Predicting County-Level COVID-19 Cases using Spatiotemporal Machine Learning: Modeling Human Interactions using Social Media and Cell-Phone Data

Behzad Vahedi (behzad@colorado.edu)
University of Colorado Boulder https://orcid.org/0000-0001-5782-3831
Morteza Karimzadeh
University of Colorado Boulder
Hamidreza Zoraghein
Social and Behavioral Science Research, Population Council

Article

Keywords: spatiotemporal machine learning, COVID-19

Posted Date: February 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-203188/v1

Version of Record: A version of this preprint was published at Nature Communications on November 8th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26742-6.

Predicting County-Level COVID-19 Cases using Spatiotemporal Machine Learning: Modeling Human Interactions using Social Media and Cell-Phone Data

Behzad Vahedi[1*], Morteza Karimzadeh[1,] Hamidreza Zoraghein[2]

[1] Department of Geography, University of Colorado Boulder; Behzad@colorado.edu; karimzadeh@colorado.edu

[2] Social and Behavioral Science Research, Population Council, New York, USA; hzoraghein@popcouncil.org

Abstract

Measurements of human interaction through proxies such as social connectedness or movement patterns have proved useful for predictive modeling of COVID-19. In this study, we first compare the power of Facebook's social connectedness with cell phone-derived human mobility for predicting county-level new cases of COVID-19. Our experiments show that social connectedness is a better proxy for measuring human interactions leading to new infections. Next, we develop a SpatioTemporal autoregressive eXtreme Gradient Boosting (STXGB) model to predict county-level new cases of COVID-19 in the coterminous US. We

evaluate the model on five weekly forecast dates between October 24 and November 28, 2020 over one- to four-week prediction horizons. Comparing our predictions with a baseline Ensemble of 32-models currently used by the CDC indicates an average 58% improvement in prediction RMSEs over two- to four-week prediction horizons, pointing to the strong predictive power of our model.

1. Introduction

Human interaction in close physical proximity is the primary cause of the transmission of highly contagious diseases such as COVID-19[1]. Measuring human interaction is therefore an important step in understanding and predicting the spread of COVID-19[2,3]. However, tracking human interactions requires rigorous contact tracing at national and regional scales which has not been implemented in the United States due to the economic, legal, and sociocultural concerns, as well as inadequate testing supplies, and insufficient national coordination[4].

As a result, researchers have adopted different proxies to track human interaction levels. One such proxy is the "Social Connectedness Index" (SCI), generated from Facebook's friendship data. SCI represents the probability that two users in a pair of regions (e.g., U.S. counties) are friends (i.e., connected) on Facebook[5]. Kuchler et al.[6] reported on the strong correlation between early

hotspots of COVID-19 outbreak and their level of social connectedness. The underlying assumption in leveraging SCI as a proxy for physical human interactions is that individuals who are socially connected on Facebook have a higher probability for physical interaction, thereby, potentially contributing to the spread of communicable diseases.

Human mobility flow, as measured by anonymized cell phone data, serves as another proxy for quantifying human interactions[7,8]. Widely used to study the spread of COVID-19, most studies incorporating cell-phone data have focused on the change in mobility within a spatial unit[9,10], while a few others have also incorporated the flow between different spatial units[11] to predict transmissions across units, albeit mostly in the early stages of the pandemic with limited evaluation data. The underlying assumption in this approach is that more movements between spatial units leads to higher interactions, and consequently, an elevated risk of disease spread.

It is unclear, however, which of these approaches—using social media connectedness versus cell phone-derived human mobility flow—is a better indicator of physical interaction within and between different regions. Furthermore, the underlying assumption in each approach may not necessarily be valid in the case of COVID-19: considering the sporadic and regional stay-at-home

orders across the United States, social connectedness may not lead to physical interaction, at least not to the same level as pre-pandemic. Similarly, given the recommended preventative measures such as mask-wearing and physical distancing[12], human flow from one location to another may not necessarily lead to physical interactions that could communicate the disease, especially in public places, where preventative measures are enforced more strictly.

In this paper, we compare the predictive power of Facebook's social connectedness index, as an example of social media proxy, with cell phone-derived human mobility data, as an example of human flow proxy, in forecasting county-level new cases of COVID-19 in the conterminous US over multiple prediction horizons. County-level prediction is more challenging than state-level prediction[13–15], yet it serves as the highest spatial resolution for national models in the U.S., since cases are aggregated and reported at the county level. Longer term County-level predictions are also essential for policy making and resource allocation.

The unique characteristics of COVID-19, including its presymptomatic and asymptomatic contagiousness, rapid spread, along with variations in regional response policies, such as inconsistent and sporadic testing and contact tracing, make forecasting the spatial patterns of this disease challenging. Researchers

have used a variety of methods including time-series autoregressive models[16–18], machine learning techniques[19–21], epidemiologic models such as SIR model and its variants[22,23], and combinations of these methods[24] for forecasting COVID-19. We experiment with five different machine learning-based spatiotemporal autoregressive algorithms to perform county-level predictions, and use the best algorithm, i.e. the one with the lowest average prediction RMSE and MAE, to compare between Facebook- and cell phone-derived features.

We compare our best model predictions against one of the most prominent collective efforts in forecasting COVID-19 in the U.S., namely, the Ensemble model developed by the "COVID-19 Forecast Hub" team[25] which is used by the Centers for Disease Control and Prevention (CDC) to report predictions of new cases and deaths in U.S. counties in one- to four-weeks ahead horizons[25,26].

Our specific contributions are as follows: (a) designing inter-county and intra-county features for spatiotemporal autoregressive machine learning of county-level new cases, (b) comparing the performance of social media connectedness (derived from Facebook) and human flow connectedness (derived from SafeGraph's cell phone data) by incorporating inter-county spatial lags for predicting county-level new COVID-19 cases, and (c) improving the long-term

prediction of county-level new cases of COVID-19 in the coterminous U.S. in comparison to a baseline Ensemble model, using an end-to-end model.

2. Results

Algorithm Selection

Five different machine learning algorithms were trained and tuned using each set of features named in the next section (details in Section 4), and tested over the last 5 weeks of our dataset (same dates as Tables 2 & 5), by holding out one week at a time for testing. Table 1 reports the average performance for each algorithm. EXtreme Gradient Boosting (XGB) performed better on unseen data compared with other tree-based ensemble algorithms and the neural networks, including Feed Forward Neural Network (FFNN) and Long Short-Term Memory (LSTM) network (Table 1). Therefore, we used XGB for developing short-term and long-term prediction models. The RMSE and MAE values reported in Table 1 are for the natural log values of [new cases per 10k population + 1], which we used as a transformed target variable in the models, given the skewed distribution of new cases (or new cases per 10k) in counties.

Table 1. Performance comparison of machine learning regressors. The best performance in each category is bolded.

Model	RMSE Train	MAE Train	RMSE Test	MAE Test
Random Forest	0.486	0.359	0.511	0.38
Stochastic Gradient Boosting	**0.438**	**0.313**	0.494	0.348
Extreme Gradient Boosting	0.441	0.316	**0.47**	**0.330**
Feed Forward Neural Network	0.524	0.391	0.566	0.438
Long Short-Term Memory	1.179	0.964	1.209	1.007

Comparing social media- and cell phone-derived features

To compare the relative strength of Facebook-derived movement and connectedness against SafeGraph-derived human mobility flows, as proxies for physical human interactions, we designed a set of intra-county and inter-county interaction features using each proxy, and incorporated each set of features separately to develop spatiotemporally lagged autoregressive prediction models of *new cases of COVID-19* (i.e. target variable). We then compared the predictions of these models against each other as well as a base model (not to be confused with the baseline model for final evaluations), all of which were trained using the XGB algorithm.

Our base model incorporates a series of socioeconomic, demographic and temperature variables, as well as *temporal* lags of the target variable in the same

county only, thus, we call it Temporal XGB (TGXB), whereas the SpatioTemporal XGB (STXGB) models , in addition to temporal lags, also incorporate intra-county movement features and spatiotemporal lags of the target variable weighted by the inter-county connectedness strength. Specifically, the spatial lags in STXGB are calculated by multiplying the target variable (natural log of weekly new cases per 10k population + 1) in "connected counties" by either (a) inter-county Facebook Social Media Connectedness Index, (in the STXGB-FB model), or (b) inter-county Flow Connectedness Index derived from SafeGraph's cell-phone movement data, forming STXGB-SG and STXGB-SGR models (described in detail in Section 4).

Table 2 and Fig. 1 present the error values of predicted new cases and new cases per 10k population in the one-week prediction horizon using the TXGB and STXGB models. The incorporation of spatiotemporal lags using county connectedness indices (in STXGB) was advantageous across the board, compared to the temporal lags only (TXGB). All variants of STXGB (-FB, -SG, and -SGR) achieved lower errors compared to TXGB. Furthermore, STXGB-FB, which uses the Facebook-derived features, outperformed all other models in average RMSEs and MAEs as well as on all forecast dates, except the fifth date when the STXGB-SG model generated slightly lower errors.

Table 2. RMSE and MAE of county-level predicted weekly new cases and new cases per 10k population. Lowest values of each error metric are highlighted. Average values across forecasting dates for each model is bold faced.

