and some SafeGraph datasets, pointing to the fact that elder and rural populations might not be fairly represented in the datasets, and confirming prior work on SafeGraph bias analysis [7]. Apple and SafeGraph also show a significant negative relation between race-*Other* (not Black, White, or Hispanic) and correlation improvements potentially pointing to the fact that other minority races/ethnicities might not be as represented in the mobility datasets as Whites, Blacks, and Hispanics. There were two instances with counter-intuitive results. First, in the Descartes dataset, race-*Other* was—unexpectedly—significantly positively associated with correlation improvements. However, when considering the interaction between race-*Other* and rurality, the significant coefficients were negative—that is, counties with high percentages of other minority races in rural settings are negatively related to correlation improvements and thus potentially not fairly represented in the Descartes dataset. Second, lower income had a significant positive coefficient for Apple—that is, counties with lower incomes were associated with higher correlation improvements. Nevertheless, the interaction term between low income and race-*Other* for the Apple dataset has a significant negative coefficient, revealing worse correlation improvements for counties with high percentages of minority races and low income, potentially due to sampling bias. It is important to clarify that despite the significance of many coefficients, the R-square values for the regressions were low, pointing to the fact that there exist other behavioral or pandemic features that could also explain correlation improvements such as masking mandates, masking behaviors, or transmission rates, among others. Nevertheless, statistics for these features at the county level and for our period of study were not accessible.

5.3 Do Correlation Improvements Change across Training Approaches?

In this section, we explore whether the improvements brought about by adding mobility data to county-based COVID-19 case prediction models are different depending on the type of training approach. Specifically, we discuss the differences in percentages of counties that benefit from adding mobility data, quantify correlation improvements, and discuss the impact of mobility data bias on county-level COVID-19 case prediction models that have been trained with STW as opposed to LTW approaches. Our main objective is to understand if using short-term training windows (which requires considerably less data and hence reduces data costs) has an impact on how COVID-19 case prediction models benefit from mobility data.

Percentage of Counties. Figure 5 shows the percentage of counties whose STW-trained COVID-19 case prediction performance improves when adding mobility data. Compared to Figure 2, we can observe that for lower lookaheads (1, 7, and 14), the percentage of counties that benefit from adding mobility data to STW-trained models is smaller than the percentage for LTW-trained models, whereas the percentage of counties remains similar for higher lookaheads (21 and 28) across both linear and ARIMAX models. To assess the statistical significance of these observations, we run a Mann-Whitney U test between the STW percentages and the LTW percentages for each lookahead, and across the 10 datasets and two predictive models (Table 2 presents the details). The distributions were found to be statistically significantly different (p-value < 0.05) for lookaheads 1 and 7, with STW-trained models having significantly smaller median percentages of counties benefiting from adding mobility data (from 16%–31%) than LTW-trained models (from 40%–43%).

Correlation Improvements. To quantify correlation improvement differences between STW- and LTW-trained models, we run a Mann-Whitney U test for each lookahead between the



(a) Results for the ARIMAX COVID-19 case prediction model

(b) Results for the Elastic Regression COVID-19 case prediction model

Fig. 5. Percentage of counties for which adding mobility data to the COVID-19 case prediction model improves the prediction performance ($ci \geq 0$). Results are for the STW approach.

Table 2. Mann-Whitney U Test between the STW and LTW Distributions of the Percentage of Counties That Benefit from Adding Mobility Data across the 10 Datasets and Two Models for Each Training Approach

Lookahead	*p*-Value	Avg. % Benefit STW	Avg. % Benefit LTW
1	0.00002	0.16473	0.40074
7	0.00004	0.31053	0.43324
14	0.08103	0.41427	0.46833
21	0.20845	0.4722	0.49895
28	0.3104	0.48988	0.50206

correlation improvement (*ci*) distributions for STW and LTW across the 10 mobility datasets and two predictive models—that is, we measure whether the improvements brought about by adding mobility data to COVID-19 case prediction models are statistically significantly different across training approaches, and by how much (distribution plots can be checked in Appendix Figures 6–9). The tests—shown in Table 3—reveal that the two distributions are statistically significantly different (p-value < 0.05) across all lookaheads, with lower median correlation improvement values for LTW-trained models than their STW-trained counterpart (0.0002–0.0442 vs. 0.0018–0.0614), and with a slightly lower maximum median *ci* value of 0.1, as opposed to STW-trained maximum $ci = 0.13$. In other words, although a higher percentage of LTW-trained counties improved their performance when adding mobility data, the median improvement range is slightly lower than their STW counterpart.

Mobility Data Bias. We repeat the regression coefficient analysis discussed in Section 5.2 for the STW training approach, and a comparison with the associations revealed for the LTW training in that section revealed similar findings. For regression models, similarly to LTW, age 65+, income,

Table 3. Mann-Whitney U Test Per Lookahead between the STW and LTW Distributions of the Correlation Improvements across the 10 Datasets and Two Models for Each Training Approach

Lookahead	p-Value	Avg. Median *ci* STW	Avg. Median *ci* LTW
1	0.00334	0.00183	0.00029
7	0.00001	0.01693	0.00738
14	0	0.0321	0.01674
21	0.00009	0.04836	0.02939
28	0.01143	0.06145	0.04425

race Black, race-*Other*, and rurality are all negatively related to correlation improvement—that is, counties with higher percentages of elder, rural, low income, or minority populations do not benefit as much from the use of mobility data, possibly due to lack of representativity of these groups in the mobility datasets. Nevertheless, the number of significant coefficients across datasets for STW is lower than for LTW—although all datasets except for Google show one or another type of bias. This result highlights that mobility data bias might be more constrained when shorter amounts of training data are used in the model.

For ARIMAX models, and similarly to LTW, we found that age 65+, ethnicity Hispanic, and race-*Other* are negatively associated with correlation improvements—that is, adding mobility data in counties with elder people and other minority race/ethnicity worsens the predictive performance when compared to non-mobility baselines, potentially revealing bias in the data collected for these groups. However, although low income and Black race are positively related to correlation improvements, interaction terms between these two factors and age 65+ showed that counties with a high percentage of elder Black population or elder low-income population were significantly negatively related to correlation improvement—that is, the mobility data associated with these groups might not be representative, thus affecting the quality of the predictions. Finally, the interaction between rurality and age 65+ partially showed that counties with higher percentages of elder rural communities did not always benefit from using SafeGraph mobility data. For further details, we encourage the reader to compare Appendix Table 10 with Appendix Table 11 (for regression models) and Appendix Tables 8 and 9 for ARIMAX models.

These analyses show that the training-testing approach impacts the number of counties that benefit from adding mobility data and creates a tradeoff: LTW improves performance for more counties albeit with smaller correlation improvements with respect to their baselines and with more bias across socio-economic and demographic variables when compared to STW, which requires less data and constitutes a more affordable approach.

5.4 Do Correlation Improvements Change across Mobility Datasets?

In this section, we explore whether the improvements brought about by adding mobility data to county-based COVID-19 case prediction models are different depending on the type of dataset. As in previous sections, we will analyze differences across the percentage of counties that benefit from adding mobility data to their predictive models, analyze differences across correlation improvements, and evaluate the role of mobility data bias on those differences. Table 4 summarizes average improvements per mobility dataset across lookaheads, training approaches, and predictive models for (1) percentage of counties and (2) correlation improvements. To evaluate statistically significant differences across datasets, we compute the Friedman test, a non-parametric test that evaluates whether median values across datasets are statistically significantly different.

Percentage of Counties. We run the Friedman test with the distribution of the percentage of counties that benefit from adding mobility data across the five lookaheads, two models, and

Table 4. Improvement Statistics Per Mobility Dataset across Training Approaches and Predictive Models

Dataset	Avg. % of Counties Improved	Avg. Median *ci* Improvement
Apple Mobility	0.38107	0.0337
Descartes Mobility	0.42670	0.02363
Google Mobility	0.53613	0.01293
(SafeGraph) Grocery Store Mobility	0.42302	0.01816
(SafeGraph) Religious Org Mobility	0.44899	0.02298
(SafeGraph) Restaurants Mobility	0.42864	0.02375
(SafeGraph) Schools Mobility	0.43075	0.02123
SafeGraph Inflow Mobility	0.42365	0.02751
SafeGraph Intraflow Mobility	0.46907	0.02720
SafeGraph Outflow Mobility	0.43595	0.02938

The percentage of counties column represents the average percentage of counties whose COVID-19 case prediction models benefit from adding mobility data, whereas the correlation improvement (ci) column quantifies the average median ci improvement over the non-mobility baseline.

two training approaches for each mobility dataset. The test was rejected ($\chi^2 = 28$ and p-value = 0.0009), thus pointing to statistically significant differences between the average percentages across datasets. The Friedman test does not identify which distribution(s) are different; however, Table 4 shows that the percentage of counties with improvements is considerably higher for the Google dataset when compared to others, possibly due to the smaller number of counties used in the analysis. Thus, we removed the Google dataset from our set of distributions and repeated the Friedman test, which did not find any statistically significant difference between datasets—that is, the percentage of counties that benefit from adding mobility data to COVID-19 prediction models does not significantly change depending on the mobility dataset except for Google (see Appendix Table 12 for further test details).

Correlation Improvements. We run the Friedman test to evaluate whether the correlation improvement brought about by different mobility datasets was statistically significantly different across datasets. The Friedman test between the median correlation improvements across models, training approaches, and lookaheads for each dataset rejected the null hypothesis ($\chi^2 = 182.7$, p-value = 0), thus pointing to statistically significant differences across datasets. Removal of individual datasets did not change the result of the tests (null hypotheses rejected), revealing that correlation improvements are significantly different across mobility datasets, with Apple and SafeGraph—which includes O-D flows—having the highest improvements across training approaches and predictive models. Nevertheless it is important to highlight that despite significant *improvements* across datasets were very modest with a maximum value of 0.033 (see Appendix Table 13 for further test details).

Mobility Data Bias. Finally, in terms of bias, all mobility datasets were associated with bias in race, age 65+, and income and rurality either as independent features or via feature interactions, with the exception of the Google dataset and SafeGraph Grocery Stores that were associated only to race, age, and income bias, but rurality did not appear to play a role in the correlation improvements (see Appendix Tables 8, 9, 10, and 11 for further details).

To summarize the analysis presented in this section, Apple and SafeGraph datasets appear to bring about very modest statistically significantly higher correlation improvements (maximum value of 0.033), but the percentage of counties and the bias identified are similar across mobility datasets.

5.5 Do Correlation Improvements Change across Predictive Models?

In this section, we explore whether the improvements brought about adding mobility data to county-based COVID-19 case prediction models are different depending on the type of predictive

Table 5. Mann-Whitney U Test between the ARIMAX and Linear Distributions of the Percentage of Counties That Benefit from Adding Mobility Data across the 10 Datasets and Two Training Approaches for Each Predictive Model

Lookahead	*p*-Value	Average % ARIMAX	Average % Linear Regression
1	0.0002	0.16672	0.39875
7	0.00512	0.32085	0.42292
14	0.00069	0.40169	0.48091
21	0.00004	0.4376	0.53354
28	0.00004	0.45166	0.54029

Table 6. Mann-Whitney U Test Per Lookahead between the ARIMAX and Linear Distributions of the Correlation Improvements across the 10 Datasets and Two Training Approaches for Each Model Type

Lookahead	*p*-Value	Average Median *ci* ARIMAX	Average Median *ci* Regression
1	0.00007	0.00196	0.00029
7	0.94738	0.01544	0.0147
14	0.00298	0.02564	0.03478
21	0.00003	0.03196	0.0612
28	0.00001	0.0381	0.08195

model. Revisiting Figures 2 and 5, we can visually observe that the percentage of counties that benefit from adding mobility data to county-level COVID-19 prediction models is larger when the models are linear regressions rather than ARIMAX. To assess the statistical significance of this difference, we run a Mann-Whitney U test for each lookahead, between the percentage of counties that benefited from adding mobility data to linear models and the percentage of counties that benefited from adding mobility data to ARIMAX models across the 10 datasets and two training approaches. The test showed that the differences between the two types of models are statistically significantly different across all five lookaheads, with the percentage of counties that benefit from adding mobility data being smaller for ARIMAX models (with values between 0.16% and 0.45%) than for linear models (0.39%–0.54%). Table 5 shows further details of the statistical test.

