

Improving COVID-19 Forecasting using eXogenous Variables

Mohammadhossein Toutiaee
hosseini@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

Shopnine Sivaraja
shopnine@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

Yuan Ke
yuan.ke@uga.edu
Department of Statistics
The University of Georgia
Athens, GA

Xiaochuan Li
xiaochuan.li@uga.edu
Department of Statistics
The University of Georgia
Athens, GA

Aishwarya Venkataraj
aishwaryavenkatraj@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

Ismailcem Arpinar
budak@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

John Miller
jamill@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

Yogesh Chaudhari
yogesh.chaudhari@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

Indrajeet Javeri
indrajeet.javeri@uga.edu
Department of Computer Science
The University of Georgia
Athens, GA

Nicole Lazar
nfl5182@psu.edu
Department of Statistics
The Pennsylvania State University
University Park, PA

ABSTRACT

In this work, we study the pandemic course in the United States by considering national and state levels data. We propose and compare multiple time-series prediction techniques which incorporate auxiliary variables. One type of approach is based on spatio-temporal graph neural networks which forecast the pandemic course by utilizing a hybrid deep learning architecture and human mobility data. Nodes in this graph represent the state-level deaths due to COVID-19, edges represent the human mobility trend and temporal edges correspond to node attributes across time. The second approach is based on a statistical technique for COVID-19 mortality prediction in the United States that uses the SARIMA model and eXogenous variables. We evaluate these techniques on both state and national levels COVID-19 data in the United States and claim that the SARIMA and MCP models generated forecast values by the eXogenous variables can enrich the underlying model to capture complexity in respectively national and state levels data. We demonstrate significant enhancement in the forecasting accuracy for a COVID-19 dataset, with a maximum improvement in forecasting accuracy by **64.58%** and **59.18%** (on average) over the GCN-LSTM

model in the national level data, and **58.79%** and **52.40%** (on average) over the GCN-LSTM model in the state level data. Additionally, our proposed model outperforms a parallel study (AUG-NN) by **27.35%** improvement of accuracy on average.

CCS CONCEPTS

- Computer Methodologies → Time-Series Analysis.

KEYWORDS

COVID-19, Graph Neural Networks, Knowledge Graph, Time-Series Analysis, SARIMAX

ACM Reference Format:

Mohammadhossein Toutiaee, Xiaochuan Li, Yogesh Chaudhari, Shopnine Sivaraja, Aishwarya Venkataraj, Indrajeet Javeri, Yuan Ke, Ismailcem Arpinar, Nicole Lazar, and John Miller. 2021. Improving COVID-19 Forecasting using eXogenous Variables. In *MileTS '21: 7th KDD Workshop on Mining and Learning from Time Series, August 14th, 2021, Singapore*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/nmnnnnnn.nnnnnnn>

1 INTRODUCTION

The outbreak of the COVID-19 pandemic from early 2020 until today has resulted in over 189M infected individuals and in over 4M deaths worldwide [1]. The ability to forecast the number of infections and deaths is vital to policymakers since they can manage healthcare resources, control disease upsurges, and take preventive actions when necessary to ensure public health safety. As the number of mortality and morbidity in the U.S. continued to rise within 2020, many states enforced the lockdown policy, practiced

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MileTS '21, August 14th, 2021, Singapore

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9999-9/18/06...\$15.00
<https://doi.org/10.1145/nmnnnnnn.nnnnnnn>

remote working, and imposed social distancing to slow the spread of COVID-19. These policies also affect individual mobility.

On the other hand, the number of cases in hospitals, ICUs, and on ventilators also increase as a result of the pandemic. Therefore, mobility and hospitalization patterns both at the national and local levels can provide useful measures for predicting the pandemic course, especially when this data is included in the pandemic analysis [12].

A great amount of research has been conducted since the beginning of the COVID-19 pandemic on forecasting the number of people affected. Early studies such as the one proposed in [4] used the SIR model for forecasting the infection rate in different countries. Other work [7] used ARIMA to predict the daily death rate in different states in the United States. [10] introduced AUG-NN model that enriches neural network models by augmentation which resulted in significant improvements in the accuracy. They reported the forecast values on the national level data.

