



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

SARS-COV-2 DISSEMINATION USING UNITED STATES COUNTY ADJACENCIES

by

David M. Wren

June 2021

Thesis Advisor:

Co-Advisor:

Second Reader:

Ruriko Yoshida
Chrysafis Vogiatzis,
University of Illinois
Ralucca Gera

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 June 2021	3. REPORT TYPE AND DATES COVERED Master's thesis		
4. TITLE AND SUBTITLE SARS-COV-2 DISSEMINATION USING UNITED STATES COUNTY ADJACENCIES			5. FUNDING NUMBERS	
6. AUTHOR(S) David M. Wren				
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) Currently, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission is increasing amongst the world's population at an alarming rate. Reducing the spread of SARS-CoV-2 is paramount for public health officials as they seek to effectively manage resources and potential population control measures such as social distancing and quarantine. By analyzing the United States' county network structure, one can model and interdict potential higher infection areas. County officials can provide targeted information, preparedness training, and increased testing in these areas. While these approaches may provide adequate countermeasures for localized areas, they are inadequate for the holistic United States. We solve this problem by collecting data on coronavirus-19 (COVID-19) infections and deaths from the Center for Disease Control and Prevention and a network adjacency structure from the United States Census Bureau. Generalized network autoregressive (GNAR) time series models have been proposed as an efficient learning algorithm for networked datasets. This thesis fuses network science and operations research techniques to univariately model COVID-19 cases, deaths, and current survivors across the United States' county network structure.				
14. SUBJECT TERMS United States, COVID-19, network science, time series analysis, forecasting, GNAR			15. NUMBER OF PAGES 167	
			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 ADJACENCIES**

David M. Wren
Major, United States Army
Bachelor's Degree, Baylor University, 2009
MBA, Syracuse University, 2016

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
June 2021**

Approved by: Ruriko Yoshida
Advisor

Chrysafis Vogiatzis
Co-Advisor

Ralucca Gera
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Currently, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission is increasing amongst the world's population at an alarming rate. Reducing the spread of SARS-CoV-2 is paramount for public health officials as they seek to effectively manage resources and potential population control measures such as social distancing and quarantine. By analyzing the United States' county network structure, one can model and interdict potential higher infection areas. County officials can provide targeted information, preparedness training, and increased testing in these areas. While these approaches may provide adequate countermeasures for localized areas, they are inadequate for the holistic United States. We solve this problem by collecting data on coronavirus-19 (COVID-19) infections and deaths from the Center for Disease Control and Prevention and a network adjacency structure from the United States Census Bureau. Generalized network autoregressive (GNAR) time series models have been proposed as an efficient learning algorithm for networked datasets. This thesis fuses network science and operations research techniques to univariately model COVID-19 cases, deaths, and current survivors across the United States' county network structure.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
1.1	Background	1
1.2	COVID-19 Modeling Approaches.	2
1.3	United States County Network	2
1.4	Data Availability and Model Adaptation	2
1.5	Results and Future Research	4
2	Literature Review	7
2.1	Network Models	9
2.2	Time Series Forecasting	13
2.3	Epidemiological Models	18
2.4	Simulations	25
2.5	Summary	28
3	Methodology	31
3.1	Data	31
3.2	Model Implementation	51
3.3	Summary	55
4	Model Results	57
4.1	Introduction	57
4.2	Daily Results	57
4.3	Weekly Results	88
4.4	Summary	120
5	Conclusions and Summary	123
5.1	Conclusions	123
5.2	Department of Defense Impact	124

5.3	Future Work	124
5.4	Summary	130
	List of References	133
	Initial Distribution List	145

List of Figures

Figure 1.1	United States COVID-19 Cases and Survivors	3
Figure 1.2	United States COVID-19 Deaths	4
Figure 3.1	United States County Degree Histogram	37
Figure 3.2	Louvain Community Detection	40
Figure 3.3	Leiden Community Detection	41
Figure 3.4	County Centralities	45
Figure 3.5	County Eccentricities	46
Figure 3.6	Power Law Distribution	47
Figure 3.7	County Clustering Coefficients	48
Figure 3.8	Edge Mixing	50
Figure 3.9	Continuous Rolling Horizon	53
Figure 3.10	Fixed Interval Rolling Horizon	53
Figure 4.1	Daily Time Interval - Binary Network - Cases Boxplot (MASE) .	60
Figure 4.2	Daily Time Interval - Binary Network - Cases - Time Series (MASE)	61
Figure 4.3	Daily Time Interval - Binary Network - Cases Boxplot (MAPE) .	62
Figure 4.4	Daily Time Interval - Binary Network - Cases Time Series (MAPE)	63
Figure 4.5	Daily Time Interval - Great Circle Network - Cases Boxplot (MASE)	65
Figure 4.6	Daily Time Interval - Great Circle Network - Cases - Time Series (MASE)	66
Figure 4.7	Daily Time Interval - Great Circle Network - Cases Boxplot (MAPE)	67

Figure 4.8	Daily Time Interval - Great Circle Network - Cases Time Series (MAPE)	68
Figure 4.9	Daily Time Interval - Binary Network - Deaths Boxplot (MASE)	70
Figure 4.10	Daily Time Interval - Binary Network - Deaths - Time Series (MASE)	71
Figure 4.11	Daily Time Interval - Binary Network - Deaths Boxplot (MAPE)	72
Figure 4.12	Daily Time Interval - Binary Network - Deaths Time Series (MAPE)	73
Figure 4.13	Daily Time Interval - Great Circle Network - Deaths Boxplot (MASE)	75
Figure 4.14	Daily Time Interval - Great Circle Network - Deaths - Time Series (MASE)	76
Figure 4.15	Daily Time Interval - Great Circle Network - Deaths Boxplot (MAPE)	77
Figure 4.16	Daily Time Interval - Great Circle Network - Deaths Time Series (MAPE)	78
Figure 4.17	Daily Time Interval - Binary Network - Survivors Boxplot (MASE)	80
Figure 4.18	Daily Time Interval - Binary Network - Survivors - Time Series (MASE)	81
Figure 4.19	Daily Time Interval - Binary Network - Survivors Boxplot (MAPE)	82
Figure 4.20	Daily Time Interval - Binary Network - Survivors Time Series (MAPE)	83
Figure 4.21	Daily Time Interval - Great Circle Network - Survivors Boxplot (MASE)	85
Figure 4.22	Daily Time Interval - Great Circle Network - Survivors - Time Series (MASE)	86
Figure 4.23	Daily Time Interval - Great Circle Network - Survivors Boxplot (MAPE)	87
Figure 4.24	Daily Time Interval - Great Circle Network - Survivors Time Series (MAPE)	88

Figure 4.25	Weekly Time Interval - Binary Network - Cases Boxplot (MASE)	91
Figure 4.26	Weekly Time Interval - Binary Network - Cases - Time Series (MASE)	92
Figure 4.27	Weekly Time Interval - Binary Network - Cases Boxplot (MAPE)	93
Figure 4.28	Weekly Time Interval - Binary Network - Cases Time Series (MAPE)	94
Figure 4.29	Weekly Time Interval - Great Circle Network - Cases Boxplot (MASE)	96
Figure 4.30	Weekly Time Interval - Great Circle Network - Cases Time Series (MASE)	97
Figure 4.31	Weekly Time Interval - Great Circle Network - Cases Boxplot (MAPE)	98
Figure 4.32	Weekly Time Interval - Great Circle Network - Cases Time Series (MAPE)	99
Figure 4.33	Weekly Time Interval - Binary Network - Deaths Boxplot (MASE)	101
Figure 4.34	Weekly Time Interval - Binary Network - Deaths Time Series (MASE)	102
Figure 4.35	Weekly Time Interval - Binary Network - Deaths Boxplot (MAPE)	103
Figure 4.36	Weekly Time Interval - Binary Network - Deaths Time Series (MAPE)	104
Figure 4.37	Weekly Time Interval - Great Circle Network - Deaths Boxplot (MASE)	106
Figure 4.38	Weekly Time Interval - Great Circle Network - Deaths Time Series (MASE)	107
Figure 4.39	Weekly Time Interval - Great Circle Network - Deaths Boxplot (MAPE)	108
Figure 4.40	Weekly Time Interval - Great Circle Network - Deaths Time Series (MAPE)	109
Figure 4.41	Weekly Time Interval - Binary Network - Survivors (MASE) . .	111

Figure 4.42	Weekly Time Interval - Binary Network - Survivors Time Series (MASE)	112
Figure 4.43	Weekly Time Interval - Binary Network - Survivors Boxplot (MAPE)	113
Figure 4.44	Weekly Time Interval - Binary Network - Survivors Time Series (MAPE)	114
Figure 4.45	Weekly Time Interval - Great Circle Network - Survivors Boxplot (MASE)	116
Figure 4.46	Weekly Time Interval - Great Circle Network - Survivors Time Series (MASE)	117
Figure 4.47	Weekly Time Interval - Great Circle Network - Survivors Boxplot (MAPE)	118
Figure 4.48	Weekly Time Interval - Great Circle Network - Survivors Time Series (MAPE)	119
Figure 4.49	Great Circle Performance Bar Plot	121

List of Tables

Table 2.1	MAPE Interpretation	17
Table 3.1	United States Counties	35
Table 3.2	Network Comparison	38
Table 3.3	Network k -cores	43
Table 3.4	Model Parameter Summary	55
Table 4.1	Daily Time Interval - Binary Network - Cases Model Summary . .	59
Table 4.2	Daily Time Interval - Great Circle Network - Cases Model Summary	64
Table 4.3	Daily Time Interval - Binary Network - Deaths Model Summary .	69
Table 4.4	Daily Time Interval - Great Circle Network - Deaths Model Summary	74
Table 4.5	Daily Time Interval - Binary Network - Survivors Model Summary	79
Table 4.6	Daily Time Interval - Great Circle Network - Survivors Model Sum- mary	84
Table 4.7	Weekly Time Interval - Binary Network - Cases Model Summary .	90
Table 4.8	Weekly Time Interval - Great Circle Distance Network - Cases Model Summary	95
Table 4.9	Weekly Time Interval - Binary Network - Deaths Model Summary	100
Table 4.10	Weekly Time Interval - Great Circle Distance Network - Deaths Model Summary	105
Table 4.11	Weekly Time Interval - Binary Network - Survivors Model Summary	110
Table 4.12	Weekly Time Interval - Great Circle Distance Network - Survivors Model Summary	115

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ABM	agent-based model
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
DHS	Department of Homeland Security
FIPS	federal information processing standard
GNAR	generalized network autoregression
IID	independent and identically distributed
MAPE	mean absolute percentage error
MASE	mean absolute scaled error
MERS	Middle East respiratory syndrome
ML	machine learning
MSEIR	metapopulation susceptible-exposed-infected-recovered
NPS	Naval Postgraduate School
OSM	Open Street Map
OSMnx	Open Street Map Networkx
PCA	principal component analysis

RMSE	root mean squared error
SARS	severe acute respiratory syndrome
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SEIR	susceptible-exposed-infected-recovered
SEIRS	susceptible-exposed-infected-recovered-susceptible
SIIR	susceptible-infected-infected-recovered
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
USNS	United States Naval Ship
USS	United States Ship
WHO	World Health Organization
XML	extensible markup language
ZIP	zone improvement plan

Executive Summary

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has shaken the very fabric of modern society. Coronavirus disease 2019 (COVID-19) has claimed the lives of millions of individuals as families struggle to provide basic necessities as economies are decimated. While scientists, researchers, and epidemiologists struggle to develop adaptive models to understand and forecast SARS-CoV-2, public health officials have grappled with uncertainty as the virus spreads amongst the population.

Due to the ubiquitous use of machine learning and openly available information, researchers are able to collaborate and develop effective models to disrupt and combat the spread of SARS-CoV-2. By collecting and analyzing COVID-19 statistics within a given area, one is able to develop a deeper, holistic understanding of the civil domain. Compartment models such as the susceptible-infected-recovered (SIR), susceptible-infected-infected-recovered (SIIR), and susceptible-exposed-infected-recovered (SEIR) provide an epistemological method while techniques such as agent-based simulations are able to model individual movements.

This thesis utilizes information from the Center for Disease Control and Prevention (CDC), United States Census Bureau (USCB), National Bureau of Economic Research, and United States of America Facts (USAFacts) to develop and model a network containing univariate data being either COVID-19 infections, deaths, or survivors. Due to the time-dependent nature of the data, this thesis utilizes two primary measures of performance: mean absolute scaled error (mean absolute scaled error (MASE)) and mean absolute percentage error (mean absolute percentage error (MAPE)). This thesis demonstrates that the data permutations and generalized network autoregression (GNAR) parameters are able to effectively model the spread of SARS-CoV-2.

Despite these promising results, data analysts must work closely with public health officials to ensure a model remains useful and relevant. Data analysts should defend the usefulness and application of their models but insist on integration with a human acumen. Cumulatively, this model provides local, state, and federal public health officials situational awareness within the United States.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

First and foremost I would like to thank my wife. She has been with me through two career changes, two master's degrees, and a deployment. She truly is my rock and I can not state how much she has been there for me and supported me. Every day is an adventure for us and this thesis concludes another chapter of our lives.

Next I would like to thank Dr. Yoshida. She, like my wife, helped me out tremendously through this process and without her this process would have been a lot more difficult.

I would also like to thank Dr. Vogiatzis and Dr. Gera for their expertise, mentorship, and guidance through this process. Thank you very much for helping me achieve this goal.

Lastly, I would like to thank the professors at the Naval Postgraduate School (NPS) for the opportunity to learn and grow during my tenure.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:

Introduction

1.1 Background

The current severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic and its subsequent coronavirus disease 2019 (COVID-19) infection is spreading across the planet at a frightening rate causing debilitating damage to the fabric of our very existence. Economies have been crippled. Health care systems and resources have been strained. Food security has been tested. Families have been torn apart as loved ones are lost (World Health Organization 2020).

However, epidemics such as as SARS-CoV-2 are not a recent phenomenon for humanity. Throughout history, diseases, viruses, and pathogens have resulted in countless deaths and instilled fear in untold others. Shamans, healers, doctors, scientists, public health officials, and researchers alike have sought to understand an outbreak to disrupt, bring under control, and ultimately eradicate the deadly plague of their time. Through their research and discipline, the modern practice of epidemiology was born.

Epidemiology is defined as “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). Epidemiology serves as a critical component for doctors and health officials to shape public health policy decisions by designing, collecting, and analyzing historical, contemporary, and emerging threats. Epidemiologists utilize a variety of tools and techniques to aid in their analytical techniques.

In this thesis, we utilize the United States county network to forecast COVID-19 infections, deaths, and survivors within each county.

1.2 COVID-19 Modeling Approaches

Due to the exponential growth of big data acquisition and the advancement of machine learning techniques in recent years, epidemiologists and data scientists are able to synchronise their efforts to create a predictive forecasting model with relative ease. Network science has certainly benefited in recent years from advancements in computational capacity and capability as researchers continue to expand the scientific study of networks and their applications. In particular, social and telecommunications networks provide researchers an opportunity to analyze an individual's connections and key influencers, respectively.

As epidemics naturally contain data that is directly related to a specified time period, time series models such as Holt-Winters or seasonal decomposition provide an additional model opportunity. The susceptible-infected-recovered (SIR) compartmental model and its variants provide researchers with an opportunity to efficiently classify individuals into categories as they receive and recover from a given pathogen. Agent-based simulations provide another alternative by modeling individual movements and a pathogen's transmission within a community.

1.3 United States County Network

As of the 2020 census, the United States is divided into 3,143 county and county equivalents each with their own adjacency structure (United States Census Bureau 2020a). As such, this thesis applies network science techniques to determine underlying characteristics and properties, which can be used to develop a holistic understanding of the country. For example, a network's centrality measures and clustering coefficient enable the network to be disintegrated and analyzed further. To better represent the United States' county network structure, adjacencies can be represented as either binary or the great circle distance between county centroids.

1.4 Data Availability and Model Adaptation

All data in this thesis are openly available from organizations such as United States of America Facts (USAFacts), the United States Census Bureau (USCB), the Center for Disease Control and Prevention (CDC), and the National Bureau of Economic Research. Figures 1.1 and 1.2 depict the total number of COVID-19 cases and survivors as well as deaths,

respectively. Although individuals have certainly contracted COVID-19 but haven't taken a diagnostic test and since recovered, this thesis defines the term "survivor" as those individuals who have tested positive for COVID-19 and recovered. The vertical axis in Figures 1.1 and 1.2 represents the total cumulative count. The horizontal axis in Figures 1.1 and 1.2 is the given time series for which data was collected. By inspection, the number of survivors nearly mirrors the number of cases in Figure 1.1. Since both data sets are cumulative counts, they continue to increase.

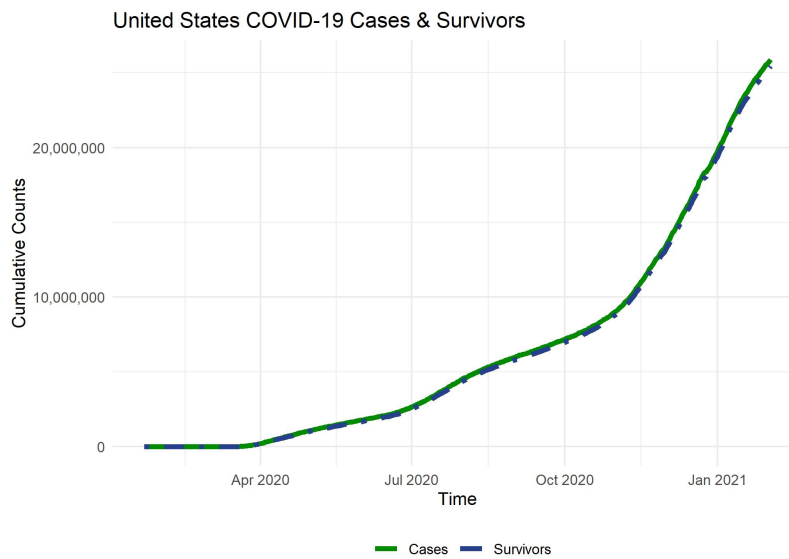


Figure 1.1. United States COVID-19 Cases and Survivors. The vertical axis represents the cumulative total survivors and the horizontal axis represents time. As we can see from the figure, COVID-19 cases and survivors climb at an exponential rate. Adapted from USAFacts (2021a) and USAFacts (2021b).

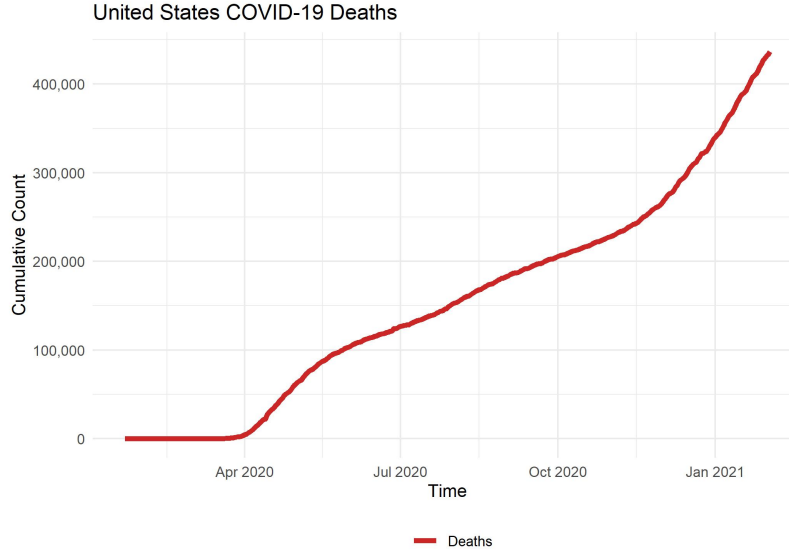


Figure 1.2. United States COVID-19 Deaths. The vertical axis represents the cumulative total deaths and the horizontal axis represents time. As we can see from the figure, COVID-19 deaths climb at an exponential rate. Adapted from USAFacts (2021b).

To model these data, this thesis employs the generalized network autoregression (GNAR) package in R (Leeming et al. 2020). Researchers created the GNAR package to accept univariate time series data across a given network structure to analyze, model, and develop a daily, weekly, monthly, quarterly, or yearly prediction (Leeming et al. 2020). In this thesis, we utilize COVID-19 cases, survivors, and deaths for each county as the univariate data while the network structure is the United States county adjacency network. We use daily and weekly time intervals, binary and great circle distance network adjacency structures, and COVID-19 cases, deaths, and survivor univariate data sets, resulting in 12 total combinations. Within each combination, we specify three individual GNAR models, which adjust the influence a given county experiences by its neighbors.

1.5 Results and Future Research

Since daily COVID-19 infections are inherently time-based data, we utilize mean absolute percentage error (MAPE) and mean absolute scaled error (MASE) as our primary performance measures to demonstrate that GNAR is able to effectively predict cases, survivors,

and deaths. We can employ a Naïve model by setting the forecast for any time period equal to the previous period's actual value to establish a baseline performance measure. Forecasting models that beat the Naïve model are said to have predictive power (Erdem 2021). Each weekly model outperforms the Naïve model and by extension, demonstrates their employment capability. Additionally, the results in Chapter 4 demonstrate two recurrent trends: the MASE and MAPE performance for the case and survivor combinations are nearly identical and the great circle distance often hinders accuracy. Despite these findings, the results illustrate that the GNAR package is able to accurately forecast COVID-19 cases, deaths, and survivors. However, adapting the GNAR package and collecting additional data could further increase each combination's accuracy.

Finally, this thesis discusses potential avenues for additional research topics. For example, we could disintegrate COVID-19 infections to individual zone improvement plan (ZIP) codes to build a network adjacency structure at a smaller level. Each compartment within the SIR model could be used as a univariate data sequence. Despite GNAR's success, this thesis proposes expanding the current univariate data limitation by modifying or creating an entirely new package that enables multivariate time series modeling. Multivariate predictors such as a county's population, demographics, and economic conditions could be used to develop a deeper understanding, create a more effective model, and ultimately develop a more accurate epidemic model to aid public policy officials.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:

Literature Review

SARS-CoV-2 is a positive-sense single-stranded ribonucleic acid virus that is contagious in humans (Machhi et al. 2020). The virus primarily spreads through respiratory droplets from speaking, coughing, or sneezing (Sanche et al. 2020). While the transmission rate is not fully known, it is estimated that each infection results in an additional 5.7 cases (Sanche et al. 2020). The first known cases of COVID-19 were reported in early December 2019 in Wuhan, China. Since then, the virus has spread and crossed international boundaries with relative ease and indiscriminately infects and kills individuals, decimates communities, and disrupts economic activity. As such, new cases can expand at an exponential rate within a community that has no prior exposure or limited medical capacity and capability. Moreover, virus mutations are very common and as such, it becomes increasingly difficult to combat with anti-viral treatments and increased public health measures (Sanju'a et al. 2010). While not an exhaustive or definitive list, the CDC identified multiple SARS-CoV-2 variants with several notable variants (Center for Disease Control and Prevention 2020).

Infectious viruses such as Variola, Influenza, SARS-CoV, and human immunodeficiency virus have often had a crippling effect throughout the history of civilization. Economic, social, and physical mobility enable individuals to seek and form larger communities, which creates the conditions for viruses to propagate and spread with consequential and often devastating results. Throughout history, these dense communities and a lack of education, awareness, and understanding enabled these viruses to form outbreaks and, as a result, proliferate unabated. The modern study of epidemiology traces its roots thousands of years as experts attempt to preserve life. While uneducated, untrained, and often unprepared in retrospect, physicians, scientists, and biologists studied and applied multiple techniques in an attempt to slow and ultimately halt a virus's or disease's progression.

A network forms as individuals, or nodes, create bonds with one another, or edges, within and external to their respective communities. In particular, "networks have attracted considerable recent attention in physics and other fields as a foundation for the mathematical representation of a variety of complex systems, including but not limited to biological and

social systems, the Internet, telecommunications, and many others” (Newman 2006). As a result, analysts and epidemiologists can apply graph theory and network science to potentially determine vital information such as a virus’s transmission rate and incubation period. Network science may also be applied through a diligent and coordinated effort to disrupt a virus’s diffusion by calculating a node’s centrality measures and notifying adjacent individuals of possible exposure and that they should self-isolate and minimize their exposure to others. Additionally, a network’s shortest path can also be used to find a path of minimum cost (or length) between two nodes (Ahuja et al. 1993). Analysts can apply the concept of a shortest path between two nodes to determine how quickly a virus may spread within a given network. Public health officials can then interdict and potentially disrupt higher infection areas by providing targeted information, preparedness training, isolation measures such as quarantining, and increase testing.

By their very nature, epidemics contain time series data. Time series data consists of a series observations for which measurements are obtained at discrete points in time and, as a result, these data will contain correlated observations in most cases (James et al. 2013). Consequently, time series data are no longer independent and identically distributed (IID). Therefore, applying traditional, supervised learning models such as simple linear regression or regression trees often under perform. However, time series forecasts such as the Holt-Winters or seasonal decomposition provide researchers alternative models. Epidemiologists can apply time series analysis to extract essential statistics and then apply time series forecasts to predict and further develop meaningful insights about a virus or pathogen.

The modern practice of epidemiology fuses studies such as mathematics, statistics, and sociology to identify, analyze, and dissect risk factors and conditions of diseases and viruses. Epidemiologists utilize models such as the compartmental Susceptible-Infected-Recovered (SIR) model to analyze an infectious disease’s progression through a population to predict future growth patterns, duration, casualties, and the subsequent impact within a community. In recent years, the ubiquitous use of computing and machine learning (ML) enables researches to quickly scale and process previously inconceivable amounts of data to improve the understanding of a populace’s health and to provide time-sensitive recommendations to public health officials for intervention (Wiemken and Kelley 2020).

Although the methodologies and models previously discussed provide adequate solutions

in an ever-changing environment, they are often deterministic and inflexible. Viruses are inherently stochastic in their nature as they evolve and adapt to their surroundings. As a result, models that incorporate randomness may provide a more accurate representation of a virus or disease during an epidemic.

While computational prowess has increased exponentially over recent years due to the number of transistors in a densely integrated circuit doubling approximately every two years, the utility and notoriety of ML models over simulations have equally benefited (Moore et al. 1965). Machine learning technology powers many aspects of modern day life and its ubiquitous nature will continue to prevail (LeCun et al. 2015). As analysts collect, collate, and analyze data, simulations afford a computationally efficient technique to implement and model a stochastic component. Moreover, embedding randomness within a simulation creates useful and simple models (Sanchez and Sanchez TO APPEAR). Although certainly beneficial, researchers do not necessarily need to know the exact parameters and characteristics of the data nor do they need to precisely know the mechanics and interactions between them. Determining these exact specifications, especially at the beginning of an epidemic, can be costly or simply infeasible. Instead, researchers and analysts can run a simulation hundreds of thousands of times with various model parameters and aggregate observations to determine the data's underlying structure.

Since the first confirmed diagnosis of COVID-19 in December 2019, the virus has infiltrated all corners of the globe with nearly every country reporting cases and deaths (World Health Organization 2021). Consequently, multiple concurrent research efforts to analyze, model, and predict SARS-CoV-2 and COVID-19 are occurring at unprecedented rates. The following models provide an overview of various adaptations.

2.1 Network Models

Although established relatively recently, network science has expanded significantly as the field draws interest from physicists, social scientists, statisticians, mathematicians, biologists, and computer scientists. Networks such as telecommunications can be depicted using nodes or vertices and the connections between them as edges. As a result, networks develop mathematical properties that can, in turn, be analyzed. These characteristics can have a significant impact on the system's underlying behavior and can be used to understand how,

for example, internet traffic may flow or how a disease spreads throughout a community (Newman 2010).

Social networks are created as individuals form relationships and consequently communities, which can be analyzed through the use of network science and graph theory. Individuals may create edges or connections through familial affiliation, professional relationships, friendship, dating, or others (Newman 2010). Metrics such as closeness centrality, eigenvector centrality, and betweenness centrality offer insight into a node's significance within the network. Closeness centrality of a given vertex, C_i , is defined by the following equation:

$$C_i = \frac{1}{\sum_{j=1}^N d(i, j)}, \quad (2.1)$$

where $d(i, j)$ is the distance between vertices i and j (Newman 2010). Betweenness centrality of a given vertex, x_i , is defined by the following equation:

$$X_i = \sum \frac{n_{st}^i}{g_{st}}. \quad (2.2)$$

where n_{st}^i is the number of $s - t$ shortest paths that i belongs to and g_{st} is the number of shortest paths (Newman 2010). Eigenvector centrality can be used to rank vertices and their importance relative to other vertices. To compute the eigenvector centrality, let $A = (a_{i,j})$ represent a graph's adjacency matrix. Then the eigenvector centrality for a given vertex x_i of node i is defined as:

$$x_i = \frac{1}{\lambda} \sum_k a_{k,i} x_k \quad (2.3)$$

where the eigenvalue $\lambda \neq 0$ is a constant (Newman 2010). These centrality measures also enable researchers to determine the relative importance of vertices.

