



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

SARS-COV-2 DISSEMINATION USING UNITED STATES COUNTY COMMUTING DATA

by

Patrick M. Urrutia

December 2021

Thesis Advisor:

Co-Advisor:

Second Reader:

Ruriko Yoshida
Chrysafis Vogiatzis,
University of Illinois
Johannes O. Royset

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE December 2021	3. REPORT TYPE AND DATES COVERED Master's thesis		
4. TITLE AND SUBTITLE SARS-COV-2 DISSEMINATION USING UNITED STATES COUNTY COMMUTING DATA			5. FUNDING NUMBERS	
6. AUTHOR(S) Patrick M. Urrutia				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (maximum 200 words) The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has impacted the world for the past year and a half, and it has been spreading to all corners of the Earth. Analyzing the dissemination of SARS-CoV-2 can allow leaders of certain areas to preemptively enact measures that could prevent the virus from spreading further. By analyzing commuting data between counties in the United States, one can create a predictive model that will allow interdiction of routes with high traffic between areas to stop the spread of the virus. At the county level, leaders can use this information to provide extra precautions, medical equipment, and testing in their area of jurisdiction. We solve this problem by obtaining data about coronavirus-19 (COVID-19) cases and deaths from the Center for Disease Control and Prevention and county commuting data from the United States Census Bureau. Then we propose to apply the generalized network autoregressive (GNAR) time series model for analyzing this network over time series data. This by-county predictive approach is broken down by state, in order to reflect more localized trends. This thesis combines time series analysis and network science to model COVID-19 cases and deaths by state.				
14. SUBJECT TERMS COVID-19, COVID, virus, PCA, principal component analysis, network, transportation, algorithm, SARS-CoV-2, R, Python, time series, mean absolute scaled error, MASE, mean absolute percentage error, MAPE, epidemiology, epidemiologist, pandemic, epidemic, commuting, commute, travel, county, ARIMA, Holt-Winters, naive model, naive, naïve model, naïve, supervised learning, generalized network autoregressive, GNAR			15. NUMBER OF PAGES 107	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**SARS-COV-2 DISSEMINATION USING UNITED STATES COUNTY
COMMUTING DATA**

Patrick M. Urrutia
Ensign, United States Navy
BS, United States Naval Academy, 2020

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
December 2021**

Approved by: Ruriko Yoshida
Advisor

Chrysafis Vogiatzis
Co-Advisor

Johannes O. Royset
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has impacted the world for the past year and a half, and it has been spreading to all corners of the Earth. Analyzing the dissemination of SARS-CoV-2 can allow leaders of certain areas to preemptively enact measures that could prevent the virus from spreading further. By analyzing commuting data between counties in the United States, one can create a predictive model that will allow interdiction of routes with high traffic between areas to stop the spread of the virus. At the county level, leaders can use this information to provide extra precautions, medical equipment, and testing in their area of jurisdiction. We solve this problem by obtaining data about coronavirus-19 (COVID-19) cases and deaths from the Center for Disease Control and Prevention and county commuting data from the United States Census Bureau. Then we propose to apply the generalized network autoregressive (GNAR) time series model for analyzing this network over time series data. This by-county predictive approach is broken down by state, in order to reflect more localized trends. This thesis combines time series analysis and network science to model COVID-19 cases and deaths by state.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
1.1	Background	1
1.2	COVID-19 Modeling Approaches.	1
1.3	United States County Commuting Network	2
1.4	Data Availability and Model Adaptation	2
1.5	Results and Future Research	2
1.6	Summary	3
2	Literature Review and Background	5
2.1	Data Description	5
2.2	Network Approaches	7
2.3	Data Analysis.	9
2.4	Generalized Network Autoregressive Processes	14
2.5	Other Approaches in COVID-19 Analysis	17
2.6	Summary	24
3	Methodology	27
3.1	Data	27
3.2	County Network Structure.	31
3.3	Model Implementation	32
3.4	Summary	37
4	Results	39
4.1	Rhode Island Results	40
4.2	Massachusetts Results	47
4.3	California Results	54
4.4	Florida Results	61
4.5	Arkansas Results	68

4.6	Summary	75
5	Conclusions and Future Work	77
5.1	Conclusions	77
5.2	Future Work	78
5.3	Summary	80
	List of References	81
	Initial Distribution List	87

List of Figures

Figure 2.1	United States Cumulative COVID-19 Cases	6
Figure 2.2	United States Cumulative COVID-19 Deaths	6
Figure 3.1	Rolling Origin	34
Figure 4.1	RI Cases MASE	42
Figure 4.2	RI Cases MAPE	43
Figure 4.3	RI Deaths MASE	45
Figure 4.4	RI Deaths MAPE	46
Figure 4.5	MA Cases MASE	49
Figure 4.6	MA Cases MAPE	50
Figure 4.7	MA Deaths MASE	52
Figure 4.8	MA Deaths MAPE	53
Figure 4.9	CA Cases MASE	56
Figure 4.10	CA Cases MAPE	57
Figure 4.11	CA Deaths MASE	59
Figure 4.12	CA Deaths MAPE	60
Figure 4.13	FL Cases MASE	63
Figure 4.14	FL Cases MAPE	64
Figure 4.15	FL Deaths MASE	66
Figure 4.16	FL Deaths MAPE	67
Figure 4.17	AR Cases MASE	70

Figure 4.18	AR Cases MAPE	71
Figure 4.19	AR Deaths MASE	73
Figure 4.20	AR Deaths MAPE	74

List of Tables

Table 2.1	MAPE Interpretation	13
Table 3.1	Number of Counties by State	30
Table 3.2	Model Parameter Summary	36
Table 3.3	State Vaccination Rates	37
Table 4.1	Rhode Island Cases Model Summary	41
Table 4.2	Rhode Island Deaths Model Summary	44
Table 4.3	Massachusetts Cases Model Summary	48
Table 4.4	Massachusetts Deaths Model Summary	51
Table 4.5	California Cases Model Summary	55
Table 4.6	California Deaths Model Summary	58
Table 4.7	Florida Cases Model Summary	62
Table 4.8	Florida Deaths Model Summary	65
Table 4.9	Arkansas Cases Model Summary	69
Table 4.10	Arkansas Deaths Model Summary	72

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ABM	agent-based model
ACS	American Community Survey
ARIMA	autoregressive integrated moving average
CDC	Center for Disease Control and Prevention
COVID-19	coronavirus disease 2019
DES	discrete event simulation
DOE	design of experiments
FIPS	federal information processing standard
GNAR	generalized network autoregression
ICU	intensive care unit
MAPE	mean absolute percentage error
MASE	mean absolute scaled error
ML	machine learning
PCA	principal component analysis
RMSE	root mean squared error
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SEIR	susceptible-exposed-infected-recovered
SEIRS	susceptible-exposed-infected-recovered-susceptible
SIR	susceptible-infected-recovered

SNA	social network analysis
STL	seasonal and trend decomposition using Loess
USAFacts	United States of America Facts
USCB	United States Census Bureau
ZIP	zone improvement plan

Executive Summary

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has impacted the world at large, ignoring all boundaries of country and society. Global supply chains have been disrupted. Economies have been crippled. Daily lives of people of all countries have changed beyond recognition as a result of restrictions. Uncertainty lingers over everyone like a haze as this virus propagates through the population. Therefore, epidemiologists have championed efforts to combat this virus through analytical techniques that create predictive models.

As a result of improvements in computing power in recent years, researchers are able to create effective models for predicting coronavirus disease 2019 (COVID-19), using them to inform the public. The collection and analysis of localized COVID-19 data allow for a refined understanding of the virus' spread through an area. For example, the susceptible-infected-recovered (SIR) model accounts for individuals who have contracted and recovered from a disease, while an agent-based simulation models individual behavior in a pandemic.

In this thesis, we utilize open-source data from United States of America Facts (USAFacts), Center for Disease Control and Prevention (CDC), and United States Census Bureau (USCB) in order to create models that incorporate not only univariate time series COVID-19 cases and deaths data but also by-county commuting data. Using the generalized network autoregression (GNAR) package in R, we combine various parameters to create three models for every state. We compare these models using two measures of performance for time series models: mean absolute scaled error (MASE) and mean absolute percentage error (MAPE).

Here we demonstrate that the GNAR package is indeed effective in predicting the spread of COVID-19. We find a model that proves that commuting data in fact helps in predicting COVID-19 cases and deaths. By working together, data analysts and epidemiologists can improve this modeling approach by using different input data among other techniques. This thesis provides public health officials and local leaders with greater situational awareness about the spread of COVID-19 through their areas of jurisdiction.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I would first like to thank my friends who went through the Operations Research master's program with me. I couldn't have done it without their help and support. I will always remember our trips to Lake Tahoe for snowboarding.

I would also like to thank Capt. (soon to be Major) Adam Dickson, a student in the Defense Systems Analysis department, who was a great friend and mentor to me during my time at NPS.

Next, I would like to thank my parents and my grandpa for always pushing me to be the best that I can be and to further my education.

Additionally, thank you to Dr. Vogiatzis and Dr. Royset, who were a tremendous help with their expertise and guidance through the thesis process.

Thank you to all of my NPS professors, who taught me to think critically.

Finally, I would like to thank Dr. Yoshida, who tirelessly worked with me in order to make sure my thesis was working properly. Her work with students will surely make a difference in the world one day.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:

Introduction

1.1 Background

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic and its subsequent coronavirus disease 2019 (COVID-19) have spread to all corners of the earth, ignoring country boundaries and threatening the safety of all people. In addition to many deaths and sickness, the pandemic has crippled the economies of the earth. Many loved ones have been lost.

Human history is rife with attempts to curb diseases from spreading, from vaccinations to water purification (Connolly and Heymann 2002). Until recent history, diseases were the leading cause of death in all wars (Connolly and Heymann 2002). Every era has experienced an epidemic of sorts, leading to many deaths and diverse methods of eradicating disease. Through these efforts, the practice of epidemiology has evolved into its current form.

Epidemiology is the study of the “occurrence and distribution of health-related events, states, and processes in specified populations, including the study of the determinants influencing such processes, and the application of this knowledge to control relevant health problems” (Porta 2014). The ability to collect data about diseases and to predict the spread of them is paramount in epidemiology. In order to do so, epidemiologists employ a variety of analytical techniques. In this thesis, we aim to create a model that will forecast COVID-19 cases and deaths by county.

1.2 COVID-19 Modeling Approaches

Recent years have shown a large increase in computing power, allowing for more advances in the areas of data collection, network science, machine learning, and simulation. These areas have directly influenced the way epidemiologists approach predicting a disease’s spread. Because of the time-varying nature of disease spread, time series algorithms such as Holt-Winters and seasonal decomposition have proven useful in the epidemiology realm. The susceptible-infected-recovered (SIR) model provides a simple yet effective way to

model individuals who have been infected with and recover from a disease. Discrete-event and agent-based simulations also model individual decision making and offer insight into whether public health measures prevent virus spread.

Despite the progress made by researchers, there does not currently exist a model using a commuting data network structure to predict COVID-19 cases and deaths. Modeling a virus's spread using data that describes travel patterns between counties could lend insight to how the virus propagates. The model and methodology from this thesis can assist local leaders in implementing public health measures in order to mitigate the spread of the virus.

1.3 United States County Commuting Network

The United States is broken up into 3,143 county equivalents as of 2020 (United States Census Bureau 2020a). These counties are connected to each other via workers commuting from one county to another, creating an underlying commuting network structure for the entire United States. In this thesis we apply network science techniques in order to describe the underlying connection between the counties and its affect on a disease's spread.

1.4 Data Availability and Model Adaptation

The data for this thesis come from open sources of organizations such as United States of America Facts (USAFacts), United States Census Bureau (USCB), and the Center for Disease Control and Prevention (CDC). We use the generalized network autoregression (GNAR) package in R to model this data (Leeming et al. 2020). This package inputs univariate time series data and a network structure and outputs a predictive model for the univariate data across the network (Leeming et al. 2020). Using a combination of parameters, we create three models to predict COVID-19 cases and deaths for five example states, resulting in 30 individual models.

1.5 Results and Future Research

We employ the mean absolute scaled error (MASE) and mean absolute percentage error (MAPE) as measures of performance for the time series models in order to demonstrate the effectiveness of the GNAR package. These measures are calculated using the naïve model,

which serves as a baseline which sets the forecast for any time period equal to the actual value of the previous time period. Forecasting models that outperform the naïve model are said to have predictive power (Erdem 2020). The results in Chapter 4 indicate that one of our models consistently outperforms the naïve model, providing evidence for its value as a forecasting model.

We end this thesis with further research into the GNAR package’s use for epidemiological prediction. For example, live traffic data could inform an updated commuting network structure. Dividing the data into zone improvement plan (ZIP) could provide a more localized forecast. Employing principal component analysis (PCA) may enable incorporation of multivariate data into the model. Finally, refining the selection of parameter values of GNAR could also improve performance and better inform public health officials.

1.6 Summary

For this research we build supervised models via the GNAR package to predict the number of cases and deaths from COVID-19 using time series data on county networks with commuting information collected from open sources of organizations such as USAFacts, USCB, and the CDC. Here, we show our models have predictive powers using rolling horizon designs, performance evaluations over the time horizon, with MASE and MAPE and we predict how commuting flows between counties influence spread of COVID-19 over states and we wish to use this model to prevent spread of a disease in future.

This thesis is organized as follows: In Chapter 2, we summarize the background of this research. More specifically we summarize how we collect the datasets. Then after we define network and how networks are applied to data analysis, we describe machine learning algorithms we apply in this research, including time series analysis models and performance measurements on time series models. Then we define the GNAR model which is a main ingredient of this research to analyze COVID-19 dataset. We end this chapter with other methodologies to analyze the spread of the COVID-19 in the U.S.. In Chapter 3, we discuss details on the structure of the COVID-19 dataset and how we constructed the county networks for each state in the U.S.. After we discuss assumptions of the GNAR model and details on the input parameters of the GNAR model on each county network, we define a rolling horizon design to measure a performance of each model on the COVID-19 cases

and deaths in each county. Then we discuss how we choose tuning parameters of a model and how we choose some states to analyze for this thesis. In Chapter 4, we discuss results from our analysis in each state. In Chapter 5, we summarize our research and we end with some future work.

CHAPTER 2:

Literature Review and Background

The COVID-19 pandemic has ravaged the world, spreading to all corners of the earth. The virus has propagated so easily because of its transmissible nature, spreading through water droplets resulting from talking, coughing, or sneezing (Sanche et al. 2020). Since the virus's first reports in December of 2019 in Wuhan, China, it has mutated multiple times and indiscriminately killed people of all countries and all economic levels. When a virus mutates, combating it with vaccines and other public health measures becomes increasingly difficult (Sanju'a et al. 2010).

2.1 Data Description

We collected the data for this thesis from multiple sources with open information about the COVID-19 pandemic. There are two principal sources of data for this thesis. First, we obtained the data about the number of cases and deaths per county, updated daily from USAFacts (USAFacts 2021b). Second, we obtained the data about the number of commuters between counties from the USCB (United States Census Bureau 2019).

2.1.1 Cases and Deaths Data

The USAFacts website aggregates the cases and deaths data from the following sources: the CDC, state-level public health agencies, and local-level public health agencies (USAFacts 2021b). For the purpose of this thesis, and according to CDC guidelines, “deaths” are only counted if COVID-19 played a significant role in causing death (USAFacts 2021b). This data contains the number of COVID-19 cases and deaths per county. Additionally, because the data are updated daily, this data is a time series data set.

Figures 2.1 and 2.2 give the cumulative COVID-19 cases and deaths for the entire United States from January 22, 2020 to September 17, 2021. The vertical axis represents the total cumulative COVID-19 cases or deaths count for the whole country. The horizontal axis is the time series for the cases or deaths data. Because both datasets describe cumulative COVID-19 data by county, they continue to increase.

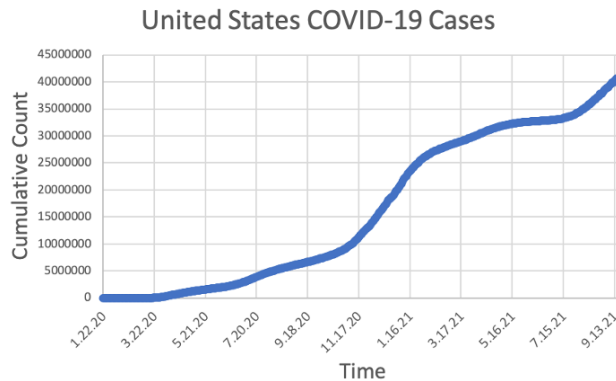


Figure 2.1. United States' COVID-19 Cases. The vertical axis represents the total cases, and the horizontal axis represents time. As we can see from the image, the rate of cases increased substantially at the beginning of 2021, and gradually tapered off. Adapted from USAFacts (2021a).