However, to assess the statistical significance of the differences in the correlation improvements (*ci*) between linear and ARIMAX models, we run a Mann-Whitney U test for each lookahead, between the *ci* values associated with adding mobility data to linear models and the *ci* values associated with adding mobility data to ARIMAX models across the 10 datasets and two training approaches. Except for lookahead 7, all other lookaheads show significantly different *ci* distributions, with larger improvements associated with linear regression models at higher lookaheads (0.00029–0.08195 for linear vs. 0.00196–0.0381 for ARIMAX). Full test details are available in Table 6.

Finally, in terms of bias differences across models, we can observe that both models suffer from similar types of bias across datasets and training approaches (i.e., income, race, age 65+, and (a little bit less frequently) rurality). Nevertheless, there is one main difference between linear regression and ARIMAX models. Linear models trained with Google mobility data (both with STW and LTW training approaches) did not identify any significant socio-economic or demographic features in the bias analysis, meaning that these features do not play a role in the performance improvement of county-level COVID-19 case linear predictions. It is important to clarify that although some bias was identified for ARIMAX trained with Google mobility data, the number of features identified was also considerably lower than for any other dataset. We posit that this could be potentially related to the smaller number of counties available (990, see

Table 1). Detailed bias numbers can be found in Appendix Tables 8 and 9 (ARIMAX) and Appendix Tables 10 and 11 (linear regressions). Overall, linear regressions appear to be a better choice, with a larger number of counties benefiting from adding mobility data, with larger correlation improvements at higher lookaheads and with similar bias.

6 DISCUSSION

Our analysis on the use of mobility data for COVID-19 case prediction has shown the heterogeneity and limitations of the benefits of mobility data inclusion. As discussed in Section 5, at most 60% of the counties improve their performance when adding mobility data to the prediction model (top value for linear regressions with the LTW approach). The median correlation improvements across lookaheads, datasets, and models are minimal for next-day predictions and modest for higher lookaheads, with 50% of the counties showing correlation improvements of at most 0.1 and 25% of the counties showing correlation improvements of at most 0.3.

Our results have shown that, even in the best case, 40% of counties would face no improvement when mobility data is added to their COVID-19 case prediction models, making the question of whether or not mobility data would be helpful to county-level decision makers essentially a coin flip. Although companies made mobility data freely available at the initial height of the COVID-19 pandemic, companies have already begun to discontinue these practices (Apple's reports are no longer available, and Google stopped releasing data as of October 15, 2022). As decision makers face the continued spread of COVID-19 and potential future diseases, they must consider whether it will be worth purchasing mobility data. Our analysis shows that purchasing county-level mobility data will not benefit many counties, and decision makers should proceed with caution accordingly.

It is also concerning that the extent to which mobility data improves predictions is in part a function of the composition of the population in the county. In fact, across most of the mobility datasets, correlation improvements were lower for counties with higher Black, Hispanic, and other non-White populations as well as low-income and rural populations. As older and minority patients have been disproportionately affected by the COVID-19 pandemic, we would hope to provide more and better resources to these groups to ameliorate the disparities. Instead, we see that mobility data could serve to entrench these disparities, providing decision makers in counties with more vulnerable populations with worse-performing models, leading to worse-informed policy decisions.

Our analyses have also demonstrated that linear regression models outperform ARIMAX models in both (1) the percentage of counties that benefit from adding mobility data and (2) the correlation improvement when using the same mobility data. These differences are statistically significant. Based on this, we recommend that decision makers favor linear regression models when interested in using interpretable models. We have also shown that the performance across datasets is quite similar, with Apple and some SafeGraph datasets having slightly superior correlation improvements, albeit still so small that in many cases will not lead to a change in the strength of the correlation. This result highlights that decision makers looking to use mobility data as a source of behavioral information should not worry about the dataset they gain access to, since all seem to similarly improve the correlation over their non-mobility baselines. Finally, we have discussed how the training-testing approach presents a tradeoff for decision makers, with LTW approaches increasing the number of counties that benefit from adding mobility data, whereas STW approaches—that require considerably less data and reduce costs—increase the correlation improvements (albeit with small values) and have lower bias.

7 LIMITATIONS AND FUTURE WORK

Although we explore many angles of the use of mobility data in COVID-19 case prediction, there are several limitations and opportunities for further research. One limitation is the reliability of our dependent variable—COVID-19 cases. Other research has shown that especially in the early stages of the pandemic, published case numbers were not reflective of the actual spread of COVID-19, in large part due to the lack of testing resources [39]. Hospitalization and death counts are more reliable than case counts as dependent variables, but they also face issues with under-counting [37] and present further complication with time lags. In future work, we will seek to leverage methods for case count correction [29] to understand the effects of using *corrected* cases on COVID-19 prediction models' accuracy and fairness. Future work should also evaluate changes in the reported findings in this article when hospitalization and death counts, instead of cases, are considered.

Our study focuses on two prediction models—linear and ARIMAX—and the findings discussed only apply to these models. Using other modeling and predictive approaches such as compartmental [44] or multi-population models [4] might reveal different results. However, it is important to clarify that, unlike the approach proposed in this article, compartmental and multi-population models rely heavily on the availability of more granular data, either in the form of individual mobility patterns or mobility data segmented by sub-groups, and that such data is not freely accessible and often comes at a significant cost.

Finally, a national-level policy maker might be interested in a unified model that learns COVID-19 trends for all counties. This would allow for inclusion of county-level variables indicative of population vulnerability directly in the model, potentially yielding more accurate results. One could also explore methods to ameliorate bias in mobility data, whether through interventions on the model or development of techniques to identify how mobility data providers can improve their data collection and transformation procedures.

A APPENDIX

Table 7 shows the average performance—measured as the average correlation between predicted and actual COVID-19 cases—across lookaheads and linear regression implementations for LTW and STW training-testing approaches, respectively. We can observe that, on average, Ridge is the best performing model for three out of the five lookaheads, and ElasticNet is the best performing model for three out of five lookaheads as well. Given their majority best performance, we select this as implementations for the analysis. Tables 8, 9, 10, 11 contains the values for the regression coefficient analysis for ARIMAX LTW, ARIMAX STW, Ridge Regression LTW and ElasticNet Regression STW respectively. Table 12 contains the non-parametric Friedman test analysis, which highlights whether the percentage of counties that benefit from adding mobility data is statistically significantly different or not. Table 13 highlights the results if the *ci* brought about by different mobility datasets were significantly different across datasets. Table 14 identifies the p, d, q values of the ARIMAX models that were used and the counts of counties that had the stated parameters.

Table 7. Average p_{corr} Value for Each Regularized Method

	STW				LTW			
Lookahead	OLS	Ridge	Lasso	ElasticNet	OLS	Ridge	Lasso	ElasticNet
1	0.965098	**0.965128**	0.962530	0.963475	0.952180	**0.952182**	0.938414	0.945211
7	0.639711	0.640578	0.644599	**0.647571**	0.572311	**0.572374**	0.548012	0.560634
14	0.453315	0.454412	0.462288	**0.465076**	0.366340	**0.366426**	0.350444	0.361948
21	0.290965	0.292186	0.308287	**0.308882**	0.214946	0.215025	0.212894	**0.218513**
28	0.152364	0.153709	**0.177438**	0.174648	0.121256	0.121314	**0.134021**	0.132422

The best value for each window and each lookahead is bolded.

Table 8. Regression Coefficient Analysis for ARIMAX LTW

Feature	Apple	Descartes	Google	(SafeGraph) Grocery Store	(SafeGraph) Religious Org
r2	0.0176	0.0157	0.0173	0.0065	0.0094
adjusted_r2	0.013	0.0117	0.0091	0.0032	0.006
const	0.0585	0.0528	0.0342	0.0411	0.0475
age_65+	−0.0493**	−0.0032	0.0078	−0.0052	−0.0022
low_income	0.0319*	−0.014	0.0097	−0.0007	0.0049
black	−0.0102	0.0343**	−0.0063	−0.0028	−0.0134
hispanic	0.0195	0.023	0.0032	0.0129	0.0021
race_other	−0.0935***	0.0313*	0.0038	−0.0184	−0.0044
rurality	−0.042*	0.0137	0.0046	−0.0093	−0.0035
age_65+:low_income	−0.039**	−0.0049	0.002	−0.002	−0.001
age_65+:black	−0.0006	−0.0042	0.017	−0.0022	−0.0062
age_65+:hispanic	−0.0238**	−0.0096	−0.0156*	−0.005	−0.0035
age_65+:race_other	0.0188	0.0003	0.0139	0.0085*	0.0024
age_65+:rurality	0.0396*	0.0026	−0.0214	0.0143	0.0052
low_income:black	0.0012	0.0259***	−0.0026	−0.0043	−0.0222***
low_income:hispanic	0.0156	0.0013	−0.0179*	0.0083	0.0026
low_income:race_other	−0.0381***	0.0129	0.0035	−0.0033	−0.0051
low_income:rurality	−0.0047	0.0053	−0.0147	−0.0011	−0.0006
black:hispanic	−0.0019	−0.0063*	−0.0055	−0.0025	−0.0077***
black:race_other	0.0019	0.0055*	−0.0023	0.0001	−0.0034
black:rurality	0.017*	−0.0022	−0.0059	0.0054	0.0088
hispanic:race_other	0.0109**	−0.0014	−0.0061	0.001	0.0036
hispanic:rurality	0.0155*	−0.0076	0.0003	0.0003	0.0094
race_other:rurality	0.0409***	−0.0223**	−0.0078	0.0075	−0.0025

Feature	(SafeGraph) Restaurants	(SafeGraph) Schools	SafeGraph Inflow	SafeGraph Intraflow	SafeGraph Outflow
r2	0.0112	0.0111	0.0119	0.0083	0.0089
adjusted_r2	0.0079	0.0078	0.0081	0.0051	0.0055
const	0.0489	5.58e−2	0.0565	0.0568	0.0582
age_65+	−0.029*	−0.0075	−0.0298*	−0.0152	−0.0343*
low_income	−0.0106	0.0106	0.0075	0.0017	0.0097
black	0.0075	−0.0104	−0.0131	0.0036	0.0139
hispanic	0.0045	−0.022	−2.06e−2	−0.0055	−0.0109
race_other	−0.0201	−0.0289	−0.0373*	−0.0378**	−0.0302*
rurality	−0.0235*	−0.014	−0.0429***	−0.014	−0.0354**
age_65+:low_income	−0.0022	0.0017	−0.0025	−0.0043	−0.0158
age_65+:black	0.0058	0.0056	0.0091	0.0101	−0.0062
age_65+:hispanic	−0.0025	−0.0007	0.0043	0.0028	0.0019
age_65+:race_other	0.0038	0.0103*	0.0102*	0.0093*	0.007
age_65+:rurality	0.0519***	0.0122	0.0519***	0.0231	0.0449**
low_income:black	0.0132*	−0.0019	−0.0086	0.0061	0.0088
low_income:hispanic	0.0065	−0.0025	−0.0138	−0.0033	−0.0057
low_income:race_other	−0.0099	−8.30e−3	−0.0099	−0.0078	−0.0022
low_income:rurality	0.0106*	−0.0017	−0.0079	0.0029	−0.0043
black:hispanic	−0.0041	−0.0003	−0.0063*	−0.0043	−0.0051*
black:race_other	0.0017	0.0003	−0.0023	0.0019	0.0031
black:rurality	0	0.002	5.70e−3	−0.0007	0.0075
hispanic:race_other	−0.0029	0.0032	−0.0012	0.0008	0.0037
hispanic:rurality	0.0067	0.022***	0.0089	0.0051	0.0097
race_other:rurality	0.0096	0.0087	0.0214	0.0226*	0.0201*

Values highlighted in green represent statistically significant positive coefficients, whereas values highlighted in red represent statistically significant negative coefficients. Significance level (p−value smaller than): ***, 0.001 **, 0.01; *, 0.05.

