



SARS-CoV-2 Dissemination Using a Network of the US Counties

Patrick Urrutia¹ · David Wren¹ · Chrysafis Vogiatzis²  · Ruriko Yoshida¹

Received: 29 November 2021 / Accepted: 2 April 2022 / Published online: 30 April 2022
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

Abstract

During 2020 and 2021, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission has been increasing among the world's population at an alarming rate. Reducing the spread of SARS-CoV-2 and other diseases that are spread in similar manners is paramount for public health officials as they seek to effectively manage resources and potential population control measures such as social distancing and quarantines. By analyzing the US county network structure, one can model and interdict potential higher infection areas. County officials can provide targeted information, preparedness training, and increase testing the researchers conclude that traditional the researchers conclude that traditional in these areas. While these approaches may provide adequate countermeasures for localized areas, they are inadequate for the holistic USA. We solve this problem by collecting coronavirus disease 2019 (COVID-19) infections and deaths from the Center for Disease Control and Prevention, and adjacency between all counties obtained from the United States Census Bureau. Generalized network autoregressive (GNAR) time series models have been proposed as an efficient learning algorithm for networked datasets. This work fuses network science and operations research techniques to univariately model COVID-19 cases, deaths, and current survivors across the US county network structure.

Keywords COVID-19 · Time series · Generalized network autoregressive processes (GNAR) · Network analysis

This article is part of the Topical Collection on *Dynamics of Disaster*.

Ruriko Yoshida is supported partially from NSF Statistics Program DMS 1916037.

✉ Chrysafis Vogiatzis
chrys@illinois.edu

Extended author information available on the last page of the article

1 Introduction

The severe acute respiratory syndrome coronavirus (SARS-CoV-2) that was first detected in late 2019 and the resultant disease (COVID-19) have completely upended the way we led our lives. The global pandemic has led to debilitating damage to our lives, economies, health care systems, and food security [1]. The virus is extremely transmissible, spreading through water droplets produced when talking, coughing, or sneezing [2]. Since the virus' first reports in 2019, it has mutated multiple times and some of these mutations have proven to spread faster and infect easier [3]. Additionally, when a virus mutates, combating it with vaccines and other public health measures becomes increasingly difficult [4]. Hence, it is in our best interest to investigate the spreading mechanisms and obtain ways to predict outbreaks, which are one of the goals of this study. The transmission of the virus has been further aided by modern society's hypermobility [5]. Early on, COVID-19 spread fast within China and other countries due to both international and national travel, through air, land, and seas.

In previous works, the relationship between human mobility and epidemics spreading has been investigated. Recently, we have had some very important works relating COVID-19 to traffic. Indicatively, we mention the fundamental work of Kraemer et al. [6] which analyzed human mobility data and traced infection metrics in the early and later stages of the COVID-19 pandemic. The results reinforce that earlier in the pandemic, strict travel restrictions are helpful and lead to easier confinement and control; later, once the outbreak is spread, travel restrictions are less useful and local measures, such as social distancing and masking, are preferable. In a second recent work [7], the authors posit that while traffic and human mobility are often the culprits for driving viral spreading, the traffic network structure is often overlooked from these studies. Hence, they propose a traffic-driven model that accounts for that; here, we also account for human mobility through the transportation network and consider the effects of traffic on edge affectthe spread of a disease.

Specifically, in this work we also investigate whether and how travel patterns affect COVID-19 dissemination; we do so by employing generalized network autoregression on a proxy of the US transportation network. We generate and use a network of all counties in the USA in an effort to forecast the spread of COVID-19 using data available for each county as well as travel patterns across counties. The remainder of the manuscript is organized as follows. First, in Sect. 2, we provide a brief literature review on models that have been put to the use to forecast COVID-19 cases and to protect the communities from its transmission. Then, in Sect. 3, we discuss our approach using the generalized network autoregression (GNAR). We also provide a description of the data that were acquired to perform the analysis in Sect. 4. Section 5 presents the computational experiments and the results we observed during our analysis. We conclude this work in Sect. 6.

2 Literature Review

Due to the impact of COVID-19 in our daily lives, a lot of research has already appeared on the analysis of the spread of the disease. That said, epidemics and pandemics such as the one caused by SARS-CoV-2 are not a recent phenomenon for humanity. As an example, the “Spanish flu” ravaged the world in the early 1900s. Shortly after the outbreak of the disease, in 1918, researchers Kermack and McKendrick published papers that presented mathematical models for predicting the number of infections in a population as a function of time: the assumption was that it is valid to split the population into smaller clusters or “compartments” when analyzing a disease’s propagation through a population [8]. Their foundational work continues to help epidemiologists model outbreaks of diseases today.

More recently, epidemiological models such as the Susceptible-Infected-Removed (Recovered) (SIR) and Susceptible-Exposed-Infected-Removed (Recovered) (SEIR) and other extensions have been put to use to model the movement of individuals from one “compartment” (i.e., Susceptible, Exposed, Infected, Removed) to the next [9]. As an example, a person may be moved from the initial state of Susceptible to the intermediate state of Infected upon exposure to and infection with a disease; later that same person may be categorized as Removed once they recover. As expected, such epidemiological models have been applied in the fight against COVID-19. These models have been largely successful, revealing their utility for policy to prevent the spread of disease.

In Cameroon, research based on SIR determined that the number of COVID-19 cases was limited due to the health precautions taken [10]. Another similar application of the SIR model originates from Saudi Arabia, where researchers analyzed the number of COVID-19 cases and deaths both with and without public health measures such as quarantine enforcement [11]. Although SIR models have been accurate enough in predicting the size of the COVID-19 outbreaks, more recent research indicates that individuals who contract the virus once can become infected again [12], necessitating a means to dynamically update the parameters of the SIR model in an effort to improve its predictive power. In [13], the authors propose time-varying these parameters to account for changes over time, using machine learning to determine exactly how to update these parameters.

Moreover, the incubation period of COVID-19 (i.e., the period during which an infected individual bears no symptoms yet can still transmit the virus to others) has proven to be an important factor in the spread of COVID-19 [14]. While asymptomatic, some recently infected individuals can unknowingly spread COVID-19, a fact that needs to be included in epidemic models [15].

Similarly to the work from Saudi Arabia, researchers in Wuhan used the SEIR model to analyze the impacts of public health measures such as quarantines and restrictions of movement [16]. Following the time-varying updates recommended in [13], researchers in Portugal dynamically adjusted the exposure rates and other parameters in order to simulate infected asymptomatic individuals who can spread the virus [17].

Another methodology that has been put to the use in the fight against COVID-19 is agent-based simulation modeling. Even before SARS-CoV-2 first appeared, researchers have been using simulation in conjunction with transit data; the insight is that population movements will critically affect the spread of diseases [18]. As far as COVID-19 is concerned, agent-based simulation models have been used to test the effect of public health mitigation efforts. As an example, in [19] using agent-based simulation models, the researchers conclude that traditional measures such as mask-wearing and social distancing, as well as lockdowns, are viable tools in the fight against COVID-19.

The work presented here is heavily motivated by the literature on diffusion processes on networks (see [20]). COVID-19 and its spread is no exception, with many works pointing to the relationship between outbreaks and population movements through the transportation network [6, 21]. Since 2020, we have seen a multitude of works investigating the network spreading dynamics in air and rail networks as well as public transit [22–25].

Finally, we discuss time series models. Autoregressive Integrated Moving Average (ARIMA) models regress a forecast value onto previous values of the time series [26]. Thus, ARIMA models seek to describe autocorrelations in the time series data [27]. In India, researchers used ARIMA to model and predict COVID-19 infections [28], with higher accuracy of other moving average and exponential smoothing models. Still in India, other research analyzed COVID-19 spreading trends using both an ARIMA and a Holt-Winters model (Holt-Winters accounts for trends and seasonality) [29]. The accuracy of the models (during the time period specified) proved very high, at 99.8%. Another example of ARIMA and Holt-Winters models comes from Jakarta [30], finding that ARIMA outperforms the other time series approaches. Last, ensemble methods include a variety of time series models; the final prediction of an ensemble model is a combination of the time series models included [31]. Such an ensemble model was put to use in Nigeria. The time series model, called Prophet, processed missing values, seasonal effects, and outliers, allowing it to perform well against other models for predicting spread [32].