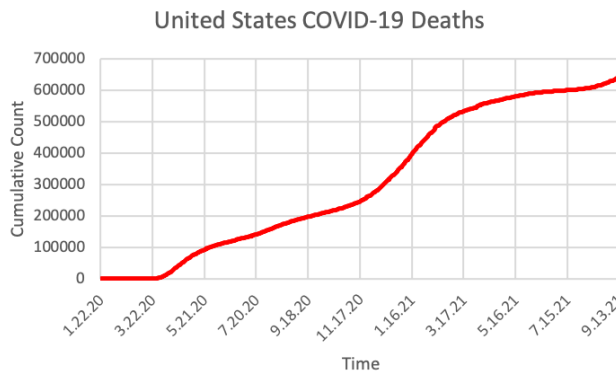


Figure 2.2. United States' COVID-19 Deaths. The vertical axis represents the total deaths, and the horizontal axis represents time. As we can see from the image, the rate of deaths increased substantially at the beginning of 2021, and gradually tapered off. Adapted from USAFacts (2021a).

2.1.2 Commuting Data

We obtained data from the USCB on the number of commuters between counties, coming from the American Community Survey (ACS), which asked respondents about their primary workplace location (United States Census Bureau 2019). When “a worker’s residence location and workplace location are coupled, a commuting flow is generated” (United States

Census Bureau 2019). Commuting flow formatted in an origin-destination pair allows one to understand the connection between one place and another (United States Census Bureau 2019). For example, Federal agencies use the commuting flow data to mark their estimate of the nation’s metropolitan and micropolitan statistical areas (United States Census Bureau 2019).

Readily available on the ACS website, we obtained this commuting data over a five-year period from 2011 to 2015. The data contains 137,451 instances of workers commuting. For the purpose of this thesis, “commuting” refers to an instance either when a worker commutes either from their home county to a different county, or when a worker commutes to work but stays in their home county.

2.2 Network Approaches

A network is defined as $\mathcal{G} = (N, E)$ where N is the set of nodes $N = \{1, \dots, n\}$ and E is the set of edges (Newman 2010). Each edge is represented as $e = (i, j) \in E$ for $i, j \in N$, and denotes a pair of nodes that are adjacent (Newman 2010).

The science of networks is very versatile and can be applied to many fields, such as biology, computer science, and statistics. Any structure of edges and nodes could be considered a network, whose mathematical properties can be analyzed. Such properties could provide deep insight into a system’s root behavior. Examples of networks include telecommunications systems, shipping routes, and disease spreading (Newman 2010).

Networks at their most basic level can describe correlations between nodes, which can be represented by counties or other similar geographic divisions. Additionally, a shortest path from one node to another also provides insight into the efficiency of a network delivering things such as goods, information, or a virus.

Objects can flow along a network through conserved spread and non-conserved spread (Newman et al. 2006). Conserved spread requires that the amount of an object entering the network remain constant and unchanging (Newman et al. 2006). Conversely, if the amount of an object flowing through the network changes as it enters and passes through, then the object spreads through the network in a non-conserved manner (Newman et al. 2006). Because a viruses spread from individual to individual, researchers more often model virus

propagation as a non-conserved spread. The SIR model, for instance, describes a virus's spreading behavior as non-conserved spread through a network of individuals.

Generalized Network Autoregression

The Generalized Network Autoregression GNAR model accepts as an input univariate time series data along with a network structure, and outputs a prediction of future observations of the time series data (Knight et al. 2020). Employing a network model, an autoregressive model, and a vector autoregressive model, the authors predicted gross domestic product data over one year for 35 countries. An autoregressive model ensures that the output variable linearly depends on the previous values of the time series data (Knight et al. 2020). A vector autoregressive model, however, takes into account lagged feedback effects between variables (Knight et al. 2020). From this study, the researchers found that model made with the GNAR function gave a lower prediction error than both the autoregressive and vector autoregressive models, performing better by 29 and 78 percent, respectively (Knight et al. 2020). This study shows evidence that the GNAR function can indeed model a network's time series autoregressive structure (Leeming et al. 2020).

Network Connectedness

Analyzing the connectedness of a network can assist in finding correlation between nodes. A network with nodes u and v is connected if there exists a path between u and v (Chartrand and Zhang 2012). Conversely, the u - v network is disconnected if there does not exist a path from u to v . Researchers have considered a network's degree as a proxy for its connectedness. They used this information to analyze correlation between countries' COVID-19 cases (So et al. 2020).

First, for comparison, the researchers constructed two correlation network graphs: one for China, and one for the rest of the world. If the correlation in the change of the number of confirmed cases between two geographical areas in the networks was greater than 0.5, then they considered an edge to be present. The presence of an edge indicated that the two areas are connected. The number of edges a country had (its degree) represented that area's connectedness. The researchers found that a more accurate prediction of risk of COVID-19 can be obtained by examining an individual country's connectedness (So et al. 2020).

Social Network Analysis

As an area of network science, social network analysis (SNA), is “a collection of methods and tools that could be used to study the relationships, interactions and communications” (Saqr and Alamro 2019). Researchers have explicitly utilized network science in analyzing COVID-19 through SNA, in which they create a network of people through tracking people who come in contact with other people within 4 meters of face-to-face distance (Firth et al. 2020). The researchers achieved their goal of minimizing the spread of the virus. SNA can assist in modeling a virus’s propagation through a network, since humans carry the virus via social interactions.

2.3 Data Analysis

In this thesis, we utilize time series data; therefore, we use time series analytical methods in order to construct and evaluate our models.

2.3.1 Time Series

The data for this thesis involves time series data, since the number of cases and deaths per county is updated daily. Thus, time series data by its very nature is not independent and identically distributed, as is data in many fields of statistical modeling, which requires those assumptions much of the time. Because time series depends on time, generating a predictive model is difficult (Cryer 2008). However, by understanding time series data, an analyst can both extract meaningful characteristics of the data and develop a predictive forecasting model that takes into account the temporal component of the data (Cryer 2008). Some characteristics of time series data are defined as follows by (Yoshida 2021) and (Hyndman and Athanasopoulos 2018).

- Trend: “Long-term (not necessarily linear) increases or decreases in the data: the long-term component of change” (Yoshida 2021).
- Seasonal (Periodic) Patterns: “Increases and decreases in data with a fixed/known period” (Yoshida 2021).
- Cycle: “Data exhibits rises and falls that are not of fixed period” (Yoshida 2021).
- Noise: “Remaining variance in the data after we’ve accounted for the components above” (Yoshida 2021).

- Seasonal Decomposition: Enables isolation of the trend-cycle (Hyndman and Athanasopoulos 2018).
- Homoscedastic: “Observed variance remains the same over time” (Yoshida 2021).
- Heteroscedastic: “Observed variance changes over time” (Yoshida 2021).
- Stationarity: “A stationary time series is one whose properties do not depend on the time when the series is observed. A stationary time series can exhibit cyclic behavior, but any series with trend, seasonality, or heteroscedasticity is non-stationary” (Yoshida 2021).

Understanding a time series dataset’s characteristics such as cycle, seasonality, and trend can provide insight into the underlying behavior of the data. The two principal methods for time series analysis are described as time domain and frequency domain (Cryer 2008). Time domain analysis focuses on properties of correlation of the data. Frequency domain, also called spectral analysis, focuses on the frequency properties.

Parametric and non-parametric techniques can be applied to time series analysis as well. Parametric analysis assumes that the data has an underlying random distribution, expressed using any number of parameters. Non-parametric techniques do not assume any underlying disposition, only estimating a dataset’s covariance, also called its spectrum (Cryer 2008).

By predicting future values based on previous observations, researchers can use time series analysis to influence policy or business decisions. Many variations of time series models exist, each inputting different parameters that describe the time series data (Cryer 2008).

2.3.2 Time Series Models

The wide variety of models that incorporate time series data allow researchers to investigate how the temporal component of data can provide insight into prediction. Examples of time series models include time series decomposition, Holt-Winters, and autoregressive integrated moving average (ARIMA). In order to evaluate their performances, we can use rolling horizons, which measure performance first by repeatedly fitting the time series model to a defined “rolling period.” The rolling horizon method then measures model’s performance in forecasting over the horizon that will be used in practice. Thus, a rolling horizon evaluates a time series model’s predictive power against known values. Intuitively, as a model’s forecast horizon increases, its predictive power decreases. Therefore, long

forecasts commonly perform worse than short forecasts.

Naïve Model

The naïve model sets all forecasts equal to the value of the last observation (Hyndman and Athanasopoulos 2018). Similarly, for data that is highly seasonal, seasonal naïve models set all forecasts equal to the value of the last observation from the same season of the year (Hyndman and Athanasopoulos 2018). Any time series model that can outperform a naïve model is said to have predictive power (Erdem 2020).

Time Series Decomposition

When time series data exhibit patterns such as trend, seasonality, and cycles, one can decompose the time series into each of these pattern categories (Hyndman and Athanasopoulos 2018). Classical decomposition assumes that the seasonal attribute of the data is constant for any given year (Hyndman and Athanasopoulos 2018). Meanwhile, X11 decomposition breaks down quarterly and monthly data (Hyndman and Athanasopoulos 2018). The most versatile method for decomposing time series is seasonal and trend decomposition using Loess (STL), which can handle any type of seasonality, even allowing it to change over time according to the user's input (Hyndman and Athanasopoulos 2018).

Autoregressive Integrated Moving Average

Autoregression describes correlation between lagged observations of the same time series (Yoshida 2021). An autoregressive model regresses a forecast value onto previous values of the time series (Yoshida 2021). Thus, ARIMA models seek to describe autocorrelations in the time series data (Hyndman and Athanasopoulos 2018). A non-seasonal ARIMA model contain three parameters, described by (Yoshida 2021) and (Hyndman and Athanasopoulos 2018):

- p : The number of autoregressive terms.
- d : The degree of first differencing involved.
- q : The number of moving-average terms.

Holt-Winters

The Holt-Winters method captures the trend, seasonal, and cyclic aspects of time series data by using exponential smoothing (Hyndman and Athanasopoulos 2018). Exponential smoothing is “a variation on moving average smoothing that weights the effect of previous observations/forecasts exponentially as you go back in time” (Yoshida 2021). Exponential smoothing adds to the old forecast an adjustment for error that occurred in the last forecast (Yoshida 2021).

Ensembles

Ensemble models combine many of the various time series models into one by taking an average of outputs from these models. The final prediction of an ensemble model is a combination of the prediction of multiple models (Kotu and Deshpande 2014). Ensemble models often perform very well for forecasting applications compared to individual models; however, they are “black box” models, meaning they are less interpretable than others (Yoshida 2021). Combining the aforementioned models into one ensemble could improve time series prediction accuracy at the cost of interpretability.

2.3.3 Time Series Performance

Measuring performance in traditional statistics often calls for measures of performance such as root mean squared error (RMSE) and adjusted R^2 . Although easily calculated, these measures are often scale-free and are generally inferior to other measures (Guo et al. 2012). For outputs of a predictive time series model, performance can be measured by MAPE and MASE. MAPE measures a model’s forecast performance, while MASE measures the absolute error of the forecast divided by the mean absolute error from the in-sample one-period naïve forecast method (Hyndman 2006). MAPE commonly falls between 0 and 1, but can be skewed outside this range if actual values are close to zero (Hyndman 2006). MASE can extend beyond 1 and is not plagued with the problems of infinite, undefined, or zero values of other accuracy measurements (Hyndman 2006). Both measurements may give insight into the predictive power of a model. MAPE and MASE are defined by the following equations:

$$\text{MAPE} = \frac{\sum_{t=1}^N \left| \frac{Y_t - F_t}{Y_t} \right|}{N}, \quad (2.1)$$

$$\text{MASE} = \frac{\sum_{t=1}^N \left| \frac{Y_t - F_t}{Y_t - Y_{t-1}} \right|}{N}, \quad (2.2)$$

where Y_t is the actual value at time t , F_t is the value of the prediction, and $Y_t - Y_{t-1}$ is the mean absolute error of the one-step naïve forecast (Hyndman 2006).

In order for a model to have predictive power, its MASE must exceed the accuracy of the respective naïve model (Lewis 1982). Additionally, a model's MAPE must be lower than 50%. Table 2.1 describes appropriate interpretations for different levels of MAPE.

Table 2.1. 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: Lewis (1982).

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

2.3.4 Time Series Research Efforts

Analyzing COVID-19 infections in India using an ARIMA(2,1,1) model, researchers found an almost exponential increase in the number of COVID-19 cases over their specified time period (Tandon et al. 2020). The ARIMA model proved more accurate than the other moving average and exponential smoothing models.

Still in India, researchers also analyzed COVID-19 trends using both an ARIMA(4,1,1) and a Holt-Winters model (Panda 2020). Akaike's Information Criteria and RMSE served as the

measures of performance for these models. The researchers obtained a maximum of 99.8% accuracy for their specified time period (Panda 2020).

Another example of ARIMA and Holt-Winters models comes from Jakarta (Sulasikin et al. 2020). Researchers modeled COVID-19 infections, finding that the ARIMA model performed better, with the higher R^2 and lower RMSE (Sulasikin et al. 2020). Their results indicated that COVID-19 infections would continually increase, allowing Jakarta's leaders to disrupt the spread through local preventative health measures (Sulasikin et al. 2020).

Finally, an ensemble model incorporating ARIMA, Holt-Winters, and additive regression was employed by Facebook to help researchers in Nigeria (Abdulmajeed et al. 2020). This ensemble model, called Prophet, processed missing values, seasonal effects, and outliers, allowing it to perform well against other models (Abdulmajeed et al. 2020). Researchers employing these techniques across the world can help leaders interdict the spread of the virus.

2.4 Generalized Network Autoregressive Processes

In this thesis, with data regarding COVID-19 cases and deaths collected and the commuting data network structure defined, we can now fit a time series network model using the Generalized Network Autoregressive Processes (GNAR) package in R Leeming et al. (2020).

Suppose we have a directed graph $\mathcal{G} = (N, E)$ where N is a node set $N = \{1, \dots, n\}$ and E is an edge set. 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 $r = 2, 3, \dots$ and $\mathcal{N}^{(1)}(i) = \mathcal{N}(\{i\})$.

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

$$w_{i,k} = \mu_{i,k} \left\{ \sum_{l \in \mathcal{N}^{(r)}(i)} \mu_{i,l} \right\}^{-1}. \quad (2.3)$$

The GNAR model uses a *covariate* for an edge affect 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

$$X_{i,t} := \sum_{j=1}^p \left(\alpha_{i,j} X_{i,t-j} + \sum_{c=1}^C \sum_{r=1}^{s_j} \beta_{j,r,c} \sum_{q \in \mathcal{N}_t^{(r)}(i)} w_{i,q,c}^{(t)} X_{q,t-j} \right) \quad (2.4)$$

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 . $\alpha_{i,j} \in \mathbb{R}$ 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 $c = 1, \dots, C$.

2.4.1 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 thesis, 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 U.S. 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 $\mu_{i,j}$ 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 $\mu_{i,j}$ are fixed over the time horizon $t > 0$. Therefore, the input of the GNAR package includes these variables. Now, the random variables which we wish to infer using the GNAR model

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

where $X_{i,n}$ is the number of COVID-19 cases of death's from COVID-19 at a county $i \in N$ at the time $t > 0$.

In this thesis, 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 . Therefore the formulation of the GNAR model becomes 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 death's from COVID-19 X_t is

$$X_{i,t} := \sum_{j=1}^p \left(\alpha_{i,j} X_{i,t-j} + \sum_{r=1}^{s_j} \beta_{j,r} \sum_{q \in \mathcal{N}_t^{(r)}(i)} w_{i,q}^{(t)} X_{q,t-j} \right) \quad (2.5)$$

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 . $\alpha_{i,j} \in \mathbb{R}$ 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 thesis, we select three combinations of tuning parameters after conducting

a model selection discussed in Section 3.3.5. 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 equation (2.3).

2.5 Other Approaches in COVID-19 Analysis

2.5.1 Epidemiological Models

The Spanish flu ravaged the world in the early 1900s. Shortly after the outbreak of the disease, in 1918, researchers W. O. Kermack and A. G. McKendrick published papers that presented mathematical models for predicting the number of infections in a population as a function of time (Kermack 1991). They postulated that it is valid to split the population into smaller “compartments” when analyzing a disease’s propagation through a population (Kermack 1991). Their foundational work continues to help epidemiologists model outbreaks of diseases today.

Susceptible-Infected-Recovered (SIR)

A simple example of an epidemiological model is the SIR model. In the SIR model, individuals linearly move from one “compartment” to the next without the ability to return to a previous compartment (Rodrigues 2016). SIR models require that a government step in to support recovery rates in their populations (Rodrigues 2016).

SIR Definitions The following terms and definitions by (Rodrigues 2016), (Ridenhour et al. 2014), and (Collins and Abdelal 2018) illustrate the SIR model.

- Susceptible, S : “Class of individuals who are susceptible to infection; this can include the passively immune once they lose their immunity or, more commonly, any newborn infant whose mother has never been infected and therefore has not passed on any immunity” (Rodrigues 2016).
- Infected, I : “In this class, the level of parasite is sufficiently large within the host and there is potential in transmitting the infection to other susceptible individuals” (Rodrigues 2016).
- Recovered, R : “Includes all individuals who have been infected and have recovered” (Rodrigues 2016).

- Infection rate, β : “The average number of people each infectious person spreads the disease to each day” (Collins and Abdelal 2018).
- Recovery rate, γ : “Inverse of the infectious period” (Ridenhour et al. 2014).
- Reproduction Rate, R_0 : “The average number of secondary infections that occurs when one infective is introduced into a completely susceptible population” (Rodrigues 2016).