Table 9. Regression Coefficient Analysis for ARIMAX STW

Feature	Apple	Descartes	Google	(SafeGraph) Grocery Store	(SafeGraph) Religious Org
r2	0.0444	0.0617	0.0745	0.0693	0.056
adjusted_r2	0.0371	0.0567	0.0601	0.0649	0.0521
const	0.0743	0.0594	0.039	0.0361	0.0417
age_65+	−0.0192	−0.0071	−0.029*	−0.0101	0.0113
low_income	0.0093	−0.0115	0.0209*	−0.0002	−0.0175***
black	0.0449***	−0.0088	0.0179	0.0128*	0.0028
hispanic	−0.0032	−0.0142	0.0123	−0.0032	−0.0138*
race_other	0.0081	0.0089	0.0017	−0.0088	−0.0087
rurality	−0.0124	0.0306**	0.0029	0.0023	0.0188**
age_65+:low_income	−0.0241*	−0.0166*	−0.044***	−0.0116*	0.0038
age_65+:black	−0.0176	−0.0108	−0.0174*	−0.0044	−0.0136***
age_65+:hispanic	−0.0169**	−0.0032	−0.0126*	0.0009	0.0069**
age_65+:race_other	−0.0136	−0.0138*	−0.0161*	−0.0007	0.0048
age_65+:rurality	0.0221	−0.0035	0.0129	0.0094	−0.0132*
low_income:black	0.0157*	−0.0203***	−0.005	−0.0006	−0.0062
low_income:hispanic	−0.001	−0.0058	−0.0011	−0.0016	−0.0014
low_income:race_other	−0.006	−0.0034	−0.0073	−0.0171***	−0.0037
low_income:rurality	0.0084	0.0259***	0.01	0.0088**	0.0142***
black:hispanic	0.0055*	0.0048*	−0.0017	0.0002	0.002
black:race_other	−0.0046*	−0.002	−0.0023	−0.0004	−0.0016
black:rurality	−0.0001	0.0035	0.0041	−0.0008	0.0112**
hispanic:race_other	−0.0008	−0.0024	0.0032	0.0006	0.0002
hispanic:rurality	0.0191**	0.0089*	−0.0004	0.0061*	0.0063*
race_other:rurality	0.0028	0.0054	0.0078	−0.005	0.0009

Feature	(SafeGraph) Restaurants	(SafeGraph) Schools	SafeGraph Inflow	SafeGraph Intraflow	SafeGraph Outflow
r2	0.0727	0.0345	0.0529	0.0323	0.0376
adjusted_r2	0.0679	0.03	0.0482	0.0274	0.033
const	0.0453	0.0479	0.0514	0.0474	0.0531
age_65+	0.0085	−0.002	0.0056	−0.0013	−0.0158
low_income	−0.0168**	−0.0049	−0.0114	−0.0033	0.0007
black	0.0003	−0.0014	0.0201*	0.0083	0.002
hispanic	−0.0058	−0.0274***	−0.0087	0.0008	−0.0211**
race_other	−0.0089	−0.0241**	−0.0273*	0.0001	−0.0104
rurality	0.0124	0.0013	0.0214**	0.0006	0.0019
age_65+:low_income	0.0028	−0.0024	−0.0119	−0.0058	−0.0145*
age_65+:black	0.0128**	−0.0023	−0.0116*	−0.0054	−0.0137**
age_65+:hispanic	0.0062*	0.0049	−0.0082*	−0.0049	−0.0011
age_65+:race_other	0.0035	0.0011	−0.0019	−0.0034	−0.003
age_65+:rurality	−0.0054	0.0024	−0.0091	0.0084	0.0182*
low_income:black	−0.0101**	0.0022	0.009*	0.0019	−0.0077
low_income:hispanic	0.005	−0.0048	−0.0064	−0.0022	−0.0063
low_income:race_other	−0.0037	−0.0079*	−0.0109**	−0.0057	−0.0092*
low_income:rurality	0.0136***	0.009**	0.0215***	0.0078*	0.0169***
black:hispanic	0.0007	0.0056***	−0.0001	−0.0001	0.0017
black:race_other	−0.0012	−0.0018	0.0017	−0.0009	−0.0032*
black:rurality	0.0133***	0.0098**	0.0086*	0.0079	0.0113**
hispanic:race_other	0.0034*	0.0023	0.0012	−0.0001	0.0007
hispanic:rurality	0.0073*	0.0184***	0.0162***	0.0064	0.0187***
race_other:rurality	0.0006	0.0138*	0.0194*	−0.0001	0.0051

Values highlighted in green represent statistically significant positive coefficients, whereas values highlighted in red represent statistically significant negative coefficients. Significance level (p−value smaller than): ***, 0.001; **, 0.01 *, 0.05.

Table 10. Regression Coefficient Analysis for Ridge Regression LTW

Feature	Apple	Descartes	Google	(SafeGraph)	(SafeGraph)
				Grocery Store	Religious Org
r2	0.0297	0.0218	0.0054	0.0195	0.0102
adjusted_r2	0.0229	0.0184	−0.001	0.0164	0.0076
const	0.101***	0.064***	0.061***	0.071***	0.063***
age_65+	−0.027	−0.017	0.01	−0.018	−0.012
low_income	0.055*	0.033*	0.024	0.032*	0.015
black	−0.067	−0.013	−0.046	−0.042*	−0.003
hispanic	−0.072*	−0.017	0.02	−0.038*	−0.017
race_other	−0.049	−0.015	−0.082	−0.041	−0.031
rurality	−0.039	−0.031	−0.003	−0.005	−0.004
age_65+:low_income	−0.033	−0.016	−0.002	−0.02	−0.013
age_65+:black	0.038	−0.028*	0.002	0.028*	0
age_65+:hispanic	−0.029	0.022*	0.001	0.021*	0.007
age_65+:race_other	−0.036	−0.003	0.026	−0.003	−0.002
age_65+:rurality	0.039	0.021	−0.026	0.002	0.012
low_income:black	−0.026	−0.036***	−0.033	−0.015	−0.006
low_income:hispanic	−0.079***	−0.021	0.01	−0.025*	−0.018
low_income:race_other	−0.011	−0.004	−0.032	−0.013	−0.008
low_income:rurality	0.002	−0.011	−0.017	0.001	0.006
black:hispanic	−0.012	−0.009*	−0.006	−0.006	0
black:race_other	0.009	−0.002	0.007	0.004	0.003
black:rurality	0.024	0.031***	0.017	0.017	0.007
hispanic:race_other	0.028***	0.007	0.013	0.004	0
hispanic:rurality	0.035*	−0.008	−0.012	0.001	0.002
race_other:rurality	0.05*	0.013	0.013	0.026	0.022

Feature	(SafeGraph)	(SafeGraph)	SafeGraph	SafeGraph	SafeGraph
	Restaurants	Schools	Inflow	Intraflow	Outflow
r2	0.0208	0.0056	0.0217	0.0154	0.0204
adjusted_r2	0.018	0.0025	0.0192	0.0131	0.0181
const	0.069***	0.058***	0.07***	0.07***	0.066***
age_65+	−0.041*	−0.01	−0.035*	−0.038*	−0.043*
low_income	0.047***	0.017	0.035**	0.037**	0.039***
black	−0.039*	0.01	−0.051**	−0.029	−0.04*
hispanic	−0.038*	−0.048*	−0.035*	−0.031*	−0.041**
race_other	−0.057**	−0.043	−0.023	−0.027	−0.033
rurality	−0.037*	0.006	−0.036*	−0.03	−0.019
age_65+:low_income	−0.04*	−0.021	−0.028*	−0.025*	0.033**
age_65+:black	0.003	−0.003	0.009	−0.011	0.011
age_65+:hispanic	0.011	0.015	0.014*	0.022***	0.014*
age_65+:race_other	−0.005	−0.005	−0.005	0.004	0.003
age_65+:rurality	0.033	−0.005	0.04*	0.03	0.029
low_income:black	−0.03**	−0.003	−0.034***	−0.031***	−0.029***
low_income:hispanic	−0.021*	−0.025*	−0.024*	−0.02*	−0.026**
low_income:race_other	−0.01	−0.004	−0.006	−0.012	−0.013
low_income:rurality	−0.004	0.008	−0.002	−0.007	0.004
black:hispanic	−0.001	−0.004	0	−0.003	0.002
black:race_other	0.001	0	0.001	0.003	0.004
black:rurality	0.024**	−0.003	0.03***	0.027**	0.016
hispanic:race_other	0.02***	0.012**	0.001	−0.005	−0.003
hispanic:rurality	0.013	0.011	0.008	0	0.01
race_other:rurality	0.042**	0.032	0.021	0.012	0.016

Values highlighted in green represent statistically significant positive coefficients, whereas values highlighted in red represent statistically significant negative coefficients. Significance level (p−value smaller than): ***, 0.001; **, 0.01; *, 0.05.