Time-series forecasting using Graph Neural Networks (GNN) has been introduced in various domains in the past, however, it is less studied in epidemic disease. For example, [15] used a Temporal Graph Convolutional Network (T-GCN) for traffic prediction while [13] used Graph Neural Networks for Stock Market Prediction. GNN based approach for COVID-19 prediction discussed in [12], generated forecast values by using spatio-temporal mobility data. Although their work is impressive, they reported COVID-19 forecasting on a very small scale, i.e. top 20 most populated counties in the United States.

In this work, we extend the current research of pandemic modeling by proposing novel approaches for predicting daily death cases in the United States on both national and state levels. We introduce a spatio-temporal graph convolutional network that can capture complex dynamics by including mobility patterns across different states and a statistical model that generates forecasts by eXogenous variables. With extensive experiments among proposed methods, we demonstrate the power of eXogenous variables combined with lagged variables within the predictive models and conclude with an analysis of eXogenous variables and their potential in monitoring virus spread.

2 DATASETS

2.1 Aggregate Mobility Data

The mobility data used in the study is obtained from COVID-19 U.S. Flows [11]. This dataset consists of dynamic human mobility patterns across the United States in the form of the daily and weekly population flows at three geographic scales: census tract, county, and state. The spatio-temporal data is obtained by analyzing, computing, and aggregating the millions of anonymous mobile phone users' visit trajectories to various places provided by "SafeGraph". To be specific, we use the daily and weekly state-to-state daily population flow [8] starting from January 19, 2020 to January 19, 2021. The data files consist of the unique identifiers, latitudes and longitudes for the origin and destination states, date, visitor flows (estimated number of visitors detected by SafeGraph between two geographic units), and the population flow (estimated population flows between two geographic units, inferred from visitor flows).

2.2 COVID-19 Dataset

The COVID-19 dataset available on GitHub comprises data about COVID-19 cases and deaths in the United States, both at the national and state levels. It also contains several eXogenous variables such as "hospitalizedCurrently", "inICUCurrently", "onVentilatorCurrently", etc.¹

The national-level data spans from January 13, 2020 to March 7, 2021, which consists of 420 days. We ignore the first 44 days due to missing values and start from February 26, 2020, when the first COVID-19 death in the U.S. was recorded. As a result, the national level time-series studied in this paper contains 376 days. We used the first 236 days as our training set and the rest 140 days as the test set. The state-level data spans from January 19, 2020 to January 19, 2021, which consists of 367 days. Similarly, we ignore the first 70 days and start from March 29, 2020 to avoid missing values. Then, the state-level time-series studied in this paper contains 297 days for each state. We used the first 185 days as the initial training set and implement a two-week ahead rolling window forecast to assess the performance of various methods.

3 METHODOLOGY

3.1 SARIMAX

Weekly seasonality of daily death increase can be observed using Auto-Correlation Function (ACF), as shown in Figure 1. Then, we propose to fit the daily death increase by a Seasonal Autoregressive Integrated Moving Average (SARIMA) [3] model defined as below

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^Dy_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t, \quad (1)$$

where y_t is a variable to forecast, i.e., the logarithm of *deathIncrease*, $t = 1, 2, \dots$, $\varphi_p(B)$ is a regular AR polynomial of order p , $\theta_q(B)$ is a regular MA polynomial of order q , $\Phi_P(B^s)$ is a seasonal AR polynomial of order P , and $\Theta_Q(B^s)$ is a seasonal MA polynomial of order Q . The differencing operator ∇^d and the seasonal differencing operator ∇_s^D eliminate the non-seasonal and seasonal non-stationarity, respectively.