The reproduction rate, R_0 of a virus critically affects a virus’s spreading pattern through a population, influencing it to either dissipate quickly or remain in the population for long periods of time. The value of R_0 serves as a “borderline between a persistence or a disease death” (Rodrigues 2016). When $R_0 < 1$ the disease will die out over time (Rodrigues 2016). When $R_0 > 1$ the disease will spread through the population as an epidemic (Rodrigues 2016). Generally, the larger the R_0 value, the more widespread and acute the epidemic will be (Rodrigues 2016). Although the simplicity of the SIR is attractive, it is limited by its inability to time-vary the infection rate parameter β and the recovery rate parameter γ (Chen et al. 2020b).

SIR Assumptions In order to use the SIR model, one must consider the following assumptions:

- First, the population must be constant, despite births, deaths, or immigration, with no new individuals entering the Susceptible compartment (Cooper et al. 2020). Logically, the effects of a disease spreading through a population vastly outweigh the effects of births, deaths, and immigration, making this assumption valid (Capitanelli 2020). In order to move from the Susceptible compartment, one must become infected (Capitanelli 2020). In order to move from the Infected compartment, one must recover (Capitanelli 2020).
- Second, all individuals have the same likelihood of contracting the virus, and their age distribution is uniformly distributed between 0 and a given country’s life expectancy (Capitanelli 2020). Each member of the population is exposed to the same amount of people each day (Capitanelli 2020).
- Third, the mixing of the three compartments is homogeneous, implying a completely connected population network in which people can interact and spread the disease (Capitanelli 2020).

SIR Equations The SIR model is described by three coupled nonlinear ordinary differential equations for the change in population of the three compartments (Rodrigues 2016):

$$\frac{dS}{dt} = -\beta SI, \quad (2.6)$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \quad (2.7)$$

$$\frac{dR}{dt} = \gamma I. \quad (2.8)$$

SIR Research Efforts One application of the SIR model comes from Cameroon, where researchers adapted it in order to analyze whether public health measures mitigated the spread of COVID-19 (Nguemdjo et al. 2020). Using the model, the researchers determined that the number of COVID-19 cases was indeed not as high because of the health precautions put into place (Nguemdjo et al. 2020). This model's success in Cameroon demonstrates its utility for predicting even on a large-scale level.

Another application of the SIR model originates in Saudi Arabia, where researchers analyzed the number of COVID-19 cases and deaths both with and without public health measures such as quarantine enforcement (Singh et al. 2020). A second modification to this model involved indicating whether an individual was confirmed dead by COVID-19 so that they cannot return to the Susceptible compartment (Singh et al. 2020). Finally, the researchers added a stochastic component for predicting the spread of the virus through the population (Singh et al. 2020). Although this model accurately predicted the size of the COVID-19 outbreaks, current research indicates that individuals who contract the virus once can become infected again (Gallagher 2021). As a result, this SIR model would need to be updated continuously in order to have predictive power.

As described above, the invariable nature of the infection rate parameter, β , and the recovery rate parameter, γ , limit the SIR model. Notwithstanding, researchers have recently time-varied these parameters to account for changes over time, using machine learning (ML) techniques to determine exactly how to vary the parameters (Chen et al. 2020b). Thus, $\beta(t)$ and $\gamma(t)$ offer a more dynamic approach to the SIR model. Using this new approach, researchers determined that public health precautions at a city level can lower the infection rate.

The SIR model allows researchers and government leaders to understand a disease's propagation through a community in spite of its simplicity. In order to further refine the SIR model's predictive ability, researchers have modified the model to include more compartments for individuals.

Susceptible-Exposed-Infected-Recovered

The aforementioned limitations of the SIR model, such as a constant population, unchanging infection rate, guaranteed recovery, and post-infection immunity are not necessarily guaranteed. In addition to vaccine deployment and other medical advances, mutations of a disease alone could affect all of these aspects of an outbreak. Differing from infections of other coronaviruses, COVID-19 infections involve a long incubation period, making the virus particularly dangerous to individuals who often come into contact with people in their daily lives (Chen et al. 2020a). As a result, an improved approach has arisen, called the susceptible-exposed-infected-recovered (SEIR) model.

SEIR Definitions and Equations An additional compartment, E , in which individuals have been infected but not become transmissible themselves, plays a large role in the SEIR model. In essence, this additional compartment represents a latency period of the disease (Hoehl et al. 2020). A new parameter, θ , is the rate of non-transmissible individuals becoming infectious. The SEIR model is described by four differential equations:

$$\frac{dS}{dt} = -\frac{\beta SI}{N}, \quad (2.9)$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \gamma E, \quad (2.10)$$

$$\frac{dI}{dt} = \theta E - \gamma I, \quad (2.11)$$

$$\frac{dR}{dt} = \gamma I, \quad (2.12)$$

where β , S , I , R , and γ are discussed in Section 2.5.1 and $N = S + E + I + R$ is the total population (Institute for Disease Modeling 2021).

SEIR Research Efforts The incubation period of a virus represents a time when an infected individual bears no symptoms yet can still transmit the virus to others (Hoehl et al. 2020).

Asymptomatic, these recently infected individuals can unknowingly assist the disease's spreading (Patil and Kotwal 2020). The SEIR model takes into account the effect of these individuals on the other compartments.

Similarly to the above example in Saudi Arabia, researchers in Wuhan, China applied the SEIR model to analyze the effect of public health measures such as quarantine initiatives and restrictions of movement (Hou et al. 2020). The researchers defined two infectious rates, β_1 and β_2 , where β_1 represented the “probability of infection per exposure when a susceptible individual has contacted an infected patient and becomes a latent exposed individual” (Hou et al. 2020). β_2 represented “the potential rate per exposure when a susceptible individual has mutual contact with an exposed individual and transmits it to another exposed individual” (Hou et al. 2020). With these refined parameters, the researchers found that quarantining latent individuals reduces the number of people they can infect with COVID-19 (Hou et al. 2020). Because of this mitigation, the number of infected individuals decreased, delaying the peak of the infection rate and reducing the critical strain on the local health care system (Hou et al. 2020).

As described in (Chen et al. 2020b), time-varied the infection rate parameter, β , and the recovery rate parameter, γ in the SIR model, researchers in Portugal dynamically adjusted the exposure rate, β , in order to simulate infected asymptomatic individuals who can spread the virus (Teles 2020). The Portuguese researcher also incorporated government intervention by varying β to reflect public health measures taken when one is notified of exposure, leading to subsequent self-quarantine (Teles 2020). Consequently, transmission rates and COVID-19 cases decreased, forecasting ahead in a set time period (Teles 2020).

2.5.2 Simulations

Simulations can contribute to understanding a stochastic system that has certain parameters, allowing researchers to measure performance. Simulations allow researchers to replicate a physical system in a computer and generate data for further analysis. This data generation process is often expensive, since increasing the number of parameters of a simulation increases computation time non linearly. However, design of experiments (DOE) allows for filtering of the number of alternatives inputted into a simulation, reducing computational time while also maintaining the underlying significance of the simulation. In particular, sim-

ulations give insight into sources of uncertainty that often go unnoticed in other numerical models.

Data regarding COVID-19 infections often is time-based, updating at marked points over time. Because of this property, discrete event simulation (DES) can model COVID-19 data. DES can “model real world systems that can be decomposed into a set of logically separate processes that autonomously progress through time” (Barrett et al. 2008). An application of DES, for example, could involve the movement of patients through a hospital system, taking into account the number of beds and length of stay of a typical patient.

Another technique of simulation is the agent-based model (ABM), which imitates human behavior and autonomy with “agents” who act independently. Their individual choices can provide insight into the behavior of the system overall. Such a technique can be applied to agents infected with COVID-19, allowing researchers to analyze the results of behavioral changes (Kerr et al. 2020). ABM allows for more integration of decision making, incorporating sociology, network science, and decision theory among other disciplines.

Simulation Research Efforts

Because of physical limits of hospitals in terms of providing care for patients in the COVID-19 pandemic, simulations of this system can greatly assist in understanding the optimal allocation of resources. Simulation can help understand problems such as reaching intensive care unit (ICU) and ventilator capacity, allowing hospitals to make changes to give care to as many patients as possible. Extrapolations from the local level to the federal level may also provide insights into the critical needs of hospitals at a given time.

Before SARS-CoV-2 existed, researchers modeled how epidemics spread in urban settings using an ABM, taking into account transportation data (Hackl and Dubernet 2019). They argue that accounting for how people move within a population is critical to understand how a disease could move through a community. Using a SIR model, they simulated the propagation of COVID-19 and employed an ABM to further model the disease in an urban setting (Hackl and Dubernet 2019).

Researchers in Cameroon, as discussed earlier, used a SIR model to analyze the effect of preventative health measures (Nguemdjo et al. 2020). The researchers first estimated the

virus parameters such as γ using their current data (Nguemdjo et al. 2020). Then they applied non-parametric techniques through bootstrapping in order to create more data points, since their original sample size was small (Nguemdjo et al. 2020). Finally, the researchers created simulations that take into account public health mitigations such as physical distancing. Their results indicate that Cameroon would have an exponential rise in COVID-19 cases.

ABM simulations easily can incorporate public health mitigations. Researchers modeled seven different public health measures such as mask wearing, social distancing, and lockdowns in order to find which ones are the most effective at minimizing the spread of the virus (Silva et al. 2020). The researchers concluded that both a conditional lockdown depending on the local infection numbers and a traditional lockdown are the most effective (Silva et al. 2020).

2.5.3 COVID-19 Prediction Efforts Using the GNAR Package

Wren (2021) utilized the GNAR package in order to predict COVID-19 cases, deaths, and survivors by county. His methodology involves various combinations of inputs into the GNAR package. First, he varies the time series data by analyzing COVID-19 cases, deaths, and survivors, which he calculates by subtracting the number of cases by the number of deaths (Wren 2021). Next, he varies whether the time series data is divided into weekly or daily intervals. Then, he varies the underlying network structure by calculating which counties are adjacent to each other, creating an adjacency matrix. Each edge in the county network represents a binary variable describing whether the counties are merely adjacent. Finally, as described in Section 2.4.1, he varies the GNAR parameters, selecting the three best-performing parameter combinations, resulting in three predictive time series models.

Using the MASE and MAPE as measures of performance, he compares his models. His thesis reports that both of the case and survivor combinations are nearly identical, due to the low death rate of the SARS-CoV-2 virus and that the great circle distance often hinders accuracy (Wren 2021). Moreover, he demonstrates that dividing the time series data into daily intervals results in inferior model performance compared to weekly intervals (Wren 2021). His results indicates that “the GNAR package is able to accurately forecast COVID-19 cases, deaths, and survivors” (Wren 2021).

In this thesis, we modify the methodology of Wren (2021) to include a different underlying

network structure, among other changes. First, the edges of the network we create represent the number of people commuting from one county to the next, in an effort to describe how the SARS-CoV-2 virus spreads as people travel for work. Next, because of Wren (2021)'s results, we only divide the time series data into weekly intervals, and we use similar GNAR parameter combinations. Finally, we only include COVID-19 cases and deaths time series data, excluding survivors.

2.6 Summary

Epidemics such as SARS-CoV-2 are difficult to model for many reasons. The parameters of some models cannot easily be time-varied or changed dynamically, as would better represent the behavior of the virus. People might ignore the recommendations of public health initiatives in spite of models proving that they in fact reduce deaths, resulting in no progress. Many aspects of epidemics are random in nature, which is difficult to account for in models. In spite of these uncertainties, researchers have been able to apply mathematical methods in order to understand the propagation of a virus through a population, aiding government leaders in pushing public health policy.

Network approaches such as GNAR, network connectedness, and SNA provide insight into how the virus flows through a system of connected people. The network framework incorporates interactions of individuals into predicting the spread of a disease.

The time series models such as ARIMA, Holt-Winters, and ensembles take into account the temporal aspect of COVID-19 data and create predictive models. Many researchers have found success using time series models in their respective countries. However, many time series models so far have proven useful only for short-term forecasting.

The compartmental epidemiological models of SIR and SEIR have also proven useful for modeling virus spreads, in spite of their simplicity. The ability to modify the model's parameters, such as time-varying, can assist in refining the model's predictive power.

Simulations allowed for the direct modeling of government health interventions, aiding in understanding their effects on the COVID-19 statistics. The ability to modify parameters using DOE has proven useful for reducing computation intensity. Simulating the possibilities of the virus's spread allows leaders to see a worst-case and best-case scenario, aiding in their

decision making process. Advances in computing power will allow simulations to become even more powerful tools of prediction.

Despite the progress made by researchers in the areas of network science, time series models, epidemiological models, and simulation, there does not currently exist a model using a commuting data network structure to predict COVID-19 cases and deaths. Using both the time series data for COVID-19 cases and deaths and the commuting data network structure, we can create a predictive model for the number of cases and deaths by county. The GNAR package inputs the time series data and the network structure in order to make a model that helps understand the relationship between COVID-19 cases and deaths amongst counties (Knight et al. 2020). The model and methodology from this thesis can assist local leaders in implementing public health measures in order to mitigate the spread of the virus.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Methodology

We obtain the data for this thesis from the USCB, USAFacts, and the CDC. USAFacts obtains their data from the CDC and updates the daily death count on their website. Manipulating this data in `Python`, we transform the data into a usable format for the GNAR package, created models, and assess them using certain metrics.

GNAR is a suitable tool for this analysis because of its incorporation of both the time series data and the underlying network structure. Although the aforementioned models in Chapter 2 are able to make predictions about time series data using a variety of parameters and assumptions, none of them have been able to incorporate the flow of people from one area to another. The network we create using this data gives a new way of describing correlation between counties. The edges of this network represent the connection between the counties, with the edge weight meaning the number of commuters. Because the virus spreads as a result of people contacting each other, this methodology is novel and can be used to better inform public health officials.

3.1 Data

The data for this thesis is entirely numerical, with no categorical predictors or response variables. No transformations are applied to the original data for this thesis' models.

3.1.1 Data Assumptions

First, the COVID-19 cases and deaths data meets the assumption of stationarity of errors because the noise of the data does not depend on the time at which the data was observed (Hyndman and Athanasopoulos 2018). Autoregressive models require stationarity of the errors, meaning that the series' variance must be constant over a long time period (Statistics Solutions 2021). Additionally, in the long term, this dataset should not contain predictable cycles or patterns (Hyndman and Athanasopoulos 2018).

The COVID-19 data is assumed to be 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 is accurate (Center for Disease Control and Prevention 2020). Additionally, the data is autoregressive, meaning that one time period's observation depends on the value of a previous period's observation, using a regression equation to predict a future value (Knight et al. 2020). In this thesis we assume no outliers are present, since they can adversely affect a model's initial creation and recommendations as they can be misleading (Frost 2020).

3.1.2 Data Limitations

The principal limitation of this data involves the constant nature of the commuting network structure (United States Census Bureau 2019). The USCB compiled this commuting data over a five 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.

Similarly, the cases and deaths data from USAFacts and the CDC do not model active COVID-19 infections. Instead, the local municipalities give a count of deaths updated daily. Therefore, we only model total COVID-19 deaths, without modeling active infections.

3.1.3 COVID-19 Cases and Deaths

One must understand how COVID-19 cases and deaths are determined and recorded in order to use this data for prediction. A COVID-19 death means that COVID-19 was confirmed as a probable cause of death (Center for Disease Control and Prevention 2020). Additionally, when confirming a COVID-19 case, laboratories must search within a subject for specific genetic material of the virus, such as proteins (Harvard University 2020). Antibody tests in blood can also provide information on whether a subject had the virus or not; however, infected people do not produce antibodies immediately (Harvard University 2020). Evidence of infection in antibody blood tests appears around three weeks after infection, not immediately (Harvard University 2020). In spite of the antibody test's lagging nature, as the pandemic continues, more testing methods have been developed that allow for quicker results.

Once a laboratory or hospital confirms an infected individual, state reporting agencies,

hospitals, healthcare providers, and laboratories must report that case to local or state health departments (Center for Disease Control and Prevention 2020). In spite of disease laws requiring this reporting to state health departments, reporting to the CDC is not mandatory (Center for Disease Control and Prevention 2020). Regardless, the CDC data team ensures integrity and regularity of the data sent to them. The CDC then sends the data back to the original reporting authority for a final confirmation or request for correction. Finally, the CDC publishes the cleaned data on its website and allows public sharing and consumption of the data (Center for Disease Control and Prevention 2020).

USAFacts is a non-profit, nonpartisan organization that provides data on such subjects as government finances and the American population (USAFacts 2020). This organization aggregates COVID-19 data from the CDC and provides two datasets for public consumption: a by-county daily cumulative total for COVID-19 deaths and a similar dataset for confirmed or probable cases. The CDC started recording data on January 22, 2020, updating the data daily. In this thesis, we use the data from January 22, 2020 to September 17, 2021, providing 604 days of data.

3.1.4 County Information

The United States is broken up into 3,143 county equivalents as of 2020 (United States Census Bureau 2020a). 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 given a unique five-digit federal information processing standard (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 United States, 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 (United States Census Bureau 2020b). However, Delaware contains the least amount of counties and Texas contains the most (United States Census Bureau 2020b). Moreover, the majority of the country’s population lives in only 143 of the 3143 counties as of 2020 (United States Census Bureau 2020b). Table 3.1 displays the number of counties

in each state.

Table 3.1. United States Counties. As we can see, the number of counties per state varies widely. Source: (The Fact File 2021).