Researchers employing these techniques across the world can help leaders interdict the spread of the virus. What we mean by this statement is to use spread prediction in a way that informs mobility policy such that threat to human life is minimized. A recent interdiction policy, motivated by COVID-19, is presented in [33]. Interestingly, the authors utilize the mobility data and a set of different network science notions on a network obtained from the districts and boroughs in New York City. Outside the context of viral spread and epidemics, researchers have investigated the idea of using betweenness centrality and extensions, such as betweenness-accessibility [34] to identify the most critical links (i.e., streets or main arteries) and nodes (i.e., zip codes, cities, or counties) whose interdiction or closure lead to better isolation of areas. While our work does not focus on interdiction, our contributions can help policy-makers identify parts of the network that are more susceptible to increases in positivity rate.

3 The Generalized Network Autoregressive Process

In this section, first we describe the Generalized Network Autoregressive Process (GNAR) [35] and the associated \mathbb{R} package [36]. Then, we present the way that we adapt the GNAR model to our problem. We also provide the different metrics that we use to evaluate the performance of time series models.

3.1 The GNAR Process

Suppose we have a directed graph $\mathcal{G} = (N, E)$ where N is a set of nodes ($N = \{1, \dots, n\}$) and E is a set of edges. Suppose we have an edge $e = (i, j) \in E$ for $i, j \in N$ and suppose a direction of e is from a node i to a node j , then we write it as $i \rightarrow j$. For any $A \subset N$ we define the *neighbor set* of A as follows:

$$\mathcal{N}(A) := \{j \in N/A \mid i \rightarrow j, \text{ for } i \in A\}.$$

The r -th stage neighbors of a node $i \in N$ is defined as

$$\mathcal{N}^{(r)}(i) := \mathcal{N}\{\mathcal{N}^{(r-1)}\} / [\{\cup_{q=1}^{r-1} \mathcal{N}^{(q)}(i)\} \cup \{i\}],$$

for $r = 2, 3, \dots$ with $\mathcal{N}^{(1)}(i) = \mathcal{N}(\{i\})$.

Under this model, we assume that we can assign a weight $\mu_{i,j}$ on an edge (i, j) . We define a distance between nodes $i, j \in N$ such that there exists an edge $(i, j) \in E$ as $d_{i,j} = \mu_{i,j}^{-1}$. Then we define

$$w_{i,k} = \frac{\mu_{i,k}}{\sum_{l \in \mathcal{N}^{(r)}(i)} \mu_{i,l}}. \tag{1}$$

The GNAR model uses a *covariate* for an edge effect in different types of nodes by an additional attribute, such as infected or not infected in an epidemiological network. Assume that a covariate takes discrete values $\{1, \dots, C\} \subset \mathbb{Z}$. Then, let $w_{i,k,c}$ be $w_{i,k}$ for a covariate c such that

$$\sum_{q \in \mathcal{N}^{(r)}(i)} \sum_{c=1}^C w_{i,q,c} = 1.$$

Now we are ready to define the generalized network autoregressive processes (GNAR) model. Suppose we have a vector of random variables in

$$X_t := (X_{1,t}, \dots, X_{n,t}) \in \mathbb{R}^n$$

which varies over the time horizon and each random variable associates with a node. For each node $i \in N$ and time $t \in \{1, \dots, T\}$ a generalized network autoregressive processes model of order $(p, [s]) \in \mathbb{N} \times (\mathbb{N} \cup \{0\})^p$ on a vector of random variables X_t is

$$X_{i,t} := \sum_{j=1}^p \left(\alpha_{ij} X_{i,t-j} + \sum_{c=1}^C \sum_{r=1}^{s_j} \beta_{j,r,c} \sum_{q \in \mathcal{N}_t^{(r)}(i)} w_{i,q,c}^{(t)} X_{q,t-j} \right) \tag{2}$$

where $p \in \mathbb{N}$ is the maximum time lag, $[s] := (s_1, \dots, s_p)$, $s_j \in \mathbb{N} \cup \{0\}$ is the maximum stage of neighbor dependence for time lag j , $\mathcal{N}_t^{(r)}(i)$ is the r th stage neighbor set of a node i at time t , and $w_{i,q,c}^{(t)} \in [0, 1]$ is the connection weight between node i and node q at time t if the path corresponds to covariate c . $\alpha_{ij} \in \mathbb{R}$ is a parameter of autoregression at lag j for a node $i \in N$ and $\beta_{j,r,c} \in \mathbb{R}$ corresponds to the effect of the r th stage neighbors, at lag j , according to a covariate $c = 1, \dots, C$.

3.2 COVID-19 Analysis Using GNAR

In order to apply the GNAR model defined in this section to the county network on COVID-19 data, we set variables as follows.

Note that the GNAR model conducts a time series analysis on the time series data on the networks. The GNAR model assumes that the topology of the network is fixed over the time horizon $t > 0$. In this research, the network is the county network $\mathcal{G} = (N, E)$, where each node $i \in N$ is a county in the particular state in the USA and we draw an edge $(i, j) \in E$ between a county $i \in N$ and a county $j \in N$ if and only if a county i has commuters traveling to a county j . A weight μ_{ij} on each edge $(i, j) \in E$ is the number of commuters from a county $i \in N$ and a county $j \in N$. The GNAR model assumes that these weights μ_{ij} are fixed over the time horizon $t > 0$. Therefore, the input of the GNAR package includes these variables. Now, what we wish to infer using the GNAR model are random variables

$$X_t := (X_{1,t}, \dots, X_{n,t}) \in \mathbb{R}^n,$$

where $X_{i,t}$ is the number of COVID-19 cases of deaths from COVID-19 at a county $i \in N$ at the time $t > 0$.

In this research, we do not have differences between all nodes, i.e., we treat all counties in N as the same type. Therefore, we ignore this covariate index c , rendering the formulation of the GNAR model as follows. For each county $i \in N$ and time $t \in \{1, \dots, T\}$ a generalized network autoregressive processes model of order $(p, [s]) \in \mathbb{N} \times (\mathbb{N} \cup \{0\})^p$ on a vector of the numbers of COVID-19 cases of deaths from COVID-19 X_t is

$$X_{i,t} := \sum_{j=1}^p \left(\alpha_{ij} X_{i,t-j} + \sum_{r=1}^{s_j} \beta_{j,r} \sum_{q \in \mathcal{N}_t^{(r)}(i)} w_{i,q}^{(t)} X_{q,t-j} \right) \tag{3}$$

where $p \in \mathbb{N}$ is the maximum time lag, $[s] := (s_1, \dots, s_p)$, $s_j \in \mathbb{N} \cup \{0\}$ is the maximum stage of neighbor dependence for time lag j , $\mathcal{N}_t^{(r)}(i)$ is the r th stage neighbor set of a county i at time t , and $w_{i,q}^{(t)} \in [0, 1]$ is the connection weight between a county i and a county q at time t . $\alpha_{ij} \in \mathbb{R}$ is a user specific parameter (tuning parameter) of autoregression at lag j for a county $i \in N$ and a user specific parameter (tuning

parameter) $\beta_{j,r} \in \mathbb{R}$ corresponds to the effect of the r th stage neighbors, at lag j . These user specific parameters are defined by a user. In this paper, we select three combinations of tuning parameters after conducting a model selection discussed in Sect. 3.2.1.

In addition, note that $w_{i,k}$ for a county $i \in N$ and its neighbor $k \in \mathcal{N}_i^{(r)}$ is computed using the Eq. (1).