The SARIMA with eXogenous factor (SARIMAX) model is an extension of the SARIMA model in (1), which has the ability to include eXogenous variables, such as hospitalization and ICU occupancy rates. The SARIMAX model can be defined as:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^Dy_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t + \sum_{i=1}^n\beta_i x_t^i, \quad (2)$$

where $\{x_t^1, \dots, x_t^n\}$ are the n eXogenous variables defined at time t with coefficients $\{\beta_1, \dots, \beta_n\}$. Further, we apply a log transformation to categorical variables.

3.2 Minimax Concave Penalty

We also consider a penalized linear regression approach to predict the daily death increase by historical data and eXogenous explanatory variables. We denote $z_t = \log y_t - \log y_{(t-7)}$ as the weekly log-return of *deathIncrease* at time t and x_t the eXogenous hospitalization variables at time t . We linearly regress z_t on $\{z_{t-h}, \dots, z_{t-(h+k-1)}\}$ and $\{x_{t-h}, \dots, x_{t-(h+k-1)}\}$, where $k = 14$

¹All code and data available at https://github.com/mh2t/COVID_Forecasting_Using_eX_Vars

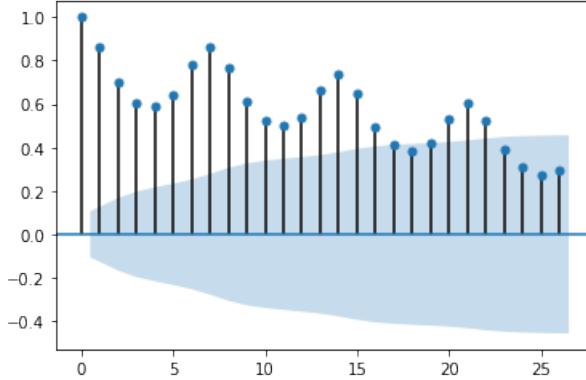


Figure 1: ACF plot of the daily death increase

and $1 \leq h \leq 14$ is a horizon parameter. The *horizon* refers to the number of days into the future for which forecast values are to be generated. Since the state level hospitalization variables are highly correlated, we first implement a sure independent screening [6] procedure to reduce the dimensionality of \mathbf{x}_t . As a result, only the top 7 exogenous variables are included in the following penalized regression model:

$$Q(\boldsymbol{\beta} | \mathbf{X}, \mathbf{z}) = \frac{1}{2N} \|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^d P_Y(\beta_j; \lambda), \quad (3)$$

where $\mathbf{z} \in \mathbb{R}^n$ is the vector of response variables, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix of all predictors and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$ is a vector of unknown regression coefficients. Besides, $P_Y(\cdot; \lambda)$ is the Minimax Concave Penalty (MCP) [14] which satisfies

$$P_Y(\beta; \lambda) = \begin{cases} \lambda|x| - \frac{\beta^2}{2\gamma}, & \text{if } |\beta| \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\beta| > \gamma\lambda. \end{cases} \quad (4)$$

3.3 Vector Auto-Regressive Model

Vector Auto-Regressive model [16] is another popular approach to model and predicts multivariate time-series. For a p dimensional response vector of interest, say:

$$\mathbf{y}_t = (y_{1,t}, y_{2,t}, \dots, y_{n,t})^T, \quad (5)$$

a vector auto-regressive model of order q , i.e. VAR(q), is defined as

$$\mathbf{y}_t = \delta + \Phi^{(0)} \mathbf{y}_{t-q} + \Phi^{(1)} \mathbf{y}_{t-q+1} + \dots + \Phi^{(q-1)} \mathbf{y}_{t-1} + \epsilon_t \quad (6)$$

where $\delta \in \mathbb{R}^n$ is an intercept vector, $\Phi^{(s)} \in \mathbb{R}^{n \times n}$ ($s = 0, \dots, p-1$) are regression coefficient matrices, and $\epsilon_t \in \mathbb{R}^n$ is an error vector.

Similar to the data pre-processing procedure described in Section 3.2, a weekly log-return has been applied to both the response vector and exogenous hospitalization variables to remove the seasonality.

3.4 Random Walk

By definition, a candidate series follows a random walk if the first differences are random (non-stationary). Random Walk (RW) is a common technique in graphical models [2], and it is widely used in webpage ranking, image segmentation, and time-series analysis.