Table 11. Regression Coefficient Analysis for ElasticNet Regression STW

Feature	Apple	Descartes	Google	(SafeGraph)	(SafeGraph)
				Grocery Store	Religious Org
r2	0.012576	0.022346	0.017356	0.008641	0.009302
adjusted_r2	0.008654	0.01869	0.0085	0.00556	0.006092
const	0.1071***	0.0708***	0.0656***	0.0591***	0.0689***
age_65+	−0.0306	−0.0208	−0.0126	−0.0064	−3.10e−3
low_income	0.0077	0.0098	−5.58e−5	−0.0054	0.0024
black	−0.0443*	0.0088	−0.0147	0.0134	−0.0199
hispanic	0.0115	−0.01	−0.0134	0.0034	−0.0215
race_other	−0.0064	−0.0273*	0.0017	−0.0111	−0.0033
rurality	−0.0684**	0.0094	−0.0231	−0.0005	−0.0218
age_65+:low_income	−0.0119	−0.0181	−0.0033	−0.0075	0.0044
age_65+:black	0.0303*	−0.0025	−0.0014	−0.0009	0.0015
age_65+:hispanic	−0.0252**	0.0056	0.0138	−0.0012	0.0016
age_65+:race_other	−0.0119	0.0117	−0.0215	0.0056	0.0079
age_65+:rurality	0.059**	0.0092	0.0288	0.008	0.0117
low_income:black	−0.0069	−0.0133*	−0.0139	0.0046	−0.0145*
low_income:hispanic	0.013	−0.0193*	−0.0028	0.0018	−0.0096
low_income:race_other	−0.0035	−0.0055	−0.0092	1.60e−3	0.0009
low_income:rurality	−0.0153	0.0134*	0.0011	0.0005	−0.004
black:hispanic	0.0064	−0.0115***	−0.0022	−0.0023	−0.0035
black:race_other	0.0031	0.0039	0.0025	4.00e−4	0.0015
black:rurality	0.0133	−0.0083	0.0105	−0.0015	0.0113
hispanic:race_other	0.0085	0.0051	−0.0012	−0.0008	−1.70e−3
hispanic:rurality	0.0189*	−0.0034	−0.0006	0.0017	0.0125*
race_other:rurality	0.0095	0.0056	0.0115	0.008	−0.0028

Feature	(SafeGraph)	(SafeGraph)	SafeGraph	SafeGraph	SafeGraph
	Restaurants	Schools	Inflow	Intraflow	Outflow
r2	0.011075	0.009231	0.008333	0.004581	0.008229
adjusted_r2	0.007945	0.006016	0.0052	0.001467	0.005037
const	0.0731***	0.1135***	0.0663***	0.061***	0.0655***
age_65+	−0.0167	0.029	−0.0089	−0.0232*	0.0038
low_income	0.0111	−0.0298*	0.0034	0.0098	−0.0035
black	−0.0171	0.0283	0.0005	−0.0036	0.0068
hispanic	−0.003	−0.0331	−0.0076	−0.0082	−0.0014
race_other	−0.0084	−0.0202	−0.0161	−0.0056	0.0259
rurality	−2.07e−2	0.0155	−0.0174	−0.0147	−0.0044
age_65+:low_income	−0.0154	0.0209	−0.0059	−0.0132	0.003
age_65+:black	0.0046	0.0034	0.0005	−0.0033	−0.0015
age_65+:hispanic	−0.0038	−0.0018	0.0043	0.0018	0.0009
age_65+:race_other	0.003	0.008	−0.0025	−0.0076*	−0.0099*
age_65+:rurality	0.02	−0.0186	0.0166	0.0285*	0.0075
low_income:black	−0.0133	0.0224*	−0.0023	−0.0049	−0.0021
low_income:hispanic	−0.0086	−0.007	−0.0005	−0.004	−0.0039
low_income:race_other	−0.0004	0.0037	−0.0055	−0.0074	−0.0009
low_income:rurality	−0.0047	0.0101	−0.0071	0.0031	−0.0036
black:hispanic	−0.0038	0.0018	−0.0013	−0.0022	−0.002
black:race_other	0.0024	−0.0004	0.0023	0.0013	−0.0001
black:rurality	0.0109	−0.0157	0.0043	0.0051	0.0014
hispanic:race_other	−0.0008	−0.0037	0.0001	−0.0011	−0.0022
hispanic:rurality	0.0044	0.022*	0.0056	0.0028	0.0016
race_other:rurality	0.006	0.0126	0.0133	0.0058	−0.0151

Values highlighted in green represent statistically significant positive coefficients, whereas values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): ***, 0.001; **, 0.01; *, 0.05.

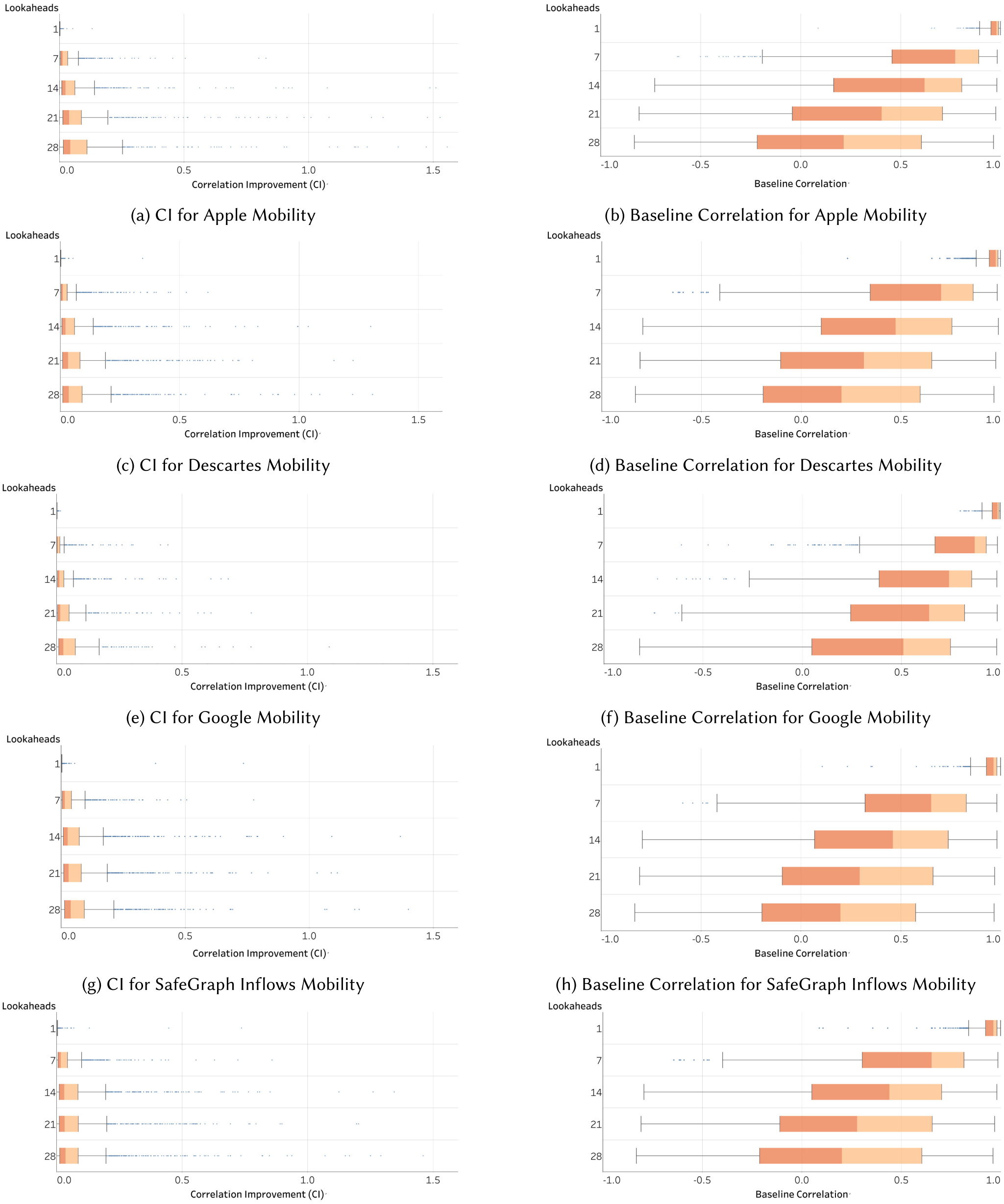
Table 12. Non-Parametric Friedman Test Analysis with Distributions Containing the Percentage of Counties Where Mobility Data Improved COVID-19 Case Prediction Models without Mobility Data across Datasets, Lookaheads, Training Approaches, and Models

Dataset Considered	All Datasets	Google Excluded
Chi-square	28.2873	9.6000
p-Value	0.0009	0.2942

Table 13. Non-Parametric Friedman Test Analysis with Distributions Containing the Median Correlation Improvement Values of Counties Where Mobility Data Improved COVID-19 Case Prediction Models without Mobility Data across Datasets, Lookaheads, Training Approaches, and Models

	Chi-Square	*p*-Value
Dataset Considered		
All datasets	182.7029	0.0
Restaurants Mobility excluded	164.0857	0.0
Religious Organization Mobility excluded	164.0603	0.0
Schools Mobility excluded	166.6444	0.0
Grocery Stores Mobility excluded	164.6127	0.0
SafeGraph Inflows Mobility excluded	164.1492	0.0
SafeGraph Outflows Mobility excluded	164.6000	0.0
SafeGraph Intraflows Mobility excluded	164.2508	0.0
Apple Mobility excluded	164.5937	0.0
Descartes Mobility excluded	164.0603	0.0
Google Mobility excluded	164.0794	0.0
Apple, Google, and SafeGraph Inflow Mobility excluded	108.6286	0.0

Each test is run with one or a few mobility datasets excluded.



(a) CI for Apple Mobility

(b) Baseline Correlation for Apple Mobility

(c) CI for Descartes Mobility

(d) Baseline Correlation for Descartes Mobility

(e) CI for Google Mobility

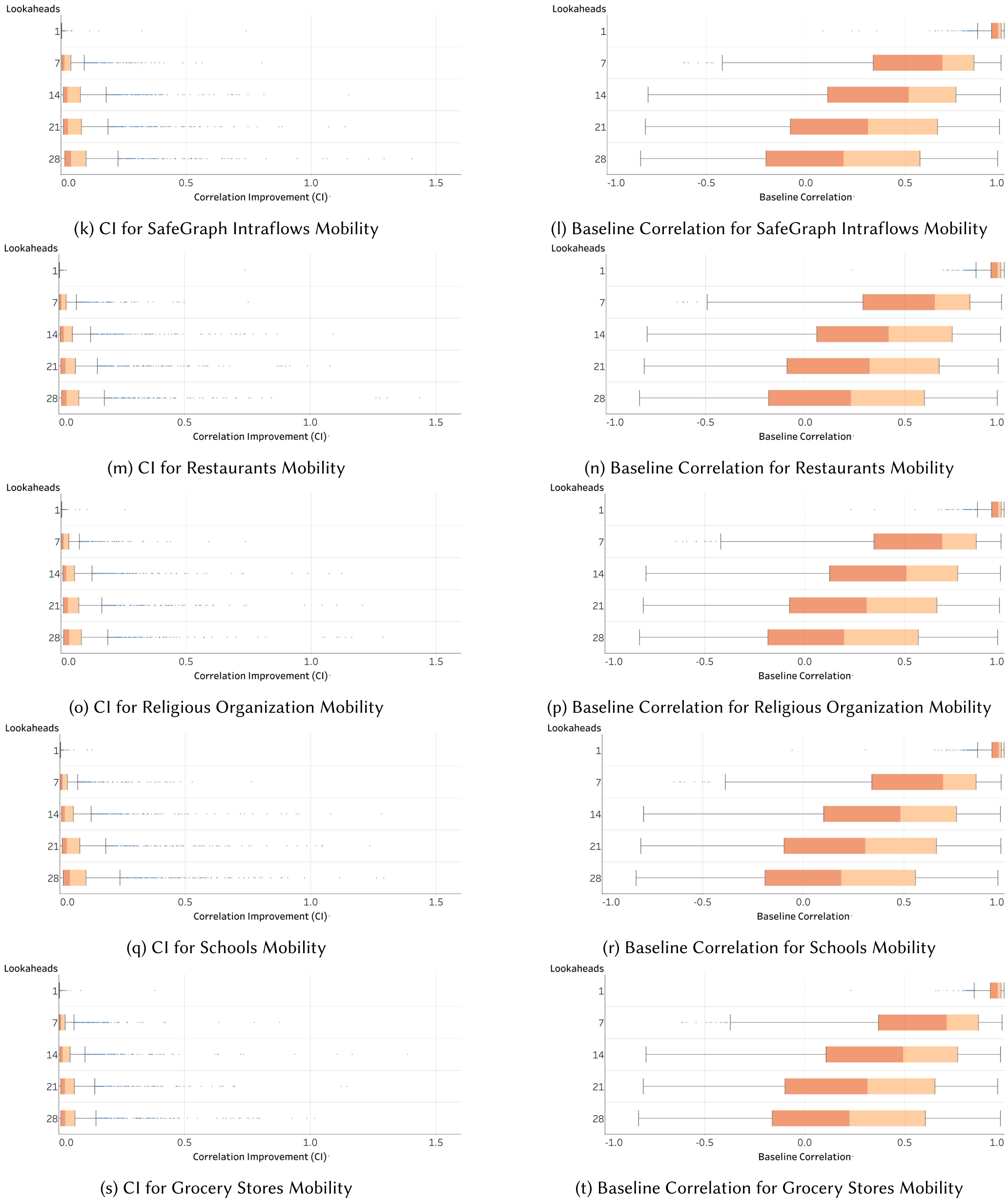
(f) Baseline Correlation for Google Mobility