Clustering and metrics such as k -core, k -clique, and k -plexes also provide other opportunities to understand a network's structure as they enable researchers to understand various cliques and communities that form within the network. The k -core is “a maximal connected

subgraph, where the elements of the subgraph are connected to at least k other elements of the same subgraph” (Csermely et al. 2013). A “ k -clique is a maximal connected subgraph having a diameter $\leq k$, where the diameter is the maximal number of links amongst the shortest path, which connect any two elements of the subgraph” (Csermely et al. 2013). A k -plex is “a maximal connected subgraph, where each of the n elements of the subgraph is linked to at least $n - k$ other elements in the same subgraph” (Csermely et al. 2013). By analyzing a network’s clustering metrics, researchers can identify and analyze tightly connected communities.

Networks, by design, provide an opportunity to analyze the most efficient, or optimal, method to facilitate the transportation of goods. For example, network flow provides an opportunity to determine the maximum amount of goods that can traverse a network (Ahuja et al. 1993). Researchers may also determine the shortest path within a given network, which can be used to determine network efficiency across a given network when delivering content such as news or a deadly pathogen.

Information, and by extension, viruses can be spread through a network through conserved spread and non-conserved spread (Newman et al. 2006). In conserved spread, the amount of content that enters the network is constant and unchanging (Newman et al. 2006). Conversely, content changes as it flows through a network in a non-conserved spread (Newman et al. 2006). Epidemiologists and researchers often apply a non-conserved spread to model a virus’s propagation throughout a network since individuals become infected and spread the disease to susceptible persons. For example, the SIR model utilizes a non-conserved approach as a virus is propagated by individuals throughout a given network.

Generalized Network Autoregression

The Generalized Network Autoregression (GNAR) model and subsequent package developed as a univariate time series model that relies upon a network’s neighbors to forecast future observations (Knight et al. 2020). The authors used and modeled gross domestic product data for 35 countries over a one year period. The authors employed not only a network model, they also utilized an autoregressive and vector autoregressive models. Autoregressive models specify the output variable should linearly depend on previous values while vector autoregressive models account for lagged feedback effects between variables (Knight et al. 2020). The GNAR model performed better than the autoregressive and vector

autoregressive models by 29 and 78 percent, respectively. As a result, the GNAR package in R demonstrated its capability to model a network's time series autoregressive structure (Leeming et al. 2020).

Network Connectedness

A given network is connected if there exists a path between node u and v (Chartrand and Zhang 2012). It follows then that if a path does not exist from u to v , then the two are disconnected. Although these definitions are generally accepted within the network science community, researchers utilized a network's degree to represent its connectedness when analyzing COVID-19 cases and the associated correlation between countries (So et al. 2020).

The authors constructed separate network graphs of both China and the rest of the world and if the correlation in the change of confirmed cases between geographical areas is greater than 0.5, an edge is present. If an edge is present between countries, the countries are then considered connected. A country's connectedness is represented by the number of connections a country has (its degree). Their results demonstrated that by analyzing a country's connectedness (degree), a more timely and accurate prediction of pandemic risk can be generated than relying upon the data itself.

Social Network Analysis

As social networks involve the carriers of SARS-CoV-2, namely persons, social network analysis (SNA) may be applied to understand and model the virus's dissemination within a given community. Researchers applied SNA and conducted contact tracing on a real social network to minimize exposure (Firth et al. 2020). As their analysis showed, the researchers were able to minimize the spread of COVID-19. However, as exposures progressed and individuals increasingly went into quarantine, nearly half of the network was in isolation at a single point in time. Once individuals were released from quarantine, outbreaks increased in magnitude. Consequently, the authors argue that contact tracing and self-isolation upon notification of potential exposure may be the most effective strategy to mitigate further exposure and disrupt virus dissemination.

Telecommunications Network

Network science can be applied to key public figures and their ability to exert significant influence over their social media platform. In-degree centrality is the number of vertices of a node from which it is adjacent to (Chartrand and Zhang 2012). Researchers analyzed the in-degree centrality of social media accounts for former Presidents Obama and Trump as well as news agencies and health organizations around the world (Yum 2020). The authors identified that these key political figures and news agencies have the ability to exert significant influence within telecommunications networks (Yum 2020). As a result, it is imperative that health organizations, news agencies, and political figures convey a similar message to ensure maximum information saturation amongst a populace.

2.2 Time Series Forecasting

COVID-19 infections are reported daily and are serially correlated to the previous day's infection rates, which creates a sequence of discrete-time data. Therefore, time series data are no longer independent and identically distributed, a key component in statistical modeling. Moreover, this temporal component is distinct from cross-sectional studies where there is no natural ordering of the data, which makes generating an accurate predictive model more difficult (Cryer 2008). By understanding and analyzing time series data, an analyst's purpose is generally two-fold: to extract meaningful statistics and other characteristics of the data, and to subsequently develop an accurate predictive forecast (Cryer 2008). The following definitions are used to describe the characteristics within time series data and defined by (Yoshida 2020) and (Hyndman and Athanasopoulos 2018).

- Trend: "Long-term (not necessarily linear) increases or decreases in the data; the long-term component of change" (Yoshida 2020).
- Seasonal (Periodic) Patterns: "Increases and decreases in data with a fixed/known period" (Yoshida 2020).
- Cycle: "Data exhibits rises and falls that are not of fixed period" (Yoshida 2020).
- Noise: "Remaining variance in the data after we've accounted for the components above (Yoshida 2020).
- Seasonal decomposition: Enables isolation of the trend-cycle (Hyndman and Athanasopoulos 2018).
- Homoscedastic: "Observed variance remains the same over time" (Yoshida 2020).

- Heteroscedastic: “Observed variance changes over time” (Yoshida 2020).
- 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 2020).

Analyzing characteristics such as trend, seasonality, and cycle can describe the data’s underlying disposition. Time series analysis can be done primarily through two methods: time domain and frequency domain (Cryer 2008). Time domain analysis is focused on the correlation properties while frequency domain, or spectral analysis, is used to analyze the frequency properties of a given time series.

Time series analysis can be conducted using traditional parametric techniques as well as non-parametric techniques. Parametric techniques assume an underlying stochastic nature, which can be described by using a small number of parameters. Conversely, non-parametric techniques assume no underlying structure and attempt to estimate the data’s covariance or spectrum (Cryer 2008).

Predictive forecasts enable researchers and analysts to determine future values based on previous observations, which can be used to shape public policy decisions or business initiatives. As with nearly any model’s creation, a model should require the minimal number of parameters that will competently embody the time series data (Cryer 2008).

2.2.1 Time Series Models

Time series models continue to be introduced and represent various stochastic processes. Models such as time series decomposition, Holt-Winters, and autoregressive integrated moving average (ARIMA) can be used to describe a time series. Rolling horizons, or rolling window, mimic splitting a data set into training and testing splits by repeatedly fitting the forecasting models to a designated “rolling period” and then measuring performance in forecasting over the horizon that will be used in practice. Rolling horizons can be used to evaluate a time series model’s performance against the known values. Forecasting performance declines rapidly over a forecast horizon and as a result, short forecasts are much more accurate than long forecasts.

Naïve Model

At its most basic level, Naïve Models offer a baseline standard. A simple Naïve Model sets the forecast value to the value of the last observation. Trend Naïve Models modify the simple Naïve Model's value by incorporating the positive or negative trend to the predicted value. Seasonal Naïve Models use the forecast from the previously observed seasonal period. Models that are able to outperform Naïve Models are said to have predictive power (Erdem 2021).

Time Series Decomposition

Time series decomposition is a very useful method, which attempts to remove and understand the seasonal, cyclical, and trend components separately. For example, classical decomposition assumes a constant seasonal component at the weekly, monthly, or quarterly pattern while X11 decomposes time series data into monthly or quarterly data (Hyndman and Athanasopoulos 2018). Conversely, seasonal and trend decomposition using Loess seasonal and trend decomposition using Loess (STL) is more robust as it is able to manage any type of seasonality and is able to change over time (Hyndman and Athanasopoulos 2018).

Autoregressive Integrated Moving Average

Since time series data is often correlated with previously lagged observations, ARIMA models can be applied to understand the data's characteristics and to forecast future observations. ARIMA models are designed to describe the autocorrelations within the time series data (Hyndman and Athanasopoulos 2018). As a result, the ARIMA model uses the data's previous observations and regresses the variable against itself. Non-seasonal ARIMA models are commonly designated ARIMA(p,d,q) as defined by Hyndman and Athanasopoulos (2018):

- p : Order of the autoregressive part.
- d : Degree of first differencing involved.
- q : Order of the moving average part.

Holt-Winters

The Holt-Winters model applies an exponential smoothing technique to the trend, seasonality, and cycle components of a time series (Hyndman and Athanasopoulos 2018). Exponential smoothing is a variation on moving average smoothing that weights the effect of previous observations/forecasts decay exponentially as you go back in time. Holt-Winters models can also be incorporated into a rolling horizon design with no predictable upward or downward trend.

Ensembles

As with cross-sectional data, ensemble models combine multiple, diverse models and aggregate each base model's prediction into one final prediction for the observed value (Kotu and Deshpande 2014). As a result, an ensemble model often outperforms individual models. However, ensemble models are often “black box” models that are difficult to understand and interpret. Creating and adapting an ensemble model using the aforementioned models may be desirable given the underlying nature of a time series.

2.2.2 Time Series Performance

While traditional statistical performance measures such as root mean squared error (RMSE) and adjusted R^2 can still be calculated, they are often not scale-free and are generally inferior (Guo et al. 2012). However, time series data can be measured using Mean Absolute Percentage Error (MAPE) and Mean Absolute Scaled Error (MASE). MAPE is the prediction accuracy for a forecasting method and MASE is a measure of the absolute error the forecast values, divided by the mean absolute error of the in-sample one-step naïve forecast (Hyndman et al. 2006). MAPE commonly varies between the values of 0 and 1 while MASE can extend beyond 1 and both provide a good sense of predictive power. MAPE and MASE are defined by the following equations:

$$\text{MAPE} = \frac{\sum_{t=1}^N \frac{Y_t - F_t}{Y_t}}{N}, \quad (2.4)$$

and

$$\text{MASE} = \frac{\sum_{t=1}^N \frac{Y_t - F_t}{Y_t - Y_{t-1}}}{N}. \quad (2.5)$$

where Y_t is the actual value, F_t is the forecast value, and $Y_t - Y_{t-1}$ is the actual value at time $t - 1$ (Hyndman et al. 2006).

For a model to have a strong predictive capability, a model must surpass two standards. First, a model's MASE needs to exceed the accuracy of its comparative Naïve model. Second, a model's MAPE needs to be at a minimum of below 50% as depicted in Table 2.1 (Lewis 1982):

Table 2.1. MAPE Interpretation. The following table provides guidance and recommendations for interpreting a given time series' MAPE. As we can see, the closer MAPE is to 0, the higher forecasting power it has. Source: Lewis (1982).

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

2.2.3 Time Series Research Efforts

Although the first known case of COVID-19 was only reported in December 2019, ample observations are needed to effectively develop and train time series models. However, researchers have been able to have promising success using ARIMA and Holt-Winters models.

Researchers employed an ARIMA(2,1,1) model to analyze COVID-19 infections in India, which depicted an exponential increase in the number of confirmed cases over the specified time period (Tandon et al. 2020). Although the MAPE is above the accepted threshold of 1, the ARIMA model outperformed its moving average and exponential smoothing counterparts.

Additionally, data scientists also analyzed trends in India by using both Holt-Winters and an ARIMA (4,1,1) model (Panda 2020). Utilizing Akaike's Information Criteria and RMSE

as the model's primary measure of performance, the authors achieved a maximum of 99.8 percent accuracy for the given time period (Panda 2020).

Although utilizing a small data set, researchers in Italy applied an ARIMA model to forecast COVID-19 cases (Benvenuto et al. 2020). The Italian model successfully predicted a decrease in COVID-19 incidences despite an increase in cases.

Researchers applied several time series techniques, Holt-Winters performed the worst when comparing RMSE while the Naïve Model performed the best (Chaurasia and Pal 2020). The authors collected approximately six months of data, which was mostly sparse in the beginning. As a result, the data set is unstable and the given model may perform better with additional observations. Researchers in Jakarta applied both ARIMA and Holt-Winters forecasts to model COVID-19 outbreaks (Sulasikin et al. 2020). The ARIMA model produced the highest R^2 and RMSE, which indicated that COVID-19 infections will continue to increase. As a result, public health officials can create and implement health measures to potentially disrupt its spread. Researchers in Nigeria created an ensemble model using ARIMA, Holt-Winters, and an additive regression model developed by Facebook known as Prophet, which is able to process missing values, strong seasonal effects, and outliers relatively well (Abdulmajeed et al. 2020). The results can guide policy makers to develop and implement population control and containment measures.

2.3 Epidemiological Models

Following the Spanish flu outbreak in 1918, W. O. Kermack and A. G. McKendrick theorized, developed, and published a series of papers describing mathematical models between 1927 and 1933 to predict the number of infections or cases throughout a population as a function of time (Kermack and McKendrick 1991). The authors theorized that when creating a mathematical model, epidemiologists can divide a population into a small number of compartments, which represent an individual's condition in relation to the virus as it is transmitted through a population over time (Kermack and McKendrick 1991). Since their work in the early 20th century, Kermack and McKendrick's theories for modern epidemiological modelling serve as a foundation and ensuing theoretical progress has been steady (Rodrigues 2016). These compartmental models enable a relatively simple method to mathematical model infectious diseases.

2.3.1 SIR

The Susceptible-Infected-Recovered (SIR) model and its variants are often used due to their simplicity and intuitiveness. The SIR model is linear in that individuals move from one compartment to the next without the ability to regress. For a SIR model to be effective, a government must maintain meticulous infection and recovery rates as they are an essential component for an accurate SIR model.

Definitions

Prior to understanding the model itself, we must first establish a common framework of definitions by (Rodrigues 2016), (Ridenhour et al. 2014), and Collins and Abdelal (2018).

- 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).

Often times unknown, R_0 is critical to determining the severity and duration of a virus. If $R_0 < 1$, then the virus is said to be self-extinguishing (Rodrigues 2016). If $R_0 = 1$, the virus is endemic but if $R_0 > 1$, the virus is said to be an epidemic (Rodrigues 2016). Despite its simplicity, a severe limitation of the traditional SIR model is its inability to modify the time-varying property of β and γ to effectively and precisely predict the virus’s trend (Chen et al. 2020c).

Assumptions

The SIR model relies upon significant assumptions to ensure an accurate representation of a given data set.

- The first major assumption is that the population is constant and no new individuals enter the Susceptible compartment (Cooper et al. 2020). Moreover, the population remains constant despite actual births or immigration (Capitanelli 2020). Since the dynamics of an epidemic are dynamic and are often much faster than the dynamics of birth and death, they are often omitted in simple compartmental models (Capitanelli 2020). As a result, the only way an individual leaves the Susceptible group is by infection (Capitanelli 2020). By extension, the only way an individual leaves the Infected group is through recovery (Capitanelli 2020).
- The next major assumption is that all individuals have the same probability to contract the virus (Capitanelli 2020). Additionally, each person is exposed to a fixed number of contacts per day, which is sufficient to spread the virus (Capitanelli 2020). However, not all of these individuals are strictly susceptible as some are undoubtedly infected themselves (Capitanelli 2020).
- The final major assumption is a fully connected population network and a homogeneous mixing of the compartments (Capitanelli 2020). Consequently, individuals from the susceptible compartment are exposed to those from the infected and recovered compartments (Capitanelli 2020). As a result, the virus propagates and continues to spread (Capitanelli 2020).

Equations

The SIR model consists of a system of three nonlinear ordinary differential equations to determine the number of individuals within each of the three compartments (Rodrigues 2016):

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

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

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

Research Efforts

Researchers in Cameroon adapted the SIR model to analyze and calculate the effects of public health measures to slow the spread of SARS-CoV-2 (Nguemdjo et al. 2020). The authors determined that had these measures not been implemented, additional infections and deaths would have occurred. As a result of the SIR model's projections, Cameroonian public health officials were able to slow the spread of the virus. Moreover, the model demonstrated that it is particularly suitable for studying a large number of subjects at the national level.

While insightful, SIR's modeling assumptions are self-limiting. However, epidemiologists created a SIR model where the susceptible population is not decreasing monotonically (Cooper et al. 2020). Rather, the susceptible population can surge over time as new virus epicenters arise in communities of varying sizes. As a result, the susceptible population can be adjusted to account for individuals within the infected compartment spreading throughout a community. The model and formulation presented by the authors demonstrated reasonably well fit model predictions for several distinct countries and geographic regions, which introduced and exhibited the effects of incorporating a surging susceptible population (Cooper et al. 2020).

The SIR model utilized by researchers in the Kingdom of Saudi Arabia adapted a similar approach to determine the total number of cases and fatalities within Saudi Arabia under no preventative measures, lockdown enforcement, and new medicine implementation (Singh et al. 2020). The authors modified the original SIR model to incorporate a "fatal with confirmation" component to ensure that recovered individuals are not susceptible to the virus (Singh et al. 2020). Additionally, this model incorporated a stochastic component to analyze and predict SARS-CoV-2 transmission within its population. The presented model accurately predicted COVID-19 peaks and sizes of the outbreaks. However, initial research has since shown that previously infected and recovered COVID-19 individuals can become infected again (Gallagher 2021). As such, this model may need to be continuously modified and adapted to provide an accurate forecast.

A limitation of the traditional SIR model is the inflexibility and constant nature of the recovery rate, γ , and transmission rate, β , parameters. However, a researcher proposed a model where both the recovery and transmission rates are time-dependent (Chen et al. 2020c). By incorporating ML techniques, the model is able to monitor and dynamically adjust the aforementioned parameters as a function of time. As a result, $\beta(t)$ and $\gamma(t)$ are used to predict the number of infected rate and recovery rate during the epidemic. The model observed that a city-wide population control measures can lower the transmission rate substantially.

Despite its simplicity, the SIR model provides relatively accurate representations of a country's populace as these models have shown. The SIR model presents analysts and researchers an opportunity to develop an understanding of the population at the national level or even a local community. However, there are several SIR model variants that create additional compartments for a population.

2.3.2 Susceptible-Exposed-Infected-Recovered

As epidemiologists develop a better understanding of a given virus, the SIR model may no longer be relevant or applicable. Assumptions such as a constant population are not necessarily realistic. Recovery and subsequent immunity may not necessarily be guaranteed. Infected individuals may exhibit asymptomatic symptoms. Immunity in recovered individuals may be temporary. Transmission rates may vary as mutations occur. Vaccine discovery and deployment can drastically alter a virus's transmission rates. While previous coronavirus diseases such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) experienced no exposed or latency stages, the incubation period for individuals with COVID-19 is extremely dangerous to close contacts (Chen et al. 2020b). As such, a divergent epidemiological approach must be adapted.

Susceptible-Exposed-Infected-Recovered Definition

The susceptible-exposed-infected-recovered (SEIR) model employs an additional exposed compartment, E , where individuals have been infected but are not yet infectious themselves. During this incubation period, secondary spread from an individual who is infected will occur later. Despite the addition of this latency period, the number of total cases does not change. An additional incubation parameter, σ , is defined as the rate of latent individuals

becoming infectious. Comparable to the SIR model, the SEIR model in a closed population is represented by four ordinary differential equations:

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

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

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

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

where β , S , I , R , γ are discussed in Section 2.3.1 and $N = S + E + I + R$ is the total population (Institute for Disease Modeling 2021).

SEIR Research Efforts

Researchers identified that once an individual contracts COVID-19, there is a subsequent incubation period in which individuals are asymptomatic (Hoehl et al. 2020). These individuals can still carry and subsequently spread SARS-CoV-2 to others and in turn, continue the virus's propagation (Patil and Kotwal 2020). As a result, the SEIR model has been adapted to replicate and analyze SARS-CoV-2's diffusion.

Researchers applied a SEIR model to Wuhan, China, to study the transmission dynamics in the aftermath of the local authorities instituting strict population control measures such as lock downs, movement restrictions, community containment, and quarantine initiatives (Hou et al. 2020). Researches employed β_1 and β_2 infectious rates, 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.” “While β_2 represents the potential rate per exposure when a susceptible individual has mutual contact with an exposed individual and transmits it to another exposed individual.” By incorporating these parameters, researchers demonstrated the SEIR model's effectiveness to predict SARS-CoV-2 diffusion (Hou et al.

2020). By reducing a latent individual's contact rates through quarantine and isolation, one can reduce the number of individuals they subsequently infect with COVID-19. Moreover, as a result of these public policies, the number of SARS-CoV-2 carriers decreases, which delays the peak infection time and reduces the strain on the local, region, and national health care system.

Epidemiologists created a SEIR model to analyze SARS-CoV-2 transmission in Portugal during the pandemic's early stages by dynamically adjusting the exposure rate, β , to simulate an exposed yet asymptomatic person's ability to spread COVID-19 (Teles 2020). The parameter β changes following notification of exposure as movement restriction and isolation imposed by the government and self-protection measures reduced transmission and subsequent COVID-19 cases. As a result, researchers modeled SARS-CoV-2 infections and forecasted the epidemic curve within a limited time period (Teles 2020).

2.3.3 Other SIR Variants

While research efforts have mostly been concentrated on adapting and optimizing SIR and SEIR models, researchers also adapted susceptible-infected-infected-recovered (SIIR) and metapopulation susceptible-exposed-infected-recovered (MSEIR) models. The SIIR model is similar to the SEIR model, except presymptomatic and asymptomatic individuals do not experience movement restriction (Tomochi and Kono 2021). The MSEIR model proposed an epidemiological spread across a species linked by a conglomerating network (Chen et al. 2020a).

Network SIR

While the SIR model assumes uniform mixing amongst a population, a network-centric SIR model provides an alternative as most individuals mix within a much narrower group (Craig et al. 2020). A network SIR model can integrate and customize dissimilarities between population groups as their interactions may differ. Populations in more rural communities may experience less contacts than individuals who live in more urban areas. As a result, those individuals who live in cities experience a higher variability in the number of daily contacts, which increases their risk exposure. As such, these individuals and communities affect “the speed of the disease spread, long-run health outcomes, and the effects of the disease on economic activity” (Craig et al. 2020).

Susceptible-Infected-Infected-Recovered

Researchers utilized a Susceptible-Infected-Infected-Recovered (SIIR) model to analyze the potential for herd immunity by modifying basic SIR and SEIR assumptions (Tomochi and Kono 2021). First, a portion of the population is presymptomatic who can infect others during the incubation. Asymptomatic individuals are able to move and contact others without restriction and consequently play a vital role in the proliferation of SARS-CoV-2. Lastly, immunity duration may be finite. By incorporating presymptomatic and asymptomatic into the base SIR model, the authors demonstrated that these persons significantly impact herd immunity (Tomochi and Kono 2021).

Metapopulation Susceptible-Exposed-Infected-Recovered

Epidemiologists provide a framework to model SARS-CoV-2 propagation across a population that is large-scale spacial regions (Chen et al. 2020a). The metapopulation concept and its subsequent Metapopulation Susceptible-Exposed-Infected-Recovered (MSEIR) refers to a group of species connected but separated by an interacting network. Researchers argue that the SIR assumption for homogeneous population mixing is no longer valid due to the complex social structure within a given location. Scientists utilized artificial intelligence and ML techniques that incorporated a sliding window approach, or rolling horizon, to describe the epidemic spread. The authors successfully predicted COVID-19 cases for a small time period but the model became increasingly unstable as time window grew (Chen et al. 2020a).

2.4 Simulations

A simulation is an algorithm implemented on a computer that represents a stochastic model and can be used to make inferences about certain measures of performance. Frequently, physical systems can not be easily replicated or experimented with an analytical model. Often, collecting data can be fiscally or computationally expensive. However, design of experiments (DOE) and subsequently simulation provides an alternative where model parameters can estimate values and experiments may be run hundreds of thousands or even millions of times at relatively low cost. Simulations also enjoy good convergence properties relative to other numerical approaches and can be used to study sources of uncertainty. Additionally, randomness can be used to create simple and useful models by identifying and

eliminating extraneous predictors and components (Sanchez and Sanchez TO APPEAR). In general, performance measures fall into two categories:

- Transient (or terminating) measures: Evaluate the system over a finite period of time.
- Steady-state measures: Evaluate the system over an infinite time horizon.

.Since COVID-19 infections increase and evolve at marked points over time, discrete event simulation (DES) can be applied. DES is a tool that can be “used to model real world systems that can be decomposed into a set of logically separate processes that autonomously progress through time” (Barrett et al. 2008). As such, DES can be applied to model a hospital’s patient care system that can assist a hospital administrator’s ability to manage patient care.

Additional simulation techniques such as agent-based model agent-based model (ABM) afford researchers an opportunity to explore and experiment with various SARS-CoV-2 and COVID-19 characteristics (Kerr et al. 2020). ABM attempts to simulate human behavior by modeling actions and interactions of autonomous agents, to assess their individual effects on the entire system. ABM combines elements of decision theory, sociology, networks and complex systems while Monte Carlo methods can be used to introduce randomness within the system.

Simulation Research Efforts

As global COVID-19 cases continue to increase, medical capacity and capabilities decrease as hospitals are designed for average patient loads and not epidemics (Cavallo et al. 2020). Therefore, effective modeling of the patient care system is needed to determine the optimal allocation of resources in best and worst case scenarios. As a result, hospital administrators can effectively manage resources such as beds, ventilators, and personnel. Additionally, epidemiologists are able to develop an understanding of the epidemic the local, regional, state and federal levels.

Prior to the discovery of SARS-CoV-2, researchers created an ABM to study the effects of epidemic spread in an urban area using transportation models (Hackl and Dubernet 2019). Researchers argue that accurately modeling human movement, behavior, and interactions is critical to understanding an epidemic’s spread throughout a population. The authors utilized a SIR compartment model to simulate the epidemic’s progression and demonstrated the

viability of constructing ABMs in urban settings (Hackl and Dubernet 2019).

Due to the small sample size, Cameroonian officials applied non-parametric techniques to create additional data points through bootstrapping in the early stages of the pandemic (Nguemdjo et al. 2020). Researchers first estimated virus parameters such as β to best describe the virus's evolution. As previously discussed, they then used additional simulation efforts to construct a compartment SIR model to understand the virus's propagation within the Cameroonian population. Using these results, they then constructed simulations to model public health initiatives such as increasing public hygiene awareness, physical distancing, and public health initiatives. These descriptive statistics demonstrated that Cameroonian officials should expect to see an exponential rise in SARS-CoV-2 transmission.

Researchers constructed an ABM simulation and analyzed SARS-CoV-2 and COVID-19's impact on local economic conditions (Kano et al. 2021). Although an atypical use of ABM, their results "show that voluntary restraint measures can help mitigate an outbreak, although it generates an economic gap between job types" (Kano et al. 2021). Their model also show that it is difficult for individuals with lower economic power to maintain isolation or limited movement. As a result, virus outbreaks may occur within these lower economic power communities as they are forced to return to their occupation, which continues virus propagation. To combat these results, early and effective population control measures can help limit the spread of the virus and thus minimize the economic impact to communities with low economic power.

Researchers also employed an ABM to study the effects of individual interventions to determine the virus's economic impact (Silva et al. 2020). Researchers modeled seven different social distancing measures no preventative measures, vertical isolation, partial isolation, mask wearing, social distancing while wearing masks, conditional lockdown, and lockdown. Results indicate that a conditional lockdown predicated upon current infection levels and a traditional lockdown proved to be the most effective in controlling the virus and minimizing deaths. However, government sponsored economic subsidies are required to lessen the economic impact caused by unemployment and subsequent recession.

2.5 Summary

Despite the various modeling approaches discussed, there are inherent difficulties in modeling epidemics such as SARS-CoV-2 transmission due to a variety of reasons. Human nature is often unpredictable. Testing supplies may be limited. Medical capacities become increasingly strained as the epidemic grows. Individuals may not comply with quarantine, isolation, or curfew orders. The very stochastic nature of epidemics imply a randomly determined outcome. The systematic approach to time series analysis to understand time series data enables researchers and analysts answer statistical, mathematical, and ultimately policy related questions as to the foreseeable future. However, the aforementioned models demonstrated that statistical and analytical methods can be applied to understand the virus's underlying structure and forecast future observations to aid public health policies.