	Model	Forecast Date	RMSE New Case Prediction	MAE New Case Prediction	RMSE New Case/10k Prediction	MAE New Case/10k Prediction
including temporal lags	Base Model (TXGB)	10/24/2020	136.255	30.894	16.084	6.134
		10/31/2020	192.993	50.91	22.319	11.176
		11/07/2020	203.678	70.689	22.899	12.907
		11/14/2020	237.113	80.1	26.45	15.104
		11/21/2020	166.855	50.684	16.384	9.611
		Average	**187.379**	**56.655**	**20.827**	**10.986**
including spatiotemporal lags	STXGB with Facebook-derived features (STXGB-FB)	10/24/2020	116.312	25.909	15.083	5.708
		10/31/2020	172.582	46.398	21.938	10.817
		11/07/2020	169.602	54.613	20.925	11.072
		11/14/2020	185.391	62.243	23.263	12.477
		11/21/2020	142.312	48.625	16.297	9.373
		Average	**157.24**	**47.557**	**19.501**	**9.89**
	STXGB with SafeGraph-derived features (STXGB-SG)	10/24/2020	120.049	27.312	15.101	5.785
		10/31/2020	195.03	50.487	22.267	11.171
		11/07/2020	193.263	62.506	21.156	11.421
		11/14/2020	203.675	68.962	24.423	13.371
		11/21/2020	140.748	48.383	16.683	9.596
		Average	**170.553**	**51.53**	**19.926**	**10.269**
	STXGB with SafeGraph-derived features-rich (STXGB-SGR)	10/24/2020	122.482	27.952	15.362	5.739
		10/31/2020	207.411	54.531	22.626	11.541
		11/07/2020	178.934	60.061	20.998	11.324
		11/14/2020	186.997	64.745	24.079	13.032
		11/21/2020	141.356	47.491	16.758	9.539
		Average	**167.436**	**50.956**	**19.965**	**10.235**

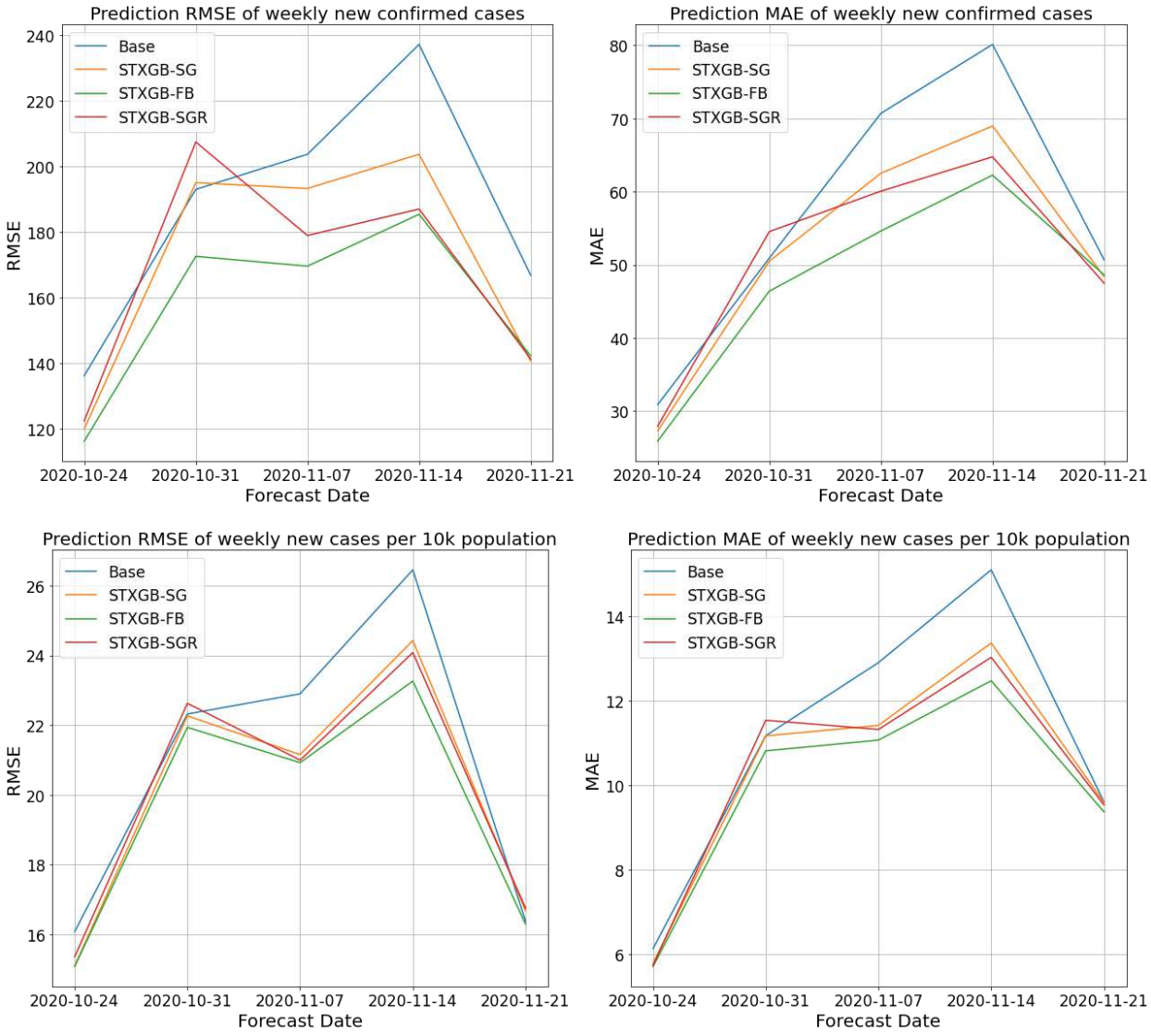

Figure 1. ***Prediction errors of the SpatioTemporal eXtreme Gradient Boosting (STXGB) models in one-week prediction horizon against a temporal autoregressive model without spatial lags (TXGB).*** *Error comparison of STXGB models with different feature sets in predicting weekly new cases (top row) and new cases per 10k population (bottom row) for one-week ahead horizon. STXGB-FB, which incorporates Facebook-derived features, including spatial lags based on Social Connectedness Index, outperforms other models. Left column: prediction RMSE. Right column: prediction MAE.*

Long-term predictions and evaluation against the COVID-19 Forecast Hub Ensemble

We compared the predictions of our best model, STXGB-FB, against the predictions of the COVID-19 Forecast Hub Ensemble of 32 models (used by the CDC in reporting forecasts of new cases[26]) over one-, two-, three-, and four-week horizons. We trained and tuned STXGB-FB for each prediction horizon separately. We then used the reported new cases by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)[27] as ground-truth to calculate RMSE and MAE of each prediction set, shown in Fig. 2 and Table 3, over varying prediction horizons across the five forecast dates. Our model considerably improves RMSEs and MAEs compared with the Ensemble model in the two-week, three-week, and four-week ahead prediction horizons, with an average 58% reduction in RMSEs and 61% reduction in MAEs.

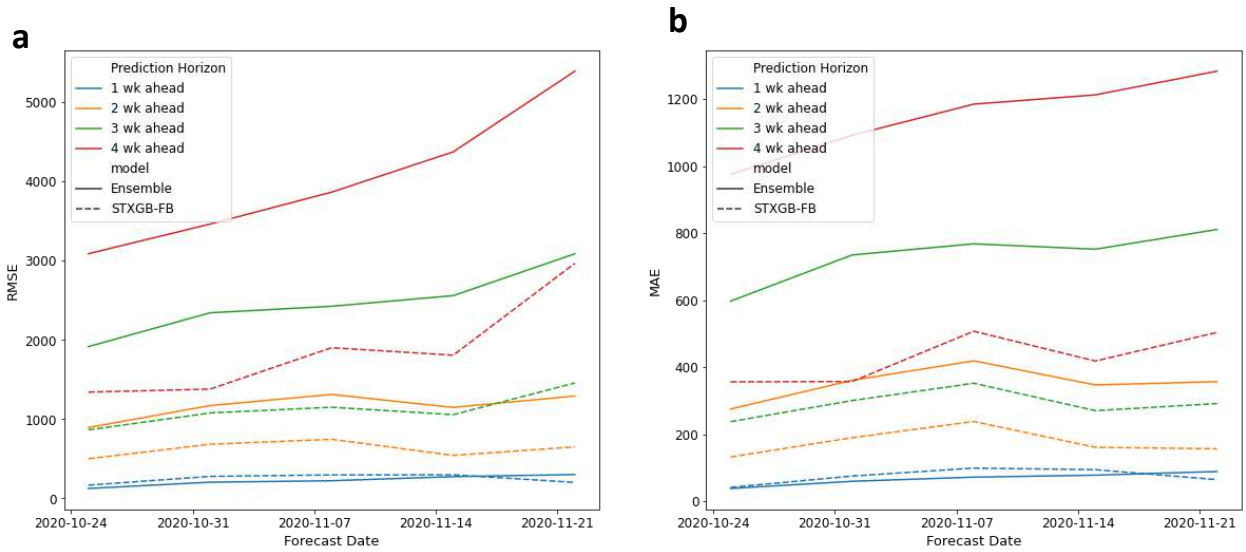

Figure 2. ***Long-term prediction error comparison.*** *Prediction errors of the STXGB-FB model (dashed line) and the Ensemble baseline (solid line) over four prediction horizons on five forecasting dates. a) Prediction RMSE and b) Prediction MAE.*

In the one-week prediction horizons, the Ensemble model outperformed our STXGB-FB model on the first three forecasting dates (Oct 24, Oct. 31, Nov. 7), but performed similarly on the Nov. 14 forecasting date. Nevertheless, the differences in the one-week horizon predictions are relatively small. STXGB-FB outperformed the Ensemble prediction on Nov. 21 across all prediction horizons. This is noteworthy since the prediction horizons (one- to four-week) on Nov. 21 overlap the post-Thanksgiving holidays in the US, which caused a surge in the number of cases[28]. In summary, out of the 20 predictions performed, the STXGB-FB model outperformed the Ensemble model in 17 of them, including in all longer than one-week prediction horizons (Table 3).

To investigate the potential of SafeGraph cell phone-derived features in long-term predictions, we generated one-, two-, three-, and four-week forecasts using the STXGB-SG model as well. This model does not perform as well as STXGB-FB, pointing to the superiority of Facebook-derived features in our models consistent with the one-week predictions (Supplementary Table 4, Supplementary Information). However, while STGXB-SG generates larger errors compared to STXGB-FB, it still outperforms the Ensemble model in long-term prediction horizons.

Table 3. Comparison of the prediction errors generated by the COVID-19 Forecast Hub Ensemble model and our STXGB-FB model in 1- to 4-week prediction horizons.

Forecast Date	Model	Prediction RMSE in Prediction Horizon			
		1 wk ahead	2 wk ahead	3 wk ahead	4 wk ahead
10/24/2020	Ensemble	126.05	894.41	1915.31	3087.32
	STXGB-FB	164.80	489.81	832.36	1157.87
	Pct. Improvement	**-30.74%**	**45.24%**	**56.54%**	**62.50%**
10/31/2020	Ensemble	205.12	1172.21	2342.40	3461.53
	STXGB-FB	247.67	554.25	892.22	1376.14
	Pct. Improvement	**-20.74%**	**52.72%**	**61.91%**	**60.24%**
11/07/2020	Ensemble	223.96	1311.84	2424.17	3864.08
	STXGB-FB	268.78	649.08	906.63	1380.22
	Pct. Improvement	**-20.01%**	**50.52%**	**62.60%**	**64.28%**
11/14/2020	Ensemble	275.24	1148.63	2558.68	4372.37
	STXGB-FB	274.35	468.53	848.68	1450.59
	Pct. Improvement	**0.32%**	**59.21%**	**66.83%**	**66.82%**
11/21/2020	Ensemble	301.87	1292.31	3086.82	5390.16
	STXGB-FB	210.64	592.83	1457.29	2646.68
	Pct. Improvement	**30.22%**	**54.13%**	**52.79%**	**50.90%**

Spatial distribution of errors

Our STXGB-FB performed worse than the baseline Ensemble model on the Oct.24, Oct. 31 and Nov. 7 forecasting dates over the one-week horizon, but considerably outperformed the Ensemble over longer-term horizons on the same forecasting dates. Further, STXGB-FB outperformed the Ensemble baseline on Nov. 14 and Nov. 21 across all prediction horizons. To find potential explanations for this inconsistency, we inspected the spatial patterns of errors. Figure 3 illustrates maps of confirmed new cases per 10k population along with prediction errors per 10k population generated by the STXGB-FB model for two forecasting dates of Oct. 31 and Nov. 7. The purple-shaded counties in the error maps are those with model underestimation of new cases, and the brown shades indicate overestimations of observed values. As can be seen in this figure, the majority of counties with high prediction errors (per 10K) are located in the rural Midwest with relatively high numbers of cases per 10k population during the November surge, albeit these are counties with fewer *total* cases compared to more populated, urban ones. It is worth noting that we use normalized (by 10K population) maps in Fig. 3, since choropleth maps would be biased by patterns of population distribution otherwise.