Sate	Counties	State	Counties
Alabama	67	Montana	56
Alaska	29	Nebraska	93
Arizona	15	Nevada	17
Arkansas	75	New Hampshire	10
California	58	New Jersey	21
Colorado	64	New Mexico	33
Connecticut	8	New York	62
Delaware	3	North Carolina	100
District of Columbia	1	North Dakota	53
Florida	67	Ohio	88
Georgia	159	Oklahoma	77
Hawaii	5	Oregon	36
Idaho	44	Pennsylvania	67
Illinois	102	Rhode Island	5
Indiana	92	South Carolina	46
Iowa	99	South Dakota	66
Kansas	105	Tennessee	95
Kentucky	120	Texas	254
Louisiana	64	Utah	29
Maine	16	Vermont	14
Maryland	24	Virginia	133
Massachusetts	14	Washington	39
Michigan	83	West Virginia	55
Minnesota	87	Wisconsin	72
Mississippi	82	Wyoming	23
Missouri	115		

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 (Broomfield County 2021). Shannon County, South Dakota renamed itself to Ogala County in 2015 out of respect to its Native American heritage (Ban 2014). 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.

3.2 County Network Structure

In this thesis, we construct a network structure from the original commuting data from the USCB using Python. The commuting data comes in the form of a dataframe 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.

In addition to time series data delineated by a certain structure, GNAR requires a similarly delineated network structure. 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.

Developing this network structure requires us to divide the commuting data matrix into states, however. The computation time for creating a commuting data matrix for the whole United States would be impractical for this dataset, since it contains 137,451 instances of workers commuting. Therefore, we have a county commuting network structure for each state.

We designed each state’s county commuting network structure to include the following for a given state:

- Workers commuting from within-state counties to within-state counties.
- 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 (Newman et al. 2006). For example, in a social network, a community could be a subgroup of friends. By dividing the network into states, we can observe “properties at the community level that are quite different from their properties at the level of the entire network” (Newman et al. 2006). Looking at the United States network as a whole may “ignore community structure and may miss many interesting features” (Newman et al. 2006). By dividing the larger county-wide network into states, we can more easily study the smaller communities in the network (Newman et al. 2006). Often, analyzing these subsets of a network can provide insight into the larger network’s behavior (Newman et al. 2006).

3.2.1 Time Series Cases and Deaths Data

Now that we created county commuting network structures for each state, we can focus on the time series cases and deaths data from USAFacts. Using Python in a similar manner as above, we filtered the cases and deaths data to only include those counties included in each state’s county commuting network structure. Therefore, we have the time series cases and deaths data for each state, including within-state commuting, out-of-state commuting, and into-state commuting.

3.3 Model Implementation

Now that we have an accurately filtered county commuting network structure and time series cases and deaths dataframe for each state, we can now feed these into the GNAR package in R in order to obtain a predictive time series network model (Leeming et al. 2020). A node in GNAR represents a single county and its time series dataset. Therefore, there are 3,143 nodes representing the COVID-19 cases and deaths from January 22, 2020 to September 17, 2021, making a total of 605 time series observations for each node.

3.3.1 GNAR Assumptions

In order to employ the GNAR package correctly, we must acknowledge some assumptions. First, the time series data set must have a corresponding underlying network structure.

Because our time series cases and deaths data and the commuting network structure are both delineated by county, this assumption is satisfied. Additionally, when creating a model, GNAR handles missing data relatively well (Knight et al. 2020). We assume our datasets are complete for each county; therefore, GNAR still fits our needs. Another assumption we satisfy is homogeneity of the data structures at each node in the network. Since each node contains the time series cases and deaths data for the same amount of days, this assumption is satisfied (Knight et al. 2020).

Moreover, the GNAR package assumes that the time series follows an autoregressive model at each node and is influenced by neighboring nodes (Knight et al. 2020). As discussed in Section 3.1.1, the data in this thesis satisfies the autoregressive assumption. Therefore, this thesis satisfies the assumptions of the GNAR package.

3.3.2 GNAR Adaptation

The GNAR package inputs both univariate time series data and an underlying network structure in order to create an overarching multivariate time series (Knight et al. 2020). The COVID-19 cases and deaths data is by nature univariate, since it only includes one variable. Univariate data most simply describes fundamental patterns and avoids multicollinearity issues of multivariate data. Thus, the nodes of this thesis' network accept only univariate data. As discussed in Section 2.4, GNAR inputs X_t , a vector of random variables which varies over the time horizon, with each random variable associating with a county in our network. In this thesis, we have two types of random variables, one for deaths and one for cases. Each county's X_t runs from $t = \text{week } 0$ to $t = \text{week } 86$.

A network includes not only nodes, but also edges. The edges of this thesis' network have weights, which GNAR incorporates into its models. As discussed in Section 2.4, the weight $\mu_{i,j}$ represents the weight along edge (i, j) in the network. The weight on each edge represents the number of commuters from one county to another. Since these edges have a directional property, going from one county to another, they are also directed.

3.3.3 GNAR Limitations

On a similar note, because GNAR only accepts univariate time series data, no multivariate analysis is possible. Multiple variables by county such as demographics, economic data, and

hospital statistics cannot fit into a GNAR predictive model. Implementing such predictors could enhance the response variable, COVID-19 cases or deaths, and result in a more accurate model overall.

3.3.4 Rolling Horizon

Rolling horizon, as discussed in Section 2.3.2, is a method 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. Rolling horizons allow one to choose both the start of the training period and the number steps ahead to predict, p . As depicted in Figure 3.1, we increase the sample size as the evaluations progress (Svetunkov 2021).

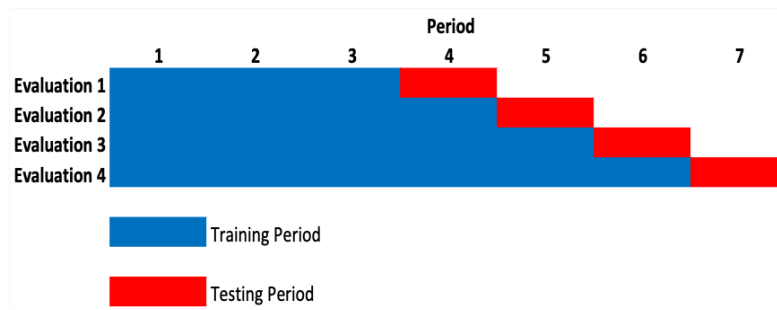


Figure 3.1. As we can see from the graph, the chosen training origin consists of periods 1,2 and 3. The testing interval, p , is one period ahead. After each evaluation, period $t - 1$ is included in the training period, where period t is the testing period.

For rolling horizon design, we first conduct an initial warm up period for train a time series model. This warm up period must be sufficiently long to allow the model to understand the data's disposition. Because of COVID-19's novel nature, the recording of its cases and deaths only started in January of 2020. Additionally, the number of cases and deaths remain sparse for the first 70 days of reporting, affecting any model's ability to predict accurately.

In order to account for a sufficiently long warm up period, we set the training origin to be the first 46 weeks of the data, ending at December 9, 2020, making a total of 322 warm up days. Thus, we include the 70 days of sparse COVID-19 reporting in the warm up period.

From January 22, 2020 to September 17, 2021, we have 604 days of data. With a training origin of 322 days, we can now fit a model for the remaining 282 days, or about 40 weeks.

Wren (2021) shows that his models that divide the time series data into weekly increments perform better than those predicting in daily increments. Therefore, we convert the daily time series data into weekly data, using the average number of deaths and cases for each week instead. Thus, the rolling horizon's training period consists of the first 46 weeks of the data. The testing period starts at week 46 and ends at week 86 of our dataset as the evaluations progress.

The testing period, p , of a rolling horizon can severely affect a model's performance. A small p results a small error for the overall performance because of a small variance in each time period for predicting near future. Conversely, a rolling horizon with a large p can produce a very large overall error since it tries to predict far into the future. To simulate a realistic prediction, we selected $p = 4$ for this thesis' weekly models. Thus, we predict 4 weeks into the future. A month's notice gives public health officials sufficient time to implement countermeasures into their local areas. Likewise, a month's forecast would also allow them to reduce restrictions if cases or deaths are on a downward trend.

A naïve model, as discussed in Section 2.3.2, serves as a baseline model from which we can measure performance of other models. Naïve models set the value of the prediction equal to the actual value from the previous time period. We incorporate naïve models into the rolling horizon design by lagging the prospective testing period by one time period. Any time series model that can outperform a naïve model is said to have predictive power (Erdem 2020).

3.3.5 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 (Leeming et al. 2020). These parameters represent the time lag, p , and the maximum stage of neighbor dependence for each of the time lags, $[s]$, as discussed in Section 2.4. 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 and betaOrder = 0. Wren (2021) conducted model selection by changing alphaOrder and betaOrder from 0 to 5 independently. He selected these parameter values because models with these values performed best (Wren 2021). Table 3.2 provides a summary of the model parameter combinations for this thesis.

Table 3.2. 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
alphaOrder = 1	alphaOrder = 0	alphaOrder = 0
betaOrder = 1	betaOrder = 1	betaOrder = 0

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.

3.3.6 State Selection

Now that we have the model framework defined, we now must choose the states on which to run our models. In order to determine if our models would perform similarly across different states, we select states of varying vaccination rates. Vaccination rates could affect the time-varying number of cases and deaths in a state, potentially leading to differing model performances. We choose the following states for our analysis: Rhode Island, Massachusetts, California, Florida, and Arkansas. Table 3.3 describes the vaccination rate (% of population) and corresponding rank out of 50 of each state we choose.

Table 3.3. State Vaccination Rates. These by-state vaccination rates help us determine which states to choose for comparison. Adapted from (Adams 2021).

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

3.4 Summary

The GNAR package enables us to input a univariate time series dataset along with an underlying network structure in order to create a predictive time series model for the univariate data. After transforming the data into a usable format for GNAR, we are able to fit three models for each prediction type (cases and deaths), giving us 6 models for each state. Our 5 states have 6 models each, making a total of 30 individual models, which we can evaluate for prediction accuracy graphically, using the MAPE and MASE as measures of performance.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:

Results

As the graphs will show, the cases and deaths time series plots mirror one another closely, with minimal differences in behavior. The MASE accuracy trends towards zero, increasing accuracy as the model learns. As discussed in Section 2.3.3, a model is considered more accurate if its MASE and MAPE are close to 0 (Hyndman 2006).

In order to determine if common transformations of data could help improve performance, we transform the county commuting network data in the following ways: log transform, square-root, and normalization. Ultimately, however, these transformations result in minimal change to the unscaled results; therefore, the models we create henceforth do not incorporate any data transformations.

All models' MAPE trends downward as the model kept predicting. The naïve models across all states started with a high MAPE at the beginning of the testing period, sometimes doubling the MAPE of all the other models. Additionally, the naïve model's MAPE appears to increase slightly across all states in the last four weeks of the testing period.

As we will see in the graphs, every state's MASE values for Models 1 and 2 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 five weeks of the testing period for all models. Models 1 and 2 perform worse than the baseline naïve model during this bimodal hump period.

All states' graphs show that Model 3 performs the best out of all of the models, since it outperforms the naïve model almost 100% of the time. Only this model, therefore, should be considered for use by epidemiologists and public health officials.

4.1 Rhode Island Results

Rhode Island, a state with few counties but a dense population that is highly vaccinated, exhibits a great deal of difference among its models. Model 3 is the best performing model in this state.

We consider three models for estimating the number of Rhode Island cases. The MASE and MAPE for the three models are summarized in Table 4.1. As seen in the second column of Table 4.1, in estimating the number of cases, Model 3 exhibits an average MASE of 0.2850, which is three times less than the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0334, which is two times less than the next best performing model, Model 1. Model 3 has the lowest mean, median, and variance in terms of MASE and MAPE. Because Model 3 varies the least and has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.1. Rhode Island Cases Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8585	0.5169	0.7696
Model 2	1.129	0.6648	1.1376
Model 3	0.2850	0.2092	0.0353
Naïve	1	1	0
MAPE			
Model 1	0.0695	0.0562	0.0013
Model 2	0.0906	0.0774	0.0019
Model 3	0.0334	0.0235	0.0007
Naïve	0.1642	0.1030	0.0205

Turning to the estimates over time, Figure 4.1 plots the MASE percentage versus time. We observe that Rhode Island's Model 3 for cases performs the best in terms of the MASE metric. Models 1 and 2 exhibit a dramatic increase towards the end of the testing period, performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks.

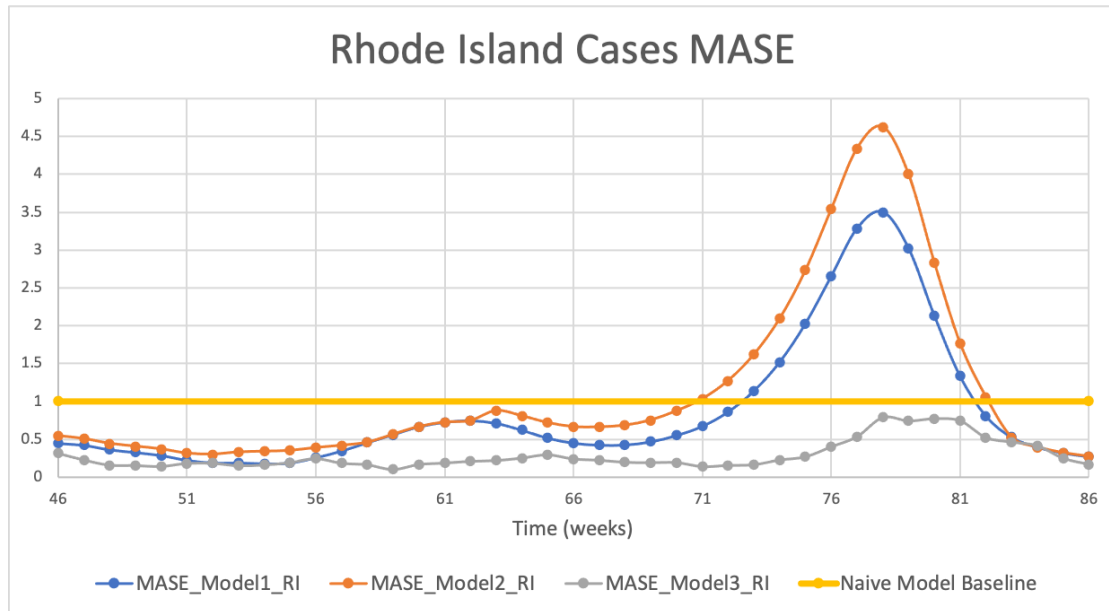


Figure 4.1. Rhode Island Cases MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 77. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.2, Rhode Island's Model 3 for cases performs the best in terms of the MAPE metric. Models 1 and 2 perform similarly through the whole testing period, performing worse than Model 3. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 62.

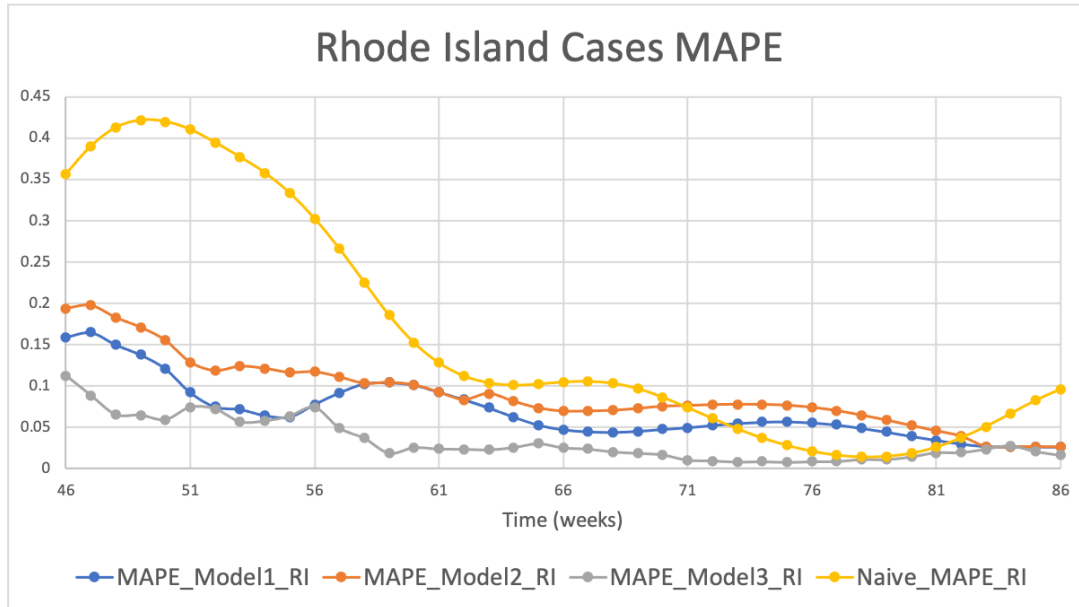


Figure 4.2. Rhode Island Cases MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 perform well against the naïve model until approximately week 71. Models 1, 2, and 3 all trend downward gradually. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

In a similar manner, we consider three models for estimating the number of Rhode Island deaths. The MASE and MAPE for the three models are summarized in Table 4.2. As seen in the second column of Table 4.2, in estimating the number of cases, Model 3 exhibits an average MASE of 0.5170, which is about one half of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0419, which is almost half of the next best performing model, Model 1. Model 3 has the lowest mean, median, and variance in terms of MASE and MAPE. Because Model 3 varies the least and has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.2. Rhode Island Deaths Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.9888	0.7898	0.6436
Model 2	1.8501	1.7844	2.0184
Model 3	0.5170	0.3698	0.0934
Naïve	1	1	0
MAPE			
Model 1	0.0615	0.0529	0.0009
Model 2	0.1073	0.1014	0.0017
Model 3	0.0419	0.0268	0.0008
Naïve	0.1201	0.0705	0.0101

Turning to the estimates over time, Figure 4.3 plots the MASE percentage versus time. We observe that Rhode Island's Model 3 for deaths performs the best in terms of the MASE metric, though it did not outperform the naïve model for about six weeks from approximately week 78 to week 83. Models 1 and 2 exhibit a dramatic increase towards the end of the testing period, performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks. Model 1 does not outperform the naïve model at any point after approximately week 61.