Network models such as social and transportation networks provide a framework for analyzing SARS-CoV-2 transmission since individuals can represent nodes and communities form around these nodes. Researchers can understand the intricacies of a given network to determine metrics such as the shortest path, centrality measures, k -core, and clustering coefficients. Understanding mathematical properties enable epidemiologists and analysts to comprehend how a virus may propagate through a network and what policy decisions should be recommended to public health officials.

Time series models such as Holt-Winters, ARIMA, and even Naïve Models also provide researchers an opportunity to understand and forecast SARS-CoV-2 diffusion amongst a population. While the initial results of these models are promising, the COVID-19 pandemic is barely in its second year and as such provides minimal data to develop meaningful insights for researchers and analysts. As a result, these models are often accurate for short-term forecasts but perform poorly in long-term forecasting.

Traditional and classically used compartmental models such as the SIR, SEIR, SIIR, MSEIR, and others provide epidemiologists a framework to evaluate epidemiological parameters such as dispersion, total number of cases, pathogen properties, or its duration. Although often simple in their nature, compartmental models provide researchers an opportunity to model a virus's behavior by approximating a small number of parameters and analyzing the potentially millions of data points (Tolles and Luong 2020).

In the early stages of the pandemic, information and data regarding SARS-CoV-2 and COVID-19 were very minimal. As a result, researchers often simulated various characteristics to develop a better understanding of the virus's characteristics. As more data became available, simulations then shifted away from initial approximations to developing forecasts and predictions, which subsequently increased their accuracy. Advances in hardware, software, and methods will continue to make simulation an increasingly valuable tool in solving real world problems.

Despite progress made by the previous studies, there currently does not exist a SARS-CoV-2 forecasting model utilizing a country's local governance adjacency network. This thesis will employ a network structure that incorporates a time series component via the GNAR package to understand the relationship between COVID-19 cases amongst the counties (Knight et al. 2020). This model can be used as another tool for local, state, and federal health officials to evaluate risk and implement public health policies to disrupt an epidemic's transmission.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Methodology

For this research, all data are openly available and can be consumed from the following sources: United States Census Bureau (USCB), National Bureau of Economic Research, and United States of America Facts (USAFacts). As each county can be identified with its own unique federal information processing standard (FIPS) code, the USCB maintains administrative accountability of each county and its associated adjacencies. The National Bureau of Economic Research calculated the great circle distance between each county using the Haversine formula (National Bureau of Economic Research 2010a). USAFacts provides daily COVID-19 cases and deaths for each county directly from the Center for Disease Control and Prevention (CDC).

3.1 Data

Since COVID-19 cases and deaths as well as the great circle distance are integers and decimals, respectively, all data in this thesis are entirely numerical with no categorical predictors or response variables. Additionally, the original number of COVID-19 cases and deaths were used and traditional data transformations such as logarithmic, square, or cube root were not applied.

3.1.1 Data Assumptions

The model's first assumption is that COVID-19 cases are stationary and as a result, do not depend on the time at which the series is observed (Hyndman and Athanasopoulos 2018). By extension, the data set does not contain any seasonal or trend components as they affect the series at different time periods. However, a time series with only cyclic behavior can be considered stationary (Hyndman and Athanasopoulos 2018). As such, the data set should not contain any predictable patterns in its long-term trend.

The next assumption is that there is no uncorrelated error as it is randomly distributed but its variance and mean are constant. If random error is present, they are assumed to be randomly distributed with a constant variance and mean zero (Statistics Solutions 2021).

The data provided by local and state officials to the CDC are accurate and complete. While COVID-19 cases and deaths may change slightly from day-to-day as human error and reporting standards may change, this thesis assumes the number of cases and deaths on any given day are accurate (Centers for Disease Control and Prevention 2020). Moreover, this thesis assumes that the data set follows an autoregressive behavior as each observation is directly linked to previous observations, which forms a regression equation to predict the next value (Brownlee 2017). Additionally, this thesis assumes that there are no outliers present in the data as they may affect model creation, analysis, conclusions, discussions, and recommendations as they can be misleading (Frost 2020).

3.1.2 Data Limitations

Data limitations may alter the results of a model. Since Hawaii is an island chain, geographically isolated from any continent and three of its five counties are not considered adjacent by the United States Census Bureau (United States Census Bureau 2010), the state and, subsequently its counties', total COVID-19 cases and deaths should remain constant at zero. Although Alaska is also geographically isolated from the continental United States, it is adjacent to Canada. However, both states are popular tourist destinations that contain airports and seaports where infections may enter and spread within their respective populations.

Local, state, and federal health officials would need to expend considerable resources to maintain an accurate count of active COVID-19 infections inside a given county. Alternatively, officials maintain a daily total count of infections and deaths. As a result, this thesis models total COVID-19 cases and deaths and is unable to accurately model active infections.

3.1.3 COVID-19 Cases and Deaths

To understand SARS-CoV-2 transmission and subsequent COVID-19 infections within the United States, one must first develop an understanding of what a positive COVID-19 case is and how the number of cases and deaths are maintained. By CDC definition, a positive COVID-19 case is a confirmed or probable case or death (Centers for Disease Control and Prevention 2020). While COVID-19 symptoms are similar to influenza, a laboratory must conduct a specialized diagnostic test and look for the specific viral proteins or virus's genetic material to confirm an active coronavirus infection (Harvard University 2020). As

additional data became available, scientists discovered that an infected person does not begin producing antibodies immediately (Harvard University 2020). Moreover, a blood antibody test may take up to three weeks to become positive and as a result, testing a recently exposed individual may not yield positive results.

Once a laboratory or hospital confirms that an individual is diagnosed as a positive COVID-19 case, state disease reporting authorities, local hospitals, healthcare providers, and laboratories must report confirmed or probable cases to local or state health departments (Centers for Disease Control and Prevention 2020). Public health officials then monitor local infection rates to identify and control potential outbreaks. While case reporting to local and state health officials is mandatory under disease laws, case notification to the CDC is voluntary (Centers for Disease Control and Prevention 2020). Once local and state officials send their data to the CDC, a separate CDC data team ensures data integrity and examines it for any irregularities. The CDC then sends its parsed data back to state and local officials to correct any inconsistencies, if necessary, and ultimately for final confirmation. Once complete, the CDC publishes the final data on its website and send it to the World Health Organization (WHO) under international health regulations (Centers for Disease Control and Prevention 2020). The CDC also provides its data openly, with additional privacy protections, to the public and is available for consumption.

USAFacts is a non-profit, nonpartisan organization that openly provides data such as government finances and the American population (USAFacts 2020). For COVID-19 data, USAFacts collects and collates information from the CDC and provides two separate data sets: a county-by-county daily tally for the total number of confirmed or probable cases and a county-by-county daily tally of deaths. The CDC first began reporting cases on January 22, 2020, and continues to update cases and deaths daily. While information continues to be compiled and reported, this thesis utilized CDC data collected from January 22, 2020, to February 2, 2021, which provided 378 days of data.

3.1.4 County Information

As of 2020, the United States of America is divided into 3,143 county and county equivalents (United States Census Bureau 2020a). The term “county” is used in 48 states while Louisiana represents its functionally equivalent administrative districts as parishes while Alaska is

comprised of boroughs and census areas. The 48 contiguous states and Washington D.C. form a network of 3,109 counties that are separate from Alaska and Hawaii. Although not connected to the continental United States' graph, Alaska's 30 counties and Hawaii's 5 counties form their own distinct networks.

Each county is assigned its own unique, five digit FIPS code where the first two numbers correspond to the state and the last three are unique to the county within the state's possession. County administrative roles vary widely between states as some counties such as those in Rhode Island are maintained merely for administrative purposes while some counties in Maryland, Missouri, Nevada, and Virginia are independent cities not belonging to one specific county yet they may function as combined city-counties.

The number of counties per state is not predetermined based upon its geographic size, population, or prominent terrain features. The state with the smallest square area, Rhode Island, has 5 counties while the state with the largest square area, Alaska, has 30. Delaware and Texas contain the fewest and most counties per state, respectively. According to the USCB, more than half of the United States' population resided in just 143 of the 3,143 counties in 2020 (United States Census Bureau 2020b). Table 3.1 arranges the total number of counties per state with the inclusion of Washington D.C. As we can see from Table 3.1, counties vary broadly by state.

Table 3.1. United States Counties. The numbers shown in this table are the number of counties in each state. Source: TheFactFile (2021).

State	Counties	State	Counties
Alabama	67	Montana	56
Alaska	30	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		

Due to their geographical arrangement, counties form a uniplex network as they become adjacent to one another (Bergs 2011). As such, one can apply aforementioned network science methods to calculate and analyze various underlying network properties. Due to their adjacent structure, one can also model the free flow of goods or SARS-CoV-2 transmission between counties.

County Adjacencies

We represent county adjacencies, or edges, through either a binary structure or the great circle distance between two county centroids. A network with a binary adjacency structure contains only 1s to represent if an edge is present or 0 if there is no edge.

In addition to a relatively simple binary adjacency matrix, one can determine the shortest path between counties using the great circle distance. The great circle distance is determined by the shortest measured distance along the surface of the sphere between two points. The National Bureau of Economic Research calculated the great circle distance from one county to every other county. Therefore, we can create a matrix that contains these values. We can then apply matrix multiplication using the binary adjacency matrix and the great circle distance matrix to produce a great circle distance adjacency matrix where zero represents no edge present and a nonzero value represents the distance between the centroid of one county to its neighbor.

County Network Structure

A graph, G , is a finite nonempty set V of objects called vertices or nodes and a set E of 2-element subsets of V called edges (Chartrand and Zhang 2012). Due to their geographic composition, the United States' counties form a simple graph with no loops or multiple edges. However, the entirety of the United States is not connected as Alaska and Hawaii are geographically separate. Despite its disconnected nature, an adjacency matrix can still be used to depict this finite graph where an edge is either present or absent between adjacent pairs of nodes, which indicates connection or disconnection respectively. In all, there are 9,294 edges in E and 3,143 nodes in V that comprise the United States' county adjacency network. "The degree of a vertex v in a graph G is the number of edges incident with a given vertex v " (Chartrand and Zhang 2012). Figure 3.1 displays a histogram of the varying county degrees within the United States. As we can see from Figure 3.1, the county degrees

are approximately normally distributed where the horizontal axis is the county degree, and the vertical axis is their frequency.

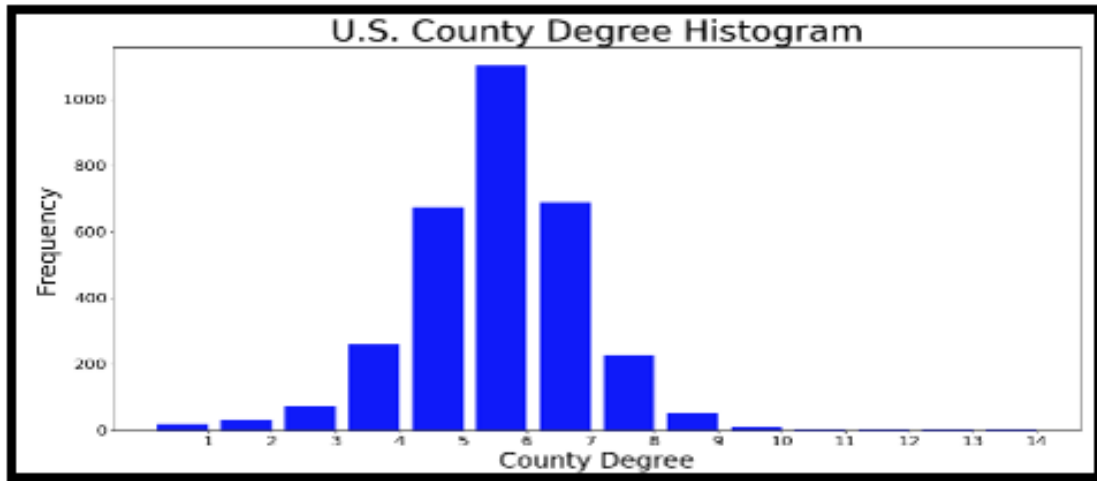


Figure 3.1. United States County Degree Histogram. The Y axis represents the degree frequency and the X axis depicts the county degree distribution. As we can see from the figure, the degree distribution is approximately normal. Adapted from United States Census Bureau (2010).

Synthetic Networks

Synthetic network models are often used as a reference model to compare a given network against and to analyze and build new, complex networks. Traditional synthetic networks such as random graphs (Erdős and Rényi 1960), small-world graphs (Watts and Strogatz 1998), scale free graphs (Barabási and Albert 1999), Configuration model (Molloy and Reed 1995), and the random geometric model (Gilbert 1961) do not provide an accurate representation of adjacent United States counties. Connectivity between random graphs is independent while the edges between United States counties are dependent. Small-world graphs have a small shortest path distance, d , that is modeled after $\log(n)$ while the United States counties do not. Scale free graphs are those whose network degree distribution is of a power law. The United States' county degree distribution is approximately normal. Similar to random graphs, adjacency and connections between configuration models and random geometric models are independent.

Conversely, triangular lattice networks are formed in Euclidean space and create a tiling

similar to the geographic layout of the United States. As a result, our real network and the triangular lattice network model share similar statistics and underlying characteristics as evidenced in Table 3.2. However, the triangular lattice network's diameter and average shortest path are higher than the original network. As a result, we can see that the network statistics provided in Table 3.2 lend credence to a large network that has few hubs, high clustering, built-in redundancy, and a sparse structure. Therefore, a triangular lattice network provides an accurate topological representation of the original network.

Table 3.2. Network Comparison. This table shows graph statistics in terms of counties and in terms of a triangular lattice. Source Adapted from United States Census Bureau (2010).

Network Metrics	United States Counties	Triangular Lattice
Nodes	3,109	3,120
Edges	9,242	9,125
Degree (Avg)	5.945	5.849
Closeness Coefficient	0.0396	0.0327
Betweenness Coefficient	0.008	0.010
Clustering Coefficient	0.435	0.410
Density	0.002	0.002
Diameter	68	78
Shortest Path (Avg)	28.823	31.429

Community Detection

Communities often appear in real networks as a group of nodes that share interests or characteristics and can be grouped into densely connected arrangements (Newman 2006). Communities can be characterized through a myriad of properties to include small-world, clustering, or the underlying structural properties of the network. As a result, these characteristics can be identified through various algorithms such as Louvain and Leiden. For social networks, a community may be a group of friends while trusted users may form a community within a cryptocurrency network.

Additionally, Newman suggests “that networks can have properties at the community level that are quite different from their properties at the level of the entire network” (Newman

2006). As a result, “analyses that focus on whole networks and ignore community structure may miss many interesting features” (Newman 2006). Moreover, researchers are able to separate the original, larger network into smaller and more easily managed communities, which act like meta-nodes (Newman 2006). As a result, these individual meta-nodes can make the original network easier to study (Newman 2006). These smaller meta nodes, or communities, often enable a deeper understanding of the larger network’s functions as the smaller meta nodes often correlate to various functions contained within the larger network (Newman 2006).

Community detection is the process to identify what nodes belong to which communities (Fortunato 2010). However, a network’s community structure is often unknown beforehand and often of unequal density and/or size. Moreover, identifying and categorizing communities within a given network can be a computationally arduous undertaking as the number of nodes, N , gets large (Hoffmann et al. 2020). Additionally, “methods for understanding what the communities mean after you find them are, by contrast, still quite primitive, and much needs to be done if we are to gain real knowledge from the output of our computer program” (Newman 2008a). As a result, efficiently calculating and subsequently understanding a network’s community structure can be a complicated endeavour.

However, the computational processing capability continues to increase as discussed in Chapter 2 provides refuge for scientists as networks become increasingly more complex. Furthermore, as researchers and mathematicians fuse more efficient algorithms with more capable computer processors, an algorithm’s processing time decreases. However, these computationally faster methods are not usually deterministic and can change the communities when the algorithm is repeated, especially when there is no clear defined community structure. While there are several algorithms for community detection, they can be categorized into overlapping and non-overlapping techniques (Yang and Leskovec 2012).

Overlapping communities are those that allow nodes to belong to one or more communities, while non-overlapping communities are those that partition the network into groups with sparser connections externally and denser connections internally (Yang and Leskovec 2012). One of the most ubiquitous methods to detect communities is through modularity (Newman and Girvan 2004). Modularity is the difference between the actual number of edges in a community and the expected number of such edges in a random graph of similar

size networks(Traag et al. 2019) As a result, modularity can range from $[-1, 1]$ where larger values indicate a stronger community structure and dense communities, which sparse connections between them.

Overlapping algorithms such as Louvain and Leiden seek to optimize modularity within a network (Traag et al. 2019). When applied to the United States county network of 3,143 nodes, the Louvain method produces 22 communities with a modularity of 0.864, while the Leiden method produces 10 communities with a modularity of 0.938. Therefore, both the Louvain and Leiden methods produce a strong community structure with few connections between communities. As evidenced in Figures 3.2 and 3.3, the Louvain method consolidated similar states while Leiden's method combined geographic regions to represent the country. Unsurprisingly, Alaska and Hawaii form their own communities due to their geographic isolation from the other 48 states. As a result, the United States county network is a strong community structure that is internally dense with sparse connections externally.

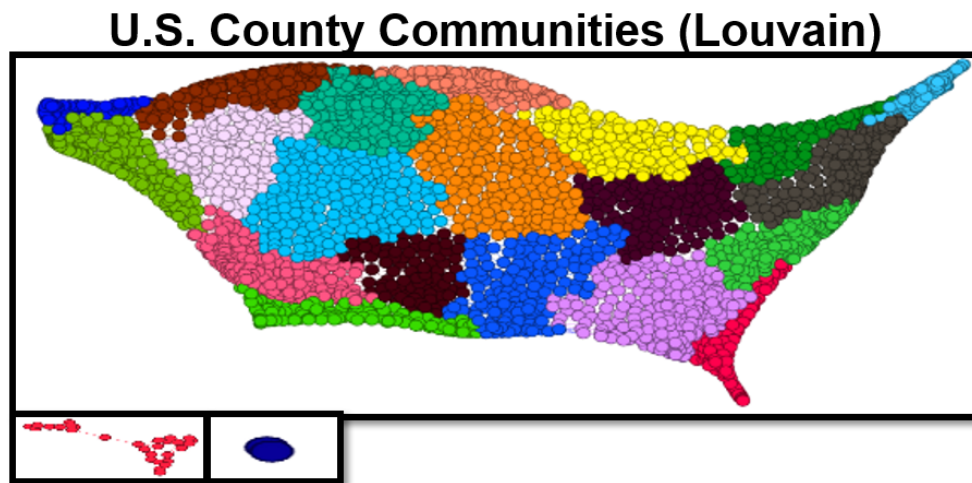


Figure 3.2. Louvain Community Detection in the United States. Adapted from United States Census Bureau (2010).

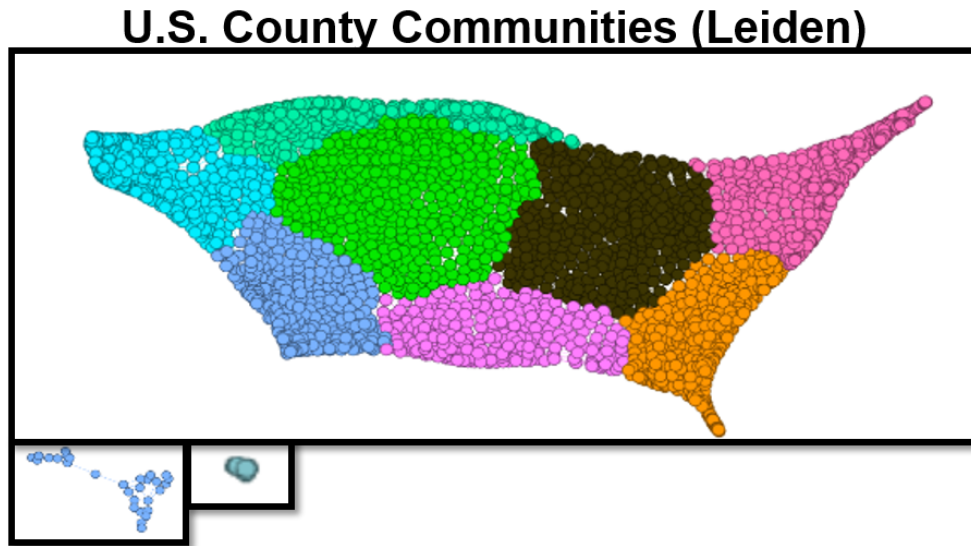


Figure 3.3. Leiden Community Detection in the United States. Adapted from United States Census Bureau (2010).

Small-World Graphs

In 1990, American author John Guare composed *Six Degrees of Separation* (Guare 1990). In his book, Guare writes that there exists a shortest path between a given individual to nearly anyone else. This small-world phenomenon, coined six degrees of separation, serves as the underlying principle behind small-world networks (Easley 2010).

In 1967, Stanley Milgram and other researchers created a social experiment to determine a shortest path between individuals in one city to another (Milgram 1967). Through their research, Milgram and his colleagues estimated that the average path length between individuals is between 5.5 and 6 (Milgram 1967). With the advent of social networking, reliable international travel, and globalization, people around the world are becoming more interconnected. Therefore, small-world theory can be examined further. Researchers at Facebook determined that of the approximate 1.9 billion active users on Facebook in 2006, users are connected to every other person by an average path length of 3.59 (Edunov et al. 2016).

A network may be considered small-world if the average distance, L , between two randomly chosen nodes, grows proportionally to the logarithm of the number of vertices, N (Watts and Strogatz 1998). Namely, $L \propto \log(N)$. The Python package `networkx` computes the small-world coefficient, ω , using the following equation:

$$\omega = \frac{L_r}{L} - \frac{C}{C_l}, \quad (3.1)$$

where C and L are the average clustering coefficient and average shortest path length of a network G (Hagberg et al. 2008). The parameters C_r and L_r are the average clustering coefficient and average shortest path length of a random graph, respectively. The parameter C_l is the average clustering coefficient of an equivalent lattice graph (Hagberg et al. 2008). While `networkx` implemented the code to generate the ω coefficient, it originated from (Humphries et al. 2006) and (Humphries and Gurney 2008).

Although detecting and creating a social network for the world's population is certainly feasible given substantial time and resources, small-world detection for the United States is a more attainable enterprise. While the given network need not be connected, the continental United States county network provides an opportunity to analyze characteristics within a network that contains ample connected nodes to determine relationships with one another. Through arduous computation, `Networkx` computed approximate value -2.9531 for ω . Values that are close to -1 imply a lattice structure and are not indicative of small-world graphs (Hagberg et al. 2008). Since `networkx` computed a value that is significantly lower than -1 and a triangular lattice network closely best approximated the county network, the United States' county network does not contain small-world properties.

Core-Periphery

As discussed in Section 2.1, the core of a network can be considered the densely connected and central nodes while the periphery are the nodes non-centrally and sparsely connected to the core (Csermely et al. 2013). The “core-periphery structure is a pervasive and a crucial characteristic of large networks” (Gera 2021b). Although k -core is calculated below for the United States' county network, k -cliques and k -plexes can also be used to represent the community structure. Hawaii and Alaska must be removed from the network to determine the core as they are not adjacent to the continental United States. As one can see from Table 3.3, the United States county network is a highly interconnected network with outlying peripheral nodes that are loosely connected to the core.

Table 3.3. Network k -cores. Numbers in this table are the number of counties in the United States with the respective k -cores. Adapted from United States Census Bureau (2010).

k -cores	United States Counties
$k = 1$	14
$k = 2$	28
$k = 3$	235
$k = 4$	2,832

Centralities

Network centrality measures are used as an estimate to determine a node's importance within a given network (Golbeck 2013). There are a number of methods to determine a node's centrality to include but not limited to: degree, closeness, betweenness, Katz, PageRank, and eigenvector. These measures provide an overview of a network's essential nodes.

Degree centrality is simply the number of edges a given vertex has (Chartrand and Zhang 2012). It follows that the higher the degree of a given node, the higher its degree centrality. However, degree centrality is not necessarily a good indicator of a node's importance globally, as it does not provide insight into its subsequent connections (Golbeck 2013). Closeness centrality quantifies how close a node v is to other nodes by calculating the shortest path length from one node to every other node (Veremyev et al. 2019). Since closeness centrality calculates the shortest path between nodes, a smaller value indicates a higher centrality for a given node.

Betweenness centrality measures how important a node is to the network's shortest path. Betweenness centrality for a vertex v is calculated by (1) selecting a pair of nodes and calculate the shortest path between them, and then (2) divide the shortest paths that contain v by the total number of shortest paths (Newman et al. 2006). Eigenvector centrality represents a node's importance to its neighbors (Newman et al. 2006). As a result, "eigenvector centrality acknowledges that not all connections are equal. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people" (Newman 2008b).

The United States' county network can be filtered to only the contiguous 48 states to ensure a connected, well-defined graph with metrics previously displayed in Table 3.2. Leading degree centrality measures represent counties that have the highest degree such as San Juan County in Utah. The leading closeness counties represent geographic centrality with Missouri, Illinois, and Arkansas representing 16, 3, and 1 of the top 20 counties, respectively. The counties with the highest betweenness centralities create approximate shortest paths in the north and south that connect the east and west coasts. The counties with the highest eigenvector centralities are the counties that exert relatively high significant influence on neighboring states with Iowan counties representing 36 of the top 40 and Minnesota consuming the remaining 4.

Figure 3.4 below depicts a box-and-whisker diagram with the closeness, betweenness, and eigenvector centralities. While there is not a clear visual distinction for closeness and betweenness centralities in Figure 3.4, a callout marker has been placed on the eigenvector data to delineate and exhibit the approximate top 40 counties. As we can see from Figure 3.4, the vertical axis depicts a given county's centrality score for closeness centrality, betweenness centrality, and eigenvector centrality measures on the horizontal axis.

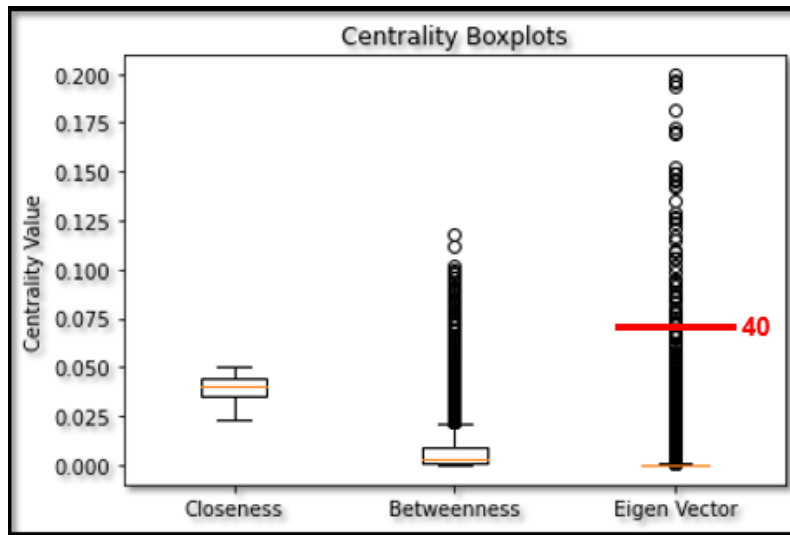


Figure 3.4. United States County Centralities. The Y axis represents the centrality score for a given centrality measure and the X axis depicts the various centrality measures. The red line is delineates the 40 counties with the highest eigenvector centrality scores. As we can see from the figure, closeness centrality has relatively low variability. Betweenness and eigenvector also have low variability but have counties that contain relatively high centralities. Adapted from United States Census Bureau (2010).

Eccentricity

The eccentricity $e(v)$ of a connected graph G is defined as the maximum distance between a vertex v_i and every other vertex v_j in G where $i \neq j$ (Chartrand and Zhang 2012). The maximum eccentricity among the vertices in G is the network's diameter while the minimum eccentricity is the network's radius (Chartrand and Zhang 2012). Since the network needs to be connected to compute its eccentricity, we remove Alaska and Hawaii from our analysis. With these states removed, the continental United States graph remains and we can then compute the eccentricity of each node and develop a histogram as depicted in Figure 3.5. As we can see from Figure 3.5, the county eccentricities are approximately normally distributed where the vertical axis is the frequency and the horizontal axis is the county eccentricities.