3.2.1 Model Parameters

The GNAR package takes in a number of parameters for its predictive time series models. For both the cases and the deaths, we adjusted two GNAR parameters to create three unique models. The first model fit applies a non-negative integer, `alphaOrder = 1`, that specifies a maximum time lag of 1 to model along with a vector of length `betaOrder = 0`, which specifies the maximum neighbor set to model at each of the time lags [36]. These parameters represent the time lag, p , and the maximum stage of neighbor dependence for each of the time lags, $[s]$, as discussed above. The second model sets `alphaOrder = 0` and `betaOrder = 1`. The third model is the default model in GNAR, with no parameter modifications, making both `alphaOrder = 0`, `betaOrder = 0`. We conduct a model selection by changing `alphaOrder` and `betaOrder` from 0 to 5 independently. Table 1 provides a summary of the model parameter combinations.

Because there are two prediction options (cases and deaths) and three model parameter selections, in total, we create 6 different combinations (e.g., Deaths - Model 1). Moreover, since we predict by state, each state has these 6 models for comparison.

3.3 Evaluation Performance

Measuring performance in traditional statistics often calls for measures of performance such as RMSE and adjusted R^2 . Although easily calculated, these measures do not measure errors in terms of the time horizon [37]. For outputs of a predictive time series model, performance can be measured by the mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE). The MAPE measures an estimated average of a model’s forecast performance over the time horizon, while the MASE measures the ratio of an estimated absolute error of the forecast divided and estimated absolute error of the naïve forecast method over the time horizon [38]. The MAPE commonly falls between 0 and 1, but can be skewed outside this

Table 1 Model Parameter Summary. We vary the value of `alphaOrder` and `betaOrder` to create three different models for the COVID-19 cases and deaths

Model 1	Model 2	Model 3
<code>alphaOrder = 1</code>	<code>alphaOrder = 0</code>	<code>betaOrder = 0</code>
<code>betaOrder = 1</code>	<code>betaOrder = 1</code>	<code>alphaOrder = 0</code>

range if actual values are close to zero [38]. The MASE is less than one if a model has smaller error than the naïve model's error and if a model has greater error than the naïve model's error, then it is greater than one [38]. The MAPE and MASE are defined by the following equations:

$$MAPE = \frac{\sum_{t=1}^N \left| \frac{Y_t - F_t}{Y_t} \right|}{N}, \quad (4)$$

$$MASE = \frac{\sum_{t=1}^N \left| \frac{Y_t - F_t}{Y_t - Y_{t-1}} \right|}{N}, \quad (5)$$

where Y_t is the observation at time t , F_t is the predicted value, and $Y_t - Y_{t-1}$ is error of the one-step naïve forecast.

In order for a model to have predictive power, its MASE must exceed the accuracy of the respective naïve model and we say it has a good forecasting if a model has the MAPE less than 0.2 [39].

When we measure the performance of each model in terms of the MASE and MAPE, we apply a *rolling horizon design* [40] in this paper. A rolling horizon design for a time series model is to assess accuracy of a time series model such that it updates the forecasted value successively using different subsets of previous and current observations, and then it takes averages of the performance of the model for different time periods.

4 Data

We obtain the data for this work from the United States Census Bureau (USCB), the United States of America Facts (USA Facts), and the Center for Disease Control and Prevention (CDC). The USA Facts obtains their data from the CDC [41] and updates the daily death count on their website [42]. Manipulating this data in Python, we transform the data into a usable format for the GNAR package, create our models, and assess them using a variety of evaluation performance metrics.

4.1 Data Description and Limitations

The data is entirely numerical, with no categorical predictors or response variables. No transformations are applied to the original data for the proposed models. We also note that the COVID-19 cases and deaths data meet the assumption of stationarity of errors because the noise of the data does not depend on the time at which the data was observed [27]. Autoregressive models require stationarity of the errors, meaning that the series' variance must be constant over a long time period [43].

Furthermore, we assume that the COVID-19 data is complete and accurate. Although human error and reporting standards affect the number of deaths and cases sometimes, on any given day, we assume the data obtained from the CDC is

accurate. Additionally, the data is autoregressive. Last, we assume the presence of no outliers [44]. The principal limitation of this data involves the constant nature of the commuting network structure [45]. The USCB compiled this commuting data over a 5-year period from 2011 to 2015, giving it a static property. We thus assume that the traffic and commuting patterns by county remain the same through the time of the COVID-19 pandemic.

4.2 County Network

The USA comprises 3,143 county or county equivalents as of 2020 [46]. 48 states use the term “county” to describe their administrative districts while Louisiana and Alaska use the terms “parishes” and “boroughs,” respectively. Each county is assigned a unique five-digit FIPS code. The first two digits represent the state’s FIPS code, while the latter three digits represent the county’s FIPS code within the state. This number serves as a uniform index for each county, facilitating county data sorting and filtering.

The number of counties per state varies widely across the USA, regardless of a state’s geographic size, population, or terrain. For example, Rhode Island, the state with the smallest land area, has 5 counties, while Alaska, the state with the largest land area, has 29 counties [47]. However, Delaware contains the least amount of counties and Texas contains the most. Moreover, the majority of the country’s population lives in only 143 of the 3143 counties as of 2020. Table 2 and Fig. 1 display the number of counties in each state.

Additionally, it is not uncommon for county information to change as time goes on. Counties can divide, merge, or rename themselves at any time, even if that time does not fall on a census year. For example, Colorado created Broomfield County from merging parts of other counties [51]. Shannon County, South Dakota, renamed itself to Ogala County in 2015 out of respect to its Native American heritage [52]. Of course, when new counties are formed or renamed, they are assigned new FIPS codes, which can complicate reporting of statistics later on. Regardless, many counties and states have protocols in place to prevent such mistakes.

In this work, we construct a network structure using the original commuting data from the USCB using Python. The commuting data comes in the form of a data frame with three columns: the county from which the individuals commute, the county to which the individuals commute, and the number of commuters. This represents a flow structure, where we can deduce how many commuters commute from one county to the next. Using Python, we transform this flow structure into a matrix format, with the row and column entries representing the “From” and “To” columns of the original data. Thus, one can easily search in this new commuting data matrix for how many commuters go from one county to the next.

In this research we divide the county network in each state. We designed the county commuting network structure for each state with the following information:

- Workers commuting from within-state counties to within-state counties.

Table 2 A list of all US counties per state (including the District of Columbia). As can be seen, the number of counties per state varies widely. Source: [48]

State	Counties
Alabama	67
Alaska	29
Arizona	15
Arkansas	75
California	58
Colorado	64
Connecticut	8
Delaware	3
District of Columbia	1
Florida	67
Georgia	159
Hawaii	5
Idaho	44
Illinois	102
Indiana	92
Iowa	99
Kansas	105
Kentucky	120
Louisiana	64
Maine	16
Maryland	24
Massachusetts	14
Michigan	83
Minnesota	87
Mississippi	82
Missouri	115
Montana	56
Nebraska	93
Nevada	17
New Hampshire	10
New Jersey	21
New Mexico	33
New York	62
North Carolina	100
North Dakota	53
Ohio	88
Oklahoma	77
Oregon	36
Pennsylvania	67
Rhode Island	5
South Carolina	46
South Dakota	66
Tennessee	95

Table 2 (continued)

Sate	Counties
Texas	254
Utah	29
Vermont	14
Virginia	133
Washington	39
West Virginia	55
Wisconsin	72
Wyoming	23

- Workers commuting from out-of-state counties to within-state counties.
- Workers commuting from within-state counties to out-of-state counties.

Dividing the network into states allows us to see more localized trends in COVID-19, instead of considering the entire country at once. States can act as “communities” in the country’s commuting network. Communities in network science are groups of nodes with similar characteristics [53].

US Counties

Counties of the United States.