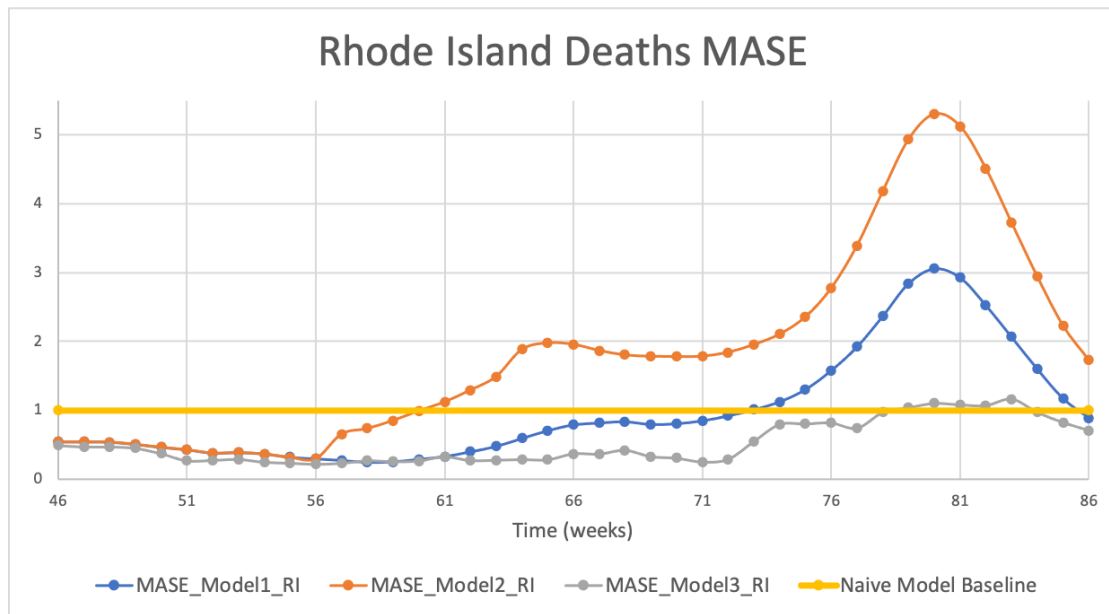


Figure 4.3. Rhode Island Deaths MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Model 2 starts to perform worse than the naïve model at approximately week 61 and continues until the end of the testing 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. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.4, Rhode Island's Model 3 for deaths performs the best in terms of the MAPE metric. Models 1 and 2 perform worse than Model 3 throughout the testing period. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. Model 2 experiences a sharp increase in MAPE at approximately week 56. The naïve MAPE only starts to perform similarly to the other models at approximately week 66.

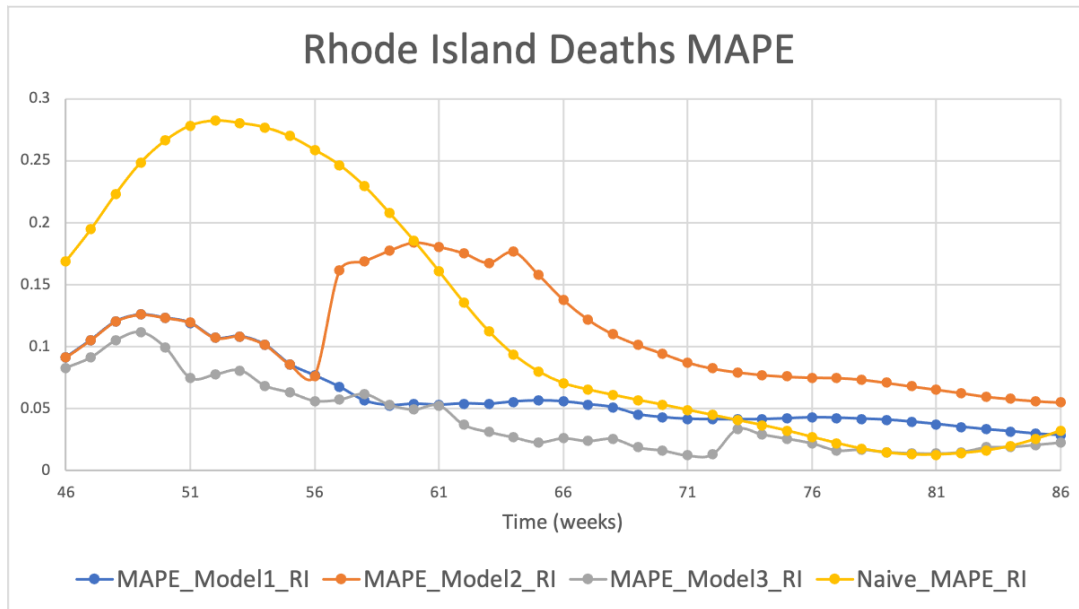


Figure 4.4. Rhode Island Deaths MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Model 2 experienced a sharp increase in MAPE at approximately week 56, trending to perform worse than the naïve model 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. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

4.2 Massachusetts Results

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.

Again, we consider three models for estimating the number of Massachusetts cases. The MASE and MAPE for the three models are summarized in Table 4.3. As seen in the second column of Table 4.3, in estimating the number of cases, Model 3 exhibits an average MASE of 0.2556, which is almost four times less than the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0327, which is about two times less than the next best performing model, Model 1. Model 3 has the lowest mean, median, and variance in terms of MASE and MAPE. Because Model 3 varies the least and has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.3. Massachusetts Cases Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8579	0.6553	0.5338
Model 2	0.9869	0.7631	0.7752
Model 3	0.2556	0.1995	0.0193
Naïve	1	1	0
MAPE			
Model 1	0.0723	0.0558	0.0013
Model 2	0.0780	0.0675	0.0012
Model 3	0.0327	0.0207	0.0008
Naïve	0.1582	0.0862	0.0204

Turning to the estimates over time, Figure 4.5 plots the MASE percentage versus time. We observe that Massachusetts' Model 3 for cases performs the best in terms of the MASE metric. Models 1 and 2 perform similarly, exhibiting a dramatic increase towards the end of the testing period and performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks.

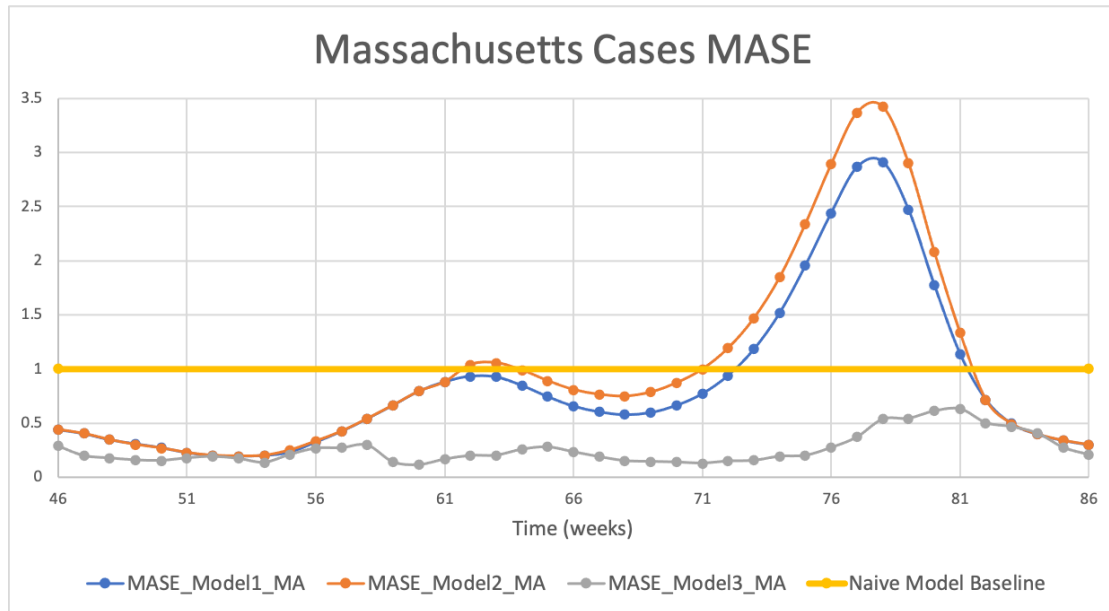


Figure 4.5. Massachusetts Cases MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 78. Models 1 and 2 perform worse than the naïve model in this time. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.6, Massachusetts' Model 3 for cases performs the best in terms of the MAPE metric. Models 1 and 2 perform similarly through the whole testing period, performing worse than Model 3. For approximately a 9 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 61.

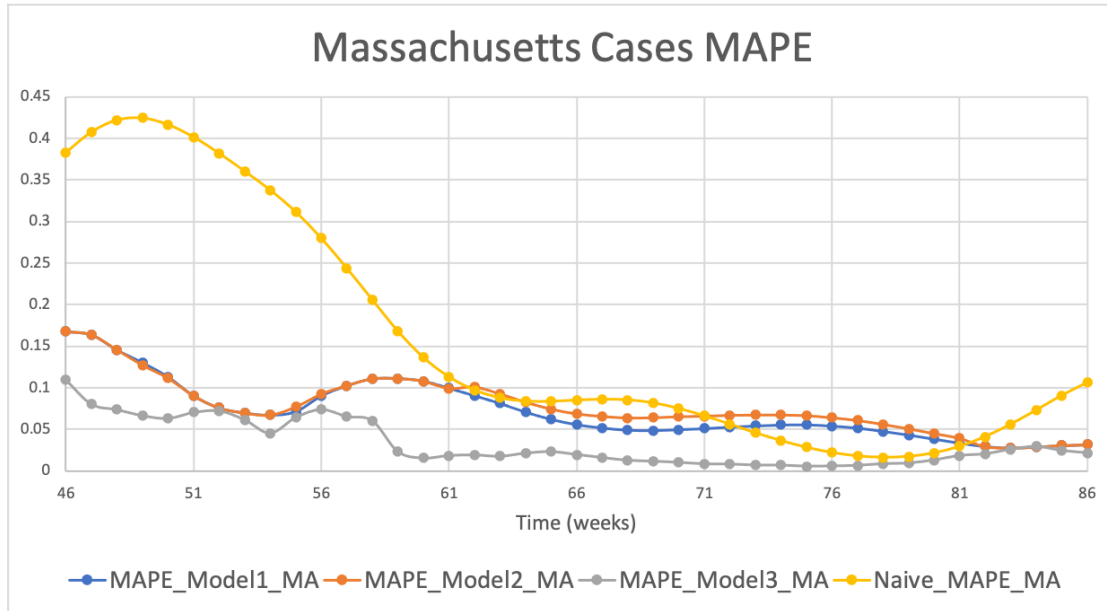


Figure 4.6. Massachusetts Cases MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely, performing worse than the naïve model from approximately week 71 to week 81. Models 1, 2, and 3 all trend downward gradually. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

In a similar manner, we consider three models for estimating the number of Massachusetts deaths. The MASE and MAPE for the three models are summarized in Table 4.4. As seen in the second column of Table 4.4, in estimating the number of cases, Model 3 exhibits an average MASE of 0.3933, which is about one half of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0467, which is almost half of the next best performing model, Model 1. Model 3 has the lowest mean, median, and variance in terms of MASE and MAPE. Because Model 3 varies the least and has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.4. Massachusetts Deaths Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8263	0.7072	0.2572
Model 2	0.8275	0.7084	0.2586
Model 3	0.3933	0.3651	0.0103
Naïve	1	1	0
MAPE			
Model 1	0.0691	0.0559	0.0015
Model 2	0.0692	0.0559	0.0015
Model 3	0.0467	0.0273	0.0014
Naïve	0.1374	0.0747	0.0136

Turning to the estimates over time, Figure 4.7 plots the MASE percentage versus time. We observe that Massachusetts' Model 3 for deaths performs the best in terms of the MASE metric. Models 1 and 2 exhibit a dramatic increase towards the end of the testing period, performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks. Models 1 and 2 follow each other closely, almost overwriting each other.

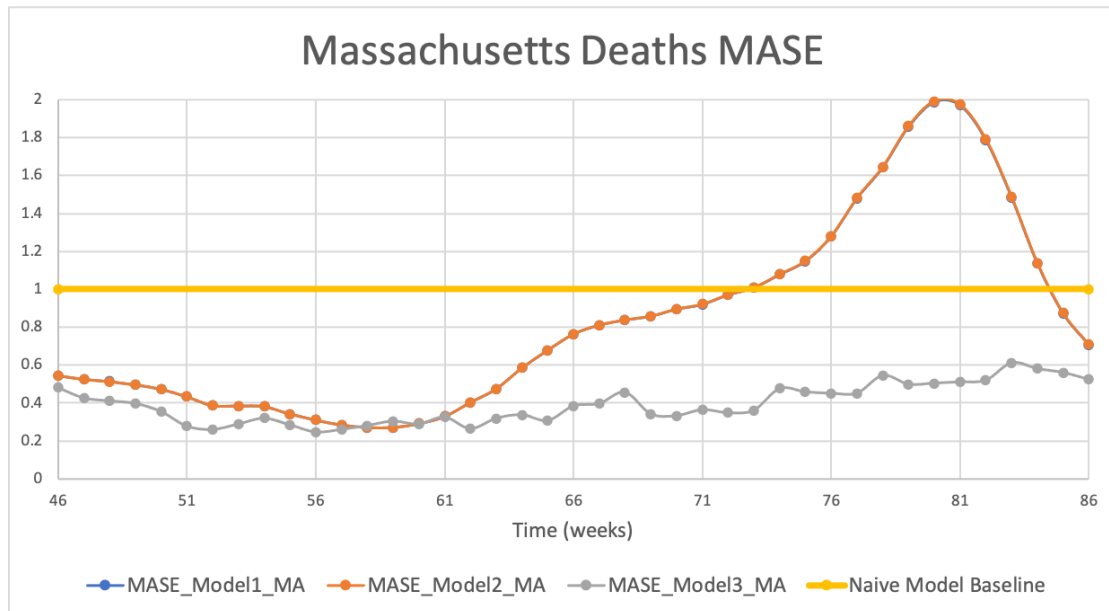


Figure 4.7. Massachusetts Deaths MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 again follow each other very closely, almost overwriting each other. Models 1 and 2 start to perform worse than the naïve model at approximately week 72 but end up performing better than at approximately week 84. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.8, Massachusetts' Model 3 for deaths performs the best in terms of the MAPE metric. Models 1 and 2 perform similarly through the whole testing period, performing worse than Model 3. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 66.

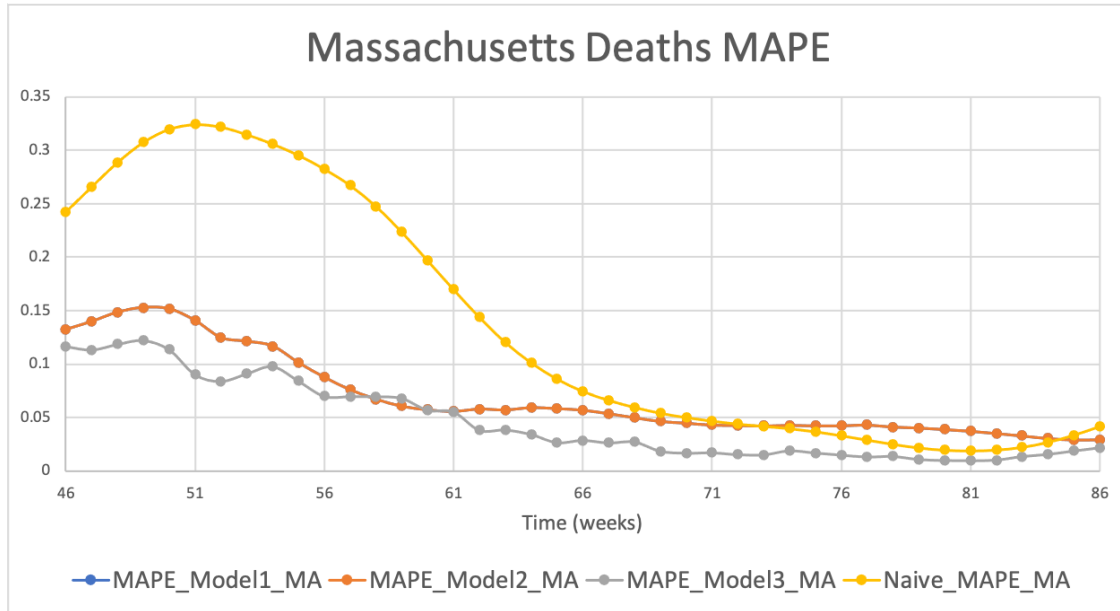


Figure 4.8. Massachusetts Deaths MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely, almost overwriting each other. They perform worse than the naïve model from approximately week 75 to week 84. All models trend gradually downwards. Adapted from (United States Census Bureau 2019) and (US-AFacts 2021a).