Figure 3 also demonstrates clusters of *apparent* underestimations in Georgia and Texas on the Oct. 31 forecasting date, followed by *apparent* overestimations in the same areas for the week after. The opposite pattern is the case for Kentucky. In the case of Georgia, the high-error clusters can almost perfectly delineate the boundary of the state. This discrepancy could be a result of lags or different policies in testing and reporting COVID-19 cases. These potential short-term lags in reporting by some states may explain why our model performs considerably better in the longer-term prediction horizons but underperforms in the one-week horizon on Oct. 31 and Nov. 7.

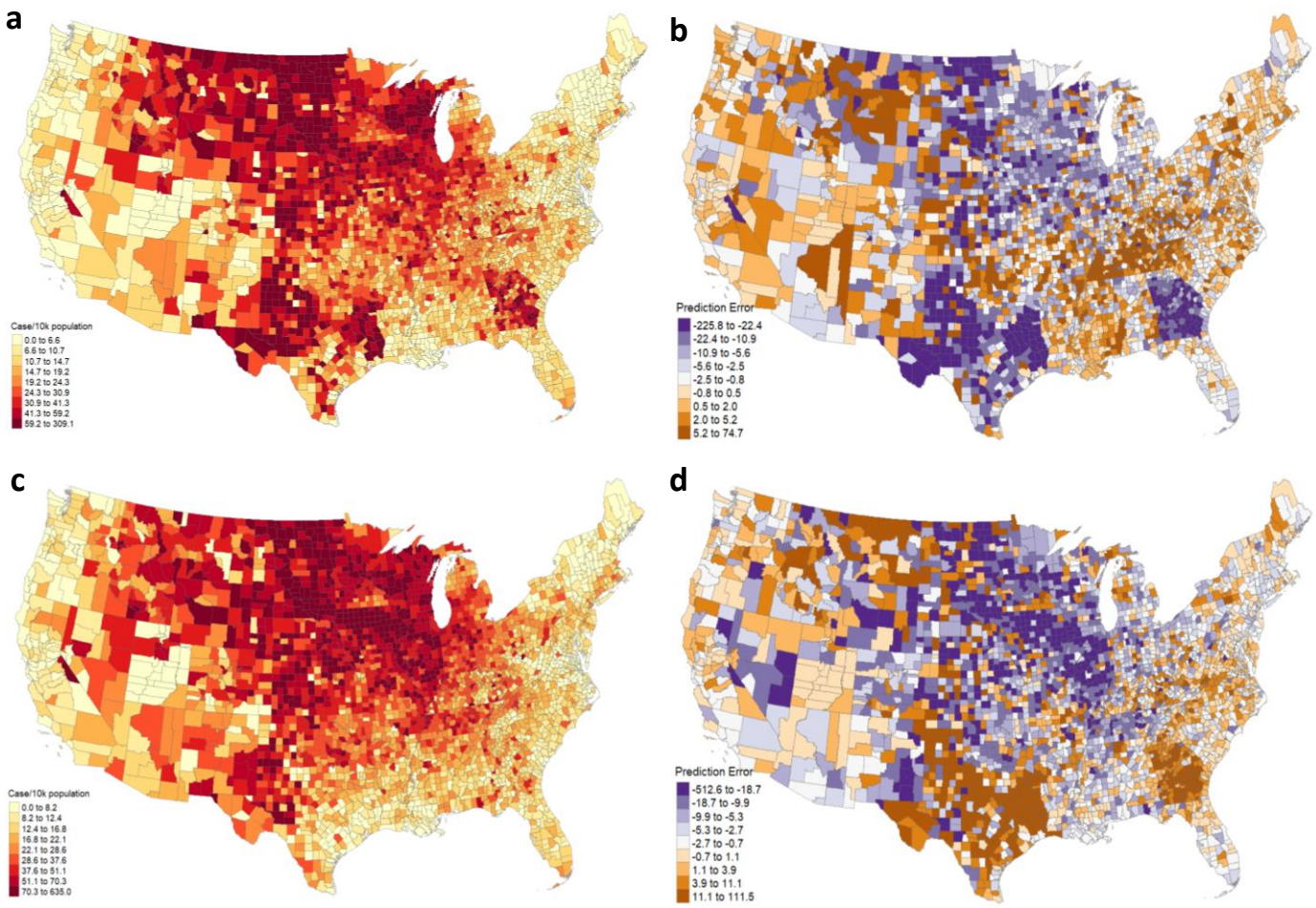

Figure 3. ***Map of COVID-19 cases per 10k population and errors in predicting them.*** *(a) number of confirmed new cases per 10k population over the week ahead of forecasting date Oct. 31, 2020 (b) prediction errors for the same forecasting date(c) number of new cases over the week ahead of Nov. 7, 2020 forecasting date (d) prediction errors for the same forecasting date. The pattern of errors in Georgia, and Texas, and Kentucky flip from Oct. 31 to Nov. 7, indicating potential lags in testing and reporting.*

The majority of counties in the U.S are rural, which are also the ones with fewer medical resources, and where social media data or cell-phone mobility data, which underlie our models, might be less representative[29–31]. To investigate our models' performance in rural-majority counties compared to the Ensemble baseline, we categorized the counties into urban- and rural- majority by

calculating an urbanization index for each county (Supplementary Information, Section B). 2391 counties (~77%) were identified as rural and 712 as urban.

We then calculated the prediction errors of the number of cases and the number of cases per 10k population for the Ensemble and STXGB-FB models in each category across four prediction horizons for the Nov. 7 forecasting date (Fig. 4). Both models generate considerably lower median errors and narrower interquartile error ranges in rural counties when predicting the total number of new cases (not normalized by population), which could be attributed to the overall higher prevalence and higher variance of COVID-19 cases in urban counties in our prediction horizons. However, the opposite is the case when predicting the number of weekly new cases *per 10k population*; both models have wider interquartile ranges in rural counties across all prediction horizons. This could be due to the overall higher prevalence of COVID-19 *per population* in rural counties during the selected prediction horizons.

As evident in Fig. 4, STXGB-FB has lower prediction errors with a narrower IQR in both urban and rural counties compared to the Ensemble model, across all prediction horizons except for the shortest one (one-week), which may be attributed to temporal fluctuations and policy variations in testing and case reporting as discussed above. The overall superior performance of STGXB-FB is

observed when predicting both the total number of new cases and new cases per 10k population, affirming the higher robustness of our model. Furthermore, the difference between the median prediction errors of STXGB-FB in urban and rural counties, when predicting the number of cases per 10k population, is smaller compared to the Ensemble model. This points to the more consistent performance of STXGB-FB in majority-rural counties, even though Facebook might not be as representative in these areas[29].

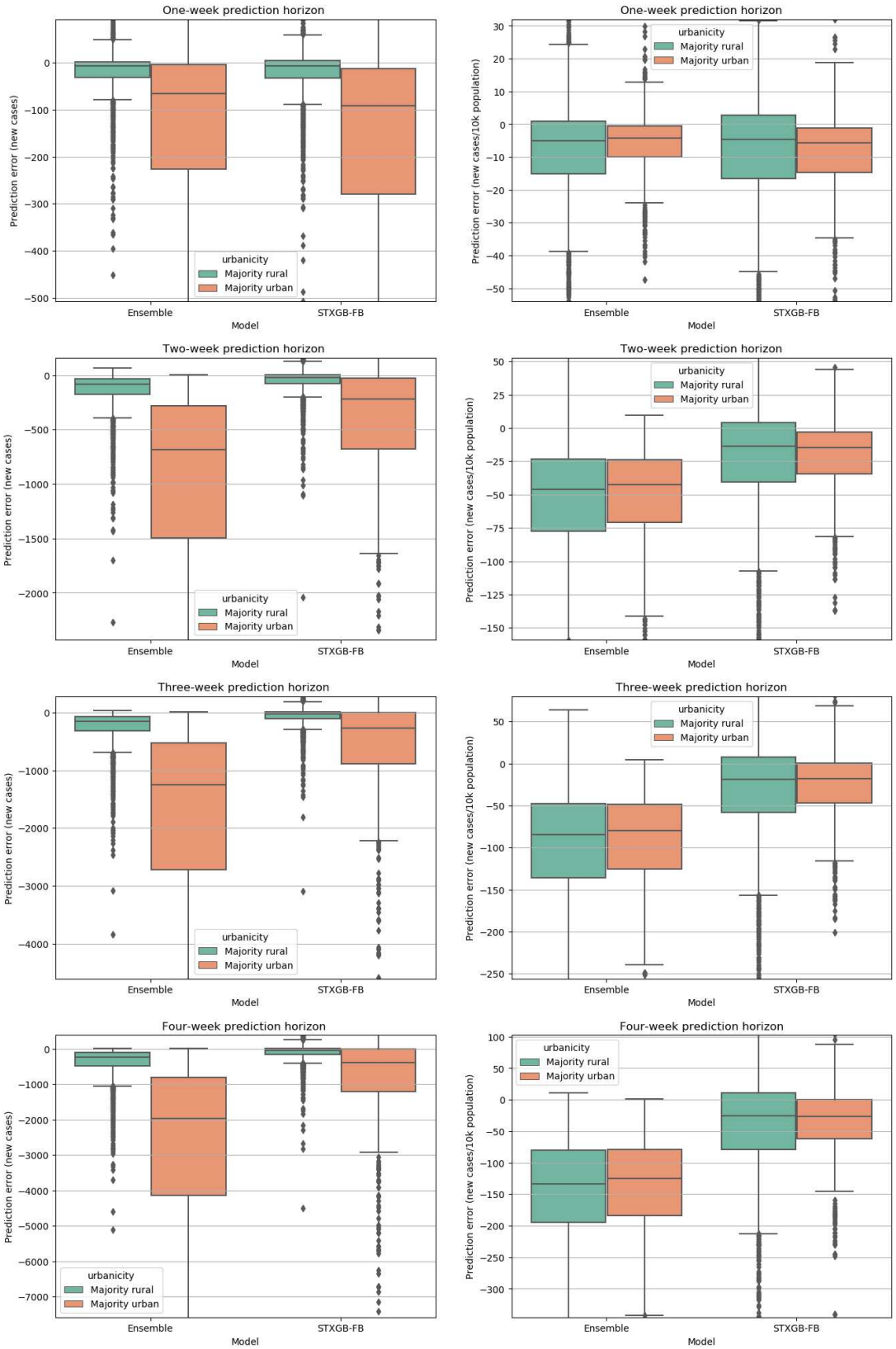

Figure 4. Prediction errors in urban vs. rural counties. *Prediction errors of the number of new cases (left column) and new cases per 10k population (right column) in rural and urban counties on the Nov. 7 forecast date across four prediction horizons. The higher and lower 3% of counties are trimmed from the plot view.*

3. Discussion

We demonstrated that incorporating (1) spatiotemporal lags using inter-county indices of connectedness and (2) intra-county measurements of movement can improve the performance of high resolution COVID-19 predictive models, especially over long-term horizons. Short-term and long-term predictions of COVID-19 cases help the federal and local governments make informed decisions such as imposing or relaxing business restrictions or planning resource allocation in response to the forecasted trends of COVID-19[32,33].