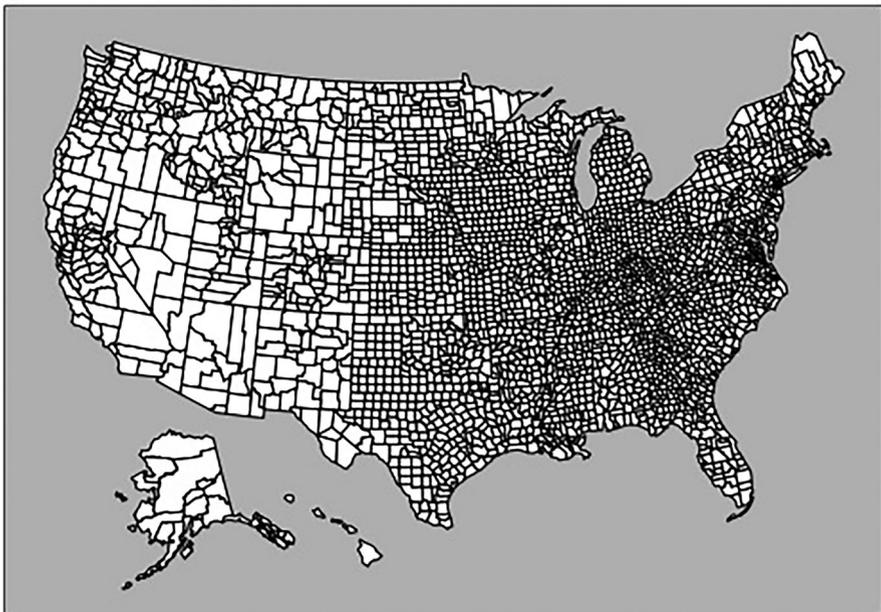


Fig. 1 All counties in the USA plotted by the `ggplot2` [49] and `usmap` [50] packages in R

Table 3 State vaccination rates as of 11/2021. The vaccination rates capture the percentage of population that is considered fully vaccinated (two doses of the appropriate vaccines, or one dose of a single dose vaccine). The rates help us determine which states to choose for our comparison. Adapted from [54]

State	Vaccination Rate (% Population)	Rank
Rhode Island	71.0	2
Massachusetts	69.8	5
California	61.4	16
Florida	59.8	22
Arkansas	48.1	43

5 Computational Experiments

As discussed earlier, GNAR takes a univariate time series dataset along with an underlying network structure in order to create a predictive time series model. After transforming the data for the GNAR model, we fit three models for each prediction type (for COVID-19 cases and COVID-19 deaths), giving us 6 models for each state. We will be selecting 5 states to showcase the results on; since the 5 states will be tested on 6 models each, this gives us a total of 30 individual models. We can evaluate the models for prediction accuracy graphically, using the MAPE and MASE as measures of performance.

In order to determine if our models would perform in a similar fashion across different states, we select a diverse array of states based on vaccination rates. Vaccination rates could affect the time-varying number of cases and deaths in a state, potentially leading to differing model performances. Hence, we choose the following states for our analysis: Rhode Island, Massachusetts, California, Florida, and Arkansas. Table 3 describes the vaccination rate (% of population) and corresponding rank out of 50 of each state we choose.

We calculate the MASE and MAPE for each GNAR model with respect to each test period within the 40-week forecast. We then calculate the mean, median, and variance of these values. In order to determine if a transformation of measurements in a given data helps in improving the performance of a model, we test transforming measurements in each county commuting network in the following ways:

1. logarithm transformation,
2. square-root transformation, and
3. normalization.

Table 4 MAPE Interpretation. The following table provides a guide for interpretation of a time series model's MAPE. The higher the MAPE, the less predictive power the model has. Source: [39]

MAPE	Interpretation
10	Highly accurate forecasting
10–20	Good forecasting
20–50	Reasonable forecasting
>50	Inaccurate forecasting

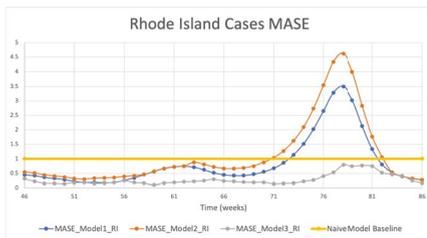
In our experiments, however, all the above transformations result in minor changes in the performance of the model; therefore, we report here the results using the original scale of measurements.

Table 5 Summary statistics of each model’s performance in MAPE and MASE over the time horizon

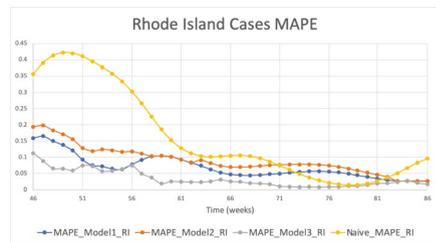
Rhode Island								
	Cases				Deaths			
	MASE		MAPE		MASE		MAPE	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Model 1	0.8585	0.7696	0.0695	0.0013	0.9888	0.6436	0.0615	0.0009
Model 2	1.129	1.1376	0.0906	0.0019	1.8501	2.0184	0.1073	0.0017
Model 3	0.285	0.0353	0.0334	0.0007	0.517	0.0934	0.0419	0.0008
Naïve	1	0	0.1642	0.0205	1	0	0.1201	0.0101
Massachusetts								
	Cases				Deaths			
	MASE		MAPE		MASE		MAPE	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Model 1	0.8579	0.5338	0.0723	0.0013	0.8263	0.2572	0.0691	0.0015
Model 2	0.9869	0.7752	0.078	0.0012	0.8275	0.2586	0.0692	0.0015
Model 3	0.2556	0.0193	0.0327	0.0008	0.3933	0.0103	0.0467	0.0014
Naïve	1	0	0.1582	0.0204	1	0	0.1374	0.0136
California								
	Cases				Deaths			
	MASE		MAPE		MASE		MAPE	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Model 1	0.9056	0.3836	0.0764	0.0014	0.801	0.2	0.0743	0.0019
Model 2	0.984	0.5088	0.083	0.0015	0.9192	0.2475	0.084	0.0019
Model 3	0.4199	0.0403	0.0492	0.0023	0.47	0.0169	0.0594	0.0028
Naïve	1	0	0.1507	0.0197	1	0	0.1447	0.0152
Florida								
	Cases				Deaths			
	MASE		MAPE		MASE		MAPE	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Model 1	0.9679	0.5556	0.0735	0.0011	0.8444	0.3151	0.0693	0.0015
Model 2	1.0033	0.598	0.0754	0.001	0.9405	0.4205	0.0743	0.0014
Model 3	0.3098	0.0272	0.0365	0.0012	0.4182	0.0119	0.0491	0.0014
Naïve	1	0	0.145	0.0177	1	0	0.1382	0.0132

Table 5 (continued)

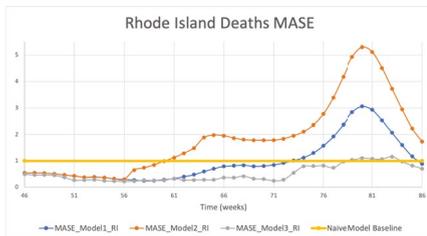
	Arkansas							
	Cases				Deaths			
	MASE		MAPE		MASE		MAPE	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Model 1	0.9679	0.5556	0.0735	0.0011	0.8573	0.3106	0.0698	0.0014
Model 2	1.0033	0.598	0.0754	0.001	0.9903	0.4565	0.0767	0.0013
Model 3	0.3098	0.0272	0.0365	0.0012	0.424	0.0141	0.0496	0.0014
Naïve	1	0	0.145	0.0177	1	0	0.138	0.0136



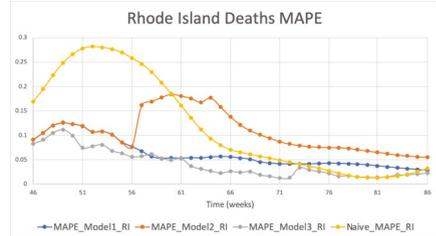
(a) Rhode Island Cases MASE. The vertical axis represents MASE. Models 1 and 2 exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 77.



(b) Rhode Island Cases MAPE. The vertical axis represents MAPE. Models 1 and 2 perform well against the naïve model until approximately week 71. Models 1, 2, and 3 all trend downward gradually.