4.3 California Results

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.

Again, we consider three models for estimating the number of California cases. The MASE and MAPE for the three models are summarized in Table 4.5. As seen in the second column of Table 4.5, in estimating the number of cases, Model 3 exhibits an average MASE of 0.4199, which is less than half of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0492, which is almost one half of the next best performing model, Model 1. Model 3 has the lowest mean and median in terms of MASE and MAPE. Because Model 3 has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.5. California Cases Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.9056	0.8205	0.3836
Model 2	0.9840	0.9102	0.5088
Model 3	0.4199	0.3480	0.0403
Naïve	1	1	0
MAPE			
Model 1	0.0764	0.0660	0.0014
Model 2	0.0830	0.0678	0.0015
Model 3	0.0492	0.0347	0.0023
Naïve	0.1507	0.0737	0.0197

Turning to the estimates over time, Figure 4.9 plots the MASE percentage versus time. We observe that California's Model 3 for cases performs the best in terms of the MASE metric. Model 3 experiences some variability at the beginning of the testing period. Models 1 and 2 perform similarly, exhibiting a dramatic increase towards the end of the testing period and performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks.

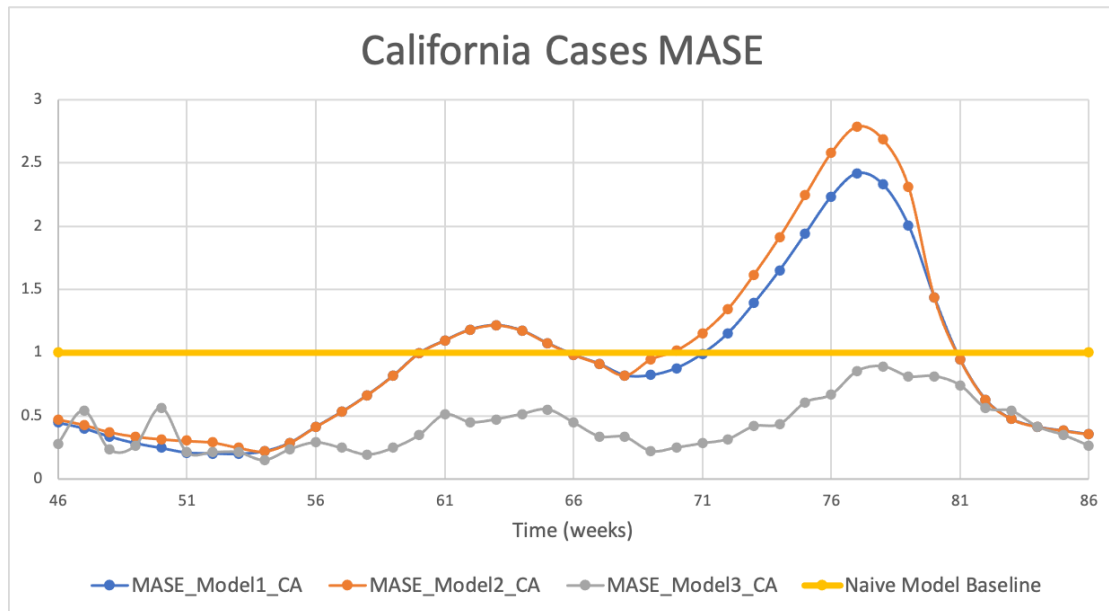


Figure 4.9. California Cases MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely and exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 77. Models 1 and 2 perform worse than the naïve model in this time. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.10, California's Model 3 for cases performs the best in terms of the MAPE metric. Model 3 experiences some variability towards the beginning of the testing period. Models 1 and 2 perform similarly through the whole testing period, performing worse than Model 3. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 61.

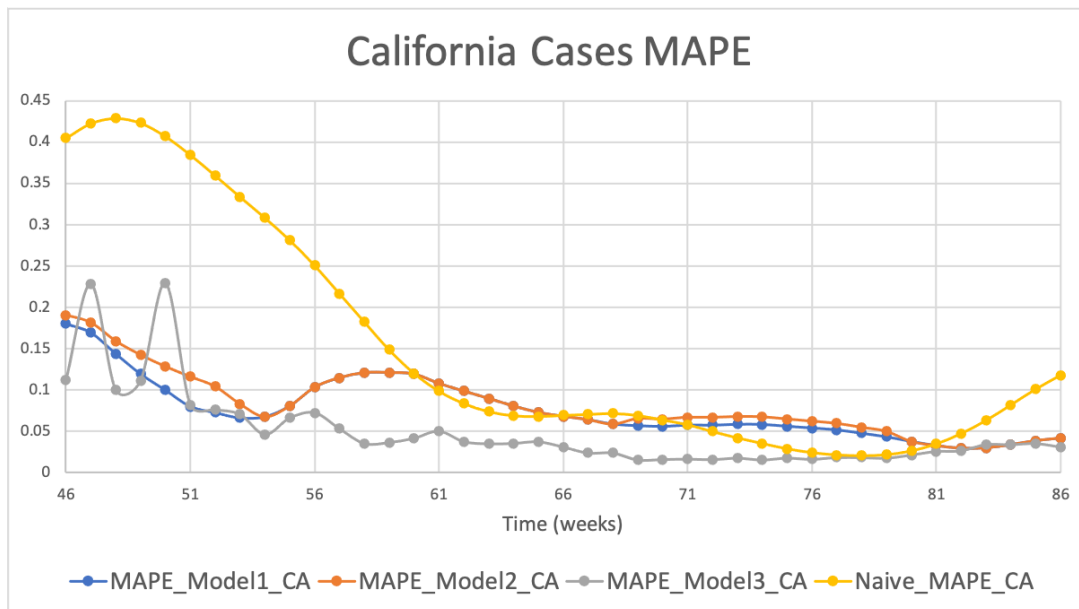


Figure 4.10. California Cases MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely, performing worse than the naïve model from approximately week 61 to week 80. Models 1, 2, and 3 all trend downward gradually. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

In a similar manner, we consider three models for estimating the number of California deaths. The MASE and MAPE for the three models are summarized in Table 4.6. As seen in the second column of Table 4.6, in estimating the number of cases, Model 3 exhibits an average MASE of 0.4700, which is about one half of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0594, which is slightly less than the next best performing model, Model 1. Model 3 has the lowest mean and median in terms of MASE and MAPE. Because Model 3 has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.6. California Deaths Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8010	0.7351	0.2000
Model 2	0.9192	0.8004	0.2475
Model 3	0.4700	0.4574	0.0169
Naïve	1	1	0
MAPE			
Model 1	0.0743	0.0615	0.0019
Model 2	0.0840	0.0777	0.0019
Model 3	0.0594	0.0393	0.0028
Naïve	0.1447	0.0770	0.0152

Turning to the estimates over time, Figure 4.11 plots the MASE percentage versus time. We observe that California's Model 3 for deaths performs the best in terms of the MASE metric. Models 1 and 2 exhibit a dramatic increase towards the end of the testing period, performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks.

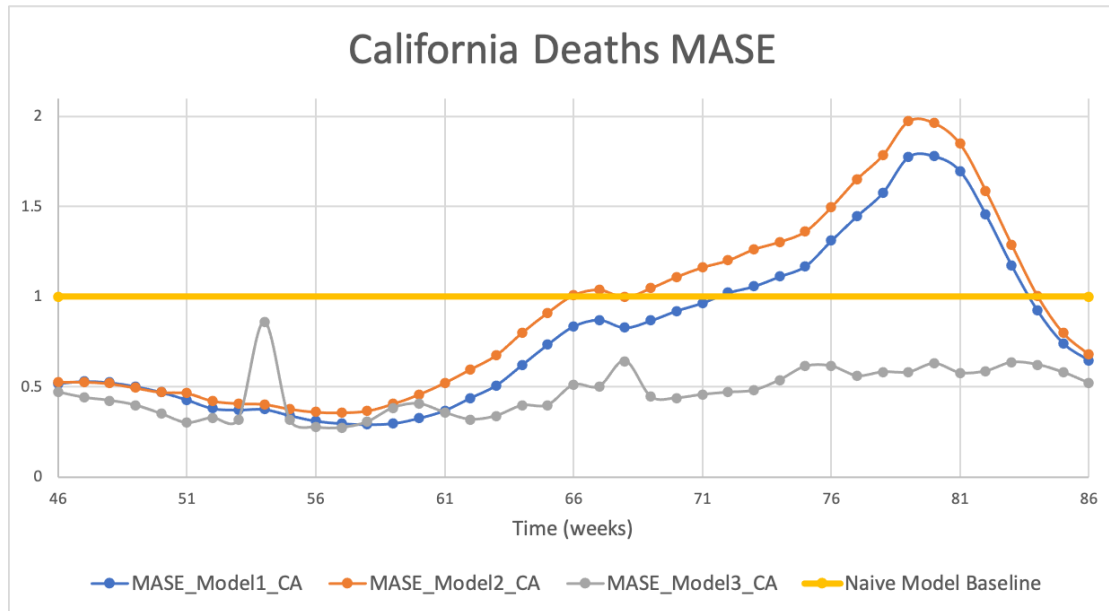


Figure 4.11. California Deaths MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 again follow each other closely. Models 1 and 2 start to perform worse than the naïve model at approximately week 70 but end up performing better than it at approximately week 84. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.12, California's Model 3 for deaths performs the best in terms of the MAPE metric. Model 3 experiences some variability towards the beginning of the testing period. Models 1 and 2 perform similarly through the whole testing period, performing worse than Model 3. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 66.

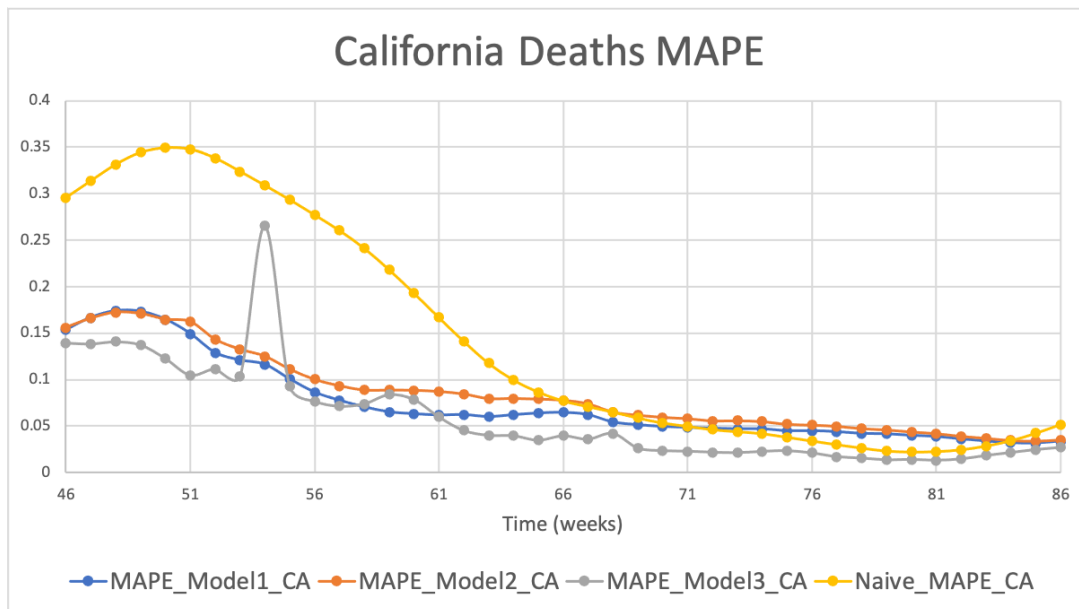


Figure 4.12. California Deaths MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely. They perform worse than the naïve model from approximately week 72 to week 84. All models trend gradually downwards. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

4.4 Florida Results

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.

Again, we consider three models for estimating the number of Florida cases. The MASE and MAPE for the three models are summarized in Table 4.7. As seen in the second column of Table 4.7, in estimating the number of cases, Model 3 exhibits an average MASE of 0.3098, which is approximately one third of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0365, which is approximately one half of the next best performing model, Model 1. Model 3 has the lowest mean and median in terms of MASE and MAPE. Because Model 3 has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.7. Florida Cases Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.9679	0.8819	0.5556
Model 2	1.0033	0.9693	0.5980
Model 3	0.3098	0.2303	0.0272
Naïve	1	1	0
MAPE			
Model 1	0.0735	0.0627	0.0011
Model 2	0.0754	0.0665	0.0010
Model 3	0.0365	0.0204	0.0012
Naïve	0.1450	0.0724	0.0177

Turning to the estimates over time, Figure 4.13 plots the MASE percentage versus time. We observe that Florida's Model 3 for cases performs the best in terms of the MASE metric. Models 1 and 2 perform similarly, exhibiting a dramatic increase towards the end of the testing period and performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks. Models 1 and 2 perform very similarly to each other throughout the testing period.

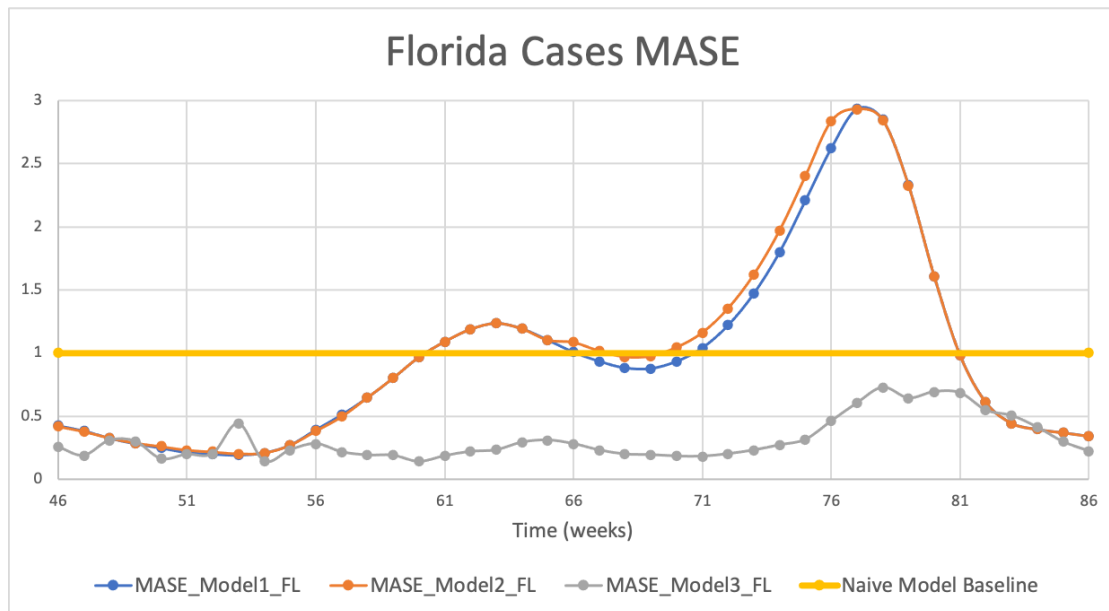


Figure 4.13. Florida Cases MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other very closely and exhibit a dramatic increase in MASE at approximately week 71, then a sharp decrease at week 77. Models 1 and 2 perform worse than the naïve model in this time. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.14, Florida's Model 3 for cases performs the best in terms of the MAPE metric. Model 3 experiences some variability towards the beginning of the testing period. Models 1 and 2 perform similarly throughout the testing period, almost overwriting each other and performing worse than Model 3 consistently. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 61.