Our results showed that using Facebook-derived features implemented in our SpatioTemporal Autoregressive eXtreme Gradient Boosted (STXGB) model generate lower prediction errors across all prediction horizons compared to SafeGraph cell phone-derived features within the same architecture (Supplementary Table 5). Notably, however, to maintain compatibility, we used a formulation similar to Facebook's Social Connectedness Index when creating a corresponding index from SafeGraph data (which we call Flow Connectedness Index, refer to Section 4). This might have had adverse effects on the predictive power of the cell-phone-derived features. Nevertheless, the resulting model performs considerably better than the Ensemble baseline in long-term predictions (Supplementary Table 4). We will investigate alternative designs of inter-county

connectedness metrics from SafeGraph mobility data in the future to ensure the utilization of the full potential of this dataset.

The superior performance of Facebook-derived features means that stronger county-level social (media) connections could lead to a higher chance of (unsafe) human interaction (e.g., house parties) and thus, COVID spread, compared to human flow from one location to another. The findings of this paper also suggest that Facebook's Social Connectedness Index can be used for successful predictive modeling of COVID-19 in data-poor countries without cell-phone movement datasets, assuming that the Facebook usage in those countries is of a comparable size and representativeness to the U.S[34]. With more than 2.5 billion active users globally, Facebook provides social connectedness for many countries. Conversely, human mobility data, to the best of our knowledge, is available in far fewer numbers of countries.

The STXGB-FB and STXGB-SG models significantly outperformed the Ensemble model currently used by the CDC in predicting county-level new cases of COVID-19 in two-, three-, and four-weeks prediction horizons, with inconsistent comparisons in the one-week horizon. Our error maps suggest that this inconsistency might be partly due to inconsistent and delayed testing and reporting by some states.

The superiority of STXGB over purely temporal models (such as our TXGB) points to the importance of incorporating both within-unit (e.g., intra-county) and between-unit (e.g., inter-county) interactions when predicting a highly contagious disease such as COVID-19. Predictive models that focus on within state boundaries are likely to underperform, because after all, the disease does spread across geographic unit borderlines.

As evident in table 1, the XGB and SGB methods performed better than other machine learning-based algorithms. A potential reason for the lower performance of FFNN and LSTM is the relatively small size of the training data (34 weeks of observation at most). Neural networks' main advantage is in their ability to learn features from data[35]; however, they require higher amounts of training data compared to tree-based models for optimizing the model's parameters. Our results are a testament to the advantages of high-performance tree-based ensemble algorithms such as XGB with more limited training data, especially if features are well-engineered. Our spatiotemporal lag features provide a template for such features to improve machine learning-based predictive modeling of infectious diseases.

It is also worth noting that between XGB and SGB, SGB generated lower average training errors when using the base and SafeGraph-derived features, but

XGB outperformed SGB in testing RMSE and MAE across all 4 models (Supplementary Table 2). This is somewhat expected, as XGB uses the second-order derivatives of the loss function for optimization, and more importantly, a regularized model formalization to control over-fitting, which is otherwise a disadvantage with regression trees[36]. This regularized model results in better performance on unseen data. To ensure consistency, we ran all regression methods 10 times; and XGB had lower testing errors compared to all other regression methods in all 10 runs.

4. Methods

This section outlines the details of feature engineering, algorithm selection, and implementation of spatiotemporal autoregressive machine learning models for predicting new cases of COVID-19 in the conterminous United States. We describe our experimental setup for comparing the predictive power of Facebook-derived features and SafeGraph's cell phone-derived mobility features (as proxies for human physical interaction) for this purpose, and the evaluation of our models against the COVID-19 Forecast Hub Ensemble baseline.

Base Features

We engineered features for machine learning that can be categorized into five groups: (A) a set of county-level demographic and socioeconomic features, (B) minimum and maximum temperatures of inhabited areas in counties, (C) temporally-lagged (i) weekly average and (ii) weekly change of incident rates (COVID-19 cases per 10k population) in each county, (D) Facebook-derived features of (i) intra-county movement measurements and (ii) exposure to COVID-19 through inter-county connectedness, and (E) SafeGraph-derived features of (i) intra-county movement measurements and (ii) exposure to COVID-19 through inter-county connectedness.

Socioeconomic, demographic, and climatic variables have been shown to be correlated with the spread of COVID-19[21,37–40], therefore we include category A and B features in all of our models to control for these factors. Section A of the supplementary information outlines the detailed methodology for generating features in these categories.

Features in category C indirectly capture population compartments of the susceptible–infected–recovered (SIR) epidemiological models[41,42]. We defined the COVID-19 incidence rate of a county as its number of cases per 10,000 population

and included a four-week lagged (t-4) weekly average of total incidence rates in each county as a feature (category C.i) to capture the Susceptible and Recovered compartments. If the feature value is small, many individuals in the unit have not yet contracted the disease; and therefore, are still *susceptible*. If the feature value is sufficiently large (level of sufficiency is learned by the model), the compartment is approaching higher levels of immunity as a whole. We use machine learning algorithms that are capable of learning such non-linear relationships. It is worth noting that vaccination rates can similarly be incorporated as a feature. However, inoculations in the U.S. started after our study period, and thus, not included here.

To account for the latency associated with the effects of temperature, new and historical incidences, and human interaction on the spread of COVID-19, we generated four temporal weekly lags of features in categories B-E (we assume that the features in category A are static during our study period). Notably, for category C, we also included (natural log-transformed values of) *change* in incidence rate($ln(\Delta\ incidence\ rat + 1)$), i.e., the subtraction of observed start-of-week incident rate from end-of-week incidence rate during the four weekly temporal-lags (t-1,…,t-4), as another set of features in all of our models (category C.ii) (more details in the supplementary information). These features conceptually capture

the Infected compartment in the SIR model, i.e., the currently infected populations in the spatial unit. Our autoregressive modeling with multiple temporal lags allows the models to learn the rate of spread in a unit, as well as the varying incubation periods of the disease in relation to the change in temperature and demographic features[43,44].

Features in categories D and E also model the SIR population compartments, but in connected counties (through either social connectedness or flow connectedness). Features in category D represent the social media proxy (of physical human interactions), whereas features in category E represent the cell phone-derived human mobility flow proxy.

Conceptualization of models with social media and cell phone features

We evaluate the predictive power of the Facebook-derived features (category D) and the SafeGraph-derived features (category E) against a *base* model by developing four different model setups (not to be confused with the final evaluations against the *baseline* Ensemble model). Here, we provide an outline of these models, with more details on the specific algorithms and features mentioned in Table 4 in the following sections.

Table 4. The complete list of features, their temporal lags, and the models in which they are used. t-1, t-2, t-3 and t-4 indicate one-, two-, three-, and four-week lags, respectively. The target variable for one

week-ahead prediction horizon on forecast date d is the number of new cases per 10k from the forecast date through the end of prediction horizon t in each county, i.e., $\Delta(\text{incidence rate})_{t,d}$. The target for the two week-horizon prediction is $\Delta(\text{incidence rate})_{t+1,d}$, three week-horizon is $\Delta(\text{incidence rate})_{t+2,d}$, and four week horizon is $\Delta(\text{incidence rate})_{t+3,d}$. Ln()in the table indicates natural logarithm, Mean () indicates weekly average, Δ indicates weekly change, i.e., difference (calculated by subtracting the value of the feature at the beginning of the week from its value at the end of the week) and Slope indicates the slope of a fitted linear regression model to the standardized daily measures of metric value as the dependent variable and standardized day of week as the independent variable.

Category	Model(s)	Variables	Temporal Lag
A- socioeconomic and demographic	All of the models	population density percentage of male population percentage of African American population percentage of Hispanic population percentage of Native American population percentage of the rural population percentage of the population with a college degree median household income percentage of the population who voted republican in 2016 presidential election	None (constant)
B- Temperature	All of the models	mean(daily minimum temperature)$_t$ mean(daily maximum temperature)$_t$	t-1, t-2, t-3, t-4
C- COVID-19 incidence rate	All of the models	Ln (Δ incidence rate $_t$+1) Ln (mean (incidence rate)$_t$ +1)	t-1, t-2, t-3, t-4 t-4
D- Facebook	-FB model	Δ SPC $_t$ mean (SPC)$_t$ mean (Stay Put)$_t$ and slope (Stay Put)$_t$	t-1, t-2, t-3, t-4 t-4 t-1, t-2, t-3, t-4

		mean (Change in Movement)$_t$ and slope (Change in Movement)$_t$	t-1, t-2, t-3, t-4
E- SafeGraph	-SG and -SGR models	Δ FPC $_t$	t-1, t-2, t-3, t-4
		mean (FPC)$_t$	t-4
		mean (% *completely_home_device_count)$_t$* and slope(% *completely_home_device_count)$_t$*	t-1, t-2, t-3, t-4
		mean(*baselined distance_traveled_from_home)$_t$* slope(*baselined distance_traveled_from_home)$_t$*	t-1, t-2, t-3, t-4
	-SGR model	mean (*baseliend median_home_dwell_time)$_t$* and slope(*baselined median_home_dwell_time)$_t$*	t-1, t-2, t-3, t-4
		mean (*baseliend % full_time_work_behavior_devices)$_t$* and slope(*baseliend % full_time_work_behavior_devices)$_t$*	t-1, t-2, t-3, t-4

The first model (base model) only includes *base features:* socioeconomic features (category A), four temporally-lagged weekly temperature features (category B), and four temporally-lagged weekly *change* in incidence rates in each county, as well as *weekly average of* incidence rates during the fourth lagged week (category C). Therefore, the base model only incorporates temporal lags of the features and the target variable in predicting new cases of COVID-19.