(c) Rhode Island Deaths MASE. The vertical axis represents MASE. Model 2 starts to perform worse than the naïve model at approximately week 61 and continues until the end of the testing period. Model 1 starts to perform worse than the naïve model at approximately week 73 and continues until approximately week 85. Model 3 did not perform better than the naïve model for about six weeks from approximately week 78 to week 83.

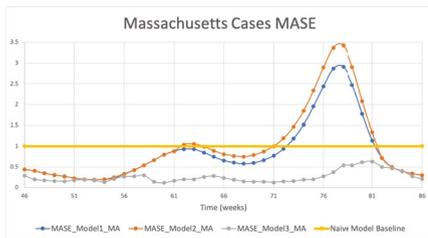


(d) Rhode Island Deaths MAPE. The vertical axis represents MAPE. Model 2 experienced a sharp increase in MAPE at approximately week 56, trending to perform worse than the naïve model until the end of the testing period. Model 2 and Model 3 stayed at a lower threshold, with Model 3 outperforming the naïve 100% of the time.

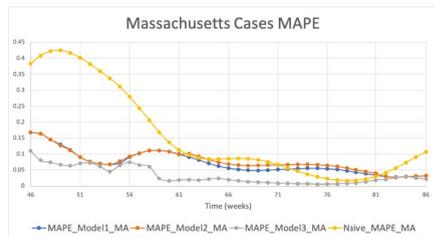
Fig. 2 The horizontal axis depicts the number of weeks since data collection began on January 22, 2020, for all plots in this figure. The four subfigures discuss the three models for the cases and deaths and the MAPE and MASE evaluation metrics for the state of Rhode Island. Adapted from [45] and [56]

The naïve models across all states started with a high MAPE at the beginning in the time horizon, sometimes doubling the MAPE of all the other models. Additionally, the MAPE for the naïve model appears to increase slightly across all states in the last 4 weeks in the time horizon. Poor performance of the naïve model in the beginning of the time horizon may be caused by a sudden increase of cases and a lack of information in the beginning of the time horizon.

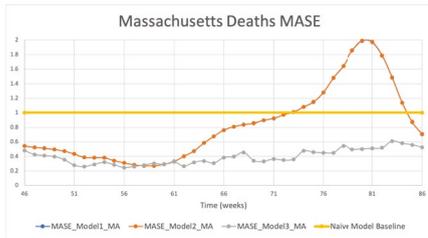
As we will see below, the MASE for Models 1 and 2 in each state exhibit a bimodal “hump”, with the largest hump centered around week 80 of our dataset. This hump then shows a sharp decrease for the last 4 weeks of the testing period for all models. Models 1 and 2 perform worse than the baseline naïve model during this bimodal hump period. The 80th week mark falls near the end of August of 2021: during that period, everywhere in the USA, the number of cases increased at a much slower pace than earlier. The timing coincides with when many people were fully



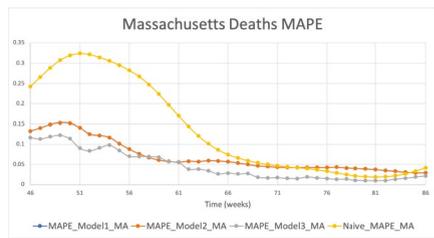
(a) Massachusetts Cases MASE. The vertical axis represents MASE. Models 1 and 2 exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 78. Models 1 and 2 perform worse than the naïve model in this time.



(b) Massachusetts Cases MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other closely, performing worse than the naïve model from approximately week 71 to week 81. Models 1, 2, and 3 all trend downward gradually.

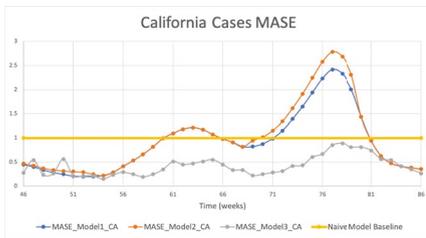


(c) Massachusetts Deaths MASE. The vertical axis represents MASE. Models 1 and 2 again follow each other very closely, almost overwriting each other. Models 1 and 2 start to perform worse than the naïve model at approximately week 72 but end up performing better than it at approximately week 84.

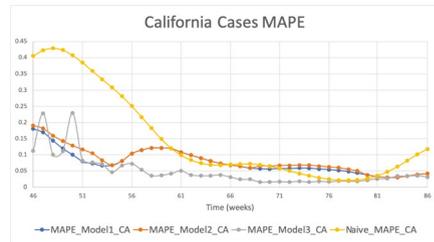


(d) Massachusetts Deaths MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other closely, almost overwriting each other. They perform worse than the naïve model from approximately week 75 to week 84. All models trend gradually downwards.

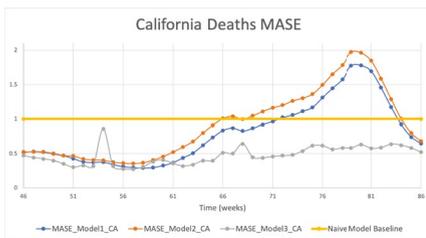
Fig. 3 The horizontal axis depicts the number of weeks since data collection began on January 22, 2020, for all plots in this figure. The four subfigures discuss the three models for the cases and deaths and the MAPE and MASE evaluation metrics for the state of Massachusetts. Adapted from [45] and [56]



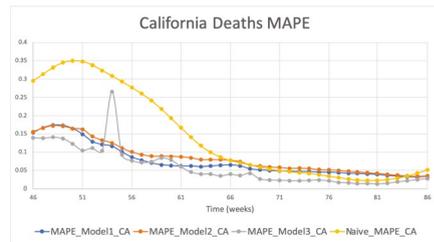
(a) California Cases MASE. The vertical axis represents MASE. Models 1 and 2 follow each other closely and exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 77. Models 1 and 2 perform worse than the naïve model in this time.



(b) California Cases MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other closely, performing worse than the naïve model from approximately week 61 to week 80. Models 1, 2, and 3 all trend downward gradually.



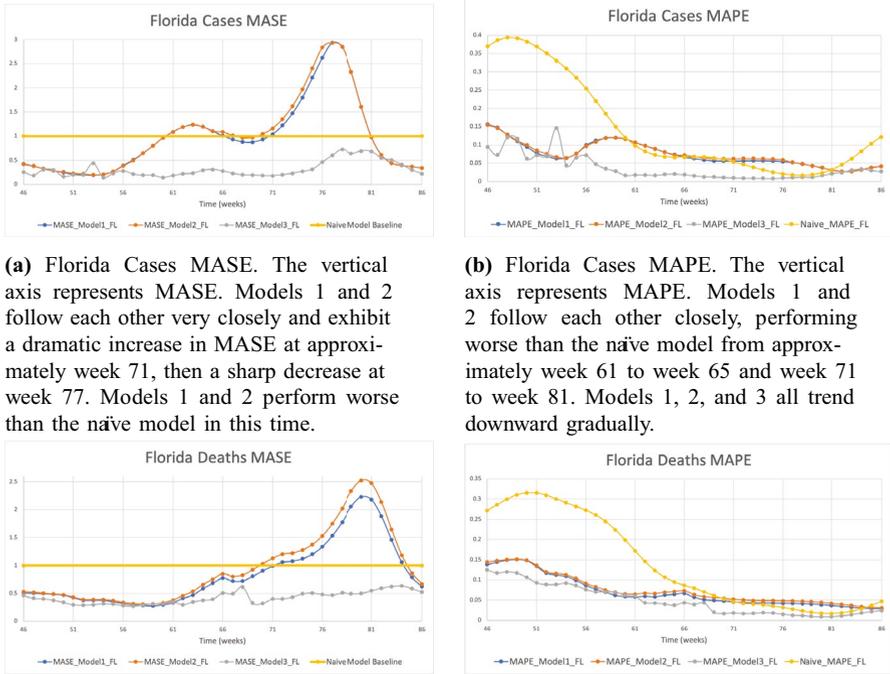
(c) California Deaths MASE. The vertical axis represents MASE. Models 1 and 2 again follow each other very closely, almost overwriting each other. Models 1 and 2 start to perform worse than the naïve model at approximately week 72 but end up performing better than it at approximately week 84.