The most practical usage of RW is in financial markets, where it states that the historical trend of a market cannot be used to forecast its future trend. Many studies showed that the RW method is applicable to most time-series data, especially when the samples have the same distribution and are independent of each other. A Gaussian Random Walk for variable y_t can be written by:

$$y_t = y_{t-1} + \epsilon_t \quad (7)$$

where ϵ_t follows Gaussian distribution.

3.5 GCN-LSTM

GCN: The main idea behind Graph Convolutional Network (GCN) models is that a given input signal (node) can be enriched via information propagation from its neighbors to improve a future prediction task. The neighbors are often defined by constructing a network of inputs where nodes and connections represent features and relations between them. This study [9] formulated an oracle definition upon the message-passing framework that was inspired by many previously proposed methods. In such settings, the spatial dependency is obtained by the 2-layer GCN architecture through the following formula:

$$f(X, A) = \sigma(\hat{A} \text{Relu}(\hat{A} X W_0) W_1) \quad (8)$$

where X denotes the feature matrix, A denotes the adjacency matrix, W_0 and W_1 respectively denote the weight matrix in the first and second layer.

Mobility Network Graph: In pandemic modeling, we usually analyze the global model through multiple time-series data obtained at the local level, which denotes the spreading dynamics in each region. The forecasting task is usually defined as a regression model that takes in a time-series of $t-k, \dots, t-1, t$ and emits a single value $t+1$. In a general case, the output can be a series of future time points $t+1, t+2, \dots$ as generated forecast values. However, the regression model requires an adjustment for modeling human mobility across regions. Mobility data forms a spatial-graph, where a region i is denoted by nodes, and every node can be connected to other nodes j, k, z, \dots , and weighted edges represent the strength of relations between the nodes.

Binary Graph: We constructed a binary adjacency matrix to feed into the trainer. This adjacency matrix has been created by the following steps: (1) the average of mobility data along the time point has been obtained, and (2) if the number of movements from origins to a destination is among the top 20%, the corresponding cell was assigned 1, and 0 otherwise. (3) Finally, the matrix was corrected to be a full-rank matrix by an orthonormal set obtained via the Gram–Schmidt process. We used `matlib` package in R for this purpose.

GCN-LSTM: Disease epidemic forecasting is a quintessential example of spatio-temporal problems for which we present a deep neural network framework that captures the number of deaths using spatio-temporal data. The task is challenging due to two main inter-twined factors: (1) the complex spatial dependency between time-series of each state, and (2) non-linear temporal dynamics with changing non-pharmaceutical interventions (NPI) such as mobility trends.

We attempt to populate a temporal knowledge graph using mobility patterns, since people moving around the regions with similar

epidemic patterns may contribute more to the forecasting process. The people traveling serve as the ground truth for training a GCN for identifying the underlying graph between sub-regions. Next, the constructed graph embedding from the GCN model is used to feed into an LSTM model to forecast the pandemic in the future. Notice that the graph embedding provides knowledge about the forecasting system, and the LSTM provides a direction for how to leverage the GCN output in the COVID-19 forecasting task.

3.6 Rolling Validation for Multiple Horizons

Classical multi-folds cross-validation approaches are not directly applicable to time-series data due to the existence of serial dependence. Instead, we follow a rolling-validation scheme in this study. To be specific, we reserve the first 60% of days in the time-series as the training set and use the rest 40% of days to make a rolling window forecast. For example, a forecast with horizon parameter $1 \leq h \leq 14$ is obtained using the model trained in the training, an input from a rolling window contains the information from $t - 13$ to time t , and an out-of-sample forecast for time $t + h$. Then, we move the rolling window forward by two weeks and repeat the above process. The two weeks ahead forecast policy is suggested by CDC. The forecast accuracy is measured by the symmetric mean absolute percentage error (sMAPE) which the smaller the better.