(g) CI for SafeGraph Inflows Mobility

(h) Baseline Correlation for SafeGraph Inflows Mobility

(i) CI for SafeGraph Outflows Mobility

(j) Baseline Correlation for SafeGraph Outflows Mobility



(k) CI for SafeGraph Intraflows Mobility

(l) Baseline Correlation for SafeGraph Intraflows Mobility

(m) CI for Restaurants Mobility

(n) Baseline Correlation for Restaurants Mobility

(o) CI for Religious Organization Mobility

(p) Baseline Correlation for Religious Organization Mobility

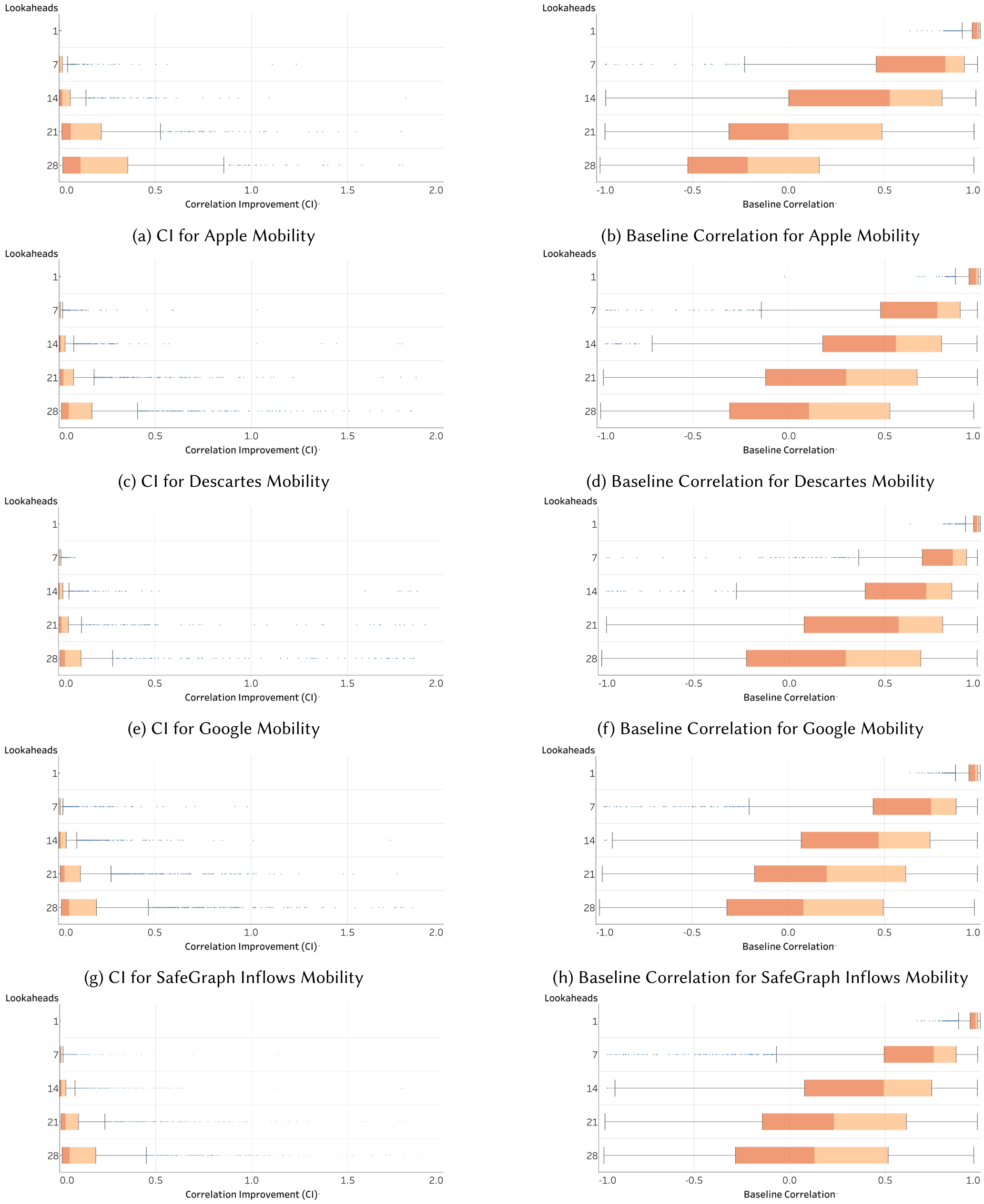
(q) CI for Schools Mobility

(r) Baseline Correlation for Schools Mobility

(s) CI for Grocery Stores Mobility

(t) Baseline Correlation for Grocery Stores Mobility

Fig. 6. Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with LTW approach using Apple Mobility data (a, b), Descartes Mobility data (c, d), Google Mobility data (e, f), SafeGraph Inflow Mobility data (g, h), SafeGraph Outflow Mobility data (i, j), SafeGraph Intraflow Mobility data (k, l), Restaurants Mobility data (SafeGraph POI) (m, n), Religious Mobility data (SafeGraph POI) (o, p), Schools Mobility data (SafeGraph POI) (q, r), and Grocery Stores Mobility data (SafeGraph POI) (s, t).



(a) CI for Apple Mobility

(b) Baseline Correlation for Apple Mobility

(c) CI for Descartes Mobility

(d) Baseline Correlation for Descartes Mobility

(e) CI for Google Mobility

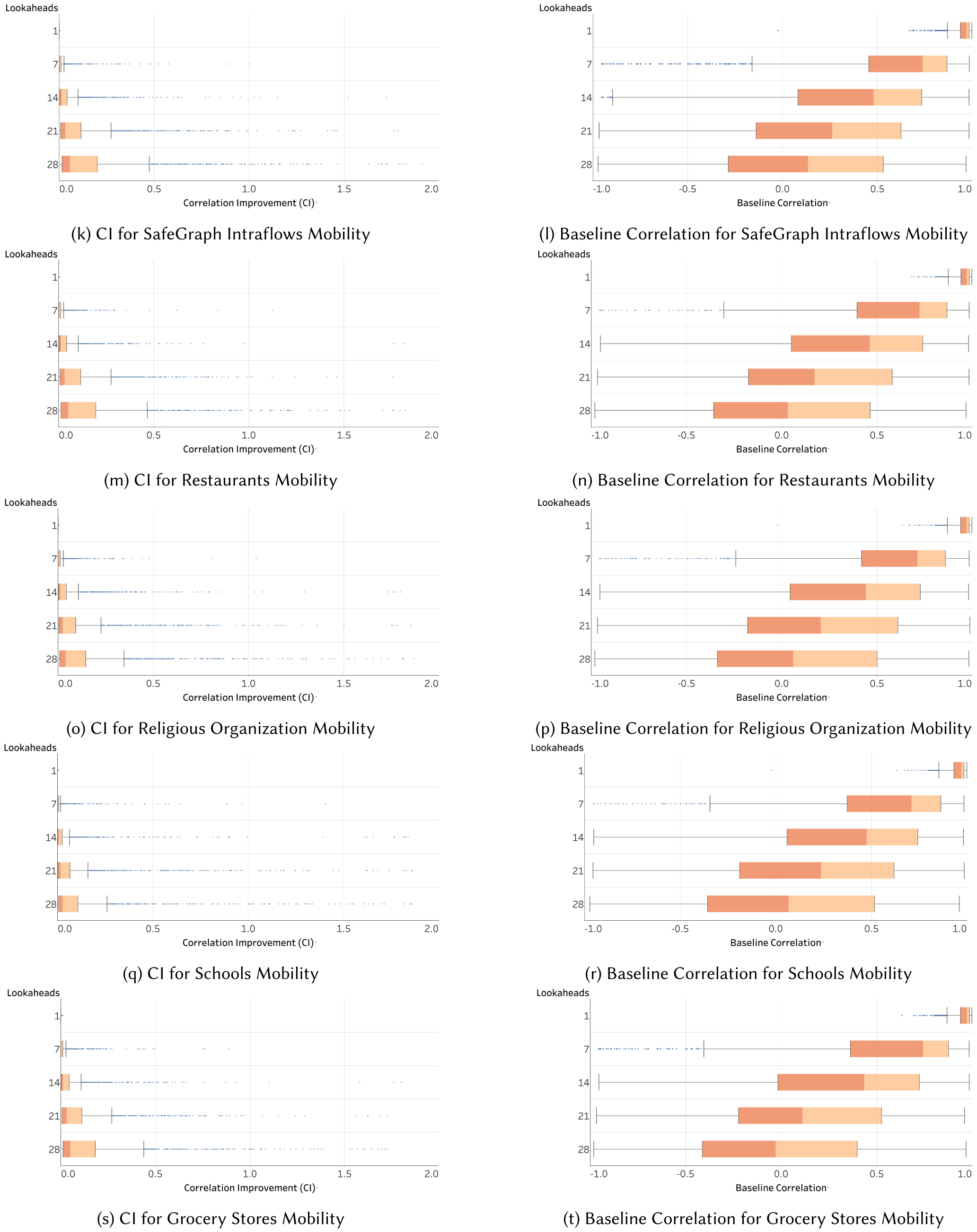
(f) Baseline Correlation for Google Mobility