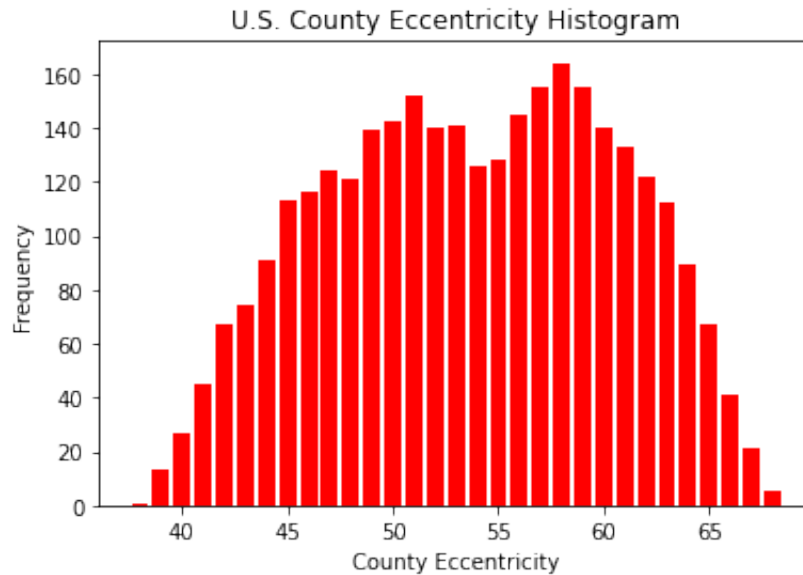


Figure 3.5. County Eccentricities. The Y axis represents the degree frequency and the X axis depicts the county eccentricities. As we can see from the figure, the eccentricity ranges from 38 to 68, which indicates a sparsely connected network. Adapted from United States Census Bureau (2010).

The minimum, mean, median, and maximum eccentricities of the continental United States graph G are 38, 53, 54, and 68, respectively. Counties that have a smaller eccentricity are states located in the center of the network such as Illinois and Iowa. Conversely, counties in Maine, California, Oregon, and Washington contain the highest eccentricity values. Counties that have eccentricities near the median and mean are located in geographically centered states such as Oklahoma and Texas. We can see that the counties form a sparsely connected network that require tens of connections to traverse.

Scale Free

A scale free network is one whose degree distribution follows a power law, that may grow due to preferential attachment (Newman et al. 2006). A power law “states that a relative change in one quantity results in a proportional relative change in another” (Glen 2016). Preferential attachment is the process where new nodes prefer linking to nodes that have a higher fitness, usually measured by high degree. However, a network that is scale free does not necessarily imply preferential attachment (Newman et al. 2006). As evidenced by Figure 3.6, the United States county network approximately follows a lognormal distribution, which implies the

network is not scale free (Newman et al. 2006). Moreover, this lognormal distribution is insufficient to imply preferential attachment. As the degree distribution increases on the horizontal axis, the $Pr(X \geq x)$ for a given function decreases on the vertical axis according to a given power law function.

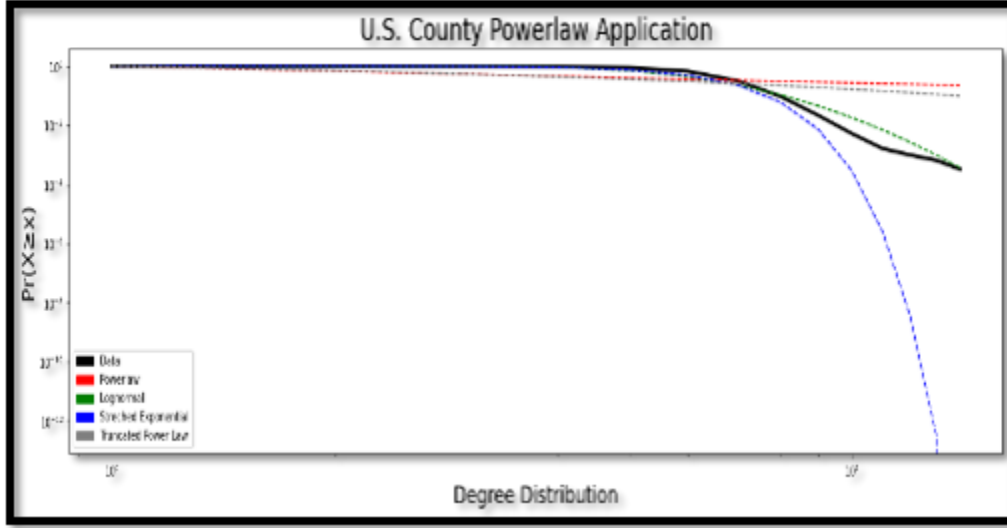


Figure 3.6. Power Law Distribution. The Y axis represents the $Pr(X \geq x)$ for a given power law function and the X axis depicts the degree distribution. As we can see from the figure, the United States county adjacency network follows a lognormal distribution. However, the lognormal distribution is insufficient to imply preferential attachment. Adapted from United States Census Bureau (2010).

Clustering Coefficient

A network's clustering coefficient “measures the average probability that two neighbors of a vertex are themselves neighbors (a measure of the density of triangles in a network)” (Gera 2021a). There are two benchmarks to calculate clustering coefficients: local clustering coefficient and global clustering coefficient (Kemper 2010). To determine a given node i 's local clustering coefficient, one needs to calculate the density at node i , which is formally defined by the following equation (Kemper 2010):

$$C_{li} = \frac{\text{Total number of linked triangles to a given node } i}{\text{Total number of triples centered on node } i}, \quad (3.2)$$

where i is a given node and l is the clustering coefficient. To determine a network's global clustering coefficient, one can take the average of all clustering coefficients (Kemper 2010).

$$C_{gi} = \frac{1}{n} \sum_{i=1} C_i, \quad (3.3)$$

where i is a given vertex and g is the global clustering coefficient.

The United States' county network's global clustering coefficient and local clustering coefficient are 0.435 and 6.094 respectively. As evidenced in Figure 3.7, although counties depicted on the horizontal axis with smaller degrees have higher clustering coefficients on the vertical axis, they also have higher variability. Vertices tend to group in communities, sharing mostly neighbors within the same community and as a result, some vertices have small/large degree based on the size of the community (Newman et al. 2006).

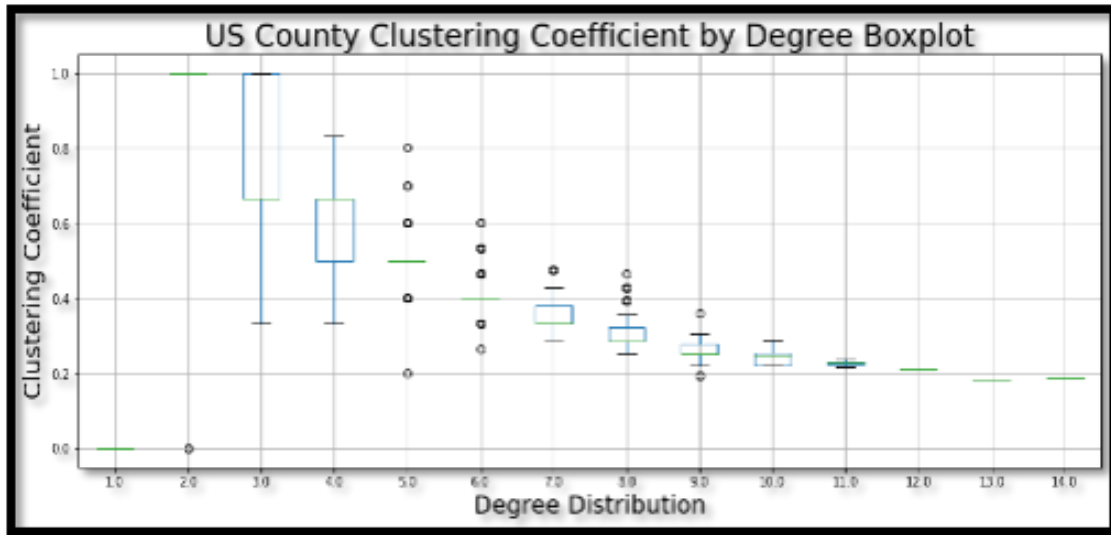


Figure 3.7. Clustering Coefficients. The Y axis represents the clustering coefficients and the X axis depicts the degree distribution. As we can see from the figure, the clustering coefficient is highly variable amongst counties that have degree three to six while the others contain less variability. Adapted from United States Census Bureau (2010).

Assortativity

Homophily, or assortative mixing, is the tendency of nodes to choose nodes with similar characteristics (Newman 2003). Conversely, disassortative mixing occurs as nodes are adjacent to those that are dissimilar. Mixing can occur due to enumerative characteristics such as gender or education. Mixing can also occur due to scalar characteristics such as age or vertex degree. Since degree is a topological property of the network, it reduces down to the Pearson correlation coefficient (Newman et al. 2006). An assortativity network by degree has a core of high degrees and a periphery of low degree while a dissasortative graph contains low degree nodes connected to nodes with high degree. Therefore, a network is assortative if there is a significant fraction of edges between same-type vertices (Zhang et al. 2012).

As illustrated previously by Table 3.2, the average degree of the United States county network is 5.945, which is further confirmed by the average neighbors calculation of 6.24. As a result, the network often creates a triangular lattice in certain geographic locations. Mild dissasortative behavior occurs primarily between nodes with degree six and those nodes (in increasing order of magnitude) with degree seven, five, and eight as illustrated by Figure 3.8. The horizontal axis is the degree distribution and the vertical axis depicts the edge frequency of a given node. Since nodes with degree seven, five, and eight are relatively close to nodes with degree six, the Pearson correlation calculation produces a coefficient just below neutral levels. Therefore, the Pearson correlation coefficient and average neighbors indicates a dissasortative mixing behavior.

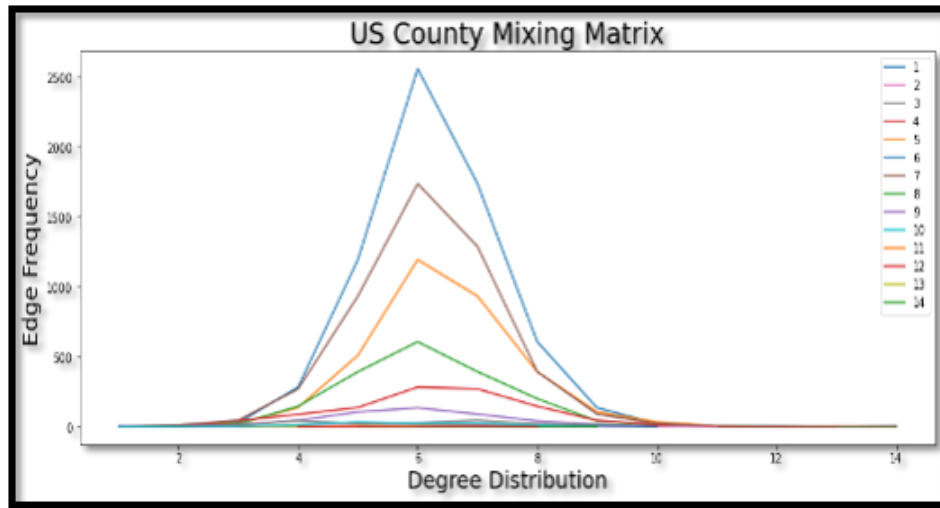


Figure 3.8. Edge Mixing. The Y axis represents the edge frequency of a given node and the X axis depicts the degree distribution. As we can see from the figure, the degree distribution represents dissortative mixing behavior. Adapted from United States Census Bureau (2010).

3.1.5 Network Limitations

Despite the network structure discussed, the United States' county network is not without its limitations. While the entirety of the network is comprised of 3,143 nodes and 9,242 edges, the network is not only disconnected as Alaska and Hawaii are geographically isolated, but three counties in Hawaii are considered not adjacent to Hawaii's other two counties. These three counties do not share any bordering terrain and subsequently any road, path, or trail infrastructure despite being accessible by air and water and are certainly governed by the state. While scale free networks do not necessarily require preferential attachment, the degrees of the U.S. county network follows an approximate lognormal distribution. The counties often self-impose system-level administrative and infrastructure constraints combined with geographic, economic, historical, and population component constraints that increase the underlying complexity of the network.

It is not uncommon for counties to merge, divide, and rename themselves. In Colorado, Broomfield County was most recently established in 2001 by merging parts of other counties (Broomfield County 2021). In 2015, Shannon County in South Dakota renamed itself to Oglala County to respect its Native American heritage (Ban 2014). When new counties are formed or renamed, they are assigned their own FIPS codes, which can initially complicate

reporting and administrative management. However, county, state, and federal governance specifies various protocols to mitigate these constraints and ensure a consistent system behavior.

3.2 Model Implementation

With data regarding COVID-19 cases and deaths collected and network adjacencies identified, we can now fit a time series network model using the GNAR package in R (Leeming et al. 2020). A node in GNAR represents one county and its associated time series data set. As a result, COVID-19 cases or deaths dating from January 22, 2020, to February 2, 2021, for each county are depicted in all 3,143 nodes as a time series of 378 observations.

3.2.1 GNAR Assumptions

The most basic assumption to successfully implement the GNAR package is that the time series data set has an underlying network structure. Since the COVID-19 data set contains time series data for each county and the counties organize into a network, this assumption is satisfied. Although GNAR is able to fit a model with missing data relatively well (Knight et al. 2020), this concern is not applicable as the data set is complete for each county. Since each county contains its own unique time series data, the data structure at node v_i is the same at any other v_j throughout the network where $i \neq j$ (Knight et al. 2020). Since node data is homogeneous throughout the network, this assumption is also satisfied.

The GNAR package assumes that the multivariate time series follows an autoregressive model at each node and is influenced by its neighboring nodes (Knight et al. 2020). This thesis previously assumed autoregressive behavior as discussed in Section 3.1.1. SARS-CoV-2 infections spread across artificial boundaries with relative ease and as a result, COVID-19 cases are likely influenced by a county's adjacent neighbors. Therefore, this modelling assumption is relevant and applicable.

3.2.2 GNAR Adaptation

The GNAR package accepts univariate time series data with an underlying network structure to create an overarching multivariate time series (Knight et al. 2020). However, each individual node v_i can be re-formulated into a simple linear model.

In its simplest form, COVID-19 cases and deaths within a given county are univariate time series data. Univariate data are often used to merely describe and identify fundamental patterns and avoid any cause and effect analysis that may occur in either bivariate or multivariate analysis. The time series data at a given vertex v are represented by total COVID-19 cases or COVID-19 deaths. A third data set can be created by subtracting the total number of COVID-19 deaths from the total number of a county's cases to potentially provide a more accurate representation of current COVID-19 survivors.

While the nodes accept univariate data, the GNAR function accepts binary or weighted edge lengths. This thesis applied two different edge weights for the county network: binary for the presence of a relationship, and great circle distance.

3.2.3 GNAR Limitations

Due to its univariate design, GNAR nodes are unable to retain individual predictors. Univariate design and model fit is the simplest analytical approach as it only contains one variable while a multivariate design could enable a more accurate model with a relevant set of predictors. County data such as population density, census statistics, unemployment, poverty, housing units, and power production could be implemented to enhance the response variable, COVID-19 cases or deaths.

Since the GNAR package assumes stationarity within the data, a user who attempts to model a time series network with a seasonal or cyclical component may achieve substandard results. Models such as Holt-Winters, ARIMA, and seasonal decomposition that are able to model seasonal data are unavailable in the GNAR package (Leeming et al. 2020). Although the data set is relatively small, COVID-19 infections could inherit seasonal, trend, and cyclical properties that cannot be modeled by GNAR in its current form.

3.2.4 Rolling Horizon

Due to the time series nature of COVID-19 infections, we can incorporate a modeling technique that updates its forecast successively using previous and current observations known as rolling horizon or rolling origin as discussed in Section 2.2.1. One can specify the start of the training period and predict a testing period p steps ahead. Additionally, we

can continue to increase the sample size as the data progresses as depicted in Figure 3.9 (Svetunkov 2021).

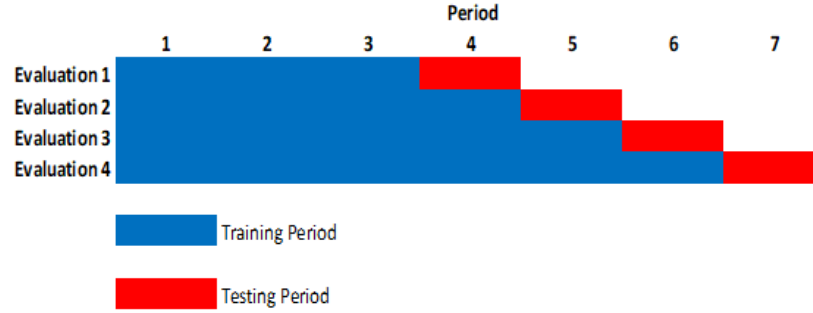


Figure 3.9. As we can see from the figure, the training origin consists of time periods 1, 2, and 3. The testing interval, p , is one period ahead. As the rolling origin progresses after each evaluation, period $t - 1$ is incorporated into the initial sample where t is the evaluated time period.

Although not employed in this thesis, one can determine a constant sample size, c , which truncates the first observation as the origin continues through the data as depicted in Figure 3.10.

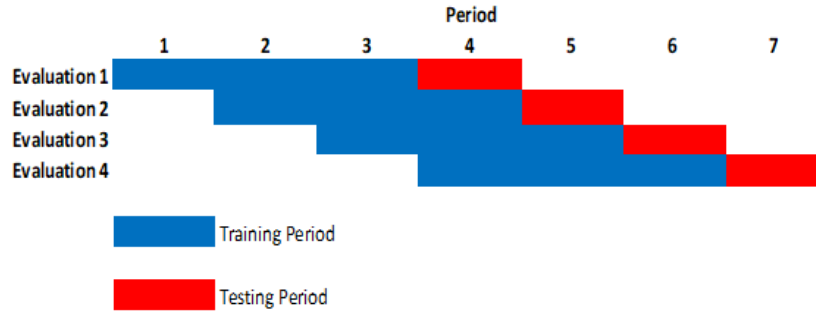


Figure 3.10. As we can see from the figure, the training period consists of time periods 1, 2, and 3. The testing interval, p , is one period ahead. As the rolling origin progresses after each evaluation, the training period is fixed at $c = 3$. As a result, the time period(s) at $t - c - p$ is truncated where t is the testing period.

For a Rolling Horizon model to be accurate, one must establish an initial latency or warm-up period, which can vary from three to five years. Due to its novel nature, COVID-19

infections within the United States have only been reported since early 2020. Moreover, the data are overwhelmingly sparse within the first 70 days of reporting, which can also impact the model's ability to understand the data and develop an accurate prediction.

To generate a capable initial starting point, we supplemented the initial 70 days of reporting with an additional training period of 120 days. As a result, we started the rolling horizon at day 190 or July 29, 2020. Since we collected data from January 22, 2020, to February 2, 2021 and we have 378 total days of data, we can now generate a model for the remaining 188 days. By extension, we can convert the daily time series data into weekly data and start the rolling horizon at week 27 and end at week 54.

Choosing the rolling horizon's testing period or, p , can significantly impact a model's performance. A relatively small p can increase a model's accuracy but it is of limited use. Conversely, a relatively large p , can be useful but often lacks predictive performance. To maintain a balanced approach, we selected time intervals $p = 30$ and $p = 4$ for the daily and weekly models, respectively. Public health officials could utilize a month's forecast to allocate resources, increase protective health measures, or increase public communication during the pandemic's height. Conversely, a month's forecast could be used to repeal restrictive population control measures if cases are on a downward trend.

As discussed in Section 1.5, a Naïve model can be used to measure baseline performance by setting the forecast for any time period equal to the previous period's actual value. Additionally, Naïve models can be incorporated into a rolling horizon by merely lagging the prospective test period by one. Forecasting models that beat the Naïve model are said to have predictive power (Erdem 2021).

3.2.5 Model Parameters

Time series data can be aggregated into weekly or even monthly count to potentially increase the model's accuracy. As previously discussed, the network adjacency can be depicted using a binary or great circle distance structure. Additionally, GNAR can model the number of total cases, total deaths, or the difference to calculate the number of survivors who are alive and COVID-19 positive. Table 3.4 provides a summary of various model parameters, which can be formed into various combinations.

Table 3.4. Model Parameter Summary. There are two Rolling Horizon Forecast model options (daily and weekly) , two Network Adjacency Structure model options (binary and great circle distance), and three Response model options (total infections, deaths, and survivors). In total, we create 12 different combinations (i.e. Daily Rolling Horizon Forecast - Binary Network Adjacency Structure - Total Infections).

Rolling Horizon Forecast	Network Adjacency Structure	Response
Daily (30 days)	Binary	Total Infections
Weekly (4 weeks)	Great Circle Distance	Total Deaths
		Survivors

As per Table 3.4, there are 12 total model combinations to choose for our modeling. Within each choice combination, we can adjust individual GNAR parameters to create three different model fits. 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 neighbour set to model at each of the time-lags (Leeming et al. 2020). The second model fit utilizes the same $\alphaOrder$ but applies a $\betaOrder = 1$. The third model fit is the default model with no parameter modifications.

3.3 Summary

Network science is an evolving field growing at an exceptional pace (Newman et al. 2006). Network science enables researchers to gain a deeper understanding of a given network. Network characteristics such as community detection, small world properties, eccentricities, and centrality measures provide insights into a network's structure.

The United States county network centrality measures enabled insight and a keen understanding of critical nodes. Closeness centrality represented geographically central counties, betweenness centrality represented the counties who had shortest paths between coasts, and eigenvector centrality represented key states that exerted influence of the network. Additionally, the Pearson correlation coefficient and average neighbors indicates dissasortative mixing.

The GNAR package enables univariate data modeling over a given network. We can forecast daily or weekly COVID-19 cases, survivors, or deaths for the United States using a binary adjacency structure or a great circle distance adjacency structure. We can use the aforementioned 12 combinations each with 3 GNAR models to create 36 individual models and evaluate their prediction accuracy. One individual model with high accuracy and prediction capability could be adapted and used by public health officials to disrupt an epidemic's spread.

CHAPTER 4: Model Results

4.1 Introduction

As Sections 4.2 and 4.3 will show, there are two reoccurring patterns. Since a county's survivor count is merely the difference between its case and death counts combined COVID-19's low mortality rate, the county's survivor count is approximately equal to its case count. Consequently, the MASE time series plots for cases and survivors across both daily and weekly time interval model combinations mirror one another.

Additionally, Sections 4.2 and 4.3 will show that MASE's accuracy increases and trends towards zero for most GNAR combinations as the model learns and subsequently forecasts. As discussed in subsection 2.2.2, MASE and MAPE values that are closer to 0 represent more accurate models (Hyndman et al. 2006). However, at approximately the 350th day, or 47th week, the MASE within some GNAR combinations decreases as it trends upwards.

As we will see in subsections 4.3.1, 4.3.2, 4.3.5, and 4.3.6, the MASE for GNAR model 3 in both cases and survivors over a weekly time interval maintain and improve their accuracy and forecasting capability thus rebuking the aforementioned trend behavior. The MASE for GNAR models 1 and 3 decrease and trend upward. Despite these trending behaviors, each model combination and their subsequent GNAR models still outperform the Naïve model. Therefore, the model combinations and GNAR models that will be depicted in Sections 4.2 and 4.3 should be considered and adapted for future use by researchers and epidemiologists.

4.2 Daily Results

Since the rate of change between a daily time interval is much smaller than compared to the weekly time interval, we experienced higher variability amongst the case, death, and survivor model combinations as evidenced in subsections 4.2.1, 4.2.2, 4.2.3, 4.2.4, 4.2.5, and 4.2.6. Despite this variability, the model combinations in subsection 4.2.3 and 4.2.4 performed exceptionally well and achieved a minimum MAPE values of approximately 10% shown in Figures 4.12 and 4.16.

As Figures 4.2, 4.6, 4.18, and 4.22 will show, cases and survivors mirror one another and MASE decreases at approximately the 350th day and trends upwards as discussed in Section 4.1. However, case and survivor model combinations experience minimum MAPE values of approximately 7% as we will see in Figures 4.4, 4.8, 4.20, and 4.24. These results demonstrated that researchers and public health officials can employ the GNAR package to model COVID-19 cases, deaths, and survivors for the United States.

4.2.1 Daily Time Interval - Binary Network - Cases

As we can see from Table 4.1, the MASE and MAPE for GNAR model 1 provide higher accuracy and forecasting capability than GNAR models 2 and 3 as it has the lowest mean and median. This is further evidenced in Figures 4.1 and 4.3. As a result, GNAR Model 1 is the most consistently accurate model for this parameter combination. Table 4.1 also depicts low variability for both MASE and MAPE within each GNAR model.

Table 4.1 also displays the MASE mean and median of each GNAR model outperforming the Naïve model. Moreover, Figure 4.2 portrays the MASE for all GNAR models outperforming the Naïve model but experiencing a significant increase towards the end of the data set as discussed in Section 4.1. This increase could be each GNAR model's failure to adapt as cases began to trend downward. However, Figure 4.4 demonstrates MAPE consistency of each GNAR model to perform at or below a 20% threshold. Additionally, MAPE continues to trend downwards for nearly the entire data set. As a result, all three GNAR models surpassed the performance standards established in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.1. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 188-day forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021a); United States Census Bureau (2010).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.4375	0.4322	0.0084
Model 2	0.5161	0.5281	0.0089
Model 3	0.5024	0.4585	0.0135
Naïve	1	1	0
MAPE			
Model 1	0.1145	0.1134	0.0007
Model 2	0.1372	0.1308	0.0010
Model 3	0.1478	0.1366	0.0022
Naïve	0.6658	0.6723	0.0019

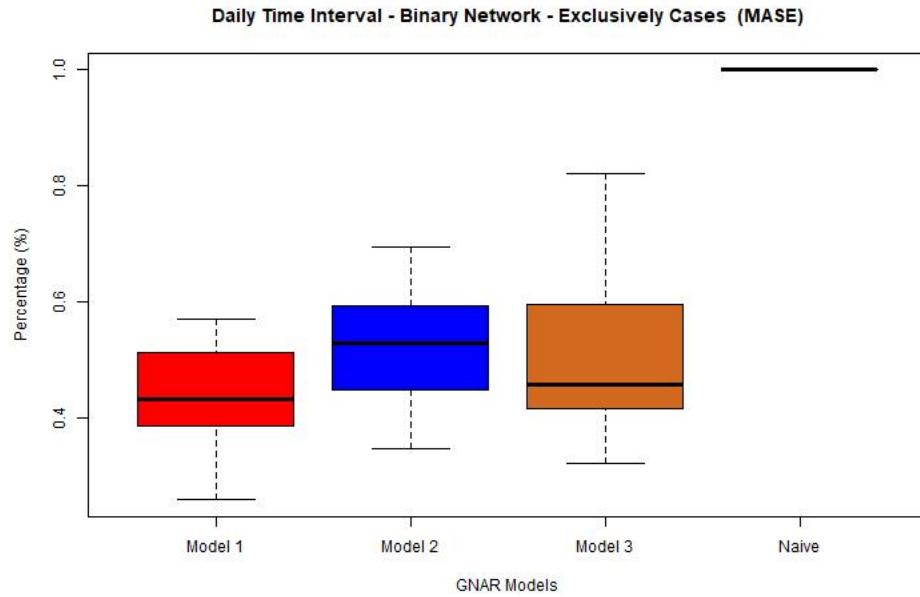


Figure 4.1. Daily Time Interval - Binary Network - Cases (MASE)
The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see from the figure above, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021a); United States Census Bureau (2010).

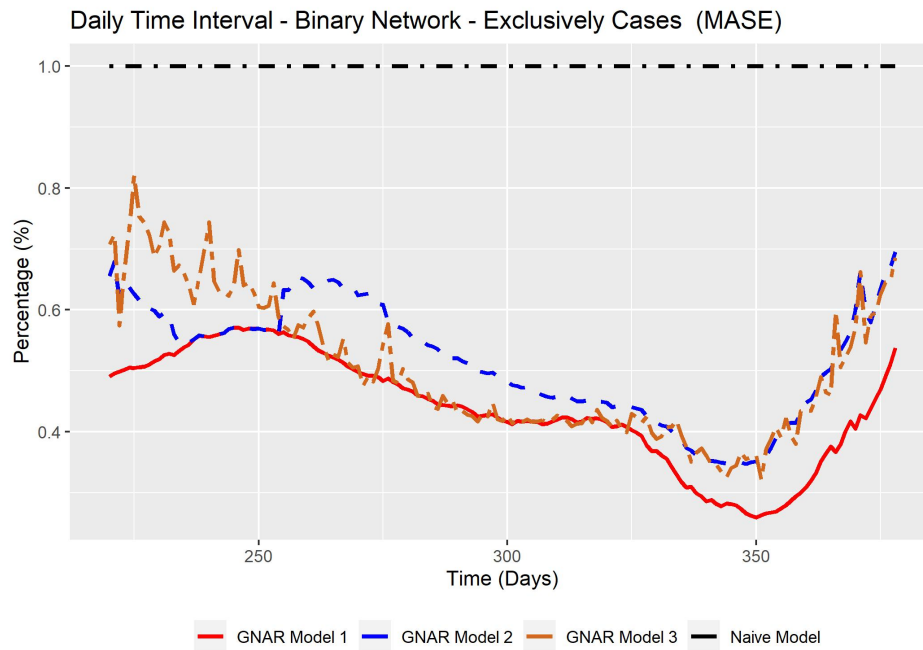


Figure 4.2. Daily Time Interval - Binary Network - Cases (MASE)
The vertical axis represents MASE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. While each GNAR model outperforms the Naïve model, MASE decreases at approximately the 350th day and trends upwards for all three models. Adapted from USAFacts (2021a); United States Census Bureau (2010).