4 EXPERIMENTS

4.1 Hyperparameters and Architectures

GCN-LSTM: The GCN-LSTM architecture used in this experiment consists of two GCN layers followed by one LSTM layer. The model is built using StellarGraph library [5]. The sizes of each of these layers vary for each horizon. The size of GCN layers falls between 10 and 32 while the size of LSTM layers falls between 150 and 300. Information from 10 previous lags is used for forecasting future instances.

SARIMAX: The eXogenous variables used in SARIMAX at both the national and state levels are "hospitalizedCurrently" and "inIcuCurrently". The specific model we fit is a SARIMAX(4, 1, 4) \times (3, 1, 1, 7) model with a constant trend.

MCP: We fit an MCP model and consider the serial dependence among daily new cases in hospitals. Each day's new hospitalized count is dependent on the previous 14 days of observations. The regularized parameters in MCP are selected by a multi-fold cross-validation. Since the number of predictors is large, a Sure Independence Screening procedure is used to remove the uninformative predictors. As a result, 7 predictors are included in the MCP model.

VAR: Similar to SARIMAX model, VAR model contains multiple predictors including "inIcuCurrently", "hospitalizedCurrently", "hospitalizedCumulative" and "onVentilatorCurrently" in addition to "deathIncrease" variable. The hyperparameters are selected by the Bayesian Information Criterion (BIC).

4.2 Baseline

We compare the performance of all the statistical models and the GCN-LSTM model with a baseline Random Walk (RW) model. For each model, we make a 14-day ahead rolling window forecast as introduced in Section 3.6. The forecast performance is measured by the sMAPE score at each horizon $1 \leq h \leq 14$.

The national-level forecast results are reported in Table 1. The SARIMAX model performs the best on this dataset with the lowest sMAPE score. RW has slightly better results than other models at horizons $h = 7$ and 14.

Table 2 shows the forecast results on the state-level data. For each state in the U.S., we make a 14-day ahead rolling window forecast with horizons $1 \leq h \leq 14$. Then, we aggregate the state-level forecast into a forecast for the national level daily death increase. The sMAPE scores on this aggregated forecast are then used to compare all models. MCP model performs the best on this experiment while RW has the worst performance in most scenarios.

Table 3 reports the sMAPE scores for different models on a set of representative states. We notice that the scores for each model are much higher than the aggregated results reported in Table 2. Also, we find that the GCN-LSTM model performs the worst when the forecasts for these representative states are considered.

4.3 Mortality Prediction Performance

In Tables 1 and 2, we compare the forecasting performance of SARIMAX with a range of models. We report the sMAPE score for the predicted "deathIncrease". Table 1 shows the performance of models over the national-level data, and Table 2 shows the performance over aggregated state-level data for the United States mortality. All performance values were reported for 14 horizons. According to Table 1, the SARIMAX model achieves the best sMAPE score at each horizon and outperforms other models on the national level forecast. The best results were achieved by evaluating several combinations, i.e. hospitalizedIncrease, hospitalizedCurrently, etc.

On the other hand, the MCP model performs best on the state level forecast which is evidenced by the results in Table 2. This trend is due to the fact that we included more lags ($k = 14$) in MCP compared to SARIMAX, and MCP tends to manage overfitting better than SARIMAX since it uses a regularized term. Although SARIMAX appears to be more accurate than MCP in forecasting national-level data, it is inferior to MCP in forecasting state-level data, owing to the fact that the global combination might not be the best option for all the states.

Further, we discovered a general pattern that the forecast results can be improved by including eXogenous variables. That may explain the success of SARIMAX and MCP in our study. Interestingly, introducing additional variables resulted in worse performance for the SARIMAX model. Table 3 compares different models for selected states.

5 CONCLUSION AND FUTURE WORK

In this paper, we evaluated and compared several statistical and machine learning models to forecast the pandemic course in the United States, using national and state levels data. We studied the effectiveness of the mobility data in the COVID-19 prediction problem regarding the accuracy and discussed the benefits of including eXogenous variables. The advantage of SARIMAX and MCP, over other competitive methods such as GCN-LSTM, is that they can incorporate eXogenous variables such as hospitalization, ICU occupancy rate, and the count of patients who require a ventilator

Table 1: Multi-Horizon (h) Rolling Forecasts for the United States (National Level Data): Competitive Models (sMAPE).