(g) CI for SafeGraph Inflows Mobility

(h) Baseline Correlation for SafeGraph Inflows Mobility

(i) CI for Outflows Mobility

(j) Baseline Correlation for SafeGraph Outflows Mobility



(k) CI for SafeGraph Intraflows Mobility

(l) Baseline Correlation for SafeGraph Intraflows Mobility

(m) CI for Restaurants Mobility

(n) Baseline Correlation for Restaurants Mobility

(o) CI for Religious Organization Mobility

(p) Baseline Correlation for Religious Organization Mobility

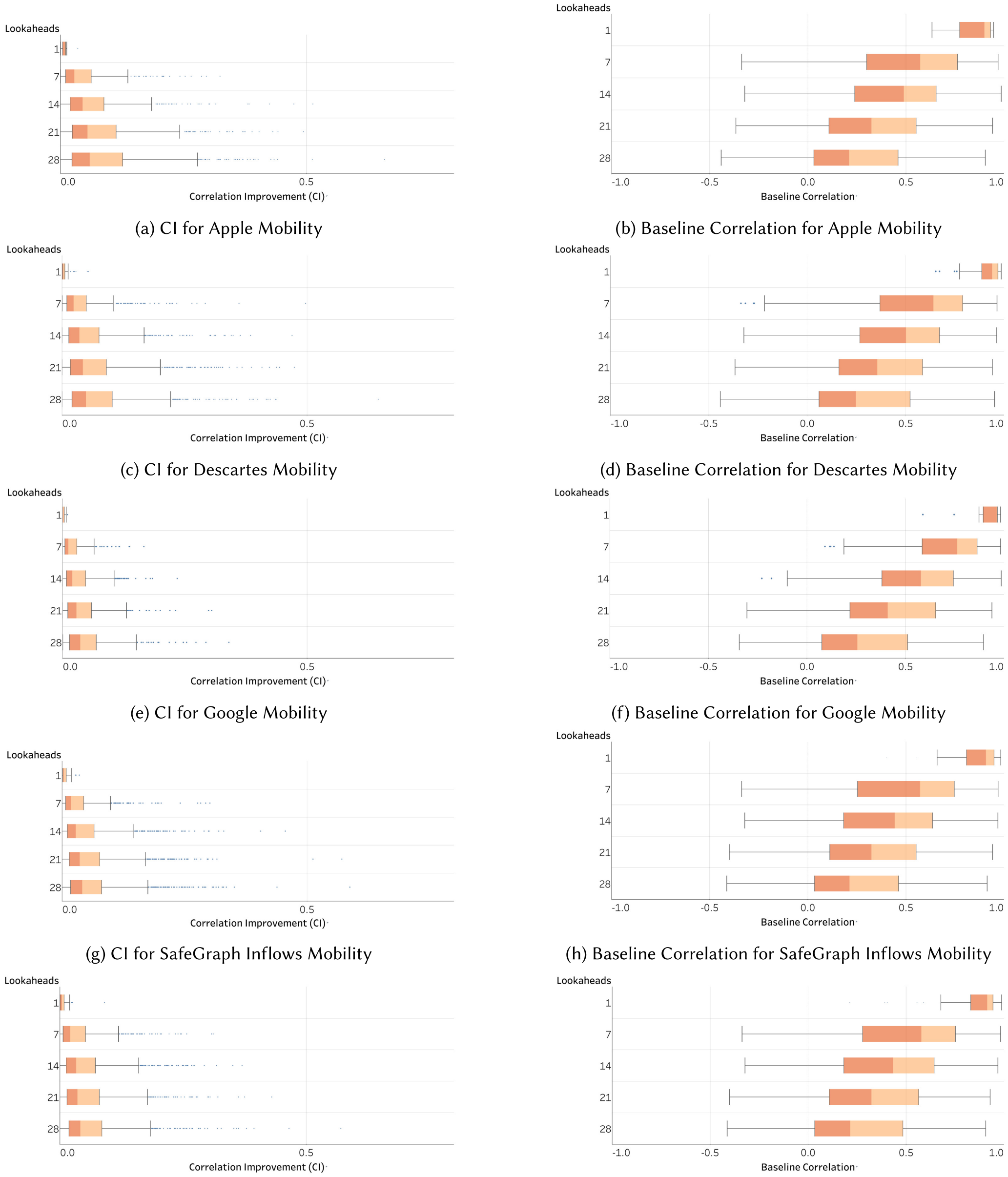
(q) CI for Schools Mobility

(r) Baseline Correlation for Schools Mobility

(s) CI for Grocery Stores Mobility

(t) Baseline Correlation for Grocery Stores Mobility

Fig. 7. Correlation improvements (*ci*) and baseline correlation distributions (non-mobility) across lookaheads for Ridge regression with the LTW approach using Apple Mobility data (a, b), Descartes Mobility data (c, d), Google Mobility data (e, f), SafeGraph Inflow Mobility data (g, h), SafeGraph Outflow Mobility data (i, j), SafeGraph Intraflow Mobility data (k, l), Restaurants Mobility data (SafeGraph POI) (m, n), Religious Mobility data (SafeGraph POI) (o, p), Schools Mobility data (SafeGraph POI) (q, r), and Grocery Stores Mobility data (SafeGraph POI) (s, t).



(a) CI for Apple Mobility

(b) Baseline Correlation for Apple Mobility

(c) CI for Descartes Mobility

(d) Baseline Correlation for Descartes Mobility

(e) CI for Google Mobility

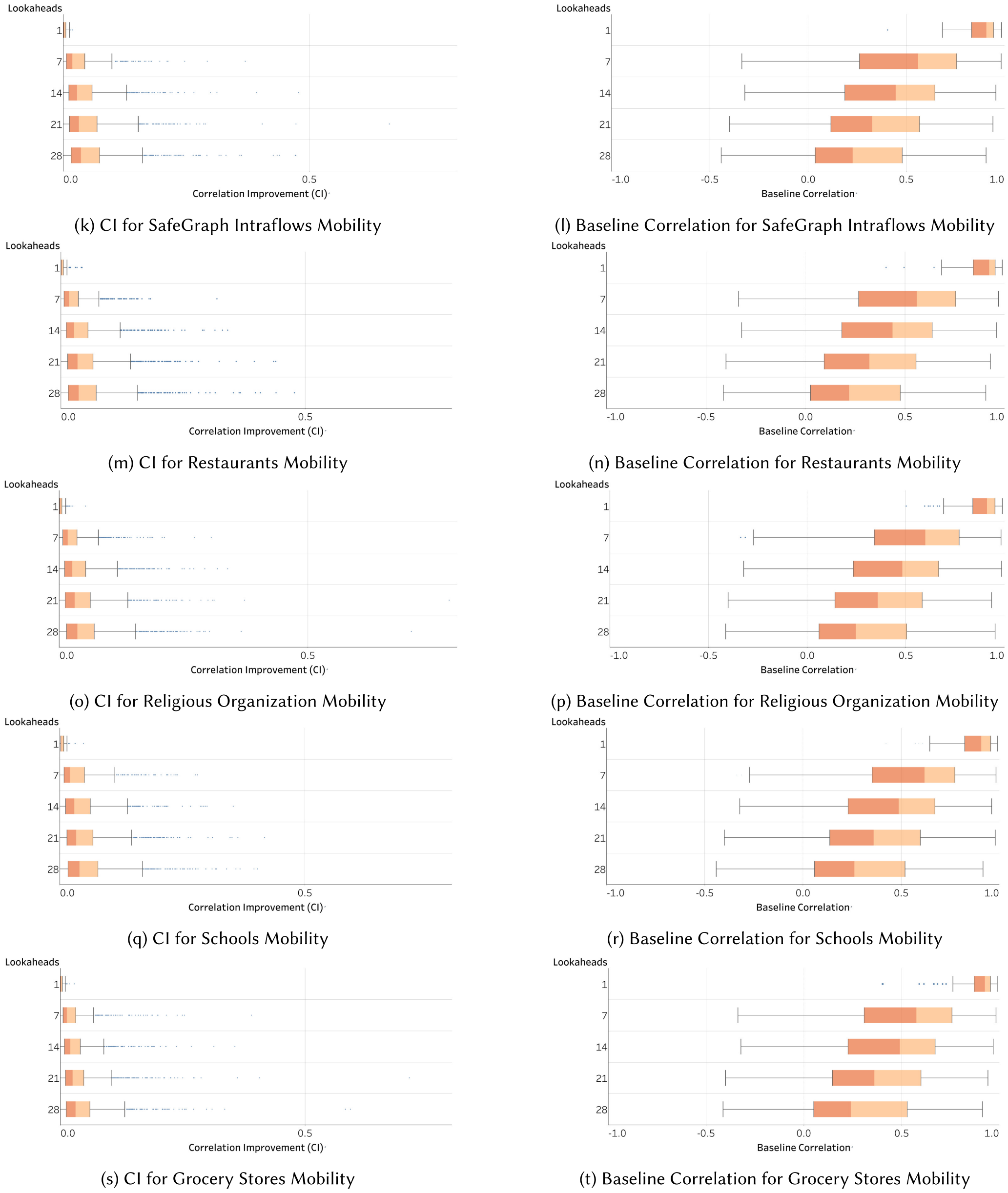
(f) Baseline Correlation for Google Mobility

(g) CI for SafeGraph Inflows Mobility

(h) Baseline Correlation for SafeGraph Inflows Mobility

(i) CI for SafeGraph Outflows Mobility

(j) Baseline Correlation for SafeGraph Outflows Mobility



(k) CI for SafeGraph Intraflows Mobility

(l) Baseline Correlation for SafeGraph Intraflows Mobility

(m) CI for Restaurants Mobility

(n) Baseline Correlation for Restaurants Mobility

(o) CI for Religious Organization Mobility

(p) Baseline Correlation for Religious Organization Mobility

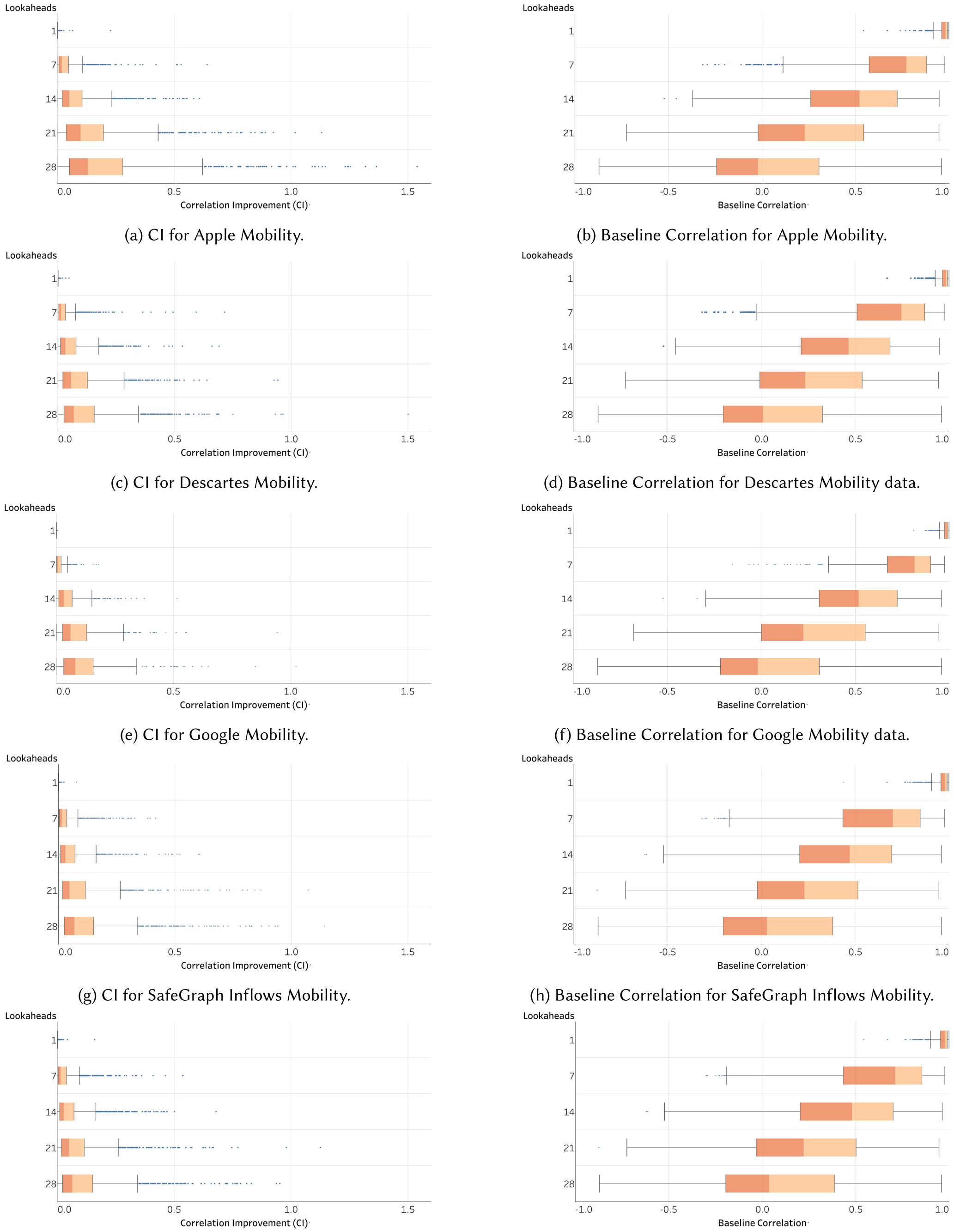
(q) CI for Schools Mobility

(r) Baseline Correlation for Schools Mobility

(s) CI for Grocery Stores Mobility

(t) Baseline Correlation for Grocery Stores Mobility

Fig. 8. Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with STW approach using Apple Mobility data (a, b), Descartes Mobility data (c, d), Google Mobility data (e, f), SafeGraph Inflow Mobility data (g, h), SafeGraph Outflow Mobility data (i, j), SafeGraph Intraflow Mobility data (k, l), Restaurants Mobility data (SafeGraph POI) (m, n), Religious Mobility data (SafeGraph POI) (o, p), Schools Mobility data (SafeGraph POI) (q, r), and Grocery Stores Mobility data (SafeGraph POI) (s, t).