(d) California Deaths MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other closely. They perform worse than the naïve model from approximately week 72 to week 84. All models trend gradually downwards.

Fig. 4 The horizontal axis depicts the number of weeks since data collection began on January 22, 2020, for all plots in this figure. The four subfigures discuss the three models for the cases and deaths and the MAPE and MASE evaluation metrics for the state of California. Adapted from [45] and [56]

vaccinated (for clarity, full vaccination in our work refers to a two-dose regimen or a single dose of an approved vaccine [55]). Hence, these big “humps” in the model performance could be caused by the effect of these vaccinations on the number of cases. Models 1 and 2 are the ones impacted by this increased number of vaccinations. This is because Model 1 and Model 2 are most affected by the numbers of cases in neighboring counties, and vaccination rates in neighbor counties are not necessarily the best predictors for the number of cases in each individual county. Overall, since Model 3 seems largely unaffected by correlations between numbers of fully vaccinated individuals in neighbor counties, Model 3 performs the best, since it outperforms the naïve model at almost all times during the time horizon. Earlier, we mentioned that full vaccination refers to individuals that have received two



(a) Florida Cases MASE. The vertical axis represents MASE. Models 1 and 2 follow each other very closely and exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 77. Models 1 and 2 perform worse than the naïve model in this time.

(b) Florida Cases MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other closely, performing worse than the naïve model from approximately week 61 to week 65 and week 71 to week 81. Models 1, 2, and 3 all trend downward gradually.

(c) Florida Deaths MASE. The vertical axis represents MASE. Models 1 and 2 follow each other closely. Models 1 and 2 start to perform worse than the naïve model at approximately week 71 but end up performing better than it at approximately week 84.

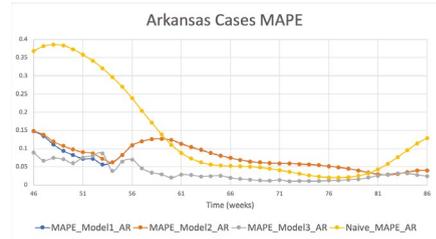
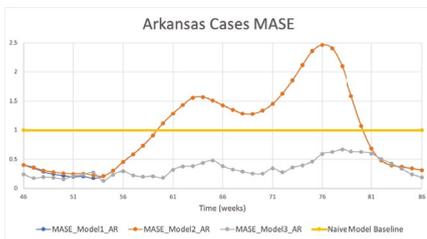
(d) Florida Deaths MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other closely. They perform worse than the naïve model from approximately week 71 to week 84. All models trend gradually downwards.

Fig. 5 The horizontal axis depicts the number of weeks since data collection began on January 22, 2020, for all plots in this figure. The four subfigures discuss the three models for the cases and deaths and the MAPE and MASE evaluation metrics for the state of Florida. Adapted from [45] and [56]

doses or a single dose of the appropriate vaccines. It would be interesting, given updated information and data, to study whether this effect changes when considering an individual as full vaccinated after having received the appropriate booster shot, also referred to as “up-to-date” individuals [55].

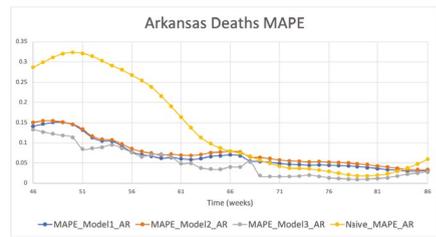
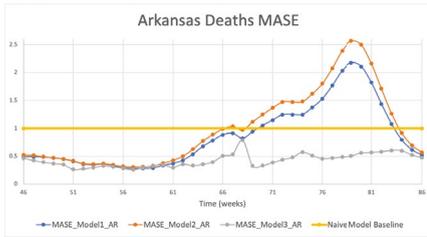
Table 5 provides a summary of the results obtained for each of the models in each of the states. More details graphical results are shown in Figs. 2, 3, 4, 5, and 6 for Rhode Island (RI), Massachusetts (MA), California (CA), Florida (FL), and Arkansas (AR), respectively. Recall that in order for a model to have predictive power, its MASE must exceed the accuracy of the respective naïve model [39]. Therefore, in order for a model to have a predictive power, MASE should be smaller than 1. Additionally, a model’s MAPE must be lower than 50%. Table 4 describes appropriate interpretations for different levels of MAPE.

As can be seen from Table 5 and Fig. 2, Rhode Island, a state with few counties but a dense population that is highly vaccinated, exhibits a great deal of difference among its models. Also from Table 5 and Fig. 3, Massachusetts, similar to Rhode Island, has a more urban population that is highly vaccinated. This state did not exhibit much difference among its models; regardless, Model 3 outperformed all of them. California, a state with a large land area and population that is highly vaccinated, did not exhibit much difference among its models. Model 3 outperformed all of them, regardless as we can see from Table 5 and Fig. 4. Florida, a state with a low vaccination rate and a large population, did not exhibit much difference among its models. Model 3 still performed the best out of all of them (Table 5 and Fig. 5).



(a) Arkansas Cases MASE. The vertical axis represents MASE. Models 1 and 2 follow each other very closely and exhibit a dramatic increase in MASE much earlier than the other states, at approximately week 56, then a sharp decrease at week 77. Models 1 and 2 perform worse than the naïve model in this time but end up performing better by week 80.

(b) Arkansas Cases MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other very closely, performing worse than the naïve model from approximately week 60 all the way to week 80. Models 1 and 2 experienced a steep increase in MAPE at week 55 and then trended downward gradually.



(c) Arkansas Deaths MASE. The vertical axis represents MASE. Models 1 and 2 follow each other closely. Models 1 and 2 start to perform worse than the naïve model at approximately week 69 but end up performing better than it at approximately week 83. This state does not experience a typical bimodal pattern of MASE as the other states.

(d) Arkansas Deaths MAPE. The vertical axis represents MAPE. Models 1 and 2 follow each other very closely. They perform worse than the naïve model from approximately week 68 to week 83. All models trend gradually downwards.

Fig. 6 The horizontal axis depicts the number of weeks since data collection began on January 22, 2020, for all plots in this figure. The four subfigures discuss the three models for the cases and deaths and the MAPE and MASE evaluation metrics for the state of Arkansas. Adapted from [45] and [56]

Finally, Arkansas, a state with a great deal of rural land and one of the lowest vaccination rates in the country, performed differently than other states. As we see in Table 5 and Fig. 6, Model 3 remains the best performing model.

All states analyzed exhibit similar behavior in the measures of performance via the MAPE and MASE. All models decrease gradually in the MAPE as the time series prediction passes through the time horizon. The MAPE for the naïve model consistently starts very high but then gradually approaches 0 along with the other models. Models 1 and 2 perform worse than the MAPE for the naïve model from approximately week 71 to week 84. Most of the states exhibited a bimodal structure in the MASE from approximately week 66 to week 84. Approximately week 70 to week 77 showed a rapid increase in the MASE across all states. Models for all states perform similarly because of the inclusion of both into-state and out-of-state travel. Likely, there are some counties that people throughout the country commute to that are included in many states, including those that we select in this work. Because Model 3 outperforms the naïve model most of the time in the time horizon for all states, this model should be considered by epidemiologists. The GNAR models on county networks with commuting information prove potentially useful for predicting COVID-19 cases and deaths.

6 Conclusion

The coronavirus pandemic has ravaged the world, killing many, and affecting the daily lives of all people. Concentrated efforts from all parts of the earth have attempted to curb this virus' spread. From mathematical models to public health policy decisions, these efforts have brought the world together in an attempt to eradicate this virus. In this work, we show that the GNAR model performs very well in predicting COVID-19 cases and deaths throughout the county network in the USA. Using the open-source data from common sources, including the USCB and USAFacts, we can create a predictive model that could better inform public health officials.