Horizon	RW	GCN	SARIMAX	SARIMA	MCP	VAR
$h = 1$	29.20	28.87	13.25	16.02	15.61	15.08
$h = 2$	48.40	33.81	15.10	19.36	17.13	16.09
$h = 3$	53.50	35.97	15.47	19.13	17.85	16.55
$h = 4$	54.20	32.26	15.21	19.26	18.02	16.53
$h = 5$	50.10	29.83	14.98	20.40	18.02	16.62
$h = 6$	32.20	36.76	14.18	20.49	17.87	16.50
$h = 7$	18.20	34.04	14.03	20.45	17.86	16.75
$h = 8$	31.30	34.29	13.52	22.70	23.03	18.61
$h = 9$	47.40	48.39	14.06	24.92	22.94	19.58
$h = 10$	52.10	54.37	14.22	25.77	23.24	19.84
$h = 11$	53.60	32.07	14.25	25.75	23.39	19.86
$h = 12$	49.90	28.39	14.52	26.27	23.16	20.10
$h = 13$	34.20	35.63	15.05	27.00	22.91	19.79
$h = 14$	24.30	32.84	15.22	26.94	22.97	20.17
Average	41.33	35.54	14.50	22.46	20.29	18.01

Table 2: Multi-Horizon (h) Rolling Forecasts for the United States (State Level Data): Competitive Models (sMAPE).

Horizon	RW	GCN	SARIMAX	SARIMA	MCP	VAR
$h = 1$	29.66	24.99	18.22	15.41	16.28	15.40
$h = 2$	50.17	34.42	18.01	16.36	16.09	15.42
$h = 3$	53.98	34.75	18.42	17.31	16.17	16.26
$h = 4$	54.30	34.42	17.53	17.84	16.38	16.32
$h = 5$	51.71	30.41	17.58	17.95	16.61	15.90
$h = 6$	33.89	30.81	18.73	17.76	16.51	16.17
$h = 7$	17.44	35.74	18.04	17.84	16.63	16.65
$h = 8$	31.54	34.62	18.64	21.21	18.17	19.65
$h = 9$	47.21	38.88	18.64	21.89	17.82	20.67
$h = 10$	51.13	44.04	18.52	22.98	19.86	21.07
$h = 11$	52.70	44.15	19.02	23.24	19.40	22.36
$h = 12$	49.94	43.34	29.08	23.62	17.98	20.36
$h = 13$	34.91	38.21	26.88	23.63	18.19	20.31
$h = 14$	23.30	44.46	23.10	23.33	18.32	20.13
Average	41.56	36.66	20.03	20.81	17.55	18.33

mask into a multivariate time-series forecast. The empirical advantages of including such exogenous variables were justified by our experiments.

Our work adds to the growing body of epidemic disease modeling with a novel approach to combine the multivariate time-series analysis with human mobility data and exogenous variables. We expect the findings in this paper can motivate future studies for selecting and incorporating important exogenous variables into time-series modeling to enhance the prediction.