(a) CI for Apple Mobility.

(b) Baseline Correlation for Apple Mobility.

(c) CI for Descartes Mobility.

(d) Baseline Correlation for Descartes Mobility data.

(e) CI for Google Mobility.

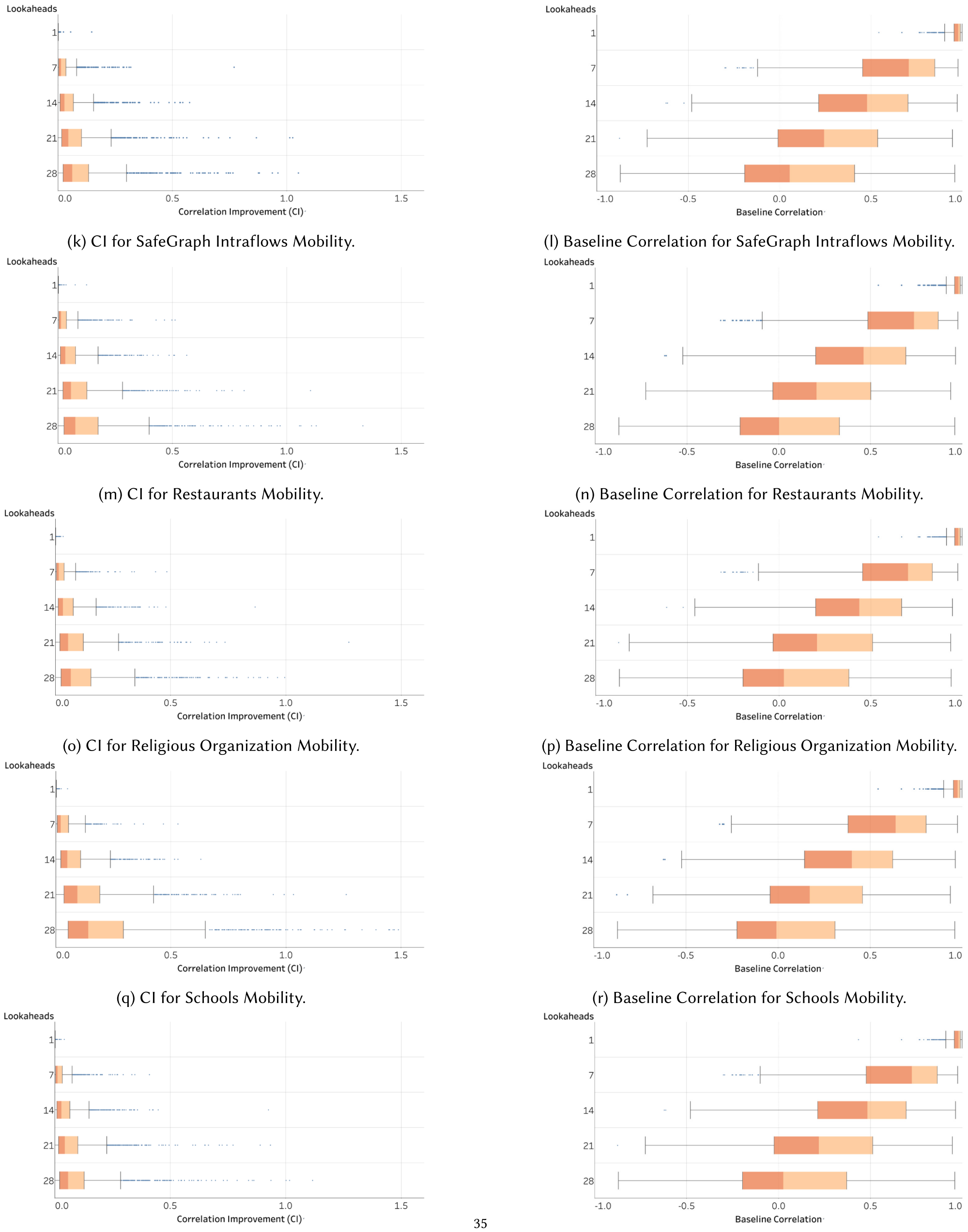
(f) Baseline Correlation for Google Mobility data.

(g) CI for SafeGraph Inflows Mobility.

(h) Baseline Correlation for SafeGraph Inflows Mobility.

(i) CI for SafeGraph Outflows Mobility.

(j) Baseline Correlation for SafeGraph Outflows Mobility.



(k) CI for SafeGraph Intraflows Mobility.

(l) Baseline Correlation for SafeGraph Intraflows Mobility.

(m) CI for Restaurants Mobility.

(n) Baseline Correlation for Restaurants Mobility.

(o) CI for Religious Organization Mobility.

(p) Baseline Correlation for Religious Organization Mobility.

(q) CI for Schools Mobility.

(r) Baseline Correlation for Schools Mobility.

(s) CI for Grocery Stores Mobility.

(t) Baseline Correlation for Grocery Stores Mobility.

Fig. 9. Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for elastic regression with the STW approach using Apple Mobility data (a, b), Descartes Mobility data (c, d), Google Mobility data (e, f), SafeGraph Inflow Mobility data (g, h), SafeGraph Outflow Mobility data (i, j), SafeGraph Intraflow Mobility data (k, l), Restaurants Mobility data (SafeGraph POI) (m, n), Religious Mobility data (SafeGraph POI) (o, p), Schools Mobility data (SafeGraph POI) (q, r), and Grocery Stores Mobility data (SafeGraph POI) (s, t).

Table 14. Summary of the Combinations of p, d, q Values Identified for the ARIMAX Models Using Grid Search and the Akaike Information Criterion

ARIMAX (p,d,q)	# Counties	ARIMAX (p,d,q)	# Counties
(0, 0, 0)	13	(4, 0, 4)	3
(0, 1, 0)	1,263	(4, 1, 0)	19
(0, 1, 1)	189	(4, 1, 1)	10
(0, 1, 2)	18	(4, 1, 2)	18
(0, 1, 3)	4	(4, 1, 3)	11
(0, 1, 4)	8	(4, 1, 4)	3
(0, 1, 5)	12	(4, 1, 5)	1
(0, 2, 0)	1	(4, 2, 0)	1
(0, 2, 1)	25	(4, 2, 1)	1
(0, 2, 2)	5	(4, 2, 2)	1
(0, 2, 3)	3	(4, 2, 3)	1
(1, 0, 0)	108	(4, 2, 4)	1
(1, 0, 1)	25	(5, 0, 0)	2
(1, 0, 2)	4	(5, 0, 1)	4
(1, 0, 3)	3	(5, 0, 2)	7
(1, 0, 4)	3	(5, 1, 0)	4
(1, 0, 5)	1	(5, 1, 1)	3
(1, 1, 0)	201	(5, 1, 2)	10
(1, 1, 1)	263	(5, 1, 3)	9
(1, 1, 2)	44	(5, 1, 4)	5
(1, 1, 3)	22	(5, 2, 2)	2
(1, 1, 4)	9	(5, 2, 3)	3
(1, 1, 5)	9	(5, 2, 5)	1
(1, 2, 1)	2	(6, 0, 3)	1
(1, 2, 2)	6	(6, 1, 3)	2
(1, 2, 3)	3	(6, 1, 4)	1
(1, 2, 4)	2	(6, 1, 5)	3
(2, 0, 0)	31	(7, 0, 2)	1
(2, 0, 1)	73	(7, 0, 3)	1
(2, 0, 2)	18	(7, 1, 0)	17
(2, 0, 3)	5	(7, 1, 1)	3
(2, 0, 4)	3	(7, 1, 2)	5
(2, 1, 0)	100	(7, 1, 3)	2
(2, 1, 1)	59	(7, 1, 4)	2
(2, 1, 2)	103	(7, 1, 5)	1
(2, 1, 3)	19	(7, 2, 1)	2
(2, 1, 4)	15	(7, 2, 3)	2
(2, 1, 5)	15	(8, 0, 0)	3
(2, 2, 1)	7	(8, 0, 2)	1
(2, 2, 3)	2	(8, 1, 0)	5
(2, 2, 4)	1	(8, 1, 1)	4
(2, 2, 5)	1	(8, 1, 2)	2
(3, 0, 0)	9	(8, 1, 3)	2
(3, 0, 1)	6	(8, 1, 4)	2
(3, 0, 2)	24	(9, 0, 0)	3
(3, 0, 3)	4	(9, 0, 1)	1
(3, 0, 4)	5	(9, 0, 3)	1
(3, 0, 5)	1	(9, 1, 1)	3
(3, 1, 0)	32	(9, 1, 2)	1
(3, 1, 1)	19	(9, 1, 4)	1
(3, 1, 2)	36	(10, 0, 0)	2
(3, 1, 3)	23	(10, 0, 4)	1
(3, 1, 4)	14	(10, 1, 1)	1
(3, 1, 5)	12	(10, 1, 2)	2
(3, 2, 0)	2	(10, 1, 3)	1
(3, 2, 1)	1	(11, 0, 0)	1
(3, 2, 2)	2	(11, 1, 0)	1
(3, 2, 5)	1	(12, 0, 3)	1
(4, 0, 0)	7	(12, 1, 0)	1
(4, 0, 1)	6	(12, 1, 1)	1
(4, 0, 2)	5	(14, 1, 0)	1
(4, 0, 3)	3	(15, 1, 0)	1

The table represents the number of counties whose p, d, q values are the ones listed in the left column.

REFERENCES

[1] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine* 18, 3 (2018), 91–93.

[2] Hamada S. Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M. Squire, and Lauren M. Gardner. 2020. Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *Lancet Infectious Diseases* 20, 11 (2020), 1247–1254.

[3] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux. 2015. Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports* 5, 1 (2015), 1–5.

[4] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589, 7840 (2021), 82–87.

[5] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David B. Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589 (2021), 82–87. https://doi.org/10.1038/s41586-020-2923-3

[6] Serina Chang, Mandy L. Wilson, Bryan Lewis, Zakaria Mehrab, Komal K. Dudakiya, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, Madhav Marathe, and Jure Leskovec. 2021. Supporting COVID-19 policy response with large-scale mobility-based modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2632–2642.

[7] Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E. Ho. 2021. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[8] David S. Curtis, Alessandro Rigolon, Dorothy L. Schmalz, and Barbara B. Brown. 2022. Policy and environmental predictors of park visits during the first months of the COVID-19 pandemic: Getting out while staying in. *Environment and Behavior* 54, 2 (2022), 487–515.