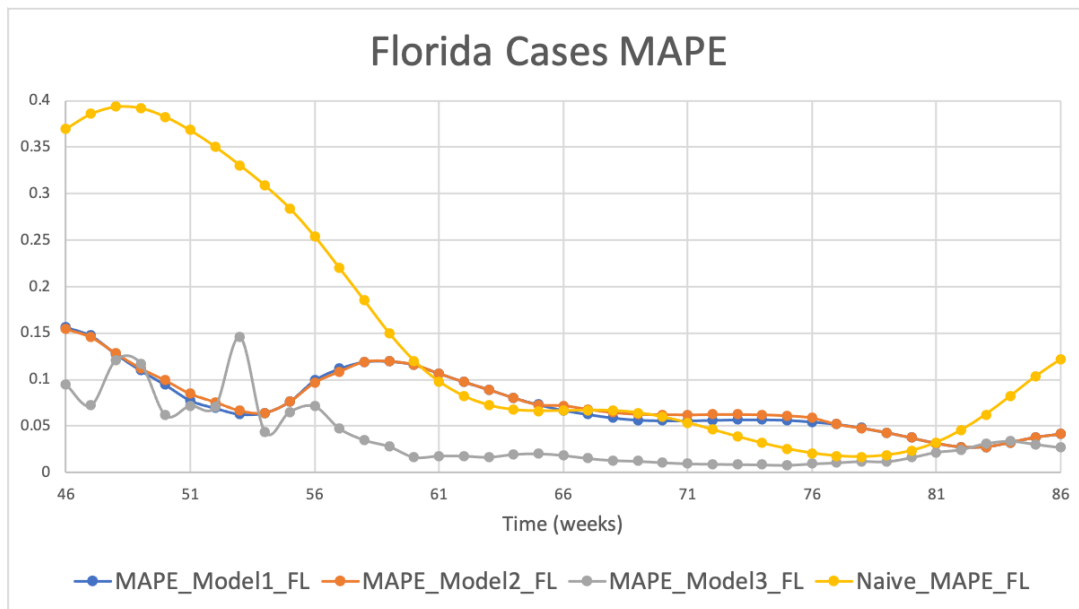


Figure 4.14. Florida Cases MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely, performing worse than the naïve model from approximately week 61 to week 65 and week 71 to week 81. Models 1, 2, and 3 all trend downward gradually. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

In a similar manner, we consider three models for estimating the number of Florida deaths. The MASE and MAPE for the three models are summarized in Table 4.8. As seen in the second column of Table 4.8, in estimating the number of cases, Model 3 exhibits an average MASE of 0.4182, which is about one half of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0491, which is slightly less than the next best performing model, Model 1. Model 3 has the lowest mean, median, and variance in terms of MASE and MAPE. Because Model 3 varies the least and has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.8. Florida Deaths Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8444	0.6806	0.3151
Model 2	0.9405	0.7515	0.4205
Model 3	0.4182	0.4013	0.0119
Naïve	1	1	0
MAPE			
Model 1	0.0693	0.0575	0.0015
Model 2	0.0743	0.0646	0.0014
Model 3	0.0491	0.0401	0.0014
Naïve	0.1382	0.0863	0.0132

Turning to the estimates over time, Figure 4.15 plots the MASE percentage versus time. We observe that Florida's Model 3 for deaths performs the best in terms of the MASE metric. Models 1 and 2 exhibit a dramatic increase towards the end of the testing period, performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks.

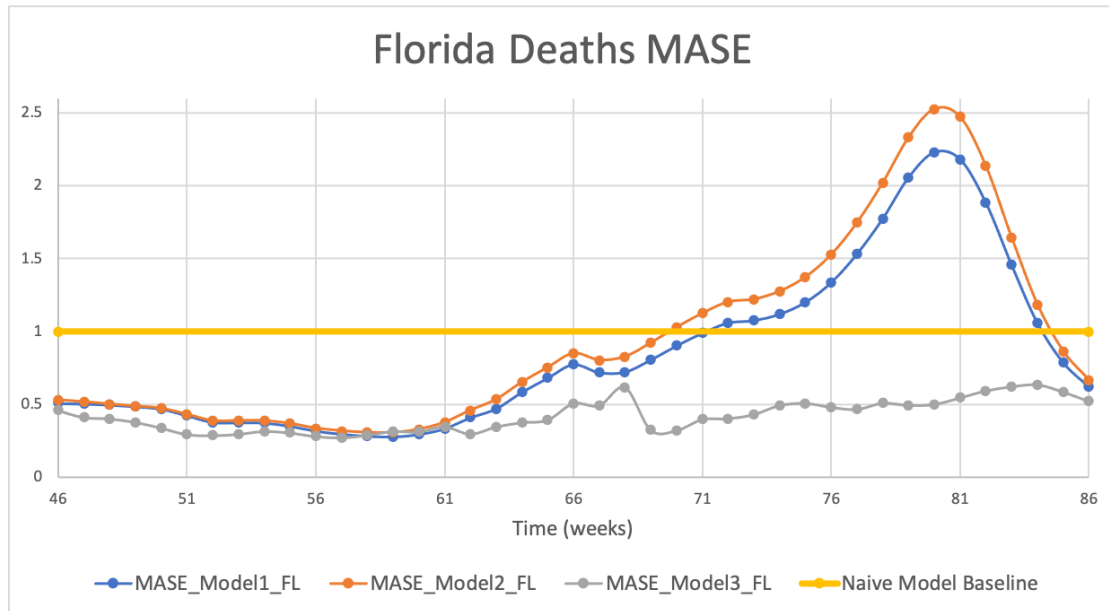


Figure 4.15. Florida Deaths MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely. Models 1 and 2 start to perform worse than the naïve model at approximately week 71 but end up performing better than it at approximately week 84. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.16, Florida’s Model 3 for deaths performs the best in terms of the MAPE metric. Models 1 and 2 perform similarly throughout the testing period, almost overwriting each other and performing worse than Model 3 consistently. For approximately a 10 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 66.

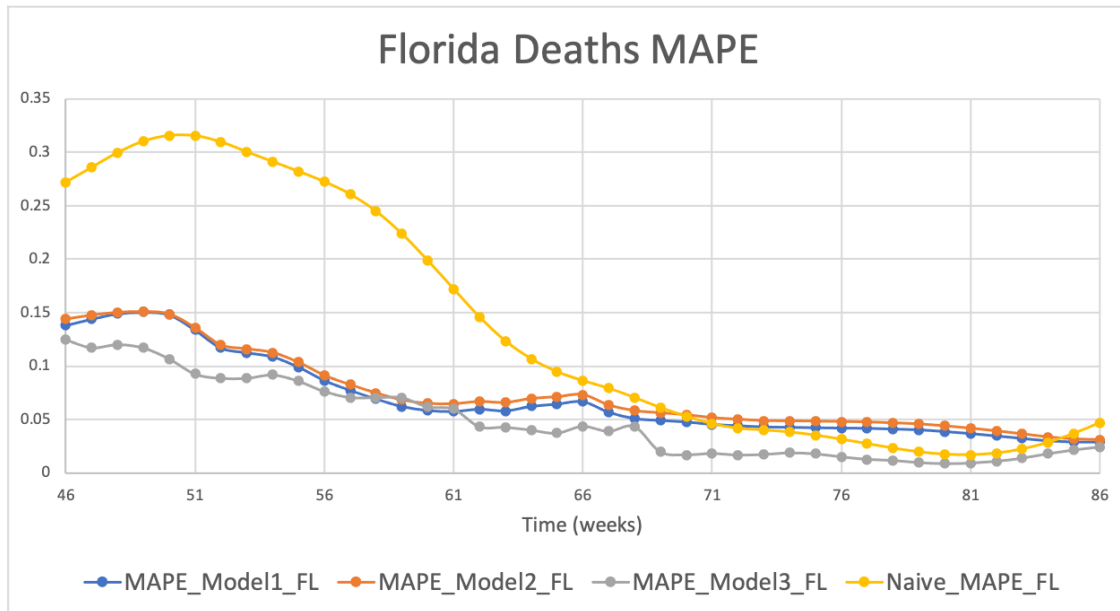


Figure 4.16. Florida Deaths MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely. They perform worse than the naïve model from approximately week 71 to week 84. All models trend gradually downwards. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

4.5 Arkansas Results

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. Model 3 remains the best performing model.

Again, we consider three models for estimating the number of Arkansas cases. The MASE and MAPE for the three models are summarized in Table 4.9. As seen in the second column of Table 4.9, in estimating the number of cases, Model 3 exhibits an average MASE of 0.3396, which is approximately one third of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0344, which is approximately one half of the next best performing model, Model 1. Model 3 has the lowest mean, median, and variance in terms of MASE and MAPE. Because Model 3 varies the least and has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.9. Arkansas Cases Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	1.0270	1.0728	0.4986
Model 2	1.0325	1.0719	0.4876
Model 3	0.3396	0.2941	0.0226
Naïve	1	1	0
MAPE			
Model 1	0.0741	0.0643	0.0011
Model 2	0.0762	0.0688	0.0011
Model 3	0.0344	0.0257	0.0006
Naïve	0.1392	0.0726	0.0170

Turning to the estimates over time, Figure 4.17 plots the MASE percentage versus time. We observe that Arkansas' Model 3 for cases performs the best in terms of the MASE metric. Models 1 and 2 perform similarly, exhibiting a dramatic increase towards the end of the testing period and performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks. Models 1 and 2 follow each other closely throughout the testing period, almost overwriting each other.

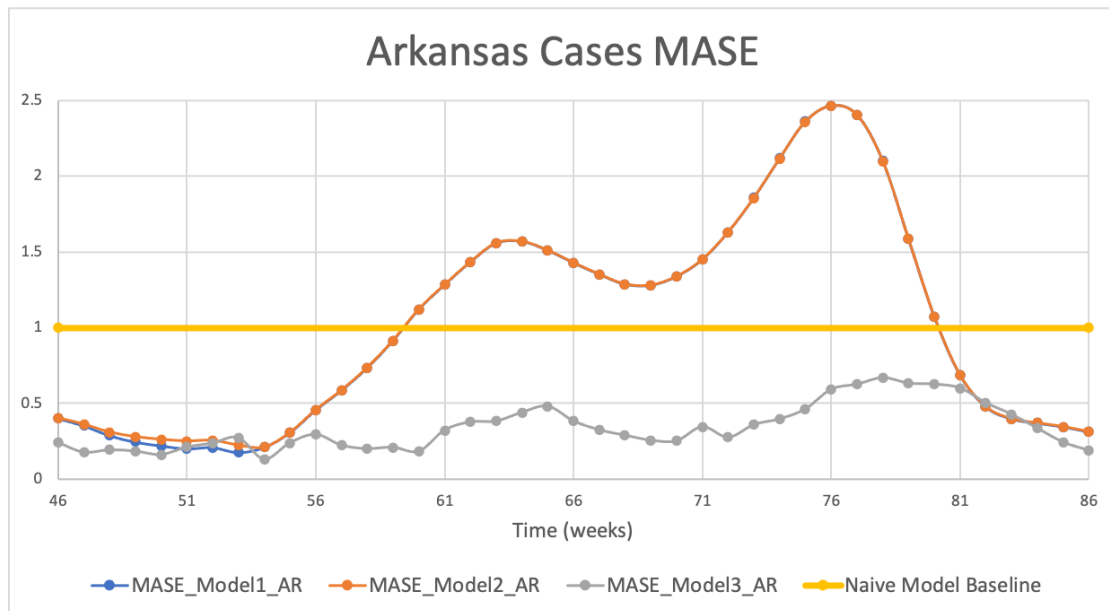


Figure 4.17. Arkansas Cases MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other very closely and exhibit a dramatic increase in MASE much earlier than the other states, 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. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.18, Arkansas' Model 3 for cases performs the best in terms of the MAPE metric. Models 1 and 2 perform similarly throughout the testing period, almost overwriting each other and performing worse than Model 3 except for a small three week period from approximately week 51 to week 54. Models 1 and 2 perform worse than the naïve model from approximately week 60 to week 80. The naïve MAPE only starts to perform similarly to the other models at approximately week 59.

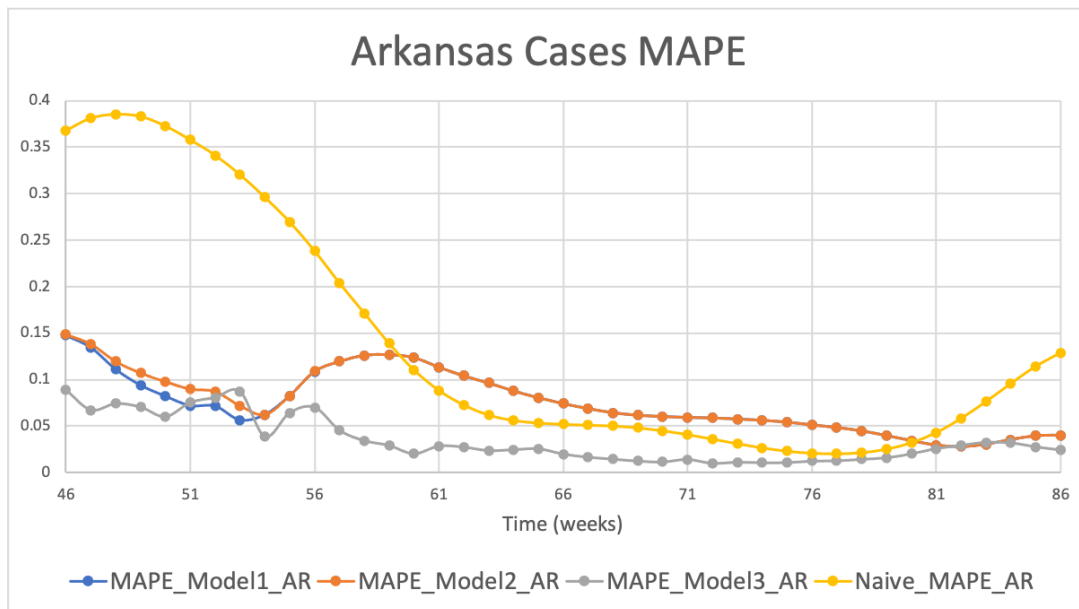


Figure 4.18. Arkansas Cases MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other very closely, performing worse than the naïve model from approximately week 60 all the way to week 80. Models 1 and 2 experienced a steep increase in MAPE at week 55 and then trended downward gradually. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

In a similar manner, we consider three models for estimating the number of Arkansas deaths. The MASE and MAPE for the three models are summarized in Table 4.10. As seen in the second column of Table 4.10, in estimating the number of cases, Model 3 exhibits an average MASE of 0.4240, which is approximately one half of the next best performing model, Model 1. The trend remains the same for the MAPE, with Model 3 exhibiting an average MAPE of 0.0496, which is slightly less than the next best performing model, Model 1. Model 3 has the lowest mean and variance in terms of MASE and MAPE. Because Model 3 has the lowest average of these two metrics, it performs the best and should be the more heavily considered model.

Table 4.10. Arkansas Deaths Model Combination Summary. 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. As discussed in Section 2.3.3, the MASE is the absolute error of the forecast divided by the mean absolute error of the naïve model (Hyndman 2006). By definition, the MASE of the naïve model is always 1 (Hyndman 2006). If a given model's MASE is less than 1, then it performs better than the naïve model. If a given model's MASE is greater than 1, then the naïve model performs better. The MAPE is the average of the error relative to actual observed number. Table 2.1 describes appropriate interpretations for different levels of MAPE. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a)

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8573	0.6743	0.3106
Model 2	0.9903	0.7731	0.4565
Model 3	0.4240	0.3954	0.0141
Naïve	1	1	0
MAPE			
Model 1	0.0698	0.0607	0.0014
Model 2	0.0767	0.0694	0.0013
Model 3	0.0496	0.0377	0.0014
Naïve	0.1380	0.0794	0.0136

Turning to the estimates over time, Figure 4.19 plots the MASE percentage versus time. We observe that Arkansas' Model 3 for deaths performs the best in terms of the MASE metric. Models 1 and 2 exhibit a dramatic increase towards the end of the testing period, performing worse than the naïve model, before gradually trending towards outperforming it in the last five weeks.

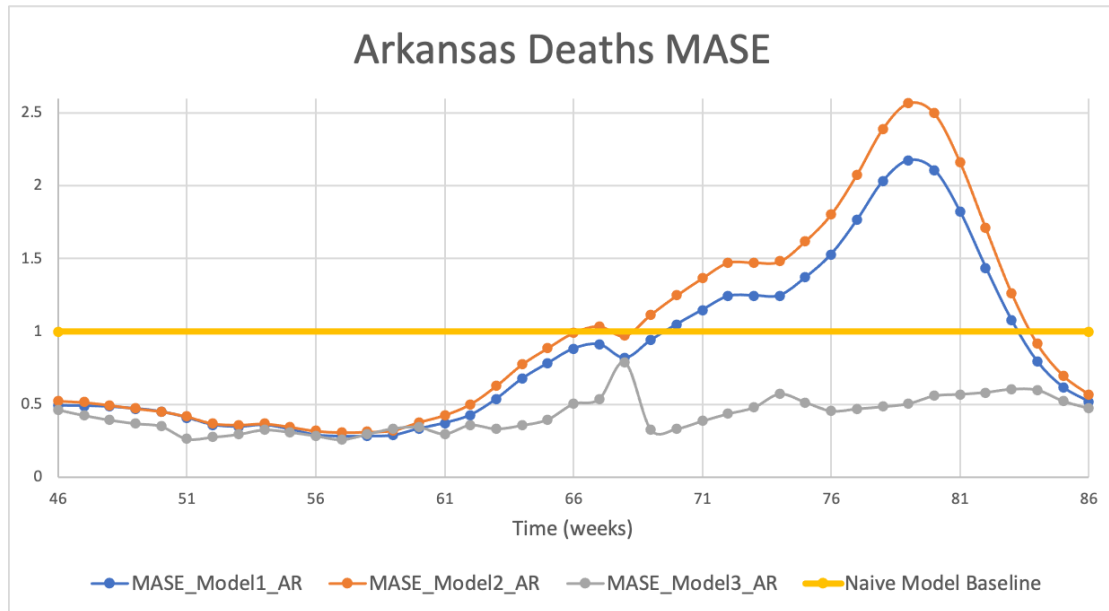


Figure 4.19. Arkansas Deaths MASE. The vertical axis represents MASE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other closely. Models 1 and 2 start to perform worse than the naïve model at approximately week 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. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

As seen in Figure 4.20, Arkansas' Model 3 for deaths performs the best in terms of the MAPE metric. Models 1 and 2 perform similarly throughout the testing period, performing worse than Model 3 consistently. For approximately a 12 week period, Models 1 and 2 perform worse than the naïve model. The naïve MAPE only starts to perform similarly to the other models at approximately week 65.