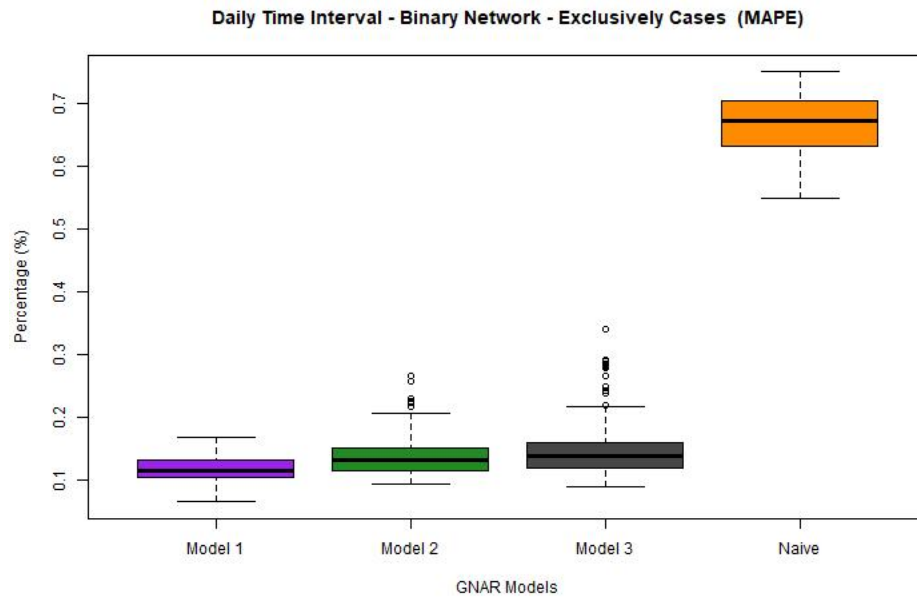


Figure 4.3. Daily Time Interval - Binary Network - Cases (MAPE)
The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see from the figure above, there is relatively low variability amongst each GNAR model, which results in more accurate forecasts. Adapted from USAFacts (2021a); United States Census Bureau (2010).

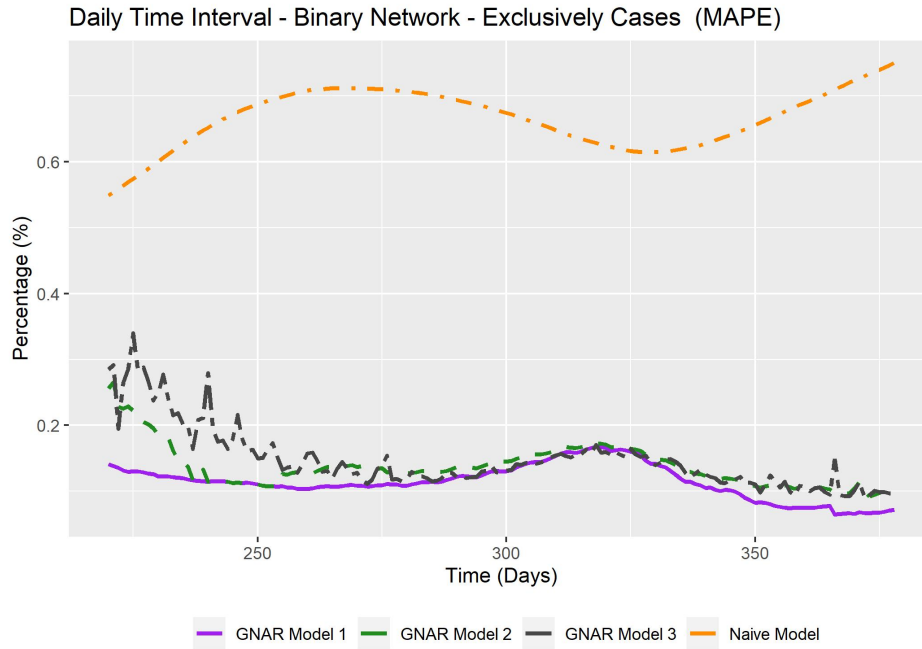


Figure 4.4. Daily Time Interval - Binary Network - Cases (MAPE)
The vertical axis represents MAPE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see from the figure above, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021a); United States Census Bureau (2010).

4.2.2 Daily Time Interval - Great Circle Network - Cases

As we can see from Table 4.2 and Figures 4.5, 4.6, 4.7, and 4.8, each GNAR model performed similar to its binary network counterpart in Section 4.2.1. Moreover, the MASE of each model outperformed the Naïve model and could be used to accurately forecast COVID-19 cases. GNAR model 1 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, model 1 is the most consistently accurate model for this parameter combination. The MASE mean and median for each GNAR model outperform the Naïve model

As Figure 4.5 portrays, MASE variability amongst the GNAR models varies, which decreases their forecasting capabilities. Moreover, Figure 4.6 portrays, the MASE for all GNAR models, which outperform the Naïve model. However, each GNAR model in Figure

4.6 experiences a significant increase towards the end of the data set as discussed in Section 4.1. Figure 4.7 displays MAPE characteristics, which contain relatively small variability thus demonstrating their forecasting capabilities. Moreover, Figure 4.8 illustrates MAPE for each GNAR model continuing to trend downwards for nearly the entire data set. As a result, all three GNAR models surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.2. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 188-day forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.4375	0.4322	0.0084
Model 2	0.5176	0.5200	0.0106
Model 3	0.5169	0.4641	0.0178
Naïve	1	1	0
MAPE			
Model 1	0.1145	0.1134	0.0007
Model 2	0.1428	0.1313	0.0019
Model 3	0.1543	0.1391	0.0031
Naïve	0.6658	0.6723	0.0019

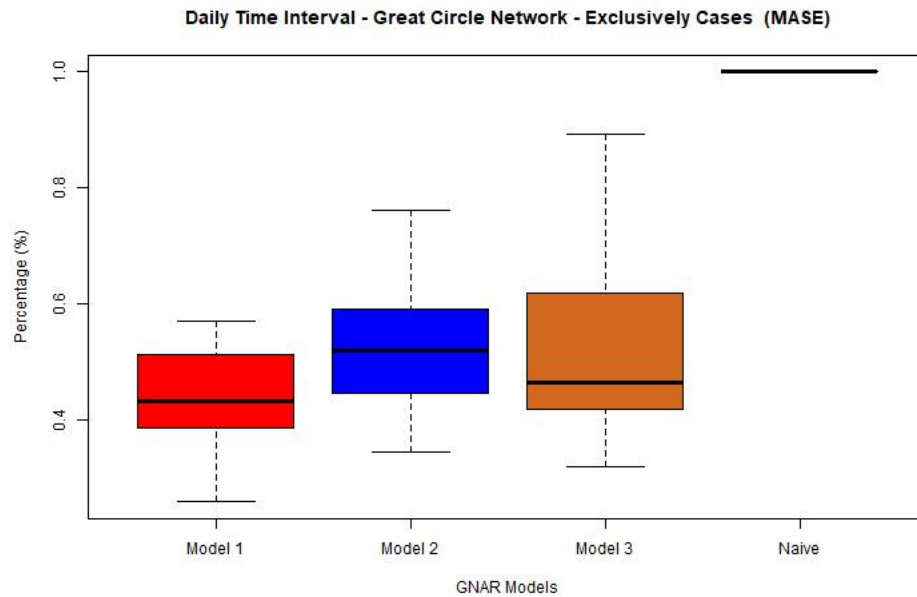


Figure 4.5. Daily Time Interval - Great Circle Network - Cases (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

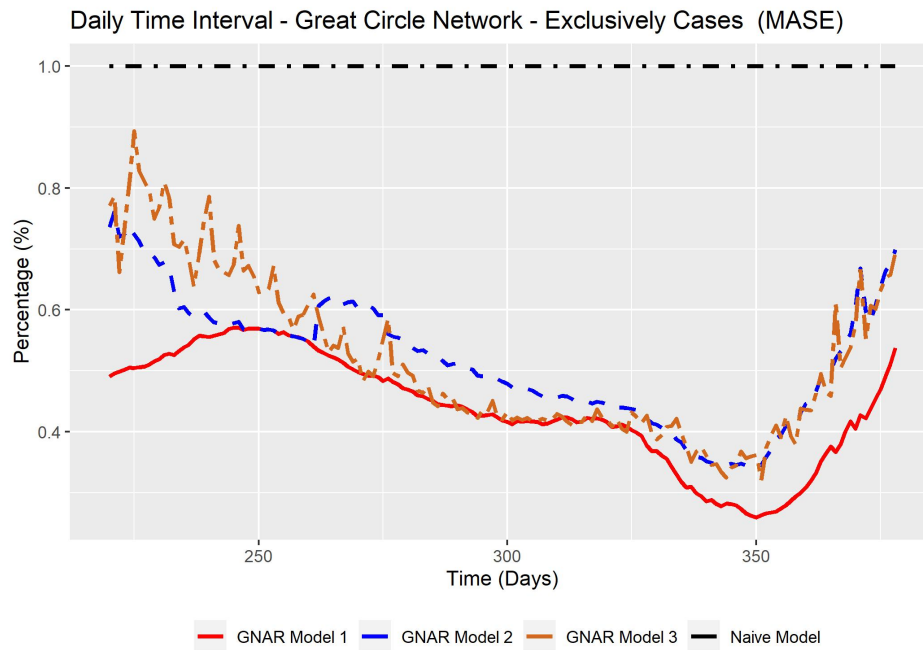


Figure 4.6. Daily Time Interval - Great Circle Network - Cases (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. While each GNAR model outperforms the Naïve model, MASE decreases at approximately the 350th day and trends upwards for all three models. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

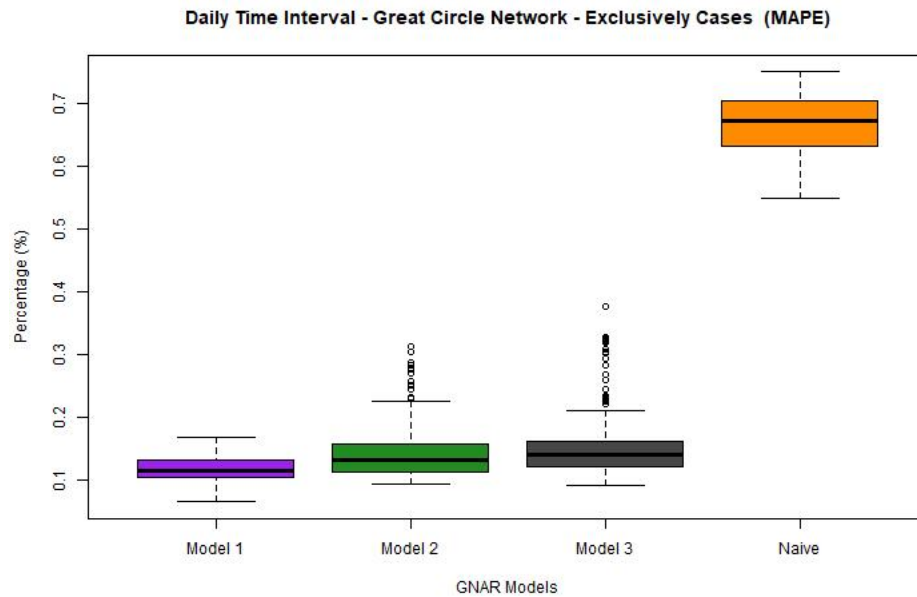


Figure 4.7. Daily Time Interval - Great Circle Network - Cases (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which results in more accurate forecasts. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

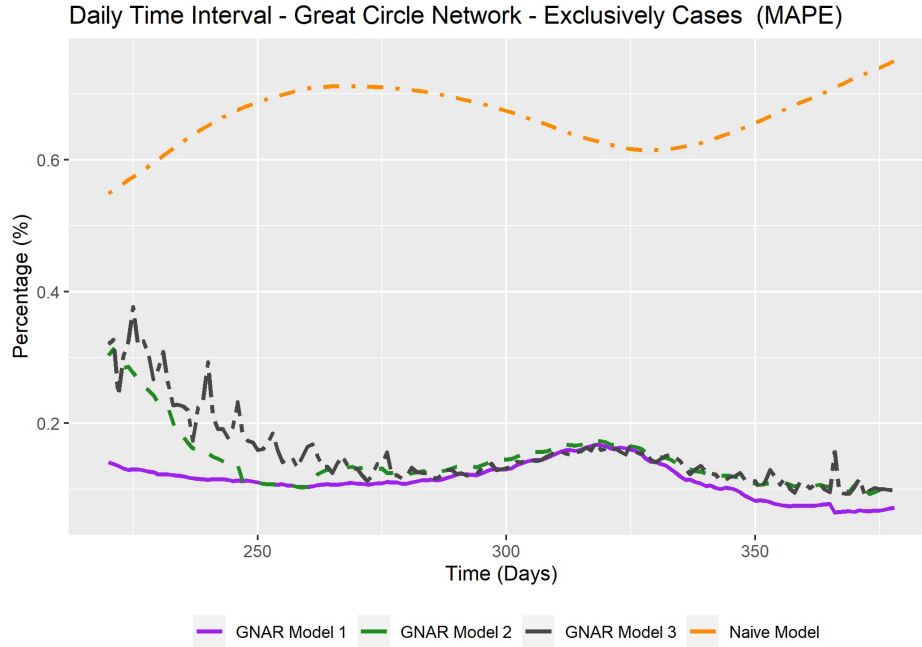


Figure 4.8. Daily Time Interval - Great Circle Network - Cases (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

4.2.3 Daily Time Interval - Binary Network - Deaths

As we can see from Table 4.3 and Figures 4.9, 4.10, 4.11, and 4.12, the MASE of each GNAR model outperformed the Naïve model and could be used to accurately forecast COVID-19 deaths. GNAR model 3 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, GNAR model 3 is the most consistently accurate model for this parameter combination.

As Figure 4.9 portrays, MASE variability amongst the GNAR models varies significantly, which decreases their forecasting capabilities. The high MASE variability is a direct result of the model learning. As the model continues to improve and increase its accuracy, MASE continues to produce lower and lower values, which are further away from the initial starting point. Moreover, Figure 4.10 portrays, the MASE for all GNAR models, which outperform

the Naïve model. However, each GNAR model in Figure 4.10 continues to trend downwards. Figure 4.11 displays MAPE characteristics, which contain relatively small variability thus demonstrating their forecasting capabilities. Moreover, Figure 4.12 illustrates MAPE for each GNAR model mirroring one another and continuing to trend downwards for nearly the entire data set. As a result, all three GNAR models surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.3. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 188-day forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); United States Census Bureau (2010).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.6173	0.5956	0.0303
Model 2	0.6157	0.5953	0.0310
Model 3	0.5932	0.5894	0.0248
Naïve	1	1	0
MAPE			
Model 1	0.1551	0.1542	0.0006
Model 2	0.1558	0.1542	0.0006
Model 3	0.1560	0.1543	0.0006
Naïve	0.7010	0.7048	0.0012

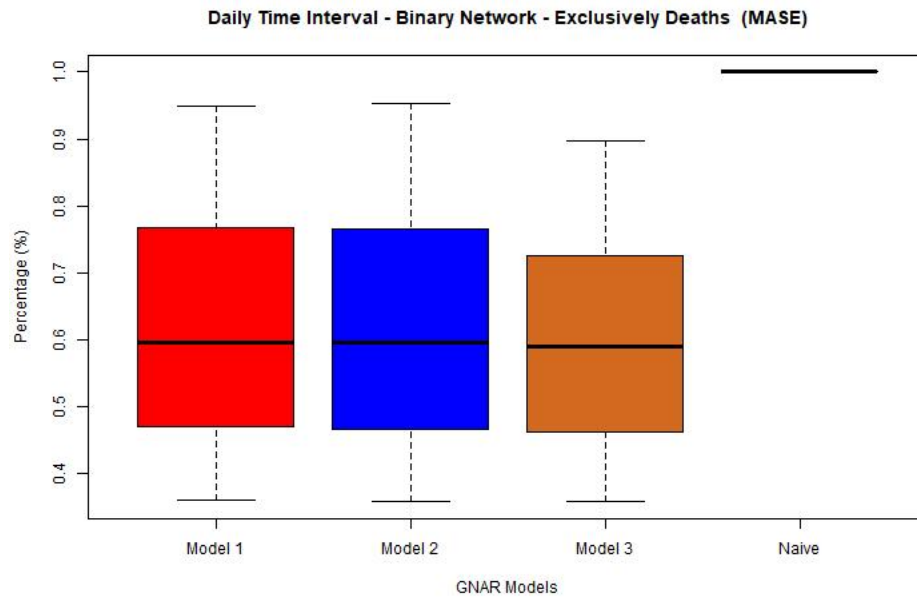


Figure 4.9. Daily Time Interval - Binary Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010).

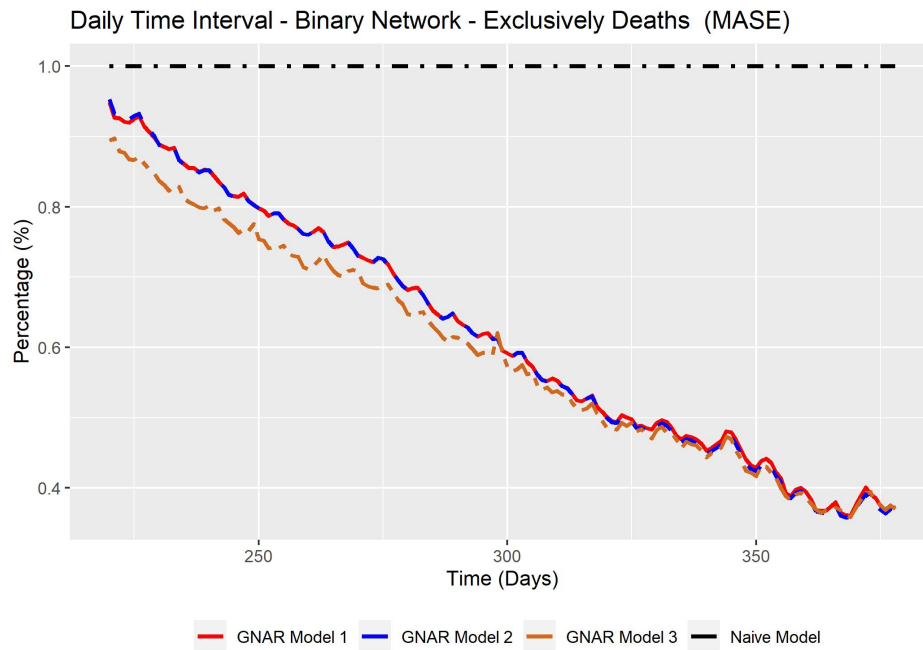


Figure 4.10. Daily Time Interval - Binary Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, each GNAR model continues to trend downwards. Given additional data, each GNAR model may continue to fall thus strengthening its predictive performance. Adapted from USAFacts (2021b); United States Census Bureau (2010).

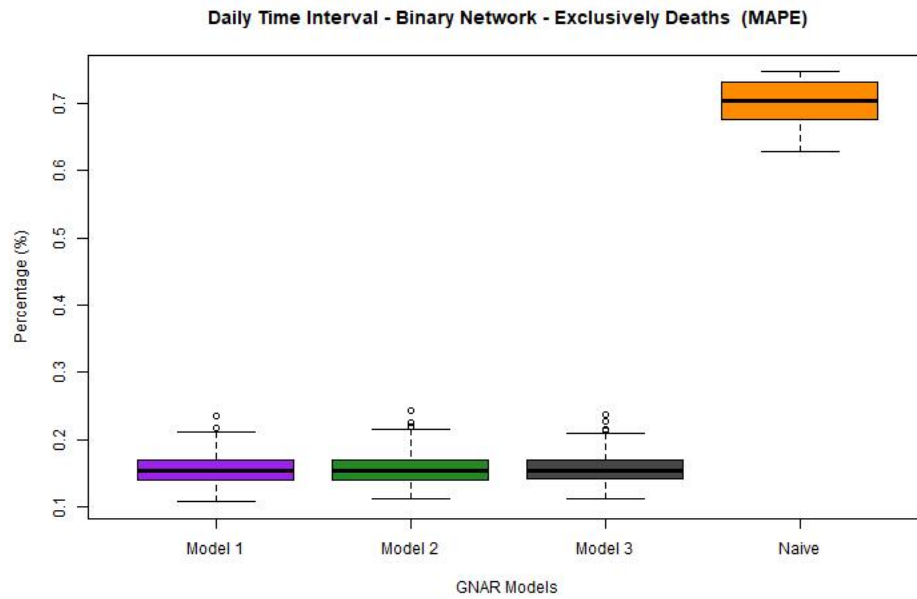


Figure 4.11. Daily Time Interval - Binary Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010).

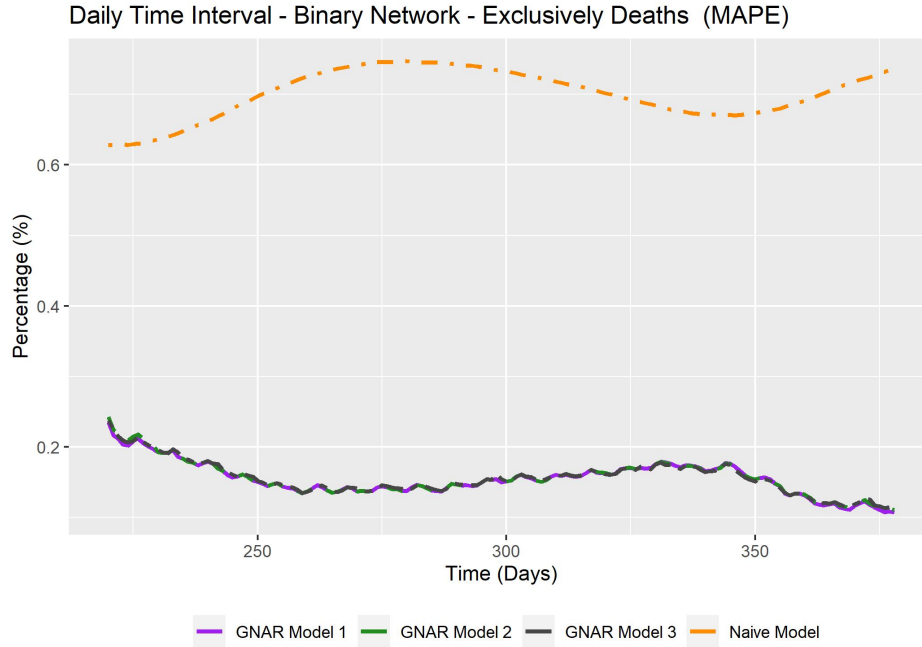


Figure 4.12. Daily Time Interval - Binary Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); United States Census Bureau (2010).

4.2.4 Daily Time Interval - Great Circle Network - Deaths

As we can see from Table 4.4 and Figures 4.13, 4.14, 4.15, and 4.16, each GNAR model performed similar to its binary network counterpart in Section 4.2.3. Moreover, the MASE of each model outperformed the Naïve model and could be used to accurately forecast COVID-19 deaths. GNAR model 3 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, model 3 is the most consistently accurate model for this parameter combination. The MASE mean and median for each GNAR model outperform the Naïve model

As Figure 4.13 portrays, MASE variability amongst the GNAR models varies significantly, which decreases their forecasting capabilities. The high MASE variability is a direct result of the model learning. As the model continues to improve and increase its accuracy, MASE

continues to produce lower and lower values, which are further away from the initial starting point. Moreover, Figure 4.14 portrays, the MASE for all GNAR models, which outperform the Naïve model. However, each GNAR model in Figure 4.14 continues to trend downwards. Figure 4.15 displays MAPE characteristics, which contain relatively small variability thus demonstrating their forecasting capabilities. Moreover, Figure 4.16 illustrates MAPE for each GNAR model mirroring one another and continuing to trend downwards for nearly the entire data set. As a result, all three GNAR models surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.4. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 188-day forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.6173	0.5956	0.0303
Model 2	0.6158	0.5955	0.0308
Model 3	0.5932	0.5899	0.0246
Naïve	1	1	0
MAPE			
Model 1	0.1551	0.1542	0.0006
Model 2	0.1553	0.1542	0.0005
Model 3	0.1557	0.1544	0.0005
Naïve	0.7010	0.7048	0.0012

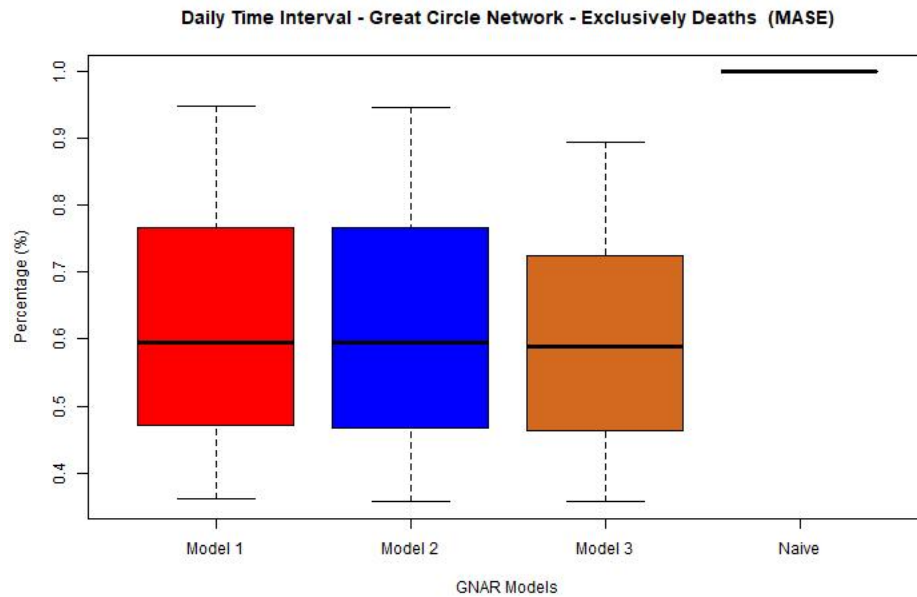


Figure 4.13. Daily Time Interval - Great Circle Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

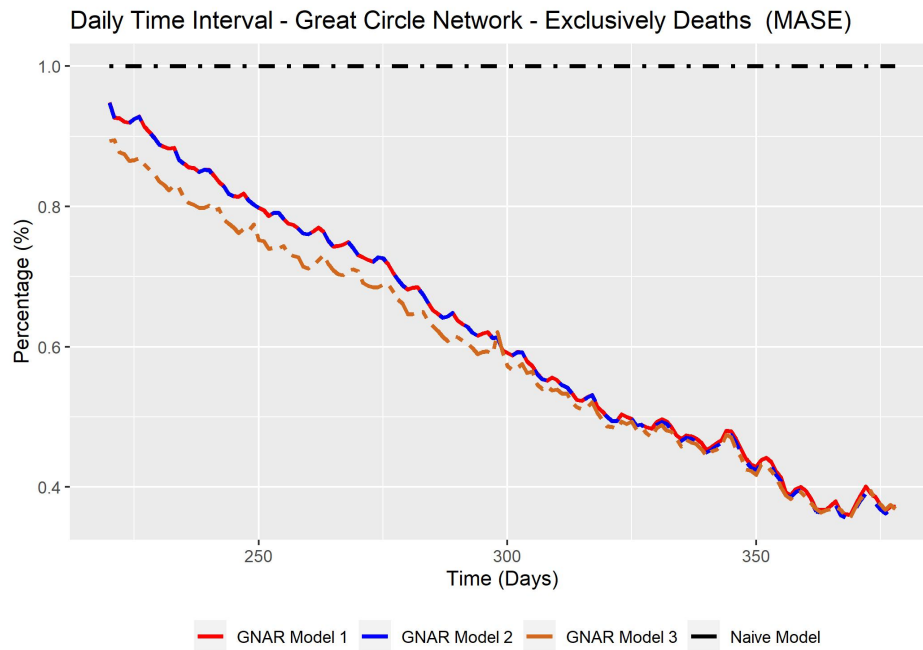


Figure 4.14. Daily Time Interval - Great Circle Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, each GNAR model continues to trend downwards. Given additional data, each GNAR model may continue to fall thus strengthening its predictive performance. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

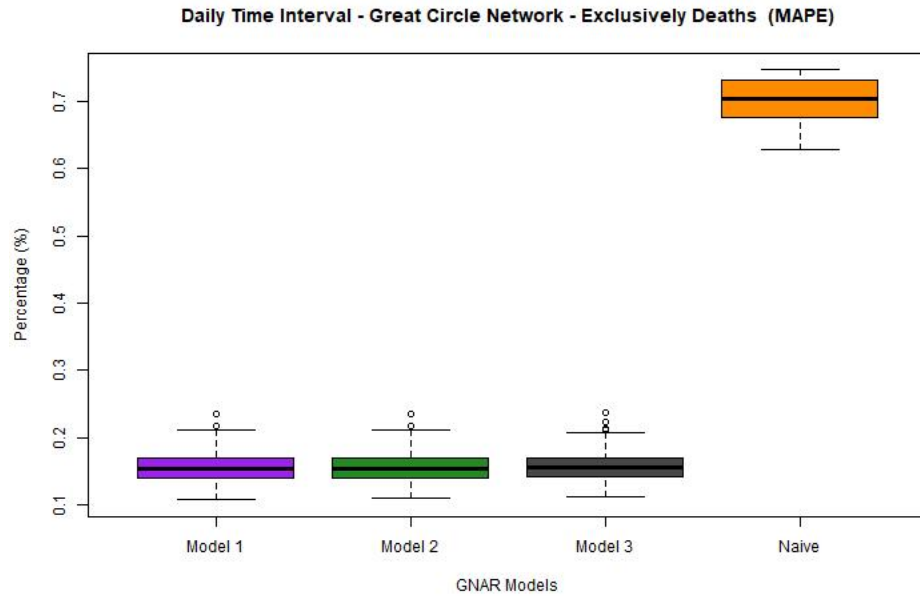


Figure 4.15. Daily Time Interval - Great Circle Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

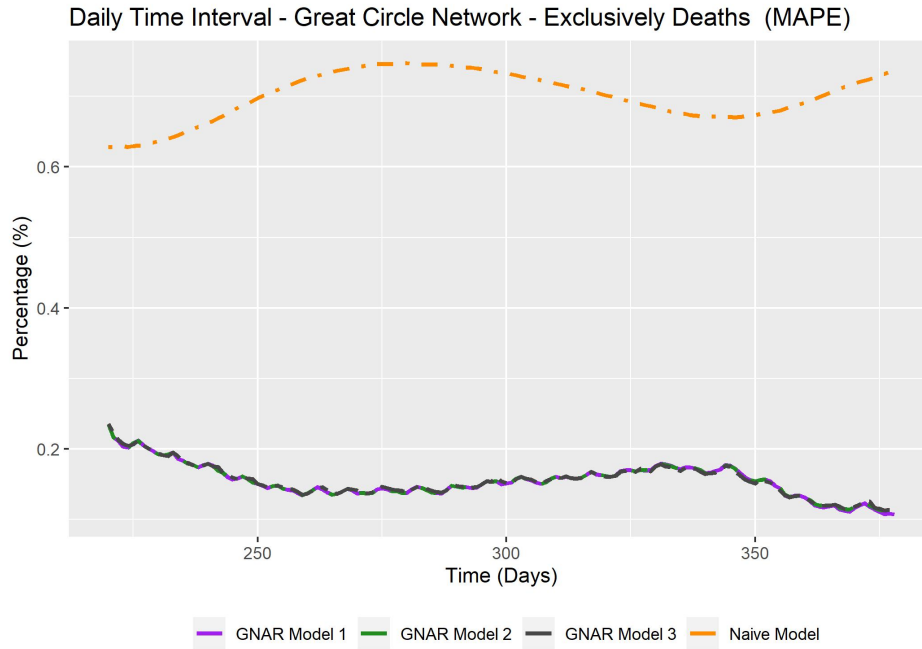


Figure 4.16. Daily Time Interval - Great Circle Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

4.2.5 Daily Time Interval - Binary Network - Survivors

As we can see from Table 4.5, the MASE and MAPE for GNAR model 1 provide higher accuracy and forecasting capability than GNAR models 2 and 3 as it has the lowest mean and median. This is further evidenced in Figures 4.17 and 4.19. As a result, GNAR Model 1 is the most consistently accurate model for this parameter combination. Table 4.5 also depicts low variability for both MASE and MAPE within each GNAR model.