For example, cell phone data is both nearly ubiquitous and surprisingly accurate [57]. Companies and organizations have been able to harness the data from commuter's cell phones using their navigation applications in order to better influence their prediction of traffic flow through an area [57]. This data is almost live, since it comes directly from the drivers as they drive along a road. This live traffic data could help describe a by-county commuting network. The network could be dynamic, changing as more data is obtained. Perhaps analyzing trends over the past few weeks in a local area could result in a more accurate and current county commuting network structure. In addition, in this paper, we assume that the traffic and commuting patterns by county remain the same through the time of the COVID-19 pandemic as of the USCB compiled this commuting data over a 5-year period from 2011 to 2015, giving it a static property. In the beginning of the pandemic, because of the lock-down in many states, we had much less traffic flows. The effect of less traffic flows between counties might be most to Model 1 and Model 2 since Model 1 and Model 2 put parameters to weight on neighbor traffic flows. However, since

Model 3 has less weight on the traffic flows between a county and its neighbor counties, Model 3 was less affected by the change of traffic flows between counties. If we had information on traffic flows between counties during the pandemic, we expect that the performances of Model 1 and Model 2 will be improved.

The data sources of this work are delineated by county, which provides data for a more localized area. However, one could further subdivide this data into zone improvement plan (ZIP) codes in order to obtain an even further refined prediction at a lower level. The CDC currently only collects data at the county level; however, with future technologies for tracking a disease's spread, the CDC could subdivide its data even further. As of November 2021, there exist 41,692 ZIP codes in the USA [58]. Since individuals are freely able to move between their ZIP codes, and since the frequency of moving between ZIP codes is likely higher on average, this subdivision of data may provide a great deal of insight into localized trends of movement of people. ZIP code analysis may demonstrate a more realistic representation of daily life and community interactions due to the relatively smaller distance between nodes.

In this paper we treated all states the same so that we ignored covariates $c = 1, \dots, C$ in Eq. (2). We might be able to use covariates $c = 1, \dots, C$ for information of low and high vaccination rates in states and it might increase performances of the models using the GNAR model to predict the number of cases.

Finally, one could find any network structure and incorporate it into the GNAR model as long as it is geographically delineated the same way as the time series data. Any data that describes a flow from one geographic area to another can be formulated into a network structure, which is a key component of a GNAR model. Comparing multiple network structures could provide insight into what is important in the dissemination of a disease. With the advent of structure centrality [59], this could be a possibly interesting avenue for extending traditional centrality metrics (see, e.g., [34]) in epidemic spreading.

Applying this methodology to other geographic areas or governance divisions could also prove useful around the world, not just the USA. Any country's municipalities, provinces, or townships could represent nodes in a network similar to the US county structure. Comparing countries of a similar geographic, climatic, and demographic makeup to the USA may especially prove insightful. One can also compare and contrast the public health policy effects in different geographic areas.

Acknowledgements We would like to thank the two anonymous reviewers, as well as the editor, for their comments that helped us improve the work from its first iteration. We would also like to take this opportunity to thank and acknowledge all workers who provided essential services during this COVID-19 pandemic.

Author Contribution PU and DW are both considered as first authors of this work and hence are listed in alphabetical order; CV is the corresponding author; RY is the supervisor of this research. All authors equally contributed to this research.

Data Availability The datasets analyzed during the current study are available from the corresponding author on reasonable request. All data are also publicly available from the United States Census Bureau [45], the United States of America Facts [56], and the Center for Disease Control and Prevention [41].

Reproducibility and Code Availability A dashboard with all of the results presented here is available through <http://chvogiat.shinyapps.io/dod2021>; this version was also presented during the Dynamics of Disasters 2021 conference in June 2021. All codes and data are available upon request.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. World Health Organization (2020) Impact of COVID-19 on people's livelihoods, their health and our food systems. World Health Organization. <https://www.who.int/news/item/13-10-2020-impact-of-COVID-19-on-people's-livelihoods-their-health-and-our-food-systems>. Accessed 19 Feb 2021
2. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R (2020) High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2 SARS-CoV-2. *Emerg Infect Dis* 26(7):1470–1477
3. Yang W, Shaman J (2021) COVID-19 pandemic dynamics in India and impact of the SARS-CoV-2 Delta (b.1.617.2) variant. medRxiv <https://doi.org/10.1101/2021.06.21.21259268>. <https://www.medrxiv.org/content/early/2021/06/25/2021.06.21.21259268>
4. Sanju'a R, Nebot MR, Chirico N, Mansky LM, Belshaw R (2010) Viral mutation rates. *J Virol* 84(19):9733–9748
5. Musselwhite C, Avineri E, Susilo Y (2020) Editorial JTH 16-the Coronavirus disease COVID-19 and implications for transport and health. *J Transp Health* 16:100853
6. Kraemer MUG, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, Open COVID-19 Data Working Group, du Plessis L, Faria NR, Li R, Hanage WP, Brownstein JS, Layan M, Vespignani A, Tian H, Dye C, Pybus OG, Scarpino SV (2020) The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* 368(6490):493–497. <https://doi.org/10.1126/science.abb4218>. <https://www.science.org/doi/abs/10.1126/science.abb4218>, <https://www.science.org/doi/pdf/10.1126/science.abb4218>
7. Wu Y, Pu C, Zhang G, Pardalos PM (2021) Traffic-driven epidemic spreading in networks: considering the transition of infection from being mild to severe. *IEEE Transactions on Cybernetics*
8. Kermack WO and McKendrick A (1991) Contributions to the mathematical theory of epidemics–I. *Bull Math Biol* 53(1). <https://doi.org/10.1007/BF02464423>
9. Rodrigues HS (2016) Application of SIR epidemiological model: new trends. *International Journal of Applied Mathematics and Informatics* 10:92–97
10. Nguemdjo U, Meno F, Dongfack A, Ventelou B (2020) Simulating the progression of the COVID-19 disease in Cameroon using SIR models. *PloS One*. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237832>. Accessed 18 Oct 2021
11. Singh D, Alanazi S, Kamruzzaman M, Alruwaili M, Alshammari N, Alqahtani S, Karime A (2020) Measuring and preventing COVID-19 using the SIR model and machine learning in smart health care. *Journal of Healthcare Engineering*. <https://doi.org/10.1155/2020/8857346>
12. Gallagher J (2021) COVID immunity: Can you catch it twice? BBC News. <https://www.bbc.com/news/health-52446965>. Accessed 18 Oct 2021
13. Chen YC, Lu PE, Chang CS, Liu TH (2020) A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering* 7(4):3279–3294
14. Hoehl S, Rabenau H, Berger A, Kortenbusch M, Cinatl J, Bojkova D, Behrens P, Böddinghaus B, Götsch U, Naujoks F (2020) Evidence of SARS-CoV-2 infection in returning travelers from Wuhan, China. *N Engl J Med* 382(13):1278–1280
15. Patil D, Kotwal S (2020) Advice guideline: clinical information for the effective management, prevention, and counseling of COVID-19 patients. *Journal of Current Pharma Research* 10(4):3894–3906
16. Hou C, Chen J, Zhou Y, Hua L, Yuan J, He S, Guo Y, Zhang S, Jia Q, Zhao C, Zhang J, Xu G, Jia E (2020) The effectiveness of quarantine of Wuhan City against the Coronavirus disease 2019 (COVID-19): a well-mixed SEIR model analysis. *J Med Virol* 92(7):841–848
17. Teles P (2020) A time-dependent SEIR model to analyse the evolution of the SARS-CoV-2 epidemic outbreak in Portugal. *Research Gate*. <https://arxiv.org/abs/2004.04735>. Accessed 18 Oct 2021
18. Hackl J, Dubernet T (2019) Epidemic spreading in urban areas using agent-based transportation models. *Future Internet* 11(4). <https://doi.org/10.3390/fi11040092>
19. Silva PC, Batista PV, Lima HS, Alves MA, Guimarães FG, Silva RC (2020) COVID-ABS: an agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals* 139. <https://doi.org/10.1016/j.chaos.2020.110088>

20. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci* 103(7):2015–2020
21. Jia JS, Lu X, Yuan Y, Xu G, Jia J, Christakis NA (2020) Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* 582(7812):389–394
22. Li T, Rong L, Zhang A (2021) Assessing regional risk of COVID-19 infection from wuhan via high-speed rail. *Transp Policy* 106:226–238
23. Mo B, Feng K, Shen Y, Tam C, Li D, Yin Y, Zhao J (2021) Modeling epidemic spreading through public transit using time-varying encounter network. *Transportation Research Part C: Emerging Technologies* 122:102893. <https://doi.org/10.1016/j.trc.2020.102893>. <https://www.sciencedirect.com/science/article/pii/S0968090X20307932>
24. Sun X, Wandelt S, Zhang A (2021) On the degree of synchronization between air transport connectivity and COVID-19 cases at worldwide level. *Transp Policy* 105:115–123. <https://doi.org/10.1016/j.tranpol.2021.03.005>. <https://www.sciencedirect.com/science/article/pii/S0967070X21000652>
25. Wu JT, Leung K, Leung GM (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-NCov outbreak originating in Wuhan, China: a modelling study. *The Lancet* 395(10225):689–697
26. Hallett AH (1986) Prediction and regulation by linear least-square methods: Peter Whittle, (Blackwell, Oxford, 1983) second revised ed. *Int J Forecast* 2(1):125–127. [https://doi.org/10.1016/0169-2070\(86\)90043-9](https://doi.org/10.1016/0169-2070(86)90043-9). <https://www.sciencedirect.com/science/article/pii/0169207086900439>
27. Hyndman R, Athanasopoulos G (2018) Forecasting: principles and practice. <https://otexts.com/fpp2/>. Accessed 18 Oct 2021
28. Tandon H, Ranjan P, Chakraborty T, Suhag V (2020) Coronavirus (COVID-19): Arima based time-series analysis to forecast near future. *Research Gate*. https://www.researchgate.net/publication/340776075_Coronavirus_COVID-19_ARIMA_based_time-series_analysis_to_forecast_near_future. Accessed 14 Oct 2021
29. Panda M (2020) Application of Arima and holt-winters forecasting model to predict the spreading of COVID-19 for India and its states. *Research Gate*. https://www.researchgate.net/publication/343000522_Application_of_ARIMA_and_Holt-Winters_forecasting_model_to_predict_the_spreading_of_COVID-19_for_India_and_its_states. Accessed 14 Oct 2021
30. Sulasikin A, Nugraha Y, Kanggrawan J, Suherman A (2020) Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta. In: 2020 IEEE International Smart Cities Conference (ISC2), Piscataway, New Jersey, pp 1–6. <https://doi.org/10.1109/ISC251055.2020.9239072>
31. Kotu V, Deshpande B (2014) Predictive analytics and data mining: concepts and practice with Rapid-Miner. Elsevier Science and Technology, San Francisco
32. Abdulmajeed K, Adeleke M, Popoola L (2020) Online forecasting of COVID-19 cases in Nigeria using limited data. *Data in Brief* 30. <https://doi.org/10.1016/j.dib.2020.105683>
33. Roy S, Biswas P, Ghosh P (2021) Effectiveness of network interdiction strategies to limit contagion during a pandemic. *IEEE Access* 9:95862–95871
34. Sarlas G, Páez A, Axhausen KW (2020) Betweenness-accessibility: estimating impacts of accessibility on networks. *J Transp Geogr* 84:102680
35. Knight M, Leeming K, Nason G, Nunes M (2020) Generalized network autoregressive processes and the GNAR package. *Journal of Statistical Software, Articles* 96(5):1–36
36. Leeming K, Nason G, Nunes M, Knight M (2020) Package ‘GNAR’. CRAN. <https://cran.r-project.org/web/packages/GNAR/GNAR.pdf>. Accessed 20 Dec 2020
37. Guo Z, Wong W, Li M (2012) Sparsely connected neural network-based time series forecasting. *Inf Sci* 193:54–71
38. Hyndman RJ et al (2006) Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting* 4(4):43–46
39. Lewis C (1982) Industrial and business forecasting methods: a practical guide to exponential smoothing and curve fitting. Butterworth-Heinemann, Oxford, United Kingdom
40. Makridakis S, Winkler RL (1983) Averages of forecasts: some empirical results. *Manag Sci* 29(9):987–996. <http://www.jstor.org/stable/2630927>
41. Center for Disease Control and Prevention (2020) FAQ: COVID-19 data and surveillance. CDC. <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html>. Accessed 25 Oct 2021
42. USAFacts (2020) About USAFacts. USAFacts. <https://usafacts.org/about-usafacts/>. Accessed 15 Feb 2021
43. Statistics Solutions (2021) Time series analysis. Statistics Solutions. <https://www.statisticssolutions.com/time-series-analysis/>. Accessed 25 Oct 2021
44. Frost J (2020) Guidelines for removing and handling outliers in data. Statistics by Jim. <https://statisticsbyjim.com/basics/remove-outliers/>. Accessed 25 Oct 2021

45. United States Census Bureau (2019) County transportation profiles | Bureau of Transportation Statistics. <https://www.census.gov/topics/employment/commuting/guidance/flows.html>. Accessed 30 Jul 2021
46. United States Census Bureau (2020) 2019 FIPS codes. <https://www.census.gov/geographies/reference-files/2019/demo/popest/2019-fips.html>. Accessed 27 Oct 2021
47. United States Census Bureau (2020) 2020 Census data. <https://www.census.gov/programs-surveys/popest/technical-documentation/file-layouts.html>. Accessed 27 Oct 2021
48. The Fact File (2021) List of U.S. States and number of counties in each. <https://thefactfile.org/us-states-counties/>. Accessed 27 Oct 2021
49. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
50. Lorenzo PD (2021) Package ‘USMAP’. <https://cran.r-project.org/web/packages/usmap/usmap.pdf>
51. Broomfield County (2021) History of Broomfield. City and county of Broomfield. <https://www.broomfield.org/386/History-of-Broomfield>. Accessed 29 Oct 2021
52. Ban C (2014) Shannon County, S.D. to be renamed Oglala Lakota County. Shannon County, S.D. <https://www.naco.org/articles/shannon-county-sd-be-renamed-oglala-lakota-county>. Accessed 2 Mar 2021
53. Newman M, Barabasi AL, Watts DJ (2006) The structure and dynamics of networks: (Princeton studies in complexity). Princeton University Press, Princeton, New Jersey
54. Adams K (2021) States ranked by percentage of population fully vaccinated: Nov. 3. Backer’s Hospital review. <https://www.beckershospitalreview.com/public-health/states-ranked-by-percentage-of-population-vaccinated-march-15.html>. Accessed 3 Nov 2021
55. Centers for Disease Control and Prevention (2022) Stay up to date with your COVID-19 vaccines. CDC. <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/stay-up-to-date.html>. Accessed 19 Mar 2022
56. USAFacts (2021a) US COVID-19 cases and deaths by state. <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>. Accessed 31 Jul 2021
57. Johnson G (2021) Cell phone data makes traffic analysis and transportation planning easier. Short Elliot Hendrickson Inc. <https://www.sehinc.com/news/cell-phone-data-makes-traffic-analysis-and-transportation-planning-easier>. Accessed 9 Nov 2021
58. United States Postal Service (2021) 42,000 ZIP codes. U.S. Postal Facts. <https://facts.usps.com/42000-zip-codes/>. Accessed 9 Nov 2021
59. Rasti S, Vogiatzis C (2021) Novel centrality metrics for studying essentiality in protein–protein interaction networks based on group structures. Networks. <https://doi.org/10.1002/net.22071>. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/net.22071>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Patrick Urrutia¹ · David Wren¹ · Chrysafis Vogiatzis²  · Ruriko Yoshida¹

Patrick Urrutia
patch.urrutia@gmail.com

David Wren
david.m.wren.mil@mail.mil

Ruriko Yoshida
ryoshida@nps.edu

¹ Department of Operations Research, Naval Postgraduate School, Monterey, USA

² Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, USA