The second model, which we identify by the "-FB" suffix (to note the inclusion of Facebook features), includes the base features as well as category D features, i.e. Facebook-derived intra-county movement features and inter-county spatiotemporal lags of the target variable, i.e. exposure to COVID-19 through

social connectedness (Social Proximity to Cases), across four temporal lags. The third model, which we identify by the "-SG" suffix, is conceptually similar to the -FB model, but with features derived from SafeGraph cell phone mobility data instead of Facebook data. Specifically, the -SG models include the base features in addition to inter-county spatiotemporal lags of the target variable, i.e. exposure to COVID-19 through human flow connectedness (which we call Flow Proximity to Cases), and a subset of category E SafeGraph-derived intra-county movement features, across four temporal lags.

To explore the full potential of the movement features provided by the SafeGraph Social Distancing Metrics (SDM) dataset, we developed a fourth model, in which two additional mobility-related measurements provided in the SDM dataset (that are least correlated with other features in Category E) are added to the -SG model. This model thus includes categories A-C and all features in category E, and is identified by the "-SGR" suffix.

Features derived from Facebook

i. Intra-county movement features

Facebook publishes the Movement Range dataset for 14 countries[45] and it includes two metrics called “Change in Movement” and “Stay Put”, each providing a different perspective on movement trends as measured by mobile devices carrying the Facebook app. The *Change in Movement* metric for each county is a measure of relative change in aggregated movement compared to the baseline of February 2nd to February 29th 2020 (excluding the February 17th 2020 President Day holiday in the US)[45]. The *Stay Put* metric measures “the fraction of the population that have stayed within a small area during an entire day”[45]. We used four temporal lags of weekly averages and slopes of each metric as a feature in our -FB model. We calculated the slopes by fitting a linear regression model to the metric value as the dependent variable and day of week as the independent variable, both transformed to standard scale N(0,1). The slope feature characterizes the overall trend in a week, as compared to the baseline period.

ii. Inter-county features and spatial lag modeling

The *intra*-county features capture the intrinsic movement-related characteristics of a county and ignore its interactions (i.e. spatial lags) with the counties to which it is *connected*. Therefore, we calculated *inter*-county metrics of connectivity as a basis for incorporating spatiotemporal lags in our models. Notably, the connectedness in this context transcends spatial connectedness in the form of mere physical adjacency.

Social connectedness Index (SCI), another dataset published by Facebook, is a measure of the intensity of connectedness between administrative units, calculated from Facebook friendship data. Social connectedness between two counties *i* and *j* is defined as[46]:

$$Social\ connectedness\ (SC)_{i,j} = \frac{FB\ Connections_{i,j}}{FB\ Users_i * FB\ Users_j} \quad (1)$$

where $FB\ Connections_{i,j}$ is the number of friendships between Facebook users who live in county *i* and those who live in county *j;* while $FB\ Users_i$ and $FB\ Users_j$ are the total number of active Facebook users in counties *i j* and, respectively. Social Connectedness is scaled to a range between 1 and 1,000,000,000 and rounded to the nearest integer to generate SCI, as published by Facebook[47]. Therefore, if the SCI value between a pair of counties is twice as large as another pair, it means the users in the first county-pair are almost twice as likely to be friends on Facebook than the second county-pair[46]. We used the latest version of the SCI dataset (at the time of our analyses), which was released in August 2020[47].

While SCI provides a measure of connectivity, our goal is to capture the spatiotemporal lags of COVID-19 cases in *county i*, i.e., the number of recent COVID-19 cases in other counties connected to county i. Using SCI, Kuchler et al.[6] created a new metric, called Social Proximity to Cases (SPC) for each county, which is a measure of the level of exposure to COVID-19 cases in *connected counties* through social connectedness. We use a slight variation of SPC, defined as follows for county *i* at time t:

$$Social\ Proximity\ to\ Cases(SPC)_{i,t} = \sum_{j} Cases\ Per\ 10k_{j,t} \times \frac{Social\ Connectedness_{i,j}}{\sum_{h} Social\ Connectedness_{i,h}} \quad (2)$$

where $Cases\ Per\ 10k_{j,t}$ is the number of COVID-19 cases per 10k population (i.e., incidence rate) in county *j* as of time *t*. For county *i*, the sums *j* and *h* are over all counties. In other words, SPC for a county *i*, in time *t,* is the average of COVID-19 incidence rates in connected counties weighted by their social connectedness to county *i*, i.e., the spatial lag of incidence rates. To the best of our knowledge, SPC data has not been published, but we were able to generate this feature using the original method[6], modified for our weekly temporal lagged features and calculated using incidence rates (cases per 10k population) rather than total number of cases. In the -FB models (Table 4), we incorporated features of weekly change (Δ) in SPC at four temporally lagged weeks (difference between the end and start of the lag week) to model the Infected SIR compartment in connected counties, as well as weekly average of SPC in the fourth lagged week (t-4), to capture the Susceptible and Recovered SIR compartments in connected counties, similar to the rational for features in category C, as explained earlier.

Features derived from SafeGraph

i. Intra-county movement features

To generate movement features from cell phone data, we used SafeGraph's SDM dataset that is "generated using a panel of GPS pings from anonymous mobile devices"[48]. The SDM dataset contains multiple mobility metrics published at the Census Block Group (CBG) level. Among these metrics, *distance_traveled_from_home* (median distance traveled by the observed devices in meters) and *completely_home_device_count* (the number of devices that did not leave their home location during a day)[48] are conceptually closest to the metrics included in the Facebook's Movement Range Dataset. We used these two features in our -SG model, which is the conceptual equivalent of the -FB model, but with cell phone-derived features instead of the Facebook-derived features (Table 4).

We included the SafeGraph's *median_home_dwell_time* (median dwell time at home in minutes for all observed devices during the time period), and *full_time_work_behavior_devices* (the number of devices that spent more than 6 hours at a location other than their home during the day)[48] in addition to the

previous two features in the -SGR model to take fuller advantage of the metrics available in the SDM dataset.

We derived *baselined* features from the SDM metrics as such: To address the potential effect of fewer cell phone observations in some CBGs, we used a Bayesian hierarchical model[49,50] with two levels (states and counties), and then smoothed the daily measurements using a seven-day rolling average to reduce the effect of outliers in the data. We then aggregated CBG-level *completely_home_device_count* and *full_time_work_behavior_devices* values up to the county level, divided by the total *device_count* in the county on the same day. For *full_time_work_behavior_devices,* we subtracted the final proportion from the February baseline of the same metric. For the *median_home_dwell_time* and *distance_traveled_from_home* variables, we calculated the weighted mean (by CBG population) of values per county, and then calculated percent of change compared to February baseline.

We used weekly averages and slopes (calculated by fitting a linear regression model to the values as the response variable and day of week as the independent variable) of these four metrics as features in our models (Table 4).

ii. Inter-county features and spatial lag modeling

Building on the conceptual structure of SCI, we derived a novel and daily inter-county connectivity index from SafeGraph's SDM dataset to quantify connectedness between counties based on the level of human flow from one county to the other (measured through cell-phone pings). We call this index "Flow Connectedness Index" (FCI). Using FCI, we then calculated a spatial lag metric that we call "Flow Proximity to Cases" (FPC) for each county. FPC captures the average of COVID-19 incidence rates in connected (by human movement) counties weighted by the FCI. Again, it is worth noting that connectedness in this sense goes beyond the physical connectivity of counties, and considers daily human movement between them as the basis for determining connectivity. The similar formulations of FCI and SCI, as well as FPC and SPC, allow for direct comparison of the two networks (i.e. FB's friendship network and SafeGraph's human flow network) in their capability to capture inter-county physical human interactions, and subsequently, to predict new COVID-19 cases.

The SafeGraph's SDM contains the number of *visits* between different CBGs. We aggregate these values to the county level to measure the daily number of devices that move (flow) between each county pair. Leveraging these flow measurements, we define Flow Connectedness Index (FCI) as:

$$Flow\ connectedness\ index(FCI)_{i,j} = \frac{Device\ flow_{i,j} + Device\ flow_{j,i}}{Device\ count_i * Device\ count_j} \quad (3)$$

where for counties i and j, $Device\ flow_{i,j}$ is the sum of visits with origin i and destination j. $Device\ count_i$ is the number of devices whose home location is in county *i*. We then scale FCI to a range between 1 to 1,000,000,000.

We defined FPC as:

$$Flow\ Proximity\ to\ Cases\ (FPC)_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} \times \frac{Flow\ Connectedness_{i,j}}{\sum_h Flow\ Connectedness_{i,h}} \quad (4)$$

Where $Cases\ Per\ 10k_{j,t}$ is the number of confirmed COVID-19 cases per 10k population in county *j* at time t, and $Flow\ Connectedness_{i,j}$ is the value of FCI between county *i* and *j*.

Facebook's social network and friendship connections do not change significantly over time, and therefore, SCI is a static index in a one-year period. Conversely, inter-county human flow from SafeGraph is dynamic and can change dramatically, even within a week. We generated daily FCI (and FPC) for each county-pair in the US to utilize the full temporal resolution of the SDM dataset. We used weekly change (Δ) of FPC for the four temporally lagged weeks, and its average only in the fourth week as features -SG and -SGR models, with the same

rationale as features in Category C and D to capture SIR compartments in connected counties (Table 4).

Model Implementation

The ultimate target variable in all of our autoregressive models is the number of new cases of COVID-19. For training and tuning the models, however, we used a transformed target variable, namely the natural log-transformed values of new cases per 10k population plus one (to avoid zero values). For reporting the model predictions, we computed the number of new cases by applying an inverse transformation, i.e. an exponential transformation minus one (Formula 5a-c). The rationale for using the log-transformed target variable, as opposed to directly predicting the weekly new cases, was to minimize skewness, and more importantly, minimize the sensitivity of the models to the population of counties. Our exploratory work did confirm that using this logged of incidence rates produced better results.

$$y_{predicted_{(i,t)}} = ln\left(\Delta\ incidence\ rate_{(i,t)} + 1\right) \quad (5.a)$$

$$\Delta\ incidence\ rate_{predicted_{(i,t)}} = e^{y_{predicted_{(i,t)}}} - 1 \quad (5.b)$$

$$\Delta\ Case_{predicted_{(i,t)}} = \left(\Delta\ incidence\ rate_{predicted_{(i,t)}}\right) * Population_i/10{,}000 \quad (5.c)$$

$\Delta\, Case_{predicted_{(i,t)}}$ in 5.c denotes the number of new cases in a county in the prediction horizon.