REFERENCES

- [1] 2021. WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int/>
- [2] David Aldous and Jim Fill. 2014. Reversible Markov chains and random walks on graphs, 2002. *Unfinished monograph, recompiled* 2002 (2014).
- [3] Nari Sivanandam Arunraj, Diane Ahrens, and Michael Fernandes. 2016. Application of SARIMAX model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems (IJORIS)* 7, 2 (2016), 1–21.
- [4] G.D. Barmparis and G.P. Tsironis. 2020. Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach. *Chaos, Solitons & Fractals* 135 (Jun 2020), 109842. <https://doi.org/10.1016/j.chaos.2020.109842>
- [5] CSIRO's Data61. 2018. StellarGraph Machine Learning Library. <https://github.com/stellargraph/stellargraph>
- [6] Jianqing Fan and Jinchi Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 5 (2008), 849–911.
- [7] Shayan Fazeli, Babak Moatamed, and Majid Sarrafzadeh. 2020. Statistical Analytics and Regional Representation Learning for COVID-19 Pandemic Understanding. [arXiv:cs.CY/2008.07342](https://arxiv.org/abs/2008.07342)
- [8] GeoDS. 2021. GeoDS/COVID19USFlows. <https://github.com/GeoDS/COVID19USFlows>
- [9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*. PMLR, 1263–1272.
- [10] Indrajeet Y Javeri, Mohammadhosseini Toutiaee, Ismailcem B Arpinar, Tom W Miller, and John A Miller. 2021. Improving Neural Networks for Time Series Forecasting using Data Augmentation and AutoML. [arXiv preprint arXiv:2103.01992](https://arxiv.org/abs/2103.01992) (2021).
- [11] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jimmeng 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.
- [12] 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](https://arxiv.org/abs/2007.03113) (2020).
- [13] Daiki Matsunaga, Toyotaro Suzumura, and Toshihiro Takahashi. 2019. Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis. [arXiv:cs.LG/1909.10660](https://arxiv.org/abs/1909.10660)
- [14] Cun-Hui Zhang et al. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* 38, 2 (2010), 894–942.
- [15] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* (2019).
- [16] Eric Zivot and Jiahui Wang. 2006. Vector autoregressive models for multivariate time series. *Modeling financial time series with S-PLUS®* (2006), 385–429.

A APPENDIX

Table 3: Multi-Horizon (h) Rolling Forecasts for Selected States: Random Walk, SARIMA, GCN-LSTM, SARIMAX, VAR and MCP (sMAPE).