Table 4.5 also displays the MASE mean and median of each GNAR model outperforming the Naïve model. Moreover, Figure 4.18 portrays the MASE for all GNAR models outperforming the Naïve model but experiencing a significant increase towards the end of the data set as discussed in Section 4.1. This increase could be each GNAR model's failure to adapt as cases began to trend downward. However, Figure 4.20 demonstrates MAPE consistency

of each GNAR model to perform at or below a 20% threshold. Additionally, MAPE continues to trend downwards for nearly the entire data set. Also discussed in Section 4.1 and evidenced throughout Table 4.5 as well as Figures 4.17, 4.18, 4.19, and 4.20, MASE and MAPE for this survivor model mirrors its counterpart in subsection 4.2.1. As a result, all three GNAR models surpassed the performance standards established in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.5. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 188-day forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.4595	0.4609	0.0099
Model 2	0.5478	0.5663	0.0121
Model 3	0.5195	0.4889	0.0141
Naïve	1	1	0
MAPE			
Model 1	0.1154	0.1139	0.0007
Model 2	0.1386	0.1360	0.0008
Model 3	0.1453	0.1323	0.0020
Naïve	0.6653	0.6724	0.0019

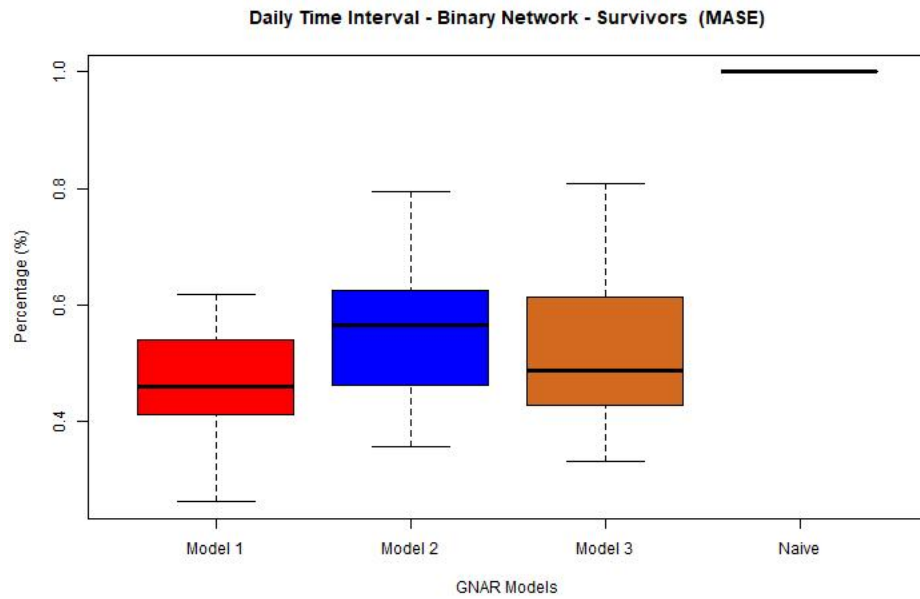


Figure 4.17. Daily Time Interval - Binary Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010).

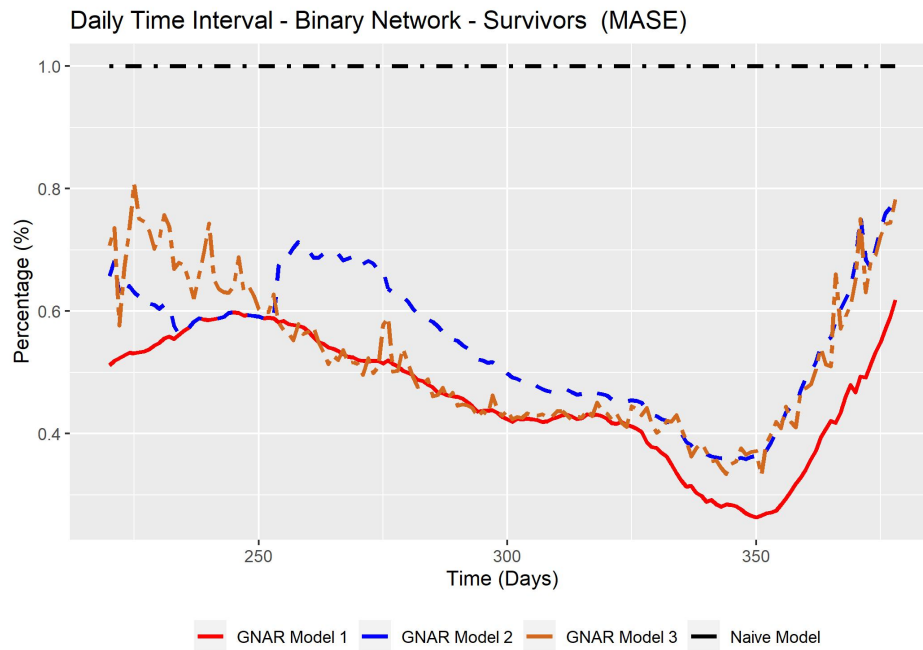


Figure 4.18. Daily Time Interval - Binary Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. While each GNAR model outperforms the Naïve model, MASE decreases at approximately the 350th day and trends upwards for all three models. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010).

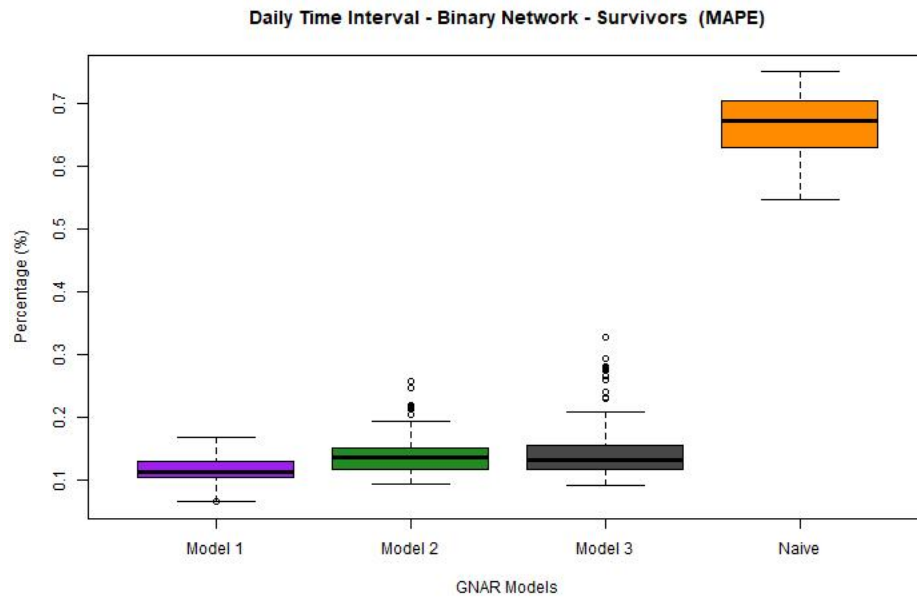


Figure 4.19. Daily Time Interval - Binary Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010).

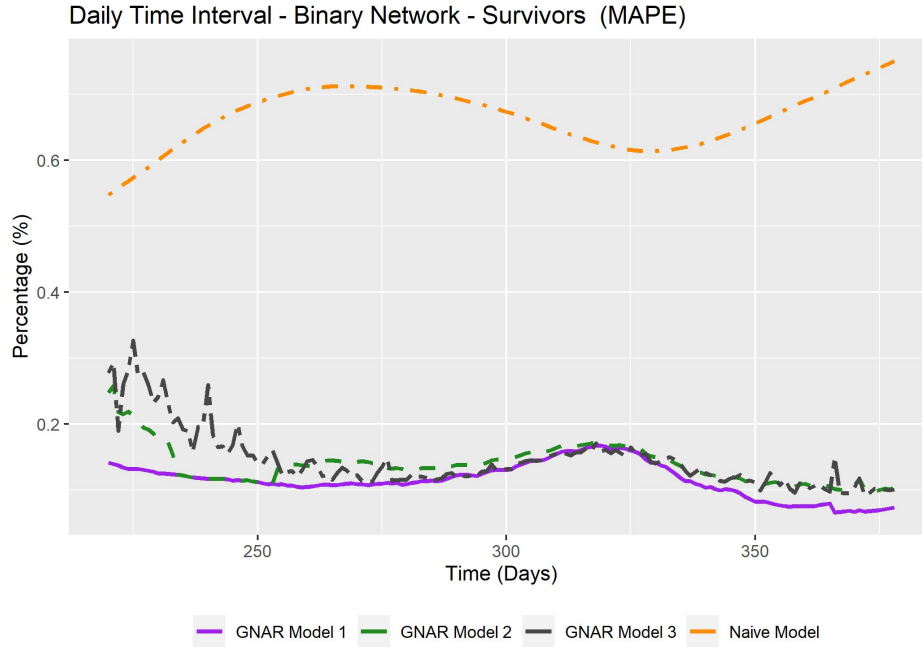


Figure 4.20. Daily Time Interval - Binary Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010).

4.2.6 Daily Time Interval - Great Circle Network - Survivors

As we can see from Table 4.6 and Figures 4.21, 4.22, 4.23, and 4.24, each GNAR model performed similar to its binary network counterpart in Section 4.2.5. The MASE of each model outperformed the Naïve model as depicted in Figure 4.22 and could be used to accurately forecast COVID-19 survivors. GNAR model 1 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, model 1 is the most consistently accurate model for this parameter combination.

As Figure 4.21 portrays, MASE variability amongst the GNAR models varies, which decreases their forecasting capabilities. However, each GNAR model in Figure 4.22 experiences a significant increase towards the end of the data set as discussed in Section 4.1. Figure 4.23 displays MAPE characteristics, which contain relatively small variability thus

demonstrating their forecasting capabilities. Moreover, Figure 4.24 illustrates MAPE for each GNAR model continuing to trend downwards for nearly the entire data set. Also discussed in Section 4.1 and evidenced throughout Table 4.6 as well as Figures 4.21, 4.22, 4.23, and 4.24, MASE and MAPE for this survivor model mirrors its counterpart in subsection 4.2.1. As a result, all three GNAR models surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 survivors.

Table 4.6. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 188-day forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.4595	0.4609	0.0099
Model 2	0.5512	0.5677	0.0133
Model 3	0.5346	0.5001	0.0186
Naïve	1	1	0
MAPE			
Model 1	0.1154	0.1139	0.0007
Model 2	0.1437	0.1371	0.0016
Model 3	0.1513	0.1357	0.0028
Naïve	0.6653	0.6724	0.0019

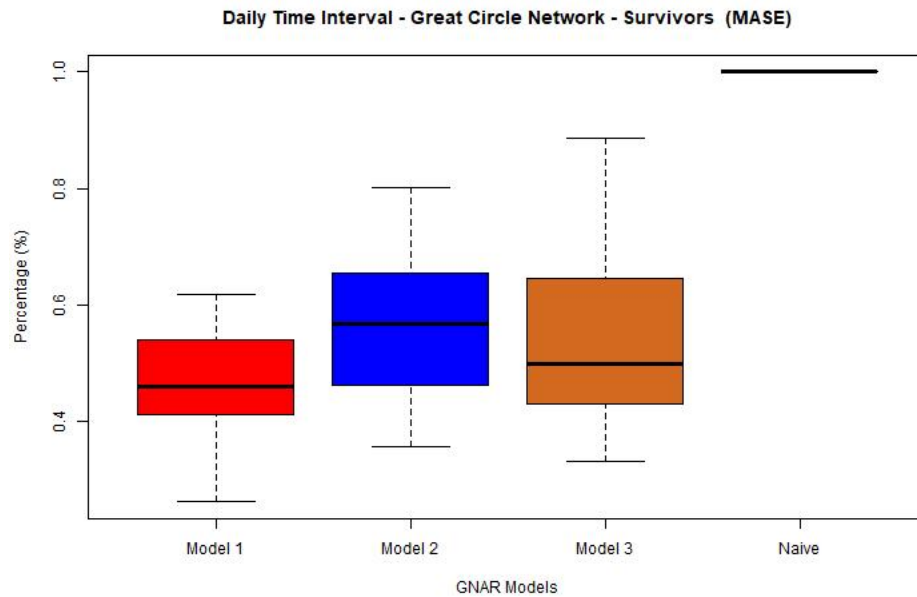


Figure 4.21. Daily Time Interval - Great Circle Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

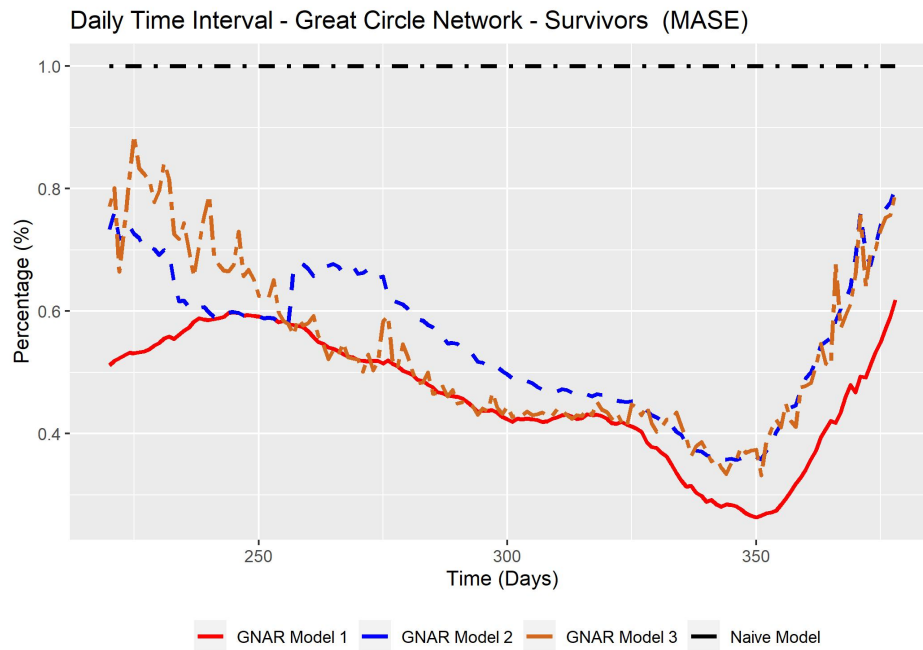


Figure 4.22. Daily Time Interval - Great Circle Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. While each GNAR model outperforms the Naïve model, MASE decreases at approximately the 350th day and trends upwards for all three models. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

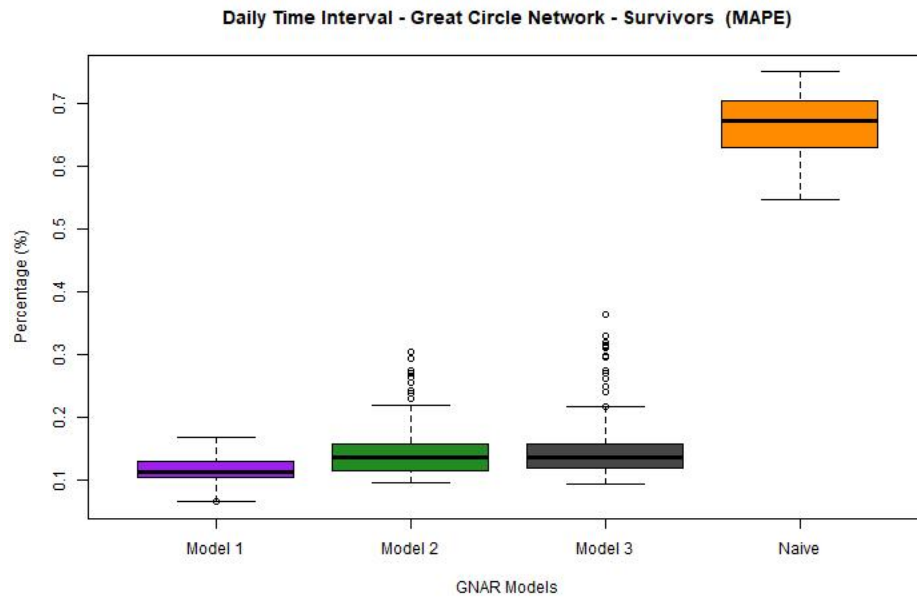


Figure 4.23. Daily Time Interval - Great Circle Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

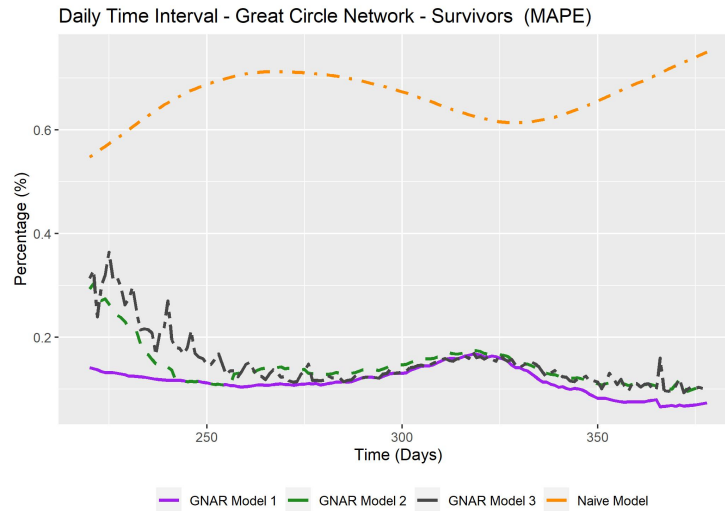


Figure 4.24. Daily Time Interval - Great Circle Network - Survivors (MAPE)
 The vertical axis represents MAPE percentage and the horizontal axis depicts the number of days since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

4.2.7 Daily Results Summary

As the results in Section 4.2 demonstrate, MASE experiences variability throughout the data set. The data points within cases and survivors are approximately equal, which results in similar MASE and MAPE characteristics. The MASE and MAPE within Sections 4.2.3 and 4.2.4 continued to trend downwards thus demonstrating their forecasting capability. While each model combination outperforms its respective Naïve model, MASE begins to trend upwards at approximately the 350th day. Ultimately, all model combinations analyzed in Section 4.2 surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases, deaths, and survivors.

4.3 Weekly Results

As discussed in Section 3.2.4, daily time series data can be converted to weekly, which can potentially increase a model's performance as the data values may increase in greater intervals. As evidenced by the following subsections, the weekly adaption proved to be

equally accurate as its daily counterpart. As a result, public health officials can effectively forecast county COVID-19 cases, infections, and survivors for each county on a weekly basis.

As Figures 4.26, 4.30, 4.42, and 4.46 will show, GNAR models 1 and 2 for both case and survivor model combinations mirror one another. Additionally, MASE for GNAR models 1 and 2 decreases at approximately the 48th week and trends upwards as discussed in Section 4.1. However, case and survivor model combinations experience minimum MAPE values of approximately 5% as we will see in Figures 4.4, 4.8, 4.20, and 4.24. These results demonstrated that researchers and public health officials can also employ the GNAR package to model COVID-19 cases, deaths, and survivors for the United States.

4.3.1 Weekly Time Interval - Binary Network - Cases

As we can see from Table 4.7, the MASE and MAPE for GNAR model 3 provide higher accuracy and forecasting capability than GNAR models 1 and 2 as it has the lowest mean and median. This is further evidenced in Figures 4.25 and 4.27. As a result, GNAR Model 3 is the most consistently accurate model for this parameter combination. Table 4.7 also depicts low variability for both MASE and MAPE within each GNAR model.

Table 4.7 also displays the MASE mean and median of each GNAR model outperforming the Naïve model. Moreover, Figure 4.26 portrays the MASE for all GNAR models outperforming the Naïve model. GNAR models 1 and 2 experience a significant increase towards the end of the data set as discussed in Section 4.1 but GNAR model 3 continues to increase its accuracy as it trends downwards. The increase for GNAR models 1 and 2 could be each GNAR model's failure to adapt as cases began to trend downward. However, Figure 4.28 demonstrates MAPE consistency of each GNAR model to perform at or below a 20% threshold. Additionally, MAPE continues to trend downwards for nearly the entire data set. As a result, all three GNAR models surpassed the performance standards established in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.7. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 27-week forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021a); United States Census Bureau (2010).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.5380	0.5392	0.0127
Model 2	0.6505	0.6592	0.0144
Model 3	0.3933	0.3910	0.0045
Naïve	1	1	0
MAPE			
Model 1	0.1315	0.1264	0.0012
Model 2	0.1642	0.1478	0.0036
Model 3	0.1170	0.1145	0.0007
Naïve	0.6751	0.6868	0.0033

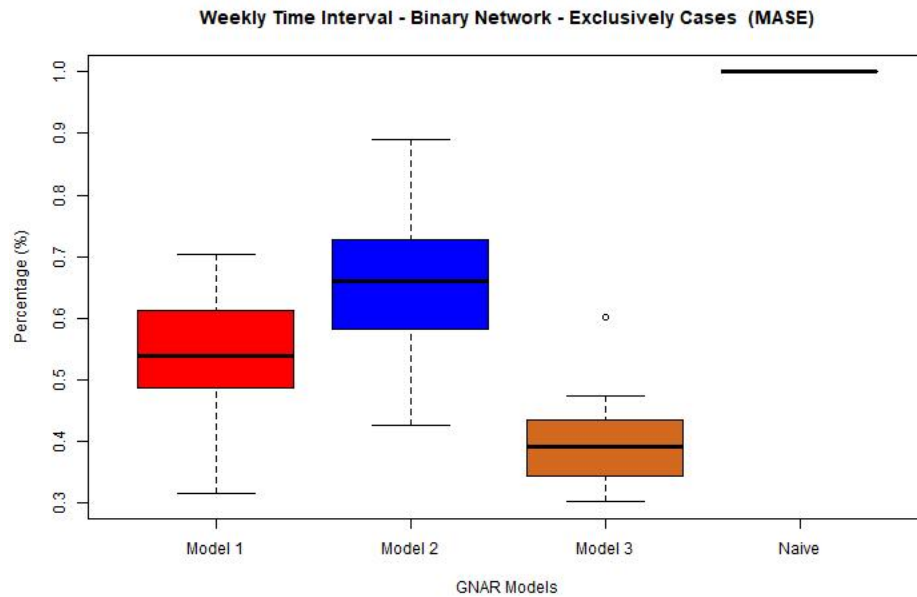


Figure 4.25. Weekly Time Interval - Binary Network - Cases (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is moderate variability amongst GNAR models 1 and 2 and low variability within GNAR model 3. Adapted from USAFacts (2021a); United States Census Bureau (2010).

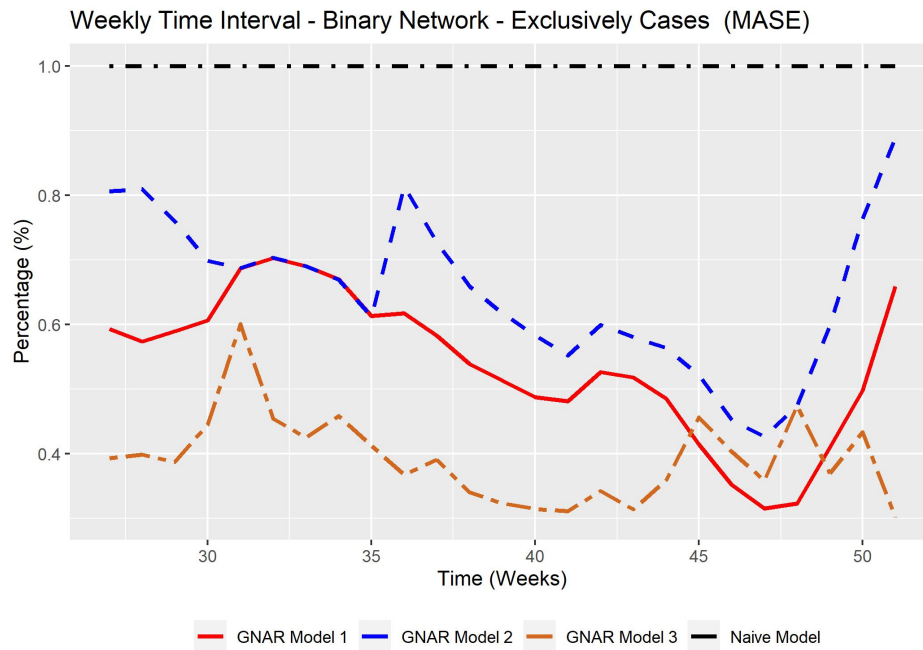


Figure 4.26. Weekly Time Interval - Binary Network - Cases (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. While each GNAR model outperforms the Naïve model, MASE decreases until approximately the 48th week and then trends upwards for GNAR models 1 and 2. GNAR model 3 continues to increase its performance as it trends downward. Adapted from USAFacts (2021a); United States Census Bureau (2010).