Our training dataset includes up to 34 training samples per county (number of total samples n=3103 x 34), with each sample holding various features in one- to four-weekly temporal lags (Table 4). The weekly calculation of features is based on weeks starting on Sundays and ending on Saturdays, with predictions also made for horizons spanning Sunday-Saturday periods as in common practice[25]. Our features, models, evaluations and comparisons are limited to the counties in the coterminous US. Table 4 summarizes the features that we used and the number of temporal lags (if any) used for each feature. All features were standardized for use in machine learning algorithms.

Our general approach to training, validation and testing of our models for different prediction horizons is similar, only, with target variables calculated separately for the specific prediction horizon. We first outline our approach for one-week ahead prediction horizons, which is used as the basis for algorithm selection and comparison of Facebook- and SafeGraph-derived features (-FB and -SG models); It is worth noting that we compare the two feature sets in longer

term predictions too (Supplementary Table 5). We then provide an overview of the implementation of the models for longer term prediction horizons.

We trained and tuned the models using randomized search and 5-fold cross validation, and tested the best tuned model for predicting new cases in the week following the forecast date (for the one-week prediction horizons, as listed in Table 5). For instance, for the forecast date of October 24th, we used features which were generated using data collected before October 24th for tuning and training. The tuned model was then used for predicting new cases in each county during the October 24th to October 31st period. We used the reported cases by the JHU CSSE. The temporally lagged features for this forecast date were generated for t-1, t-2, t-3 and t-4 weekly lags, namely, the weeks ending on Oct. 24, Oct. 17, Oct. 10, and Oct. 3 respectively.

For the next forecast date, October 31, the training size increased by one week (per county), and the target week was also shifted by one week. Table 5 summarizes the forecast dates, one-week ahead prediction horizons, and training data size. The data used in generating these features spans a period from March 29 to November 28, 2020 to cover the temporal lags, and the target variable is collected through December 12, 2020 for the evaluation of four-week ahead predictions on the last forecast date. More details on cross validation,

hyperparameters, and evaluation are presented in the supplementary information (Section C).

Table 5. Summary of training and testing data used for machine learning algorithm selection, as well as comparison of -FB and -SG models in short-term predictions

Forecast Date	One-week prediction Horizon	Training data start date	Training data end date	# training samples per county
2020-10-24	2020-10-25 to 2020-10-31	2020-03-29	2020-10-24	30
2020-10-31	2020-11-01 to 2020-11-07	2020-03-29	2020-10-31	31
2020-11-07	2020-11-08 to 2020-11-14	2020-03-29	2020-11-07	32
2020-11-14	2020-11-15 to 2020-11-21	2020-03-29	2020-11-14	33
2020-11-21	2020-11-22 to 2020-11-28	2020-03-29	2020-11-21	34

We experimented with five different supervised machine learning regression algorithms, namely Random Forest[51] (RF), Stochastic Gradient Boosting[52] (SGB), eXtreme Gradient Boosting[36,53] (XGB), Feed Forward Neural Network[54] (FFNN), and Long Short-Term Memory[55] (LSTM) network to build the autoregressive machine learning models with features described in Table 4. We evaluated the models using the dates listed in Table 5. Results are presented in Table 1. The

details of hyperparameter candidates and specific architectures are presented in the supplementary information (Section C).

Comparing Facebook-derived features with SafeGraph-derived features

Since the XGB algorithm performed best (Table 1), we chose it as the selected machine learning algorithm, and trained the base, -FB, -SG, and -SGR models using the XGB algorithm to predict new cases of COVID-19 in short-term (one week) and long-term (two to four weeks) prediction horizons. To name a specific model in this article, we use a prefix that denotes the type of lag included in the model features (i.e. T for temporal or ST for spatiotemporal), followed by the name of the algorithm (XGB), followed by a suffix denoting the features included in the model, namely, -FB, -SG, and -SGR. Thus, TXGB (Temporal eXtreme Gradient Boosting) denotes the model that is built using the XGB algorithm and includes the base, Temporally lagged features; and STXGB-FB (SpatioTemporal eXtreme Gradient Boosting) denotes the model that includes Facebook-derived features (and thus, spatiotemporal lags) and is built using XGB.

We evaluated the performance of TXGB, STXGB-FB, STXGB-SG, and STXGB-SGR by comparing the RMSE and MAE scores of the predictions against the observed numbers of new cases in the corresponding prediction horizon (results in Table 2).

We also tested the -FB and -SG models for all prediction horizons across all forecast dates (results in Supplementary Table 5).

Re-tuning and re-training all different variations of the STXGB model for each forecast date and prediction horizon resulted in considerably improved predictions compared to the baseline model (Table 4 and Supplementary Tables 4 and 5). Each STXGB model was tuned and trained on a regular desktop machine (with a 6 core Ryzen 5 3600X CPU and 64GB of RAM) in approximately 12-13 minutes for a single prediction horizon, and thus, in almost one hour for all of the four prediction horizons.

Evaluation against the COVID-19 Forecast Hub Ensemble baseline

In addition to the one-week short-term predictions, we performed long-term predictions of new COVID-19 cases in two-, three-, and four-week ahead prediction horizons. We only used the STXGB algorithm to develop long-term prediction models since it outperformed other algorithms in short-term predictions (see Section 2). We used the same set of features for long-term predictions, with modifications on the target variable to reflect different prediction horizons. For instance, the two-, three-, and four-week ahead horizons

of the Forecast date Oct. 24, were Oct 24 to Nov. 7, Oct. 24 to Nov. 14, and Oct 24 to Nov. 21, respectively.

The model for each horizon was trained and validated separately using the same training data and approach described in the previous Section (Table 5), and was tested on two, three and four weeks of unseen data, respectively, for each horizon. We evaluated the models' predictions by comparing them against the predictions generated by the COVID-19 Forecast Hub's Ensemble model as well as the ground-truth values of new cases derived from JHU CSSE COVID-19 reports. Additionally, we compared the long-term predictions of STXGB-FB and STXGB-SG (Supplementary Table 5).

Data Availability

All of the datasets used in this study are publicly available (at the time of writing this manuscript). We created socioeconomic features from the 5-year survey data--between 2014-2018--provided by the American Community Survey (ACS) and available at IPUMS National Historical GIS portal (https://www.nhgis.org/). Daily maximum and minimum temperature surfaces of the U.S. published by the NOAA are available at https://ftp.cpc.ncep.noaa.gov/GIS/GRADS_GIS/GeoTIFF/TEMP/. We used the

cumulative confirmed COVID-19 cases published by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) to generate COVID-related features. Facebook's Social Connectedness Index (SCI) database is available at https://dataforgood.fb.com/tools/social-connectedness-index/ and the movement range dataset can be found at https://data.humdata.org/dataset/movement-range-maps. Finally, the instructions for accessing SafeGraph's Social Distancing Metrics dataset is available at https://docs.safegraph.com/docs/social-distancing-metrics.

Code Availability

All code necessary for the replication of our results is available for reviewers upon request. The code will be published publicly on GitHub under MIT license upon acceptance of this article.

Summary of Contributions

Morteza Karimzadeh conceptualized the project, designed the features and contributed 30% of data processing and implementation, and contributed equally to writing. Behzad Vahedi conducted the majority of data processing, implementation, literature review, and contributed equally to writing. Hamidreza

Zoraghein contributed to the study design and 10% of implementation and writing.

Competing interests

The authors do not report any competing interests.

References

1. Chu, D. K. *et al.* Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet* **395**, 1973–1987 (2020).
2. Chiu, W. A., Fischer, R. & Ndeffo-Mbah, M. L. State-level needs for social distancing and contact tracing to contain COVID-19 in the United States. *Nat. Hum. Behav.* **4**, 1080–1090 (2020).
3. Eames, K. T. D. & Keeling, M. J. Contact tracing and disease control. *Proc. R. Soc. Lond. B Biol. Sci.* **270**, 2565–2571 (2003).
4. Clark, E., Chiao, E. Y. & Amirian, E. S. Why Contact Tracing Efforts Have Failed to Curb Coronavirus Disease 2019 (COVID-19) Transmission in Much of the United States. *Clin. Infect. Dis.* (2020) doi:10.1093/cid/ciaa1155.
5. Bailey, M., Cao, R., Kuchler, T. & Stroebel, J. The Economic Effects of Social Networks: Evidence from the Housing Market. *J. Polit. Econ.* **126**, 2224–2276 (2018).
6. Kuchler, T., Russel, D. & Stroebel, J. *The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook*. *J. Urban Econ.*, 103314 (2020).

7. Büchel, K. & Ehrlich, M. v. Cities and the structure of social interactions: Evidence from mobile phone data. *J. Urban Econ.* **119**, 103276 (2020).

8. Kang, Y. *et al.* Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. *Sci. Data* **7**, 390 (2020).

9. Bullinger, L. R., Carr, J. B. & Packham, A. *COVID-19 and Crime: Effects of Stay-at-Home Orders on Domestic Violence*. https://www.nber.org/papers/w27667 (2020) doi:10.3386/w27667.

10. Killeen, B. D. *et al.* A County-level Dataset for Informing the United States' Response to COVID-19. Preprint at https://arxiv.org/abs/2004.00756 (2020).

11. Gao, S., Rao, J., Kang, Y., Liang, Y. & Kruse, J. Mapping county-level mobility pattern changes in the United States in response to COVID-19. (2020).

12. Sen-Crowe, B., McKenney, M. & Elkbuli, A. Social distancing during the COVID-19 pandemic: Staying home save lives. *Am. J. Emerg. Med.* **38**, 1519–1520 (2020).

13. Unwin, H. J. T. *et al.* State-level tracking of COVID-19 in the United States. *Nat. Commun.* **11**, 6189 (2020).

14. Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D. The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci.* **117**, 16732–16738 (2020).

15. Zhou, Y. *et al.* A Spatiotemporal Epidemiological Prediction Model to Inform County-Level COVID-19 Risk in the United States. *Harv. Data Sci. Rev.* (2020) doi:10.1162/99608f92.79e1f45e.

16. Singh, R. K. *et al.* Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model. *JMIR Public Health Surveill.* **6**, e19115 (2020).

17. Dansana, D. *et al.* Global Forecasting Confirmed and Fatal Cases of COVID-19 Outbreak Using Autoregressive Integrated Moving Average Model. *Front. Public Health* **8**, (2020).