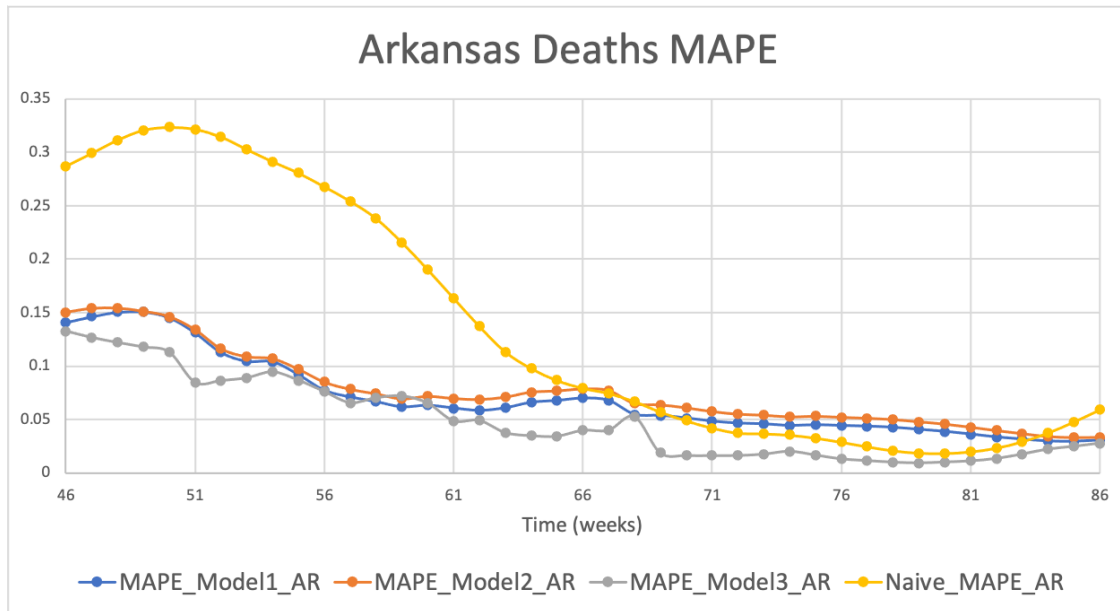


Figure 4.20. Arkansas Deaths MAPE. The vertical axis represents MAPE percentage, and the horizontal axis depicts the number of weeks since data collection began on January 22, 2020. Models 1 and 2 follow each other very closely. They perform worse than the naïve model from approximately week 68 to week 83. All models trend gradually downwards. Adapted from (United States Census Bureau 2019) and (USAFacts 2021a).

4.6 Summary

All states tested exhibit similar behavior in the measures of performance MAPE and MASE. All models decrease gradually in MAPE as the time series prediction passes through time. The naïve MAPE consistently starts very high but then gradually approaches 0 along with the other models. Models 1 and 2 perform worse than the naïve MAPE from approximately week 71 to week 84. Most of the states exhibited a bimodal structure in MASE from approximately week 66 to week 84. Approximately week 70 to week 77 showed a rapid increase in MASE across all states.

All states' models 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 thesis.

Because Model 3 outperformed the naïve model almost 100% of the time in all states, this model should be considered by epidemiologists. The GNAR package has proven potentially useful for predicting COVID-19 cases and deaths.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:

Conclusions and Future Work

5.1 Conclusions

The GNAR package has proven to be useful in predicting COVID-19 cases and deaths throughout the county network in the United States. Using the open-source data from the USCB and USAFacts, we can create a predictive model that could better inform public health officials. Future research in this area involving live traffic data, PCA, smaller geographical subdivisions, new network structures, and seasonality could contribute to the creation of an even more accurate model.

5.1.1 Data Analysis

To be able to predict the spread of COVID-19 is to save lives. The virus will continue to proliferate among the world's population, and predictive models can allow leaders to interdict its spread.

This thesis demonstrates that a network describing commuting patterns across United States counties can indeed be helpful in predicting COVID-19 cases and deaths. The ability to incorporate the flow of people from one county to another indeed reflects how the virus would spread from person to person, and thus allows us to feasibly include it in predictive models.

5.1.2 GNAR Model

Model 3, as discussed in Section 4.6, performs the best out of all the models, with its parameters `alphaOrder` and `betaOrder` = 0. This model consistently outperforms the naïve model, meaning it has predictive power. Model 3 should be considered for use by epidemiologists in predicting COVID-19 across the counties, informing public health officials as well.

5.2 Future Work

Although Model 3 in this thesis performed well, we can add more information to a GNAR in order to increase predictive power on spread of COVID-19.

5.2.1 Network Modifications

The underlying network structure is a critical part of a GNAR model's construction. Modifying a county commuting network shown in this thesis can be possibly improved the performance of a future GNAR model.

Live Traffic Data

Cell phone data is both nearly ubiquitous and surprisingly accurate (Johnson 2021). 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 (Johnson 2021). 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.

ZIP Code Analysis

The data sources of this thesis are delineated by county, which provides data for a more localized area. However, one could further subdivide this data into 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 United States (United States Postal Service 2021). 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” (Wren 2021).

Other Network Structures

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.

5.2.2 Principal Component Analysis

Unsupervised learning techniques may provide further insight by incorporating multivariate data from each county. PCA, for example, intakes a set of predictors $X_1, X_2, \dots, X_n$ instead of using a traditional response variable, Y (James et al. 2013). This technique reduces dimensionality but still describes multiple features of the data.

We can use PCA to generate the first principal component, which is a normalized linear combination of the selected predictors that contains the largest sample variance (James et al. 2013). Similarly, the second principal component is a linear combination of those predictors that has “the maximal variance out of all linear combinations that are uncorrelated with” the first principal component (James et al. 2013). Thus, these components are orthogonal. Therefore, with this way, we can create a new variable using a linear combination of multivariate data which shows the maximum variance in two dimensions.

In this way, we can intake by-county data on such subjects as public health, demographics, economics, and vaccinations and incorporate them into our prediction by reducing dimensionality in PCA. Potentially the principal components generated could act as a multiplicative weight on each county’s time series cases and deaths data.

5.2.3 New Geographic Areas

Applying this methodology to other geographic areas or governance divisions could prove useful around the world, not just the United States. Any country’s municipalities, provinces, or townships could represent nodes in a network similar to the United States county structure.

Comparing countries of a similar geographic, climatic, and demographic makeup to the United States may especially prove insightful. One can also compare and contrast the public health policy effects in different geographic areas.

5.2.4 GNAR Refinement

Finally, refining parameter choices in GNAR can result in a more accurate model. One can vary the additional GNAR parameters such as latency and neighbor weight and compare the performance of more model combinations (Leeming et al. 2020). In this thesis, for example, we specify only a one-day latency period for the `alphaOrder` parameter, but one could vary it for a longer duration such as seven or fourteen days. Experimenting with the `betaOrder` parameter, which describes the influence of county neighbors, could also potentially improve performance.

5.3 Summary

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. Among these efforts, the GNAR package has indeed proven its potential utility in informing the public about the future spread of the virus, helping to stop not only the virus but also the loss of life as a result.

List of References

- 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>.
- Adams K (2021) States ranked by percentage of population fully vaccinated: Nov. 3. Backer's Hospital Review. Accessed November 3, 2021, <https://www.beckershospitalreview.com/public-health/states-ranked-by-percentage-of-population-vaccinated-march-15.html>.
- Ban C (2014) S.D. to be renamed Oglala Lakota County. Shannon County, S.D.. Accessed October 29, 2021, <https://www.naco.org/articles/shannon-county-sd-be-renamed-oglala-lakota-county>.
- Barrett J, Jayaraman B, Patel D, Skolnik J (2008) A SAS-based solution to evaluate study design efficiency of phase I pediatric oncology trials via discrete event simulation. *Comput Methods Programs Biomed* 90(3):240–250.
- Broomfield County (2021) History of Broomfield. City and county of Broomfield. Accessed October 29, 2021, <https://www.broomfield.org/386/History-of-Broomfield>.
- Capitanelli A (2020) Modeling the spread of diseases. Towards Data Science. Accessed Oct 15, 2021, <https://towardsdatascience.com/modeling-the-spread-of-diseases-821fc728990f>.
- Center for Disease Control and Prevention (2020) Faq: COVID-19 data and surveillance. CDC. Accessed October 25, 2021, <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html>.
- Chartrand G, Zhang P (2012) *A First Course in Graph Theory* (Dover Publications, New York).
- Chen T, Rui J, Wang Q, Zhao Z, Cui J, Lin L (2020a) A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious diseases of poverty* 9(1):1–8.
- Chen Y, Lu P, Liu T (2020b) A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering* 7(4):3279–3294.
- Collins J, Abdelal N (2018) Spread of disease. Accessed October 15, 2021, <https://calculate.org.au/wp-content/uploads/sites/15/2018/10/spread-of-disease.pdf>.

- Connolly M, Heymann D (2002) Deadly comrades: war and infectious diseases. *The Lancet* 360:23–24.
- Cooper I, Mondal A, Antonopoulos C (2020) A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals* 139, <https://doi.org/10.1016/j.chaos.2020.110057>.
- Cryer J (2008) *Time Series Analysis: With Applications in R* (Springer, New York).
- Erdem K (2020) Finding correlations in time series data. Kemal Erdem. Accessed October 13, 2021, <https://erdem.pl/2020/06/finding-correlations-in-time-series-data>.
- Firth J, Hellewell J, Klepac P, Kissler S, Kucharski A, Spurgin L (2020) Using a real-world network to model localized covid-19 control strategies. *Nature Medicine* 26(10):1616–1622, ISSN 1078-8956, URL <http://dx.doi.org/10.1038/s41591-020-1036-8>.
- Frost J (2020) Guidelines for removing and handling outliers in data. Statistics by Jim. Accessed Oct 25, 2021, <https://statisticsbyjim.com/basics/remove-outliers/>.
- Gallagher J (2021) Covid immunity: Can you catch it twice? BBC News. Accessed Oct 18, 2021, <https://www.bbc.com/news/health-52446965>.
- Guo Z, Wong W, Li M (2012) Sparsely connected neural network-based time series forecasting. *Information Sciences* 193:54–71.
- 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>.
- Harvard University (2020) If you’ve been exposed to the coronavirus. Harvard University. Accessed Oct 26, 2021, <https://www.health.harvard.edu/diseases-and-conditions/if-youve-been-exposed-to-the-coronavirus>.
- Hoehl S, Rabenau H, Berger A, Kortenbusch M, Cinatl J, Bojkova D, Behrens P, Bödinghaus B, Götsch U, Naujoks F (2020) Evidence of SARS-CoV-2 infection in returning travelers from Wuhan, China. *New England Journal of Medicine* 382(13):1278–1280.
- 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. *Journal of Medical Virology* 92(7):841–848.
- Hyndman R (2006) Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting* 4(4):43–46.

- Hyndman R, Athanasopoulos G (2018) *Forecasting: principles and practice*. <https://otexts.com/fpp2/>.
- Institute for Disease Modeling (2021) SEIR and susceptible-exposed-infected-recovered-susceptible (SEIRS) models. Accessed Oct 18, 2021, <https://docs.idmod.org/projects/emod-hiv/en/latest/model-seir.html>.
- James G, Witten D, Hastie T, Tibshirani R (2013) *Introduction to Statistical Learning* (Springer, New York).
- Johnson G (2021) Cell Phone Data Makes Traffic Analysis and Transportation Planning Easier. Short Elliot Hendrickson Inc. Accessed November 9, 2021, <https://www.sehinc.com/news/cell-phone-data-makes-traffic-analysis-and-transportation-planning-easier>.
- Kermack A WO and McKendrick (1991) Contributions to the mathematical theory of epidemics—i. *Bulletin of Mathematical Biology* 53(1), <https://doi.org/10.1007/BF02464423>.
- Kerr CC, Stuart RM, Mistry D, Abeysuriya RG, Hart G, Rosenfeld K, Selvaraj P, Núñez RC, Hagedorn B, George L, Izzo A, Palmer A, Delport D, Bennette C, Wagner B, Chang S, Cohen JA, Panovska-Griffiths J, Jastrzębski M, Oron AP, Wenger E, Famulare M, Klein DJ (2020) Covasim: an agent-based model of covid-19 dynamics and interventions. *medRxiv* URL <http://dx.doi.org/10.1101/2020.05.10.20097469>.
- Knight M, Leeming K, Nason G, Nunes M (2020) Generalized network autoregressive processes and the gnar package. *Journal of Statistical Software* 96(5):1–36.
- Kotu V, Deshpande B (2014) *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner* (Elsevier Science and Technology, San Francisco).
- Leeming K, Nason G, Nunes M, Knight M (2020) Package ‘GNAR’. CRAN. Accessed December 20, 2020, <https://cran.r-project.org/web/packages/GNAR/GNAR.pdf>.
- Lewis C (1982) *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting* (Butterworth-Heinemann, Oxford, United Kingdom).
- Newman M (2010) *Networks: An Introduction* (Oxford University Press, London).
- Newman M, Barabasi A, Watts D (2006) *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)* (Princeton University Press, Princeton, New Jersey).
- Nguemdjo U, Meno F, Dongfack A, Ventelou B (2020) Simulating the progression of the COVID-19 disease in Cameroon using SIR models. *plos one*. Accessed October 18, 2021, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237832>.

- 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. Accessed Oct 14, 2021, 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.
- 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.
- Porta M (2014) *A Dictionary of Epidemiology* (Oxford University Press, New York).
- Ridenhour B, Kowalik J, Shay D (2014) Unraveling r_0 : considerations for public health applications. *American Journal of Public Health* 104(2), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3935673/>.
- Rodrigues H (2016) Application of sir epidemiological model: new trends. *International Journal of Applied Mathematics and Informatics* 10:92–97.
- Sanche S, Lin Y, 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. *Emerging Infectious Diseases* 26(7):1470–1477.
- Sanju'a R, Nebot M, Chirico N, Mansky L, Belshaw R (2010) Viral mutation rates. *Journal of Virology* 84(19):9733–9748.
- Saqr M, Alamro A (2019) The role of social network analysis as a learning analytics tool in online problem based learning. *BMC Medical Education* 19(160), <https://doi.org/10.1186/s12909-019-1599-6>.
- Silva P, Batista P, Lima H, Alves M, Guimarães F, Silva R (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>.
- 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>.
- So M, Tiwari A, Chu A, Tsang J, Chan J (2020) Visualizing covid-19 pandemic risk through network connectedness. *International Journal of Infectious Diseases* 96:558–561.

- Statistics Solutions (2021) Time series analysis. Statistics Solutions. Accessed October 25, 2021, <https://www.statisticssolutions.com/time-series-analysis/>.
- 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. *2020 IEEE International Smart Cities Conference (ISC2)*, 1–6 (Piscataway, New Jersey), <https://doi.org/10.1109/ISC251055.2020.9239072>.
- Svetunkov I (2021) Rolling origin. CRAN. Accessed November 2, 2021, <https://cran.r-project.org/web/packages/greybox/vignettes/ro.html>.
- Tandon H, Ranjan P, Chakraborty T, Suhag V (2020) Coronavirus (covid-19): Arima based time-series analysis to forecast near future. Research Gate. Accessed Oct 14, 2021, https://www.researchgate.net/publication/340776075_Coronavirus_COVID-19_ARIMA_based_time-series_analysis_to_forecast_near_future.
- Teles P (2020) A time-dependent SEIR model to analyse the evolution of the SARS-CoV-2 epidemic outbreak in Portugal. Research Gate. Accessed Oct 18, 2021, <https://arxiv.org/abs/2004.04735>.
- The Fact File (2021) List of U.S. states and number of counties in each. Accessed October 27, 2021, <https://thefactfile.org/us-states-counties/>.
- United States Census Bureau (2019) County transportation profiles | bureau of transportation statistics. Accessed July 30, 2021, <https://www.census.gov/topics/employment/commuting/guidance/flows.html>.
- United States Census Bureau (2020a) 2019 FIPS codes. Accessed October 27, 2021, <https://www.census.gov/geographies/reference-files/2019/demo/popest/2019-fips.html>.
- United States Census Bureau (2020b) 2020 census data. Accessed October 27, 2021, <https://www.census.gov/programs-surveys/popest/technical-documentation/file-layouts.html>.
- United States Postal Service (2021) 42,000 ZIP codes. U.S. Postal Facts. Accessed November 9, 2021, <https://facts.usps.com/42000-zip-codes/>.
- USAFacts (2020) About USAFacts. Harvard University. Accessed Oct 26, 2021, <https://usafacts.org/about-usafacts/>.
- USAFacts (2021a) US COVID-19 cases and deaths by state. Accessed July 31, 2021, <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>.
- USAFacts (2021b) Detailed methodology and sources: COVID-19 data. Accessed July 31, 2021, <https://usafacts.org/articles/detailed-methodology-covid-19-data/>.

Wren D (2021) *SARS-COV-2 Dissemination Using United States County Adjacencies*. Master's thesis, Naval Postgraduate School, Monterey, CA, <http://hdl.handle.net/10945/67833>.

Yoshida R (2021) Time series analysis. Lecture, Advanced Data Analysis, Feb 21, Department of Operations Research, Naval Postgraduate School, Monterey, CA.

Initial Distribution List

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California