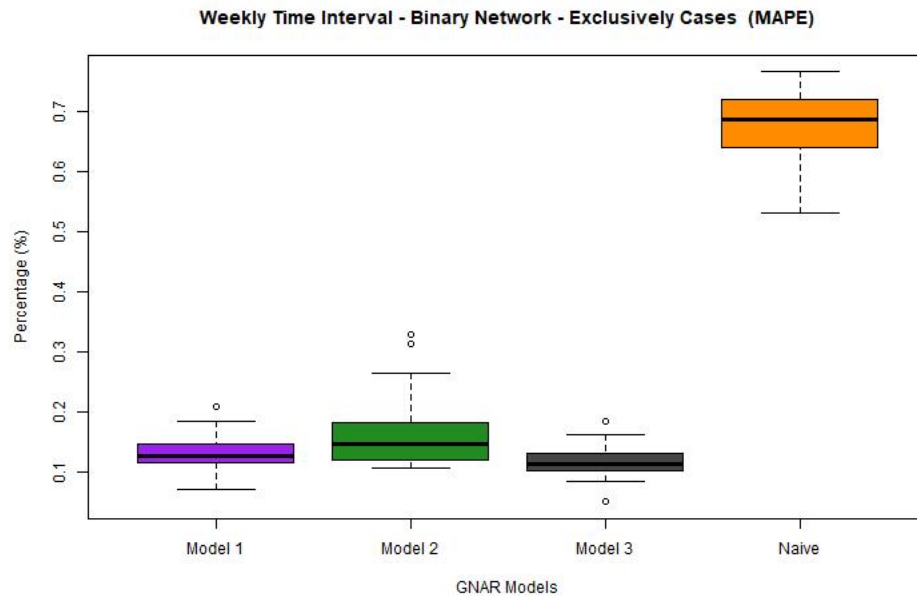


Figure 4.27. Weekly Time Interval - Binary Network - Cases (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021a); United States Census Bureau (2010).

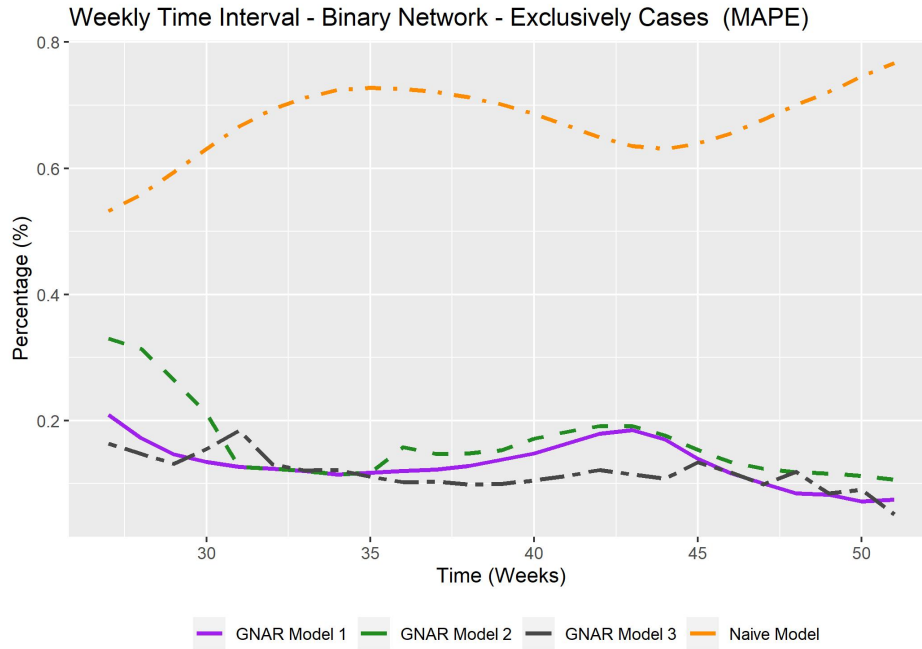


Figure 4.28. Weekly Time Interval - Binary Network - Cases (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021a); United States Census Bureau (2010).

4.3.2 Weekly Time Interval - Great Circle Distance Network - Cases

As we can see from Table 4.8 and Figures 4.29, 4.30, 4.31, and 4.32, each GNAR model performed similar to its binary network counterpart in Section 4.3.1. Moreover, the MASE of each model outperformed the Naïve model as illustrated in Figure 4.30 and could be used to accurately forecast COVID-19 cases. GNAR model 3 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, model 3 is the most consistently accurate model for this parameter combination.

As Figure 4.29 portrays, MASE variability amongst the GNAR models varies, which decreases their forecasting capabilities. GNAR models 1 and 2 experience a significant increase towards the end of the data set as discussed in Section 4.1 but GNAR model 3 continues to increase its accuracy as it trends downwards. Figure 4.31 displays MAPE char-

acteristics, which contain relatively small variability thus demonstrating their forecasting capabilities. Moreover, Figure 4.32 illustrates MAPE for each GNAR model continuing to trend downwards for nearly the entire data set. As a result, all three GNAR models surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.8. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 27-week forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.5380	0.5392	0.0127
Model 2	0.6473	0.6077	0.0210
Model 3	0.3983	0.3928	0.0055
Naïve	1	1	0
MAPE			
Model 1	0.1315	0.1264	0.0012
Model 2	0.1705	0.1278	0.0065
Model 3	0.1189	0.1150	0.0008
Naïve	0.6751	0.6868	0.0033

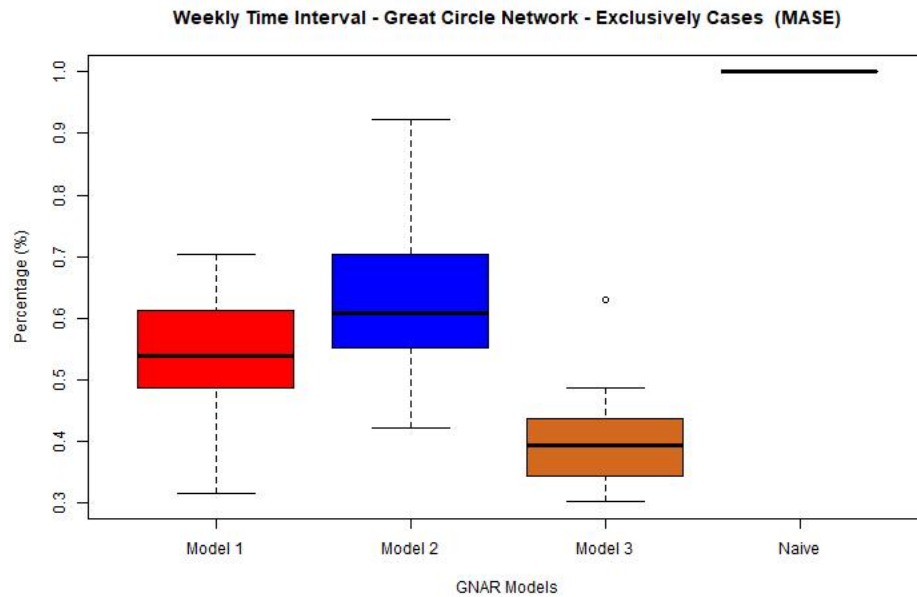


Figure 4.29. Weekly Time Interval - Great Circle Network - Cases (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is moderate variability amongst GNAR models 1 and 2 and low variability within GNAR model 3. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

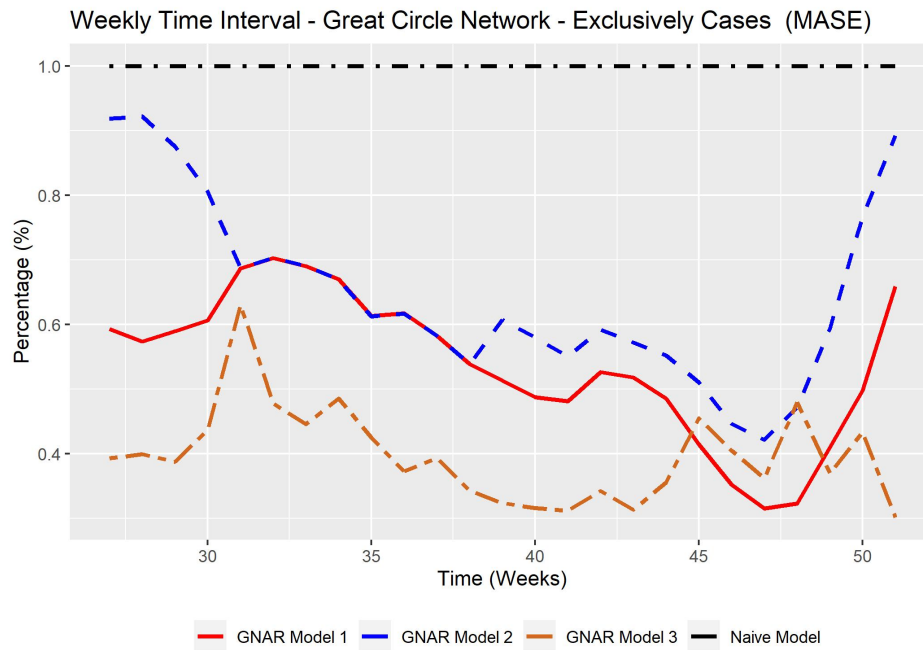


Figure 4.30. Weekly Time Interval - Great Circle Network - Cases (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. While each GNAR model outperforms the Naïve model, MASE decreases until approximately the 48th week and then trends upwards for GNAR models 1 and 2. GNAR model 3 continues to increase its performance as it trends downward. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

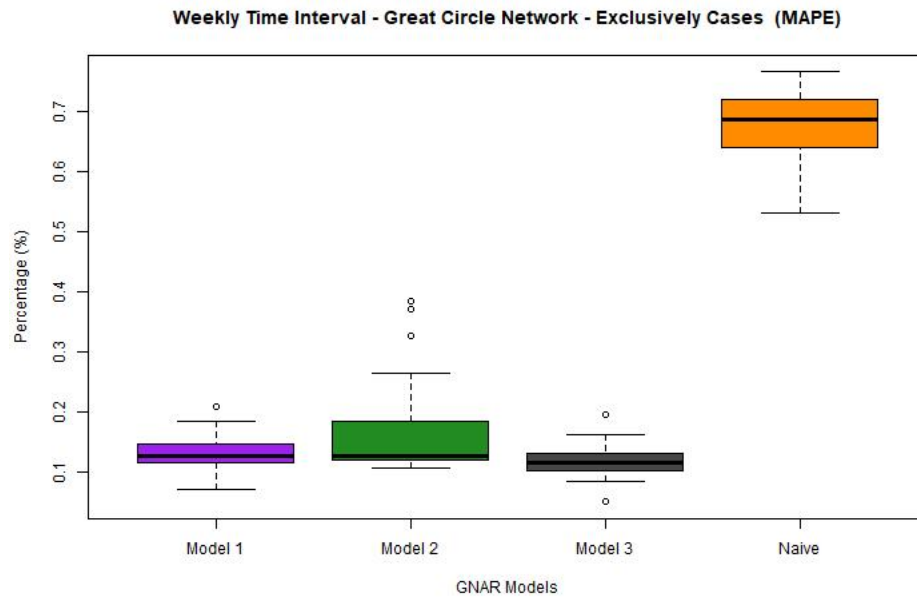


Figure 4.31. Weekly Time Interval - Great Circle Network - Cases (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

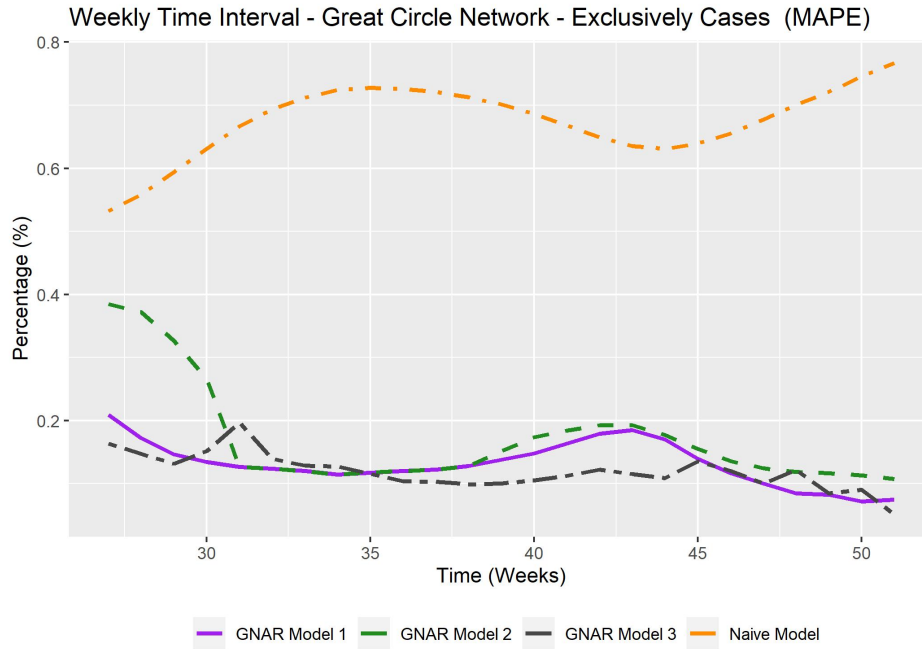


Figure 4.32. Weekly Time Interval - Great Circle Network - Cases (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021a); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

4.3.3 Weekly Time Interval - Binary Network - Deaths

As we can see from Table 4.9, the MASE and MAPE for GNAR model 3 provide higher accuracy and forecasting capability than GNAR models 1 and 2 as it has the lowest mean and median. This is further evidenced in Figures 4.33 and 4.35. As a result, GNAR Model 3 is the most consistently accurate model for this parameter combination. Table 4.9 also depicts low variability for both MASE and MAPE within each GNAR model.

Table 4.9 also displays the MASE mean and median of each GNAR model outperforming the Naïve model. The high MASE variability displayed in Figure 4.33 is a direct result of the model learning. As the model continues to improve and increase its accuracy, MASE continues to produce lower and lower values, which are further away from the initial starting point. Moreover, Figure 4.34 portrays the MASE for all GNAR models outperforming the

Naïve model and continuing to decrease as they trend downward. Additionally, Figure 4.36 demonstrates MAPE consistency of each GNAR model to perform at or below a 20% threshold and it continues to trend downwards for nearly the entire data set. As a result, all three GNAR models surpassed the performance standards established in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.9. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 27-week forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); United States Census Bureau (2010).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8014	0.7471	0.0705
Model 2	0.8033	0.7467	0.0715
Model 3	0.6728	0.6661	0.0129
Naïve	1	1	0
MAPE			
Model 1	0.1798	0.1741	0.0014
Model 2	0.1803	0.1741	0.0015
Model 3	0.1766	0.1675	0.0015
Naïve	0.7080	0.7178	0.0015

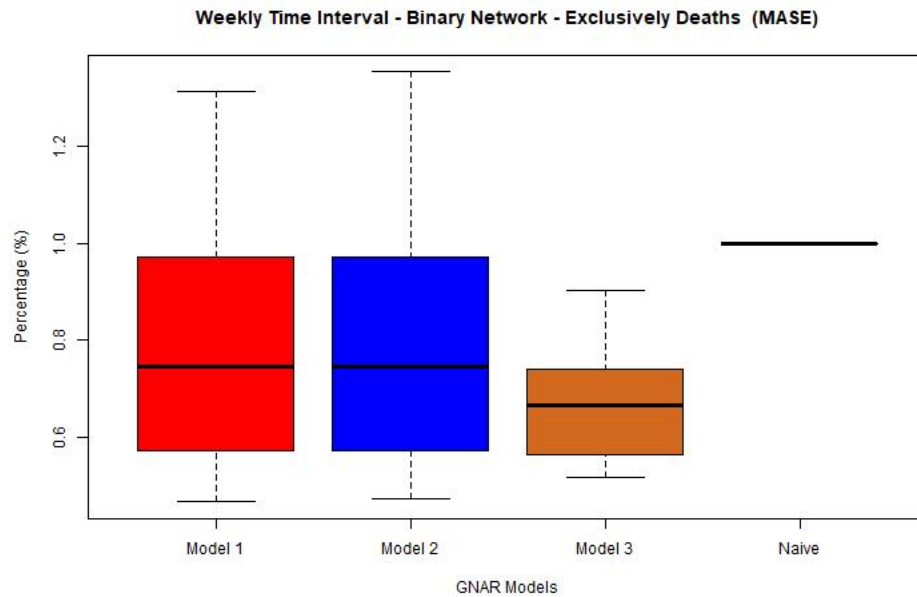


Figure 4.33. Weekly Time Interval - Binary Network - Deaths (MASE)
The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively high variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010).

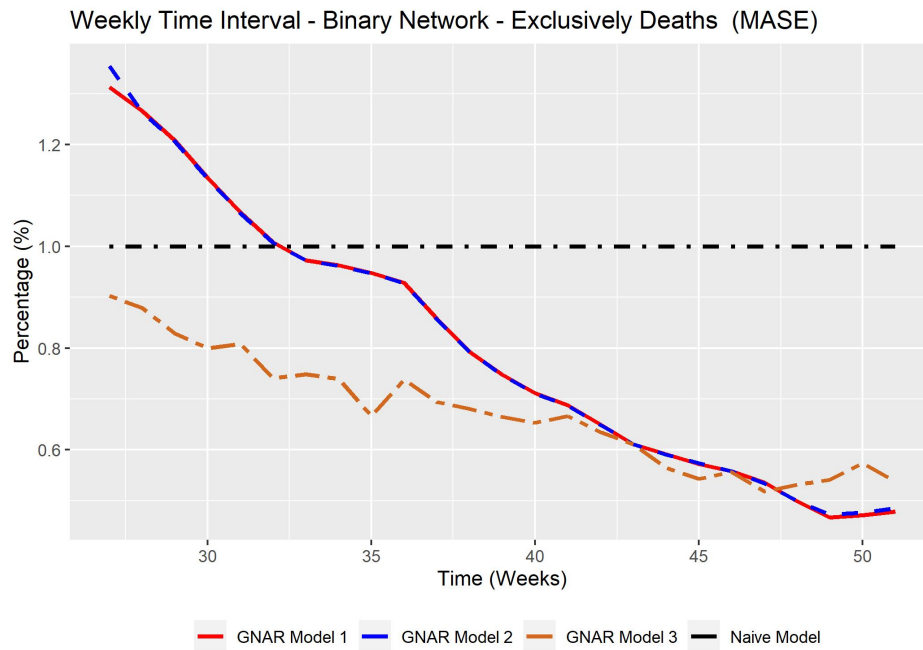


Figure 4.34. Weekly Time Interval - Binary Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, each GNAR model continues to increase its accuracy as it trends downward. Adapted from USAFacts (2021b); United States Census Bureau (2010).

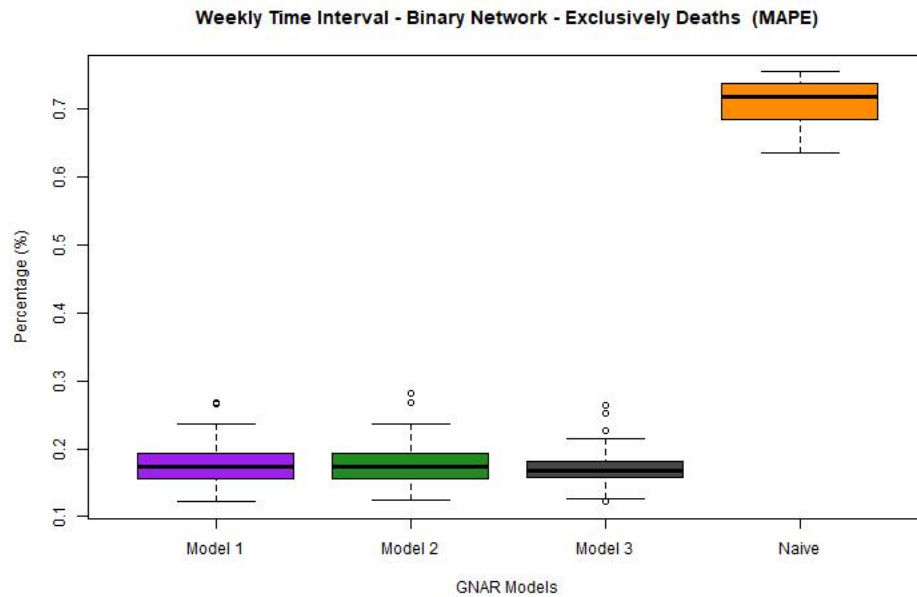


Figure 4.35. Weekly Time Interval - Binary Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010).

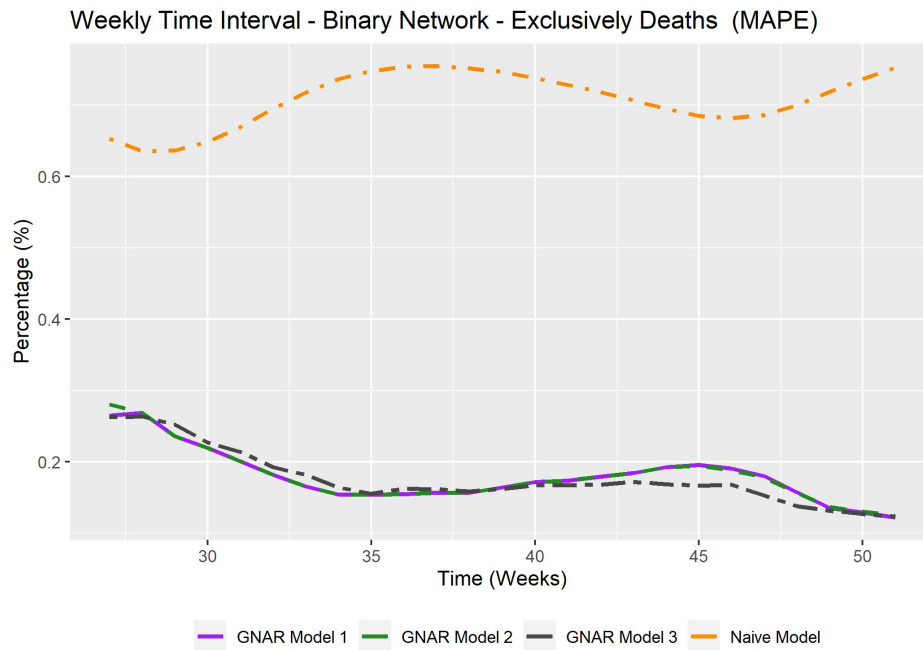


Figure 4.36. Weekly Time Interval - Binary Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); United States Census Bureau (2010).

4.3.4 Weekly Time Interval - Great Circle Distance Network - Deaths

As we can see from Table 4.10 and Figures 4.37, 4.38, 4.39, and 4.40, each GNAR model performed similar to its binary network counterpart in Section 4.3.3. Moreover, the MASE of each model outperformed the Naïve model as illustrated in Figure 4.38 and could be used to accurately forecast COVID-19 cases. GNAR model 3 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, model 3 is the most consistently accurate model for this parameter combination.

As Figure 4.37 portrays, MASE variability amongst the GNAR models varies significantly, which decreases their forecasting capabilities. The high MASE variability displayed in Figure 4.37 is a direct result of the model learning. As the model continues to improve and increase its accuracy, MASE continues to produce lower and lower values, which are

further away from the initial starting point. Figure 4.34 portrays the MASE for all GNAR models outperforming the Naïve model and continuing to decrease as they trend downward. Additionally, Figure 4.36 demonstrates MAPE consistency of each GNAR model to perform at or below a 20% threshold and it continues to trend downwards for nearly the entire data set. As a result, all three GNAR models surpassed the performance standards established in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.10. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 27-week forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.8014	0.7471	0.0705
Model 2	0.8013	0.7470	0.0700
Model 3	0.6735	0.6679	0.0132
Naïve	1	1	0
MAPE			
Model 1	0.1798	0.1741	0.0014
Model 2	0.1798	0.1741	0.0014
Model 3	0.1770	0.1679	0.0016
Naïve	0.7080	0.7178	0.0015

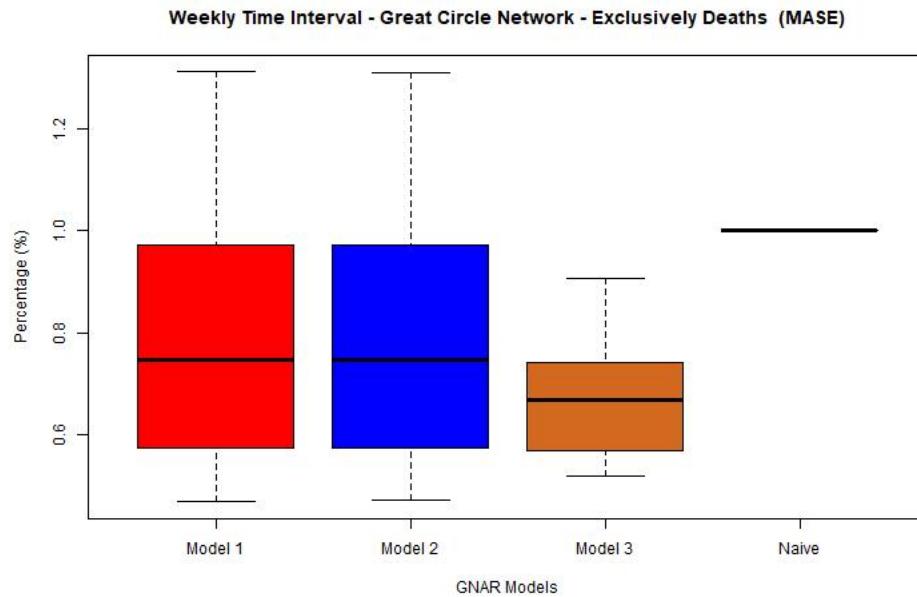


Figure 4.37. Weekly Time Interval - Great Circle Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is moderate variability amongst each GNAR model, which decreases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

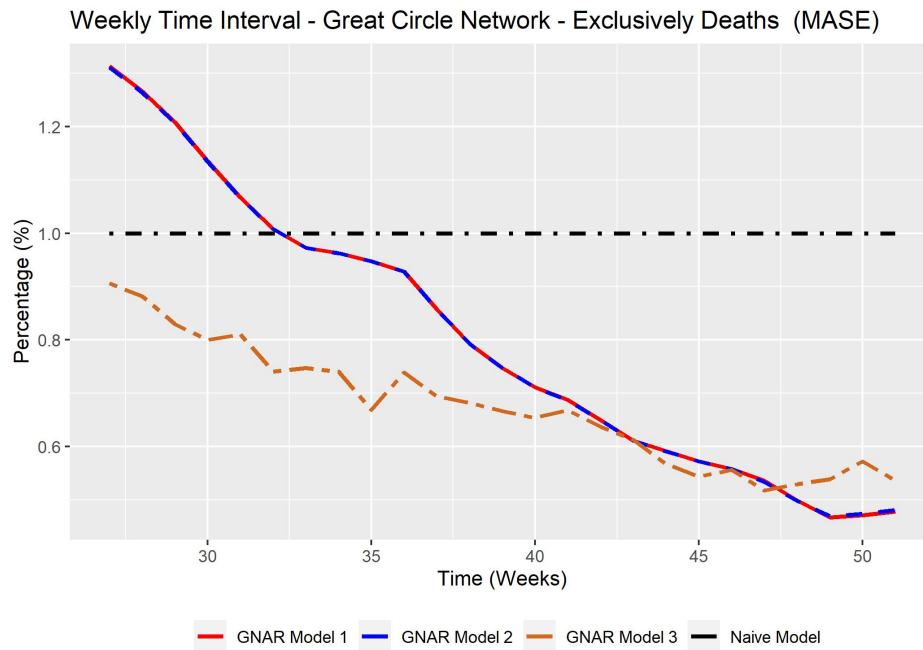


Figure 4.38. Weekly Time Interval - Great Circle Network - Deaths (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, each GNAR model continues to increase its accuracy as it trends downward. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

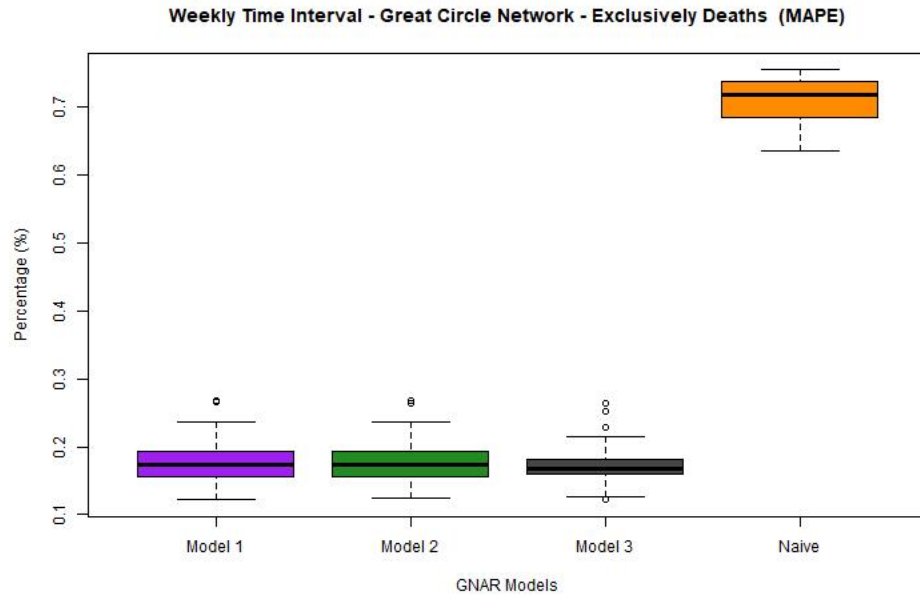


Figure 4.39. Weekly Time Interval - Great Circle Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

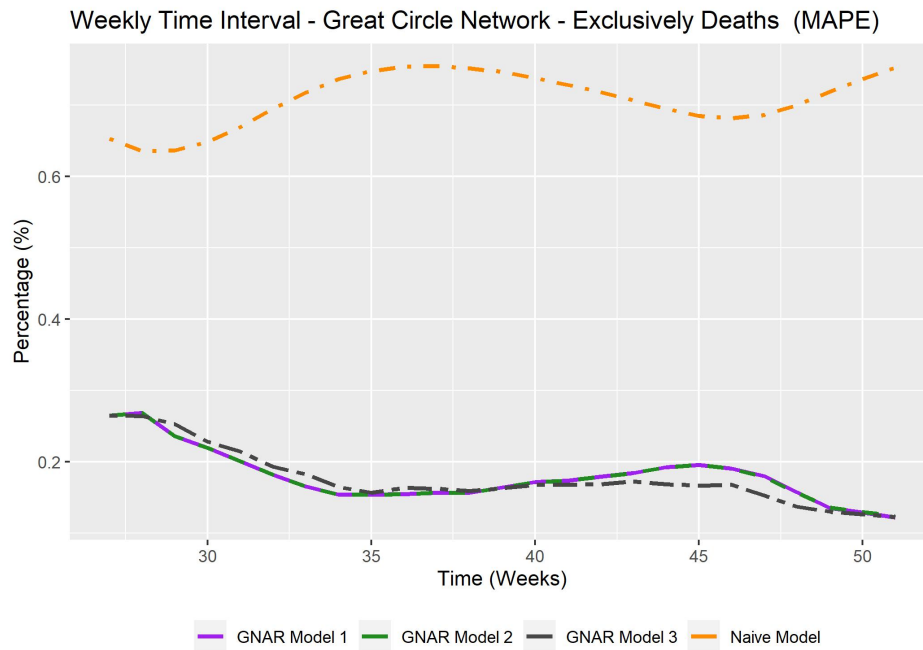


Figure 4.40. Weekly Time Interval - Great Circle Network - Deaths (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); United States Census Bureau (2010); National Bureau of Economic Research (2010b).