18. Xiang, J. *et al.* Impacts of the COVID-19 responses on traffic-related air pollution in a Northwestern US city. *Sci. Total Environ.* **747**, 141325 (2020).

19. Zeroual, A., Harrou, F., Dairi, A. & Sun, Y. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos Solitons Fractals* **140**, 110121 (2020).

20. Jo, H., Kim, J., Huang, T.-C. & Ni, Y.-L. condLSTM-Q: A novel deep learning model for predicting Covid-19 mortality in fine geographical Scale. Preprint at https://arxiv.org/abs/2011.11507 (2020).

21. Mollalo, A., Rivera, K. M. & Vahedi, B. Artificial Neural Network Modeling of Novel Coronavirus (COVID-19) Incidence Rates across the Continental United States. *Int. J. Environ. Res. Public. Health* **17**, 4204 (2020).

22. Buchwald, A. G., Adams, J., Bortz, D. M. & Carlton, E. J. Infectious Disease Transmission Models to Predict, Evaluate, and Improve Understanding of COVID-19 Trajectory and Interventions. *Ann. Am. Thorac. Soc.* **17**, 1204–1206 (2020).

23. Reiner, R. C. *et al.* Modeling COVID-19 scenarios for the United States. *Nat. Med.* **27**, 94–105 (2021).

24. Santosh, K. C. COVID-19 Prediction Models and Unexploited Data. *J. Med. Syst.* **44**, 170 (2020).

25. Ray, E. L. *et al.* Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. Preprint at https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1 (2020).

26. CDC. Cases, Data, and Surveillance. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html (Retrieved on Jan 30, 2020).

27. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).

28. Epidemiologists Urge A Cautious Christmas, After Thanksgiving Surge in Some States. *NPR.org* https://www.npr.org/sections/health-shots/2020/12/21/948809129/epidemiologists-urge-a-cautious-christmas-after-thanksgiving-surge-in-some-state (Retrieved on Jan 30, 2020).

29. Rama, D., Mejova, Y., Tizzoni, M., Kalimeri, K. & Weber, I. Facebook Ads as a Demographic Tool to Measure the Urban-Rural Divide. in *Proceedings of The Web Conference 2020* 327–338 (Association for Computing Machinery, 2020). doi:10.1145/3366423.3380118.

30. Wang, Q., Phillips, N. E., Small, M. L. & Sampson, R. J. Urban mobility and neighborhood isolation in America's 50 largest cities. *Proc. Natl. Acad. Sci.* **115**, 7735–7740 (2018).

31. Coston, A. *et al.* Leveraging Administrative Data for Bias Audits: Assessing Disparate Coverage with Mobility Data for COVID-19 Policy. Preprint at https://arxiv.org/abs/2011.07194 (2020).

32. Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C. & Vasilakis, C. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *Eur. J. Oper. Res.* **290**, 99–115 (2021).

33. CDC. Coronavirus Disease 2019 (COVID-19). *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/more/masking-science-sars-cov2.html (Retrieved on Jan 30, 2020).

34. Wyche, S. P., Schoenebeck, S. Y. & Forte, A. 'Facebook is a luxury': an exploratory study of social media use in rural Kenya. in *Proceedings of the 2013 conference on Computer supported cooperative work* 33–44 (Association for Computing Machinery, 2013).

35. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

36. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.

37. Bashir, M. F. *et al.* Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci. Total Environ.* **728**, 138835 (2020).

38. Mollalo, A., Vahedi, B. & Rivera, K. M. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Sci. Total Environ.* **728**, 138884 (2020).

39. Khalatbari-Soltani, S., Cumming, R. C., Delpierre, C. & Kelly-Irving, M. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J Epidemiol Community Health* **74**, 620–623 (2020).

40. Li, H. *et al.* Air pollution and temperature are associated with increased COVID-19 incidence: A time series study. *Int. J. Infect. Dis.* **97**, 278–282 (2020).

41. Toda, A. A. Susceptible-Infected-Recovered (SIR) Dynamics of COVID-19 and Economic Impact. Preprint at https://arxiv.org/abs/2003.11221(2020).

42. Volz, E. & Meyers, L. A. Susceptible–infected–recovered epidemics in dynamic contact networks. *Proc. R. Soc. B Biol. Sci.* **274**, 2925–2934 (2007).

43. Lauer, S. A. *et al.* The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Ann. Intern. Med.* **172**, 577–582 (2020).

44. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet* **395**, 507–513 (2020).

45. Protecting privacy in Facebook mobility data during the COVID-19 response. *Facebook Research* https://research.fb.com/blog/2020/06/protecting-privacy-

in-facebook-mobility-data-during-the-covid-19-response/ (Retrieved on Jan 30, 2020).

46. Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. Social Connectedness: Measurement, Determinants, and Effects. *J. Econ. Perspect.* **32**, 259–280 (2018).

47. Social Connectedness Index Methodology. *Facebook Data for Good* https://dataforgood.fb.com/docs/social-connectedness-index-methodology/ (Retrieved on Jan 30, 2020).

48. Social Distancing Metrics. *SafeGraph* https://docs.safegraph.com/docs/social-distancing-metrics (Retrieved on Jan 30, 2020).

49. Gelman, A. *et al. Bayesian Data Analysis, Third Edition*. (CRC Press, 2013).

50. Devine, O. J., Louis, T. A. & Halloran, M. E. Empirical Bayes Methods for Stabilizing Incidence Rates before Mapping. *Epidemiology* **5**, 622–630 (1994).

51. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

52. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

53. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).

54. Fine, T. L. *Feedforward Neural Network Methodology*. (Springer Science & Business Media, 2006).

55. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

Figures

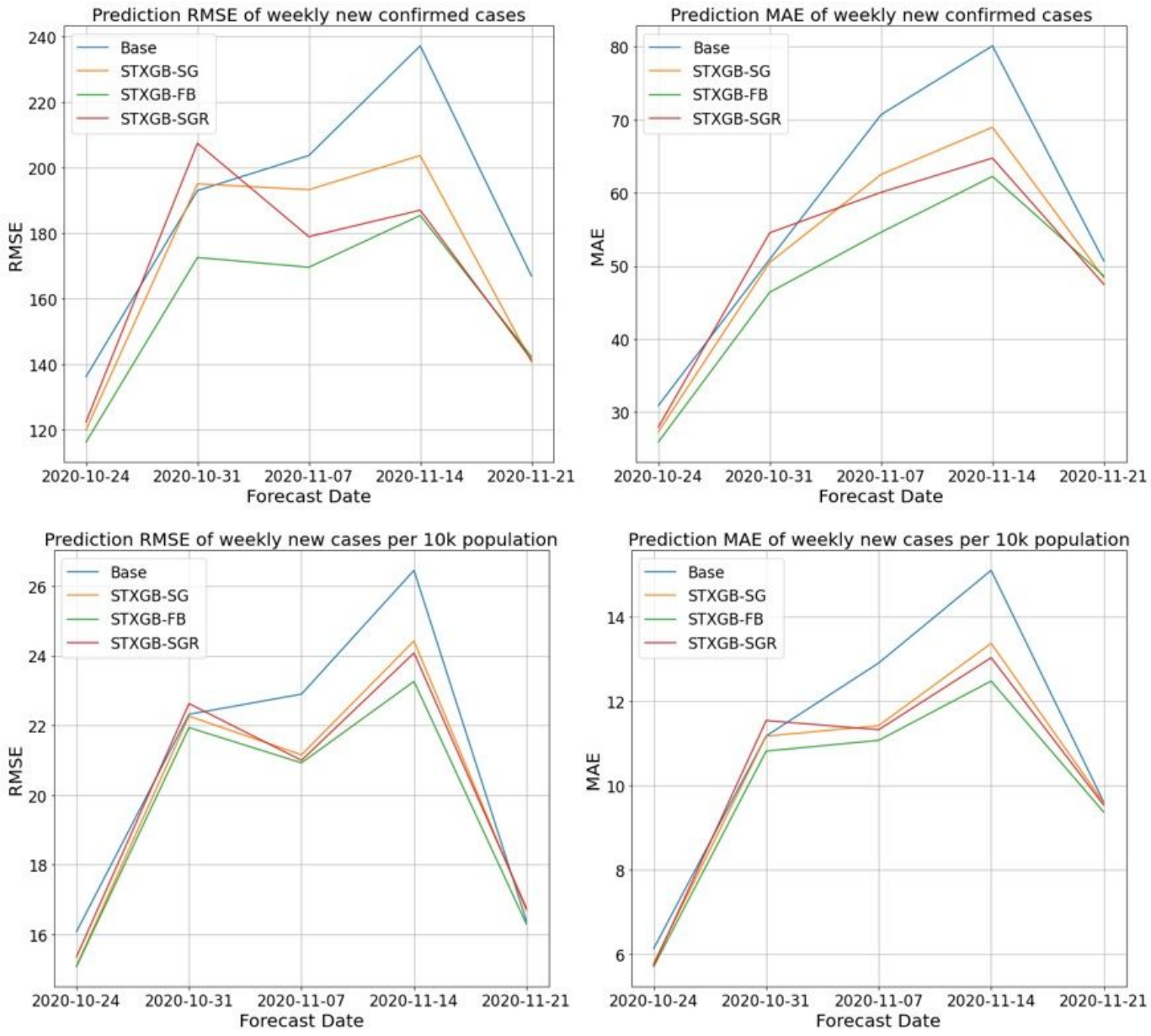

Figure 1

Prediction errors of the SpatioTemporal eXtreme Gradient Boosting (STXGB) models in one-week prediction horizon against a temporal autoregressive model without spatial lags (TXGB). Error comparison of STXGB models with different feature sets in predicting weekly new cases (top row) and new cases per 10k population (bottom row) for one-week ahead horizon. STXGB-FB, which incorporates Facebook-derived features, including spatial lags based on Social Connectedness Index, outperforms other models. Left column: prediction RMSE. Right column: prediction MAE.