[9] Tiago Tiburcio da Silva, Rodrigo Francisquini, and Mariá C. V. Nascimento. 2021. Meteorological and human mobility data on predicting COVID-19 cases by a novel hybrid decomposition method with anomaly detection analysis: A case study in the capitals of Brazil. *Expert Systems with Applications* 182 (2021), 115190.

[10] Luzhao Feng, Ting Zhang, Qing Wang, Yiran Xie, Zhibin Peng, Jiandong Zheng, Ying Qin, Muli Zhang, Shengjie Lai, Dayan Wang, Zijian Feng, Zhongjie Li, and George F. Gao. 2021. Impact of COVID-19 outbreaks and interventions on influenza in China and the United States. *Nature Communications* 12, 1 (2021), 1–8.

[11] Vanessa Frias-Martinez, Victor Soto, Jesus Virseda, and Enrique Frias-Martinez. 2012. Computing cost-effective census maps from cell phone traces. In *Proceedings of the Workshop on Pervasive Urban Applications*.

[12] Vanessa Frias-Martinez and Jesus Virseda. 2013. Cell phone analytics: Scaling human behavior studies into the millions. *Information Technologies & International Development* 9, 2 (2013), 35–50.

[13] Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. 2010. Socio-economic levels and human mobility. In *Proceedings of the Qual Meets Quant Workshop (QMQ'10)*. 1–6.

[14] Vanessa Frias-Martinez, Jesus Virseda, and Aldo Gomero. 2012. Mobilizing education: Evaluation of a mobile learning tool in a low-income school. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 441–450.

[15] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. 2022. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. *Scientific Reports* 12, 1 (2022), 1–18.

[16] Cheng Fu, Grant McKenzie, Vanessa Frias-Martinez, and Kathleen Stewart. 2018. Identifying spatiotemporal urban activities through linguistic signatures. *Computers, Environment and Urban Systems* 72 (2018), 25–37.

[17] Santi García-Cremades, Juan Morales-García, Rocío Hernández-Sanjaime, Raquel Martínez-España, Andrés Bueno-Crespo, Enrique Hernández-Orallo, José J. López-Espín, and José M. Cecilia. 2021. Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. *Scientific Reports* 11, 1 (2021), 1–16.

[18] Oliver Gatalo, Katie K. Tseng, Alisa Hamilton, Gary Lin, and Eili Y. Klein, for the CDC MInD-Healthcare Program. 2021. Associations between phone mobility data and COVID-19 cases. *Lancet Infectious Diseases* 21, 5 (2021, E111. https://doi.org/10.1016/s1473-3099(20)30725-8

[19] Jay Ghurye, Gautier Krings, and Vanessa Frias-Martinez. 2016. A framework to model human behavior at large scale during natural disasters. In *Proceedings of the 2016 17th IEEE International Conference on Mobile Data Management (MDM'16)*, Vol. 1. IEEE, Los Alamitos, CA, 18–27.

[20] Grace Guan, Yotam Dery, Matan Yechezkel, Irad Ben-Gal, Dan Yamin, and Margaret L. Brandeau. 2021. Early detection of COVID-19 outbreaks using human mobility data. *PLoS One* 16, 7 (2021), e0253865.

[21] Marco Hernandez, Lingzi Hong, Vanessa Frias-Martinez, Andrew Whitby, and Enrique Frias-Martinez. 2017. *Estimating Poverty Using Cell Phone Data: Evidence from Guatemala*. World Bank Policy Research Working Paper 7969. World Bank, Washington, DC.

[22] Lingzi Hong, Enrique Frias-Martinez, and Vanessa Frias-Martinez. 2016. Topic models to infer socio-economic maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[23] Lingzi Hong and Vanessa Frias-Martinez. 2020. Modeling and predicting evacuation flows during Hurricane Irma. *EPJ Data Science* 9, 1 (2020), 29.

[24] Lingzi Hong, Cheng Fu, Paul Torrens, and Vanessa Frias-Martinez. 2017. Understanding citizens' and local governments' digital communications during natural disasters: The case of snowstorms. In *Proceedings of the 2017 ACM Web Science Conference*. 141–150.

[25] Xiao Hou, Song Gao, Qin Li, Yuhao Kang, Nan Chen, Kaiping Chen, Jinmeng Rao, Jordan S. Ellenberg, and Jonathan A. Patz. 2021. Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences* 118, 24 (2021), e2020524118.

[26] Rob J. Hyndman and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. OTexts.

[27] Cornelia Ilin, Sébastien Annan-Phan, Xiao Hui Tai, Shikhar Mehra, Solomon Hsiang, and Joshua E. Blumenstock. 2021. Public mobility data enables COVID-19 forecasting and management at local and global scales. *Scientific Reports* 11, 1 (2021), 1–11.

[28] Sibren Isaacman, Vanessa Frias-Martinez, and Enrique Frias-Martinez. 2018. Modeling human migration patterns during drought conditions in La Guajira, Colombia. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–9.

[29] Kathleen M. Jagodnik, Forest Ray, Federico M. Giorgi, and Alexander Lachmann. 2020. Correcting under-reported COVID-19 case numbers: Estimating the true scale of the pandemic. *medRxiv*. Retrieved September 1, 2023 from https://www.medrxiv.org/content/10.1101/2020.03.14.20036178v2

[30] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jinmeng Rao, and Jake Kruse. 2020. Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic. *Scientific Data* 7, 1 (2020), 1–13.

[31] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. 2020. Examining COVID-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113* (2020).

[32] Nishant Kishore, Aimee R. Taylor, Pierre E. Jacob, Navin Vembar, Ted Cohen, Caroline O. Buckee, and Nicolas A. Menzies. 2022. Evaluating the reliability of mobility metrics from aggregated mobile phone data as proxies for SARS-CoV-2 transmission in the USA: A population-based study. *Lancet Digital Health* 4, 1 (2022), e27–e36.

[33] Cheng-Pin Kuo and Joshua S. Fu. 2021. Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. *Science of the Total Environment* 758 (2021), 144151.

[34] Zakaria Mehrab, Aniruddha Adiga, Madhav V. Marathe, Srinivasan Venkatramanan, and Samarth Swarup. 2022. Evaluating the utility of high-resolution proximity metrics in predicting the spread of COVID-19. *ACM Transactions on Spatial Systems and Algorithms* 8, 4 (2022), Article 26, 51 pages.

[35] Behnam Nikparvar, Md. Mokhlesur Rahman, Faizeh Hatami, and Jean-Claude Thill. 2021. Spatio-temporal prediction of the COVID-19 pandemic in US counties: Modeling with a deep LSTM neural network. *Scientific Reports* 11, 1 (2021), 1–12.

[36] Nicola Perra. 2021. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Physics Reports* 913 (2021), 1–52.

[37] Troy Quast and Ross Andel. 2021. Excess mortality associated with COVID-19 by demographic group: Evidence from Florida and Ohio. *Public Health Reports* 136, 6 (2021), 782–790.

[38] Chotirat Ann Ralanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das. 2005. Mining time series data. In *Data Mining and Knowledge Discovery Handbook*. Springer, 1069–1103.

[39] Peter Richterich. 2020. Severe underestimation of COVID-19 case numbers: Effect of epidemic growth rate and test restrictions. *medRxiv*. Retrieved September 1, 2023 from https://www.medrxiv.org/content/10.1101/2020.04.13.20064220v1

[40] Weston C. Roda, Marie B. Varughese, Donglin Han, and Michael Y. Li. 2020. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling* 5 (2020), 271–281.

[41] Alberto Rubio, Vanessa Frias-Martinez, Enrique Frias-Martinez, and Nuria Oliver. 2010. Human mobility in advanced and developing economies: A comparative analysis. In *Proceedings of the 2010 AAAI Spring Symposium Series*.

[42] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[43] Frank Schlosser, Vedran Sekara, Dirk Brockmann, and Manuel Garcia-Herranz. 2021. Biases in human mobility data impact epidemic modeling. *arXiv preprint arXiv:2112.12521* (2021).

[44] Deshun Sun, Xiaojun Long, and Jingxiang Liu. 2022. Modeling the COVID-19 epidemic with multi-population and control strategies in the United States. *Frontiers in Public Health* 9 (2022), 751940.

[45] Zander S. Venter, Adam Sadilek, Charlotte Stanton, David N. Barton, Kristin Aunan, Sourangsu Chowdhury, Aaron Schneider, and Stefano Maria Iacus. 2021. Mobility in blue-green spaces does not predict COVID-19 transmission: A global analysis. *International Journal of Environmental Research and Public Health* 18, 23 (2021), 12567.

[46] Marcos R. Vieira, Enrique Frias-Martinez, Petko Bakalov, Vanessa Frias-Martinez, and Vassilis J. Tsotras. 2010. Querying spatio-temporal patterns in mobile phone-call databases. In *Proceedings of the 2010 11th International Conference on Mobile Data Management*. IEEE, Los Alamitos, CA, 239–248.

[47] Lijing Wang, Xue Ben, Aniruddha Adiga, Adam Sadilek, Ashish Tendulkar, Srinivasan Venkatramanan, Anil Vullikanti, Gaurav Aggarwal, Alok Talekar, Jiangzhuo Chen, Bryan Lewis, Samarth Swarup, Amol Kapoor, Milind Tambe, and Madhav Marathe. 2020. Using mobility data to understand and forecast COVID19 dynamics. *medRxiv*. Retrieved September 1, 2023 from https://www.medrxiv.org/content/10.1101/2020.12.13.20248129v1

[48] Amy Wesolowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. 2012. Quantifying the impact of human mobility on malaria. *Science* 338, 6104 (2012), 267–270.

[49] Jiahui Wu, Saad Mohammad Abrar, Naman Awasthi, Enrique Frias-Martinez, and Vanessa Frias-Martinez. 2022. Enhancing short-term crime prediction with human mobility flows and deep learning architectures. *EPJ Data Science* 11, 1 (2022), 53.

[50] Jiahui Wu, Saad Mohammad Abrar, Naman Awasthi, and Vanessa Frías-Martínez. 2023. Auditing the fairness of place-based crime prediction models implemented with deep learning approaches. *Computers, Environment and Urban Systems* 102 (2023), 101967.

[51] Jiahui Wu, Enrique Frias-Martinez, and Vanessa Frias-Martinez. 2021. Spatial sensitivity analysis for urban hotspots using cell phone traces. *Environment and Planning B: Urban Analytics and City Science* 48, 9 (2021), 2517–2519.

[52] Neo Wu, Xue Ben, Bradley Green, Kathryn Rough, Srinivasan Venkatramanan, Madhav Marathe, Paul Eastham, Adam Sadilek, and Shawn O'Banion. 2020. Predicting onset of COVID-19 with mobility-augmented SEIR model. *medRxiv*. Retrieved September 1, 2023 from https://www.medrxiv.org/content/10.1101.2020.07.27.20159996v2

[53] Nazar Zaki and Elfadil A. Mohamed. 2021. The estimations of the COVID-19 incubation period: A scoping reviews of the literature. *Journal of Infection and Public Health* 14, 5 (2021), 638–646.

[54] Choujun Zhan, Yufan Zheng, Zhikang Lai, Tianyong Hao, and Bing Li. 2021. Identifying epidemic spreading dynamics of COVID-19 by pseudocoevolutionary simulated annealing optimizers. *Neural Computing and Applications* 33, 10 (2021), 4915–4928.

Received 1 February 2023; revised 9 June 2023; accepted 18 July 2023