State	CA						GA						IL					
	Horizons	RW	SR	GCN	SRX	VAR	MCP	RW	SR	GCN	SRX	VAR	MCP	RW	SR	GCN	SRX	VAR
$h = 1$	54.10	38.60	55.63	35.63	42.02	35.10	68.80	54.40	58.74	59.54	51.87	53.13	41.60	27.40	61.76	30.49	32.72	32.99
$h = 2$	68.50	39.40	57.86	38.02	41.06	34.42	82.00	51.50	62.68	57.52	51.93	55.08	53.00	29.00	64.67	29.25	31.63	32.73
$h = 3$	76.20	40.80	64.93	36.44	41.11	33.93	87.00	52.80	68.37	59.93	51.09	52.49	54.90	30.10	68.11	30.05	31.67	32.43
$h = 4$	75.90	41.70	68.76	37.50	41.06	33.70	87.60	52.60	70.51	56.92	51.06	53.57	55.10	31.00	70.67	29.44	31.38	32.97
$h = 5$	66.10	41.70	68.31	35.31	41.32	34.34	83.30	52.20	70.14	56.42	51.18	53.45	52.10	32.30	71.32	29.17	31.07	32.84
$h = 6$	55.60	40.60	65.53	38.44	41.40	34.98	69.10	51.60	69.05	56.36	51.13	53.62	43.20	31.70	70.93	30.76	31.57	31.94
$h = 7$	41.40	40.90	60.42	40.54	41.48	34.73	48.80	52.50	62.67	54.57	51.54	53.77	31.00	31.40	70.04	30.60	31.47	31.78
$h = 8$	59.60	47.30	55.34	37.00	42.62	41.94	69.70	56.30	59.24	61.46	61.13	63.82	40.40	35.20	70.57	30.52	34.45	36.89
$h = 9$	68.80	47.90	58.54	37.55	43.28	42.44	79.70	56.80	63.7	61.21	60.27	60.75	51.10	37.00	73.42	31.34	33.21	34.82
$h = 10$	74.70	49.90	66.69	42.21	43.44	42.46	82.40	57.60	67.85	61.98	60.85	60.81	54.50	38.10	75.41	31.89	33.47	39.67
$h = 11$	74.50	50.60	68.88	36.79	43.40	43.02	83.70	57.30	72.01	58.91	60.37	62.04	56.20	38.50	76.99	29.48	34.41	36.63
$h = 12$	67.80	51.40	67.8	39.08	43.75	43.10	79.50	57.00	73	59.55	60.56	59.43	52.70	39.90	78.53	30.19	33.87	35.09
$h = 13$	57.40	49.70	64.91	37.79	43.93	42.87	66.10	55.80	71.01	59.46	60.45	60.58	42.40	41.10	78.42	31.59	34.64	35.03
$h = 14$	48.60	50.40	59.42	34.93	44.37	43.80	55.50	56.20	62.24	58.79	60.74	61.46	36.20	39.80	77.51	32.84	35.68	34.83
Average	63.51	45.06	63.07	37.66	42.45	38.63	74.51	54.61	66.52	58.76	56.01	57.43	47.46	34.46	72.03	30.54	32.95	34.33
State	TX						NY						PA					
Horizons	RW	SR	GCN	SRX	VAR	MCP	RW	SR	GCN	SRX	VAR	MCP	RW	SR	GCN	SRX	VAR	MCP
$h = 1$	58.80	29.00	36.01	27.88	28.19	29.53	26.50	25.70	35.04	20.76	24.06	24.62	57.20	46.40	70.86	45.86	46.30	42.96
$h = 2$	76.40	32.60	57.29	25.92	27.50	29.66	25.20	28.20	36.56	20.47	24.46	25.79	69.70	47.20	74.13	46.85	44.87	42.62
$h = 3$	84.30	33.30	70.94	32.02	27.76	29.43	24.50	32.80	37.84	20.25	24.56	25.45	78.50	47.00	77.8	44.13	45.24	42.97
$h = 4$	84.30	33.00	74.1	26.16	27.58	28.57	27.30	37.50	39.69	21.92	24.28	24.93	79.00	46.60	80.19	44.50	46.56	41.88
$h = 5$	75.60	32.60	74.4	34.15	27.86	28.39	26.00	40.40	41.71	22.99	24.65	25.08	75.80	46.90	80.9	45.30	46.23	41.69
$h = 6$	60.70	32.60	62.65	36.06	27.59	28.42	30.00	28.70	42.72	23.61	24.41	25.80	61.60	46.30	80.57	44.05	46.26	42.61
$h = 7$	27.20	32.70	47.52	28.22	27.36	28.43	27.90	31.90	44.08	23.31	25.22	27.32	44.30	46.80	79.94	43.93	45.99	42.14
$h = 8$	57.60	39.40	38.89	33.59	33.47	31.64	31.10	40.10	45.72	33.50	31.00	30.09	55.10	51.30	80.92	46.75	51.03	51.61
$h = 9$	74.20	41.40	58.74	30.72	32.79	32.58	33.40	44.20	47.83	31.02	31.54	29.33	69.30	52.70	82.98	46.59	51.03	55.01
$h = 10$	79.90	42.20	72.86	30.80	32.76	32.26	34.20	49.40	47.65	42.86	30.72	27.41	72.50	52.90	84.92	44.69	49.29	57.48
$h = 11$	79.60	41.90	76.34	30.62	32.68	32.37	37.60	55.00	49.62	29.57	30.49	30.28	76.00	53.60	87.59	46.79	48.18	55.89
$h = 12$	72.00	41.60	76.26	41.85	32.58	32.11	37.60	57.80	52.95	28.17	30.20	32.02	71.40	55.80	87.83	46.22	49.90	54.43
$h = 13$	60.30	41.50	64.47	35.27	32.52	32.17	40.30	46.30	53.56	28.31	30.89	32.96	61.80	54.60	87.33	44.91	48.56	52.90
$h = 14$	33.80	41.40	50.34	27.48	32.13	32.25	42.00	51.30	55.03	32.02	31.59	32.92	48.90	55.10	87.83	45.80	46.71	54.30
Average	66.05	36.80	61.49	31.48	30.20	30.56	31.69	40.66	45.00	27.05	27.72	28.14	65.79	50.23	81.70	45.46	47.58	48.46