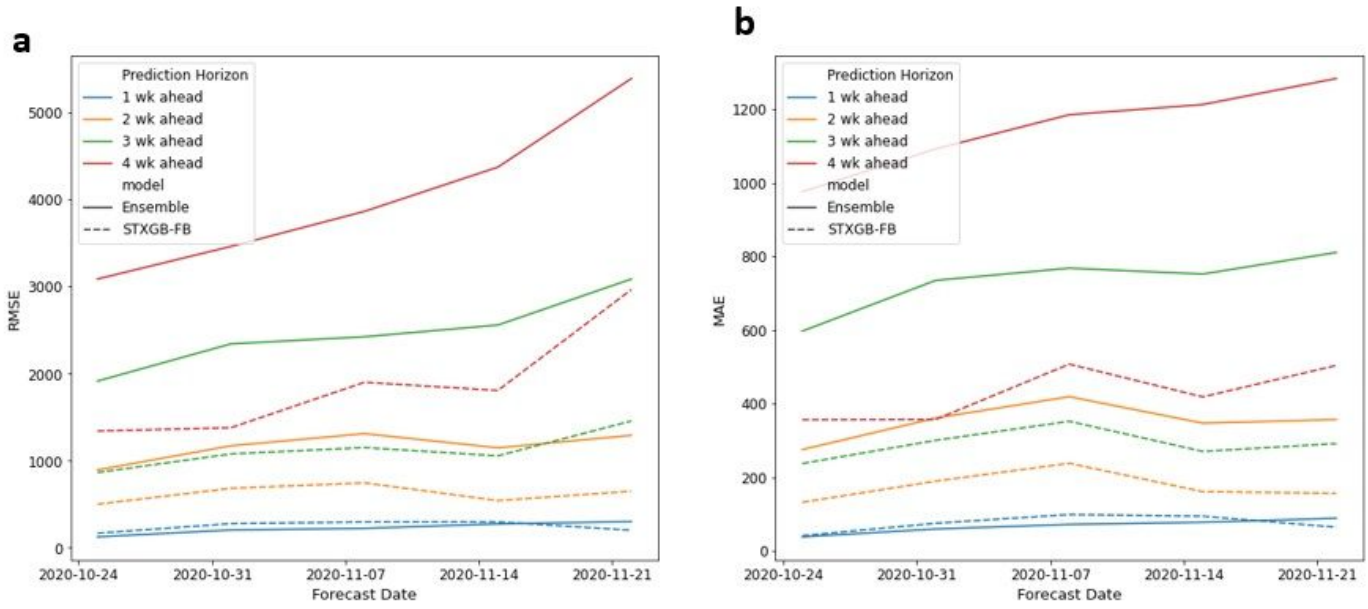

Figure 2

Long-term prediction error comparison. Prediction errors of the STXGB-FB model (dashed line) and the Ensemble baseline (solid line) over four prediction horizons on five forecasting dates. a) Prediction RMSE and b) Prediction MAE.

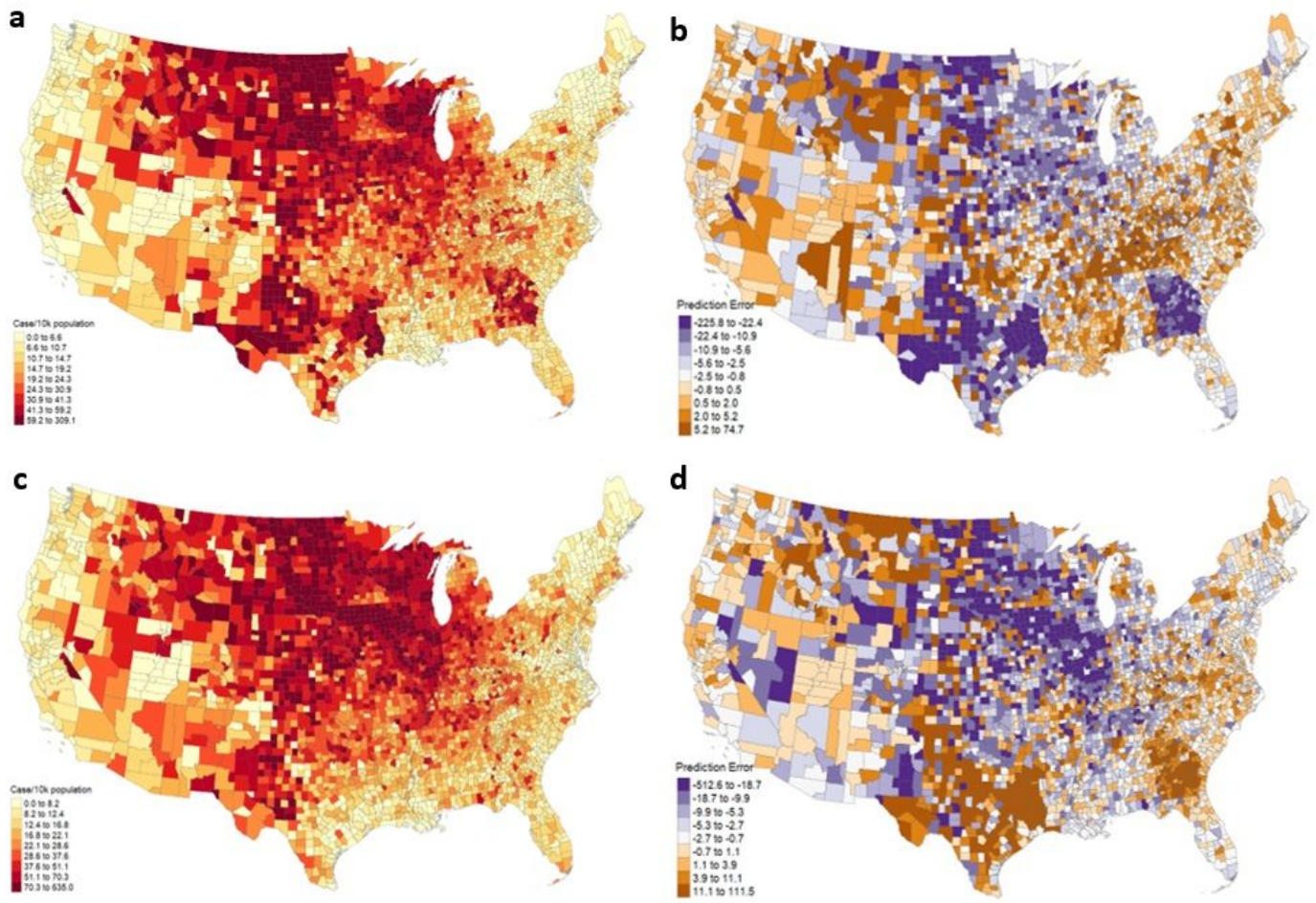

Figure 3

Map of COVID-19 cases per 10k population and errors in predicting them. (a) number of confirmed new cases per 10k population over the week ahead of forecasting date Oct. 31, 2020 (b) prediction errors for the same forecasting date(c) number of new cases over the week ahead of Nov. 7, 2020 forecasting date (d) prediction errors for the same forecasting date. The pattern of errors in Georgia, and Texas, and Kentucky flip from Oct. 31 to Nov. 7, indicating potential lags in testing and reporting.

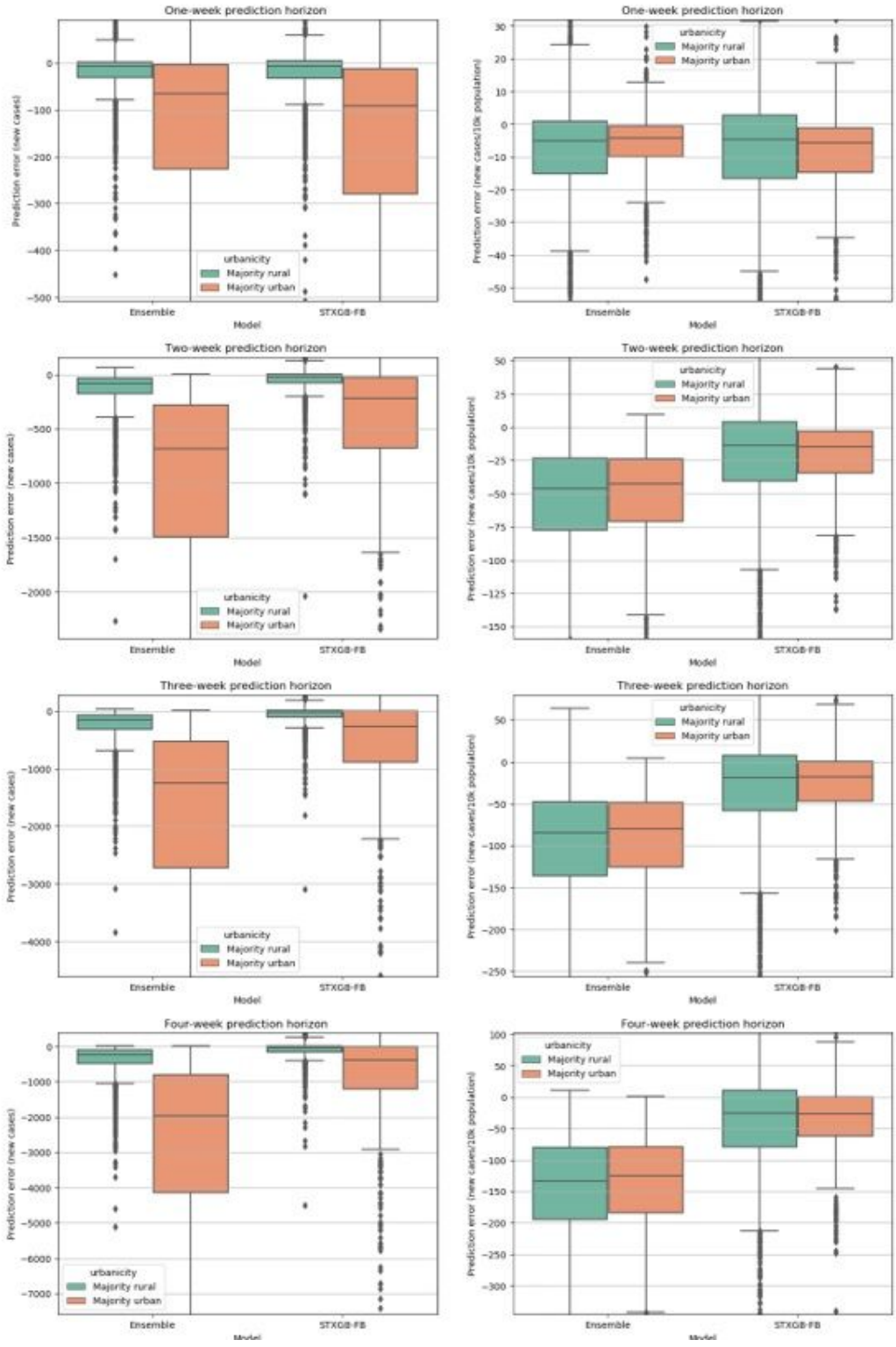

Figure 4

Prediction errors in urban vs. rural counties. Prediction errors of the number of new cases (left column) and new cases per 10k population (right column) in rural and urban counties on the Nov. 7 forecast date across four prediction horizons. The higher and lower 3% of counties are trimmed from the plot view.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SPCandFPCSupplementaryinformationsubmissionready.docx