4.3.5 Weekly Time Interval - Binary Network - Survivors

As we can see from Table 4.11, the MASE and MAPE for GNAR model 3 provide higher accuracy and forecasting capability than GNAR models 1 and 2 as it has the lowest mean and median. This is further evidenced in Figures 4.41 and 4.43. As a result, GNAR Model 3 is the most consistently accurate model for this parameter combination. Table 4.11 also depicts low variability for both MASE and MAPE within each GNAR model.

Table 4.11 also displays the MASE mean and median of each GNAR model outperforming the Naïve model. Moreover, Figure 4.42 portrays the MASE for GNAR models 1 and 3 outperforming the Naïve model. GNAR models 1 and 2 experience a significant increase towards the end of the data set as discussed in Section 4.1 and GNAR model 2 exceeds the Naïve model's performance threshold. However, GNAR model 3 continues to increase

its accuracy as it trends downwards. The increase for GNAR models 1 and 2 could be their failure to adapt as cases began to trend downward. However, Figure 4.44 demonstrates MAPE consistency of each GNAR model to perform at or below a 20% threshold. Additionally, MAPE continues to trend downward for nearly the entire data set. As a result, GNAR models 1 and 3 surpassed the performance standards established in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.11. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 27-week forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.5796	0.6021	0.0212
Model 2	0.7099	0.7333	0.0304
Model 3	0.4113	0.4028	0.0066
Naïve	1	1	0
MAPE			
Model 1	0.1321	0.1280	0.0012
Model 2	0.1646	0.1524	0.0033
Model 3	0.1188	0.1156	0.0000
Naïve	0.6746	0.6858	0.0033

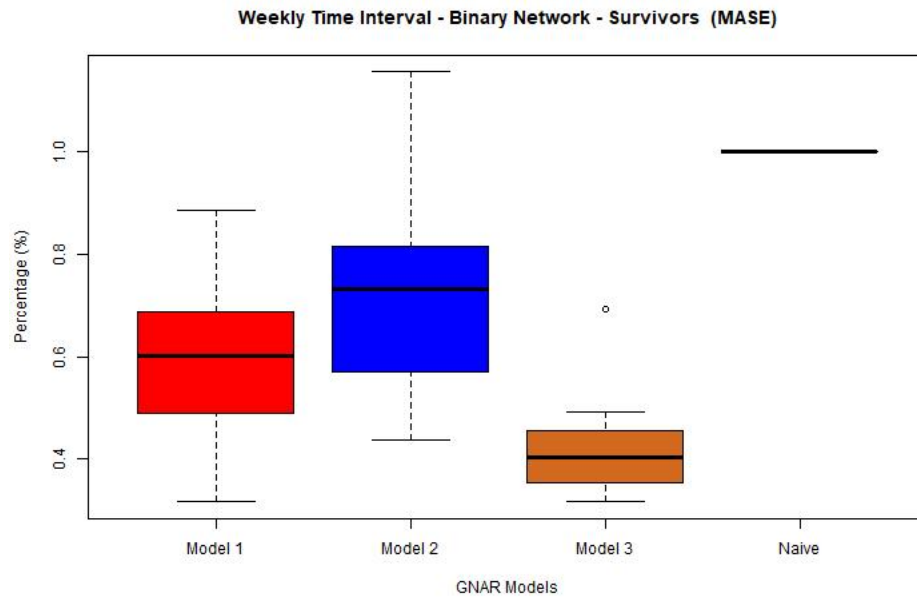


Figure 4.41. Weekly Time Interval - Binary Network - Survivors Boxplot (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is moderate variability amongst GNAR models 1 and 2 and low variability within GNAR model 3. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010).

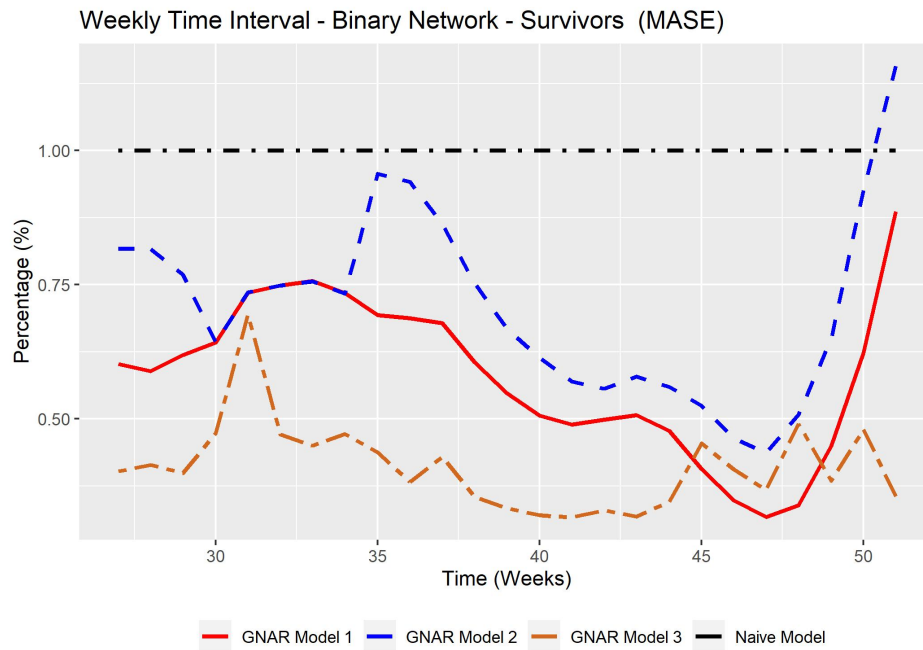


Figure 4.42. Weekly Time Interval - Binary Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. While GNAR models 1 and 3 outperform the Naïve model, GNAR model 2 exceeds the Naïve model's performance threshold and is no longer acceptable for modeling. GNAR model 3 continues to increase its performance as it trends downward. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010).

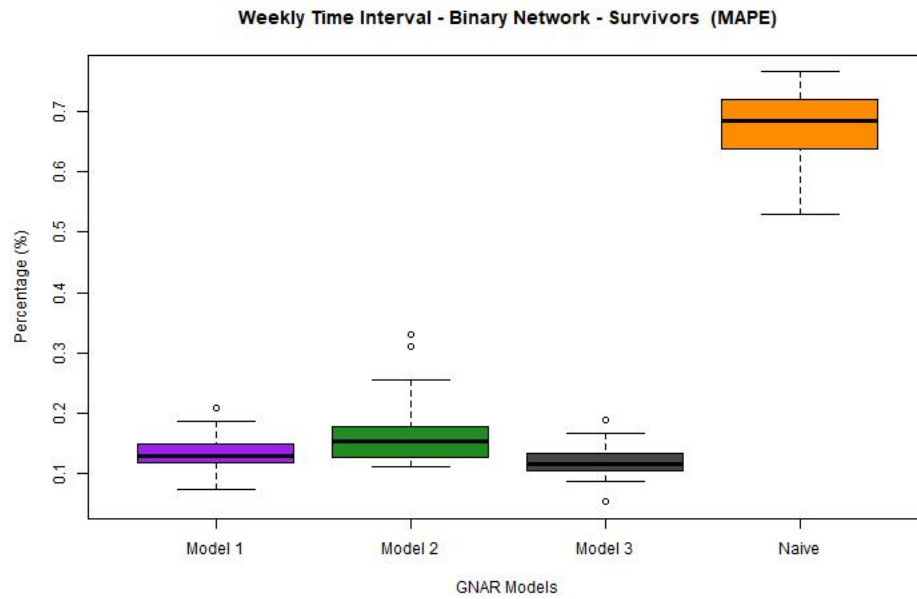


Figure 4.43. Weekly Time Interval - Binary Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010).

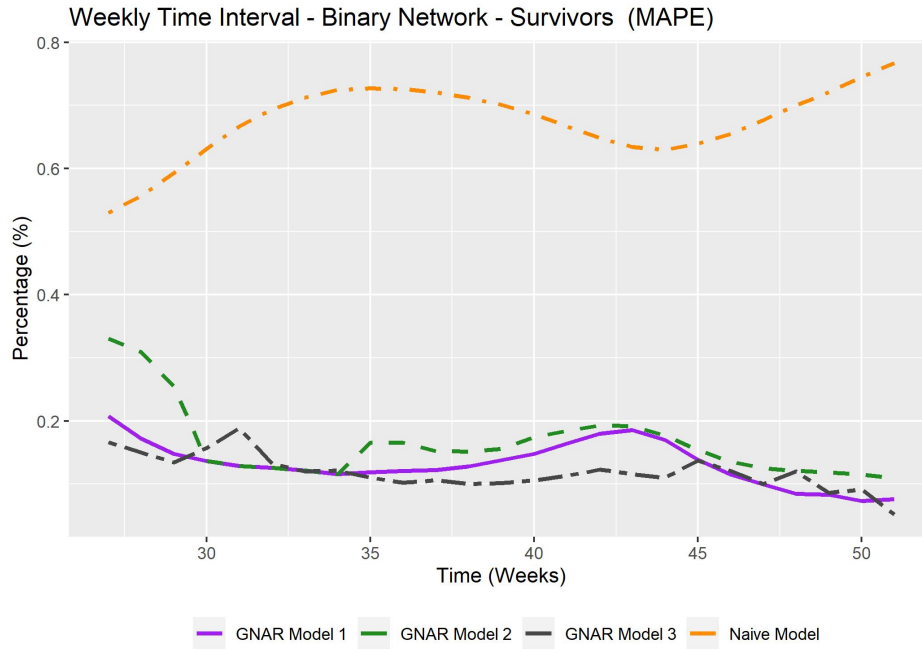


Figure 4.44. Weekly Time Interval - Binary Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010).

4.3.6 Weekly Time Interval - Great Circle Distance Network - Survivors

As we can see from Table 4.12 and Figures 4.45, 4.46, 4.47, and 4.48, each GNAR model performed similar to its binary network counterpart in Section 4.3.5. Moreover, the MASE of each model outperformed the Naïve model as illustrated in Figure 4.46 and could be used to accurately forecast COVID-19 cases. GNAR model 3 had the highest accuracy as it had the lowest mean, median, and variance for both MASE and MAPE. As a result, model 3 is the most consistently accurate model for this parameter combination.

As Figure 4.45 portrays, MASE variability amongst the GNAR models varies, which decreases their forecasting capabilities. GNAR models 1 and 2 experience a significant increase towards the end of the data set as discussed in Section 4.1 but GNAR model 3

continues to increase its accuracy as it trends downwards. Figure 4.47 displays MAPE characteristics, which contain relatively small variability thus demonstrating their forecasting capabilities. Moreover, Figure 4.48 illustrates MAPE for each GNAR model continuing to trend downwards for nearly the entire data set. As a result, all three GNAR models surpass the performance standards outlined in subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases.

Table 4.12. Model Combination Summary. We can calculate MASE and MAPE for each GNAR model with respect to each test period within the 27-week forecast. We can then calculate the mean, median, and variance of these values. As discussed in subsection 2.2.2, MASE is the average ratio between the Naïve model and each model. By definition, MASE of the Naïve model is always 1 (Hyndman et al. 2006). If MASE is less than 1, then the model which we compare against the Naïve model performs better than the Naïve model. If MASE is more than 1, then the Naïve model performs better. Intuitively, MAPE is the average of the error rate relative to the actual observed number as discussed in subsection 2.2.2. Table 2.1 provides guidance for interpreting MAPE performance. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010); National Bureau of Economic Research (2010b).

GNAR Model	Mean	Median	Variance
MASE			
Model 1	0.5796	0.6021	0.0212
Model 2	0.7171	0.7335	0.0332
Model 3	0.4205	0.4028	0.0100
Naïve	1	1	0
MAPE			
Model 1	0.1321	0.1280	0.0012
Model 2	0.1746	0.1493	0.0059
Model 3	0.1208	0.1162	0.0000
Naïve	0.6746	0.6858	0.0033

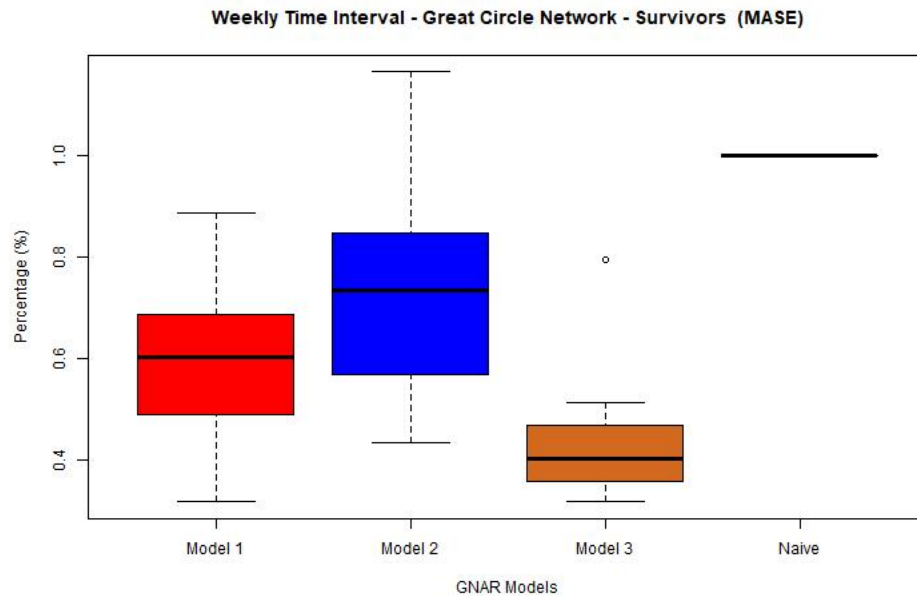


Figure 4.45. Weekly Time Interval - Great Circle Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts each GNAR model. As we can see, there is moderate variability amongst GNAR models 1 and 2 and low variability within GNAR model 3. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010); National Bureau of Economic Research (2010b).

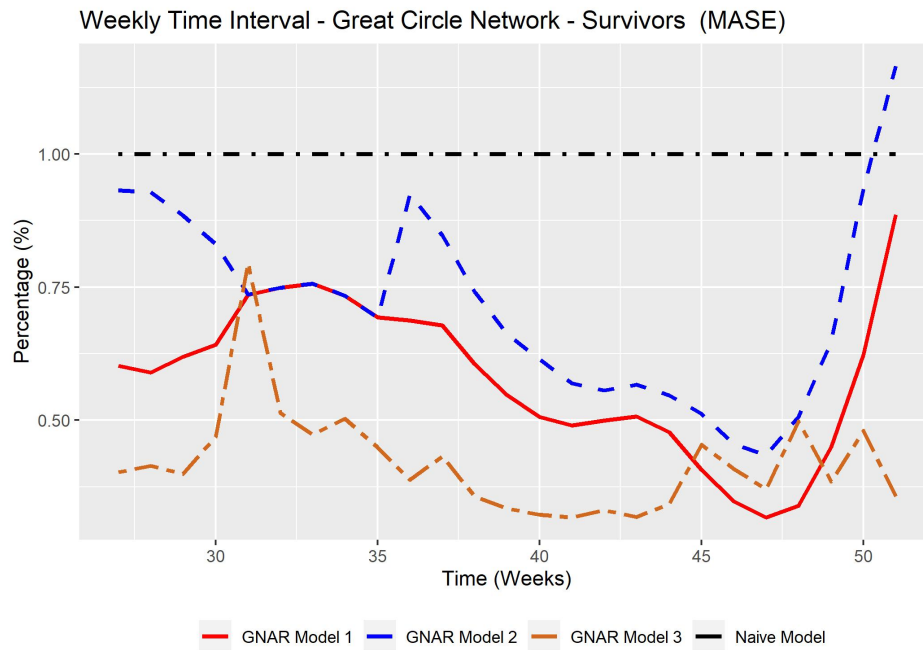


Figure 4.46. Weekly Time Interval - Great Circle Network - Survivors (MASE). The vertical axis represents MASE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. While GNAR models 1 and 3 outperform the Naïve model, GNAR model 2 exceeds the Naïve model's performance threshold and is no longer acceptable for modeling. GNAR model 3 continues to increase its performance as it trends downward. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010); National Bureau of Economic Research (2010b).

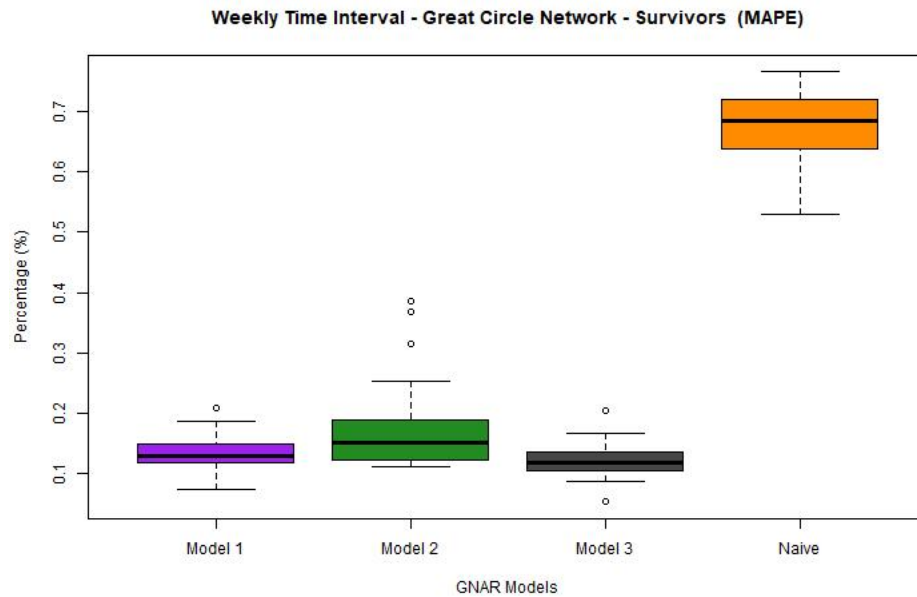


Figure 4.47. Weekly Time Interval - Great Circle Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts each GNAR model. As we can see, there is relatively low variability amongst each GNAR model, which increases their predictive capability. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010); National Bureau of Economic Research (2010b).

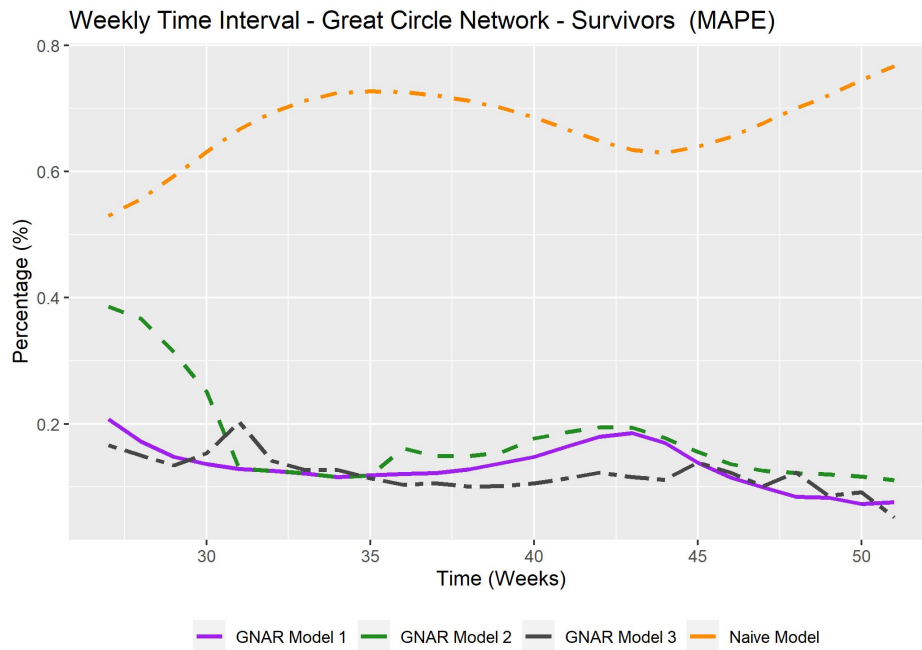


Figure 4.48. Weekly Time Interval - Great Circle Network - Survivors (MAPE). The vertical axis represents MAPE percentage and the horizontal axis depicts the number of weeks since data collection began on 22JAN21. As we can see, MAPE continues to decrease for each GNAR model as it trends downward. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010); National Bureau of Economic Research (2010b).

4.3.7 Weekly Results Summary

As the results in Section 4.3 demonstrated, MASE experiences high variability throughout the data set. The data points within cases and survivors are approximately equal, which results in similar MASE and MAPE characteristics. The MASE and MAPE within Sections 4.3.3 and 4.3.4 continued to trend downwards thus demonstrating their forecasting capability. MASE exceeded the Naïve model performance threshold for GNAR model 2 within both case and survivors. Additionally, MASE for both GNAR models 1 and 2 continue to decrease until approximately the 48th day and begin to trend upwards. Ultimately, nearly every model combination analyzed in Section 4.3 surpass the performance standards outlined in

subsection 2.2.2 and could effectively be applied to forecast COVID-19 cases, deaths, and survivors.

4.4 Summary

As we have demonstrated, a GNAR model can be used to successfully model COVID-19 infections, deaths, and survivors for daily and weekly time intervals. MASE and MAPE vary within both model combinations and their respective GNAR models. Despite this variability, all models demonstrated their capability to successfully model COVID-19 cases, deaths, and survivors.

4.4.1 Great Circle Distance Performance

Although the binary network and great circle distance network represent their edges differently, the accuracy percent difference is negligible amongst the various model combinations. Figure 4.49 displays the results by computing the percent difference between similar combinations' (e.g. daily-binary-cases and daily-great circle-cases) mean MASE statistics. The percent difference was also calculated between similar combinations' mean MAPE values. The same process was applied for similar combinations' median values. Finally, these percent difference results were aggregated and categorized into "no difference," "worse," or "better" based upon their associated value.

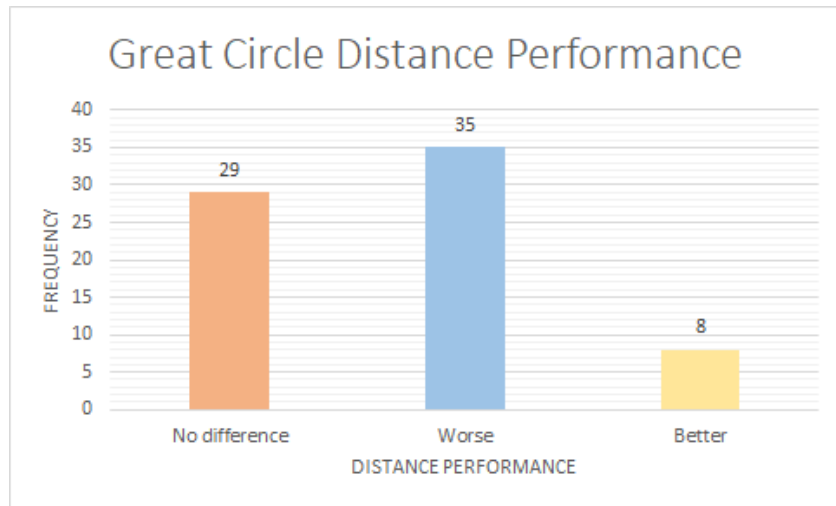


Figure 4.49. Great Circle Performance Bar Plot. The vertical axis represents the frequency while the horizontal axis displays whether or not the accuracy remained the same, performed better, or performed worse. As we can see, incorporating the great circle distance increased the accuracy of MASE or MAPE only eight times. Adapted from USAFacts (2021b); USAFacts (2021a) United States Census Bureau (2010); National Bureau of Economic Research (2010b).

As we can see from Figure 4.49, the model's performance remained constant or in fact decreased 64 out of 72 times, or 89%. The model's best performance increased occurred between the weekly case combinations in Sections 4.3.1 and 4.3.2 with accuracy increasing by nearly 14%. However, most of the remaining seven positive instances experienced modest gains of approximately 2%. Of the 35 decreased performance occurrences, most were negligible ranging from approximately 0.5% to 6%. As a result, the network structure could be adjusted to a binary framework to simplify calculations.

4.4.2 Case and Survivor Similarities

As we have seen, the MASE and MAPE performance for cases and survivors is comparable for both daily and weekly combinations. These values are unsurprising given COVID-19's relatively low fatality rate (Johns Hopkins University 2021). Moreover, public health officials could use this data to forecast the survivors within their given county.

4.4.3 Daily and Weekly Performance

The daily models demonstrated consistent behavior for the case and survivor combinations. After approximately 48 weeks of data reporting, MASE for the cases and survivors began to increase. Since COVID-19 survivors is dependent on the number of cases, its MASE increased as well. MASE for every daily case and survivor GNAR model decreased until approximately 350th day then increased. MASE for GNAR models 1 and 2 for weekly case and survivor model combinations decreased until approximately 350th day then increased. However, GNAR model 3 continues to increase its MASE accuracy as it continued to trend downward. The mean and median of the four weekly case and survivor combinations were lower than their daily counter part.

Conversely, the various death model combinations performed well as both MASE and MAPE continued to decrease as they trended downward. The mean and median of the two daily combinations were lower than their weekly counterpart. As a result, deaths should be modeled on a daily time interval while cases and survivors should be modeled weekly.

CHAPTER 5:

Conclusions and Summary

5.1 Conclusions

The GNAR package demonstrated its ability to model COVID-19 infections within the United States' counties. Open and accurate CDC data reporting enabled various combinations to effectively model cases, deaths, and survivors. However, future research through additional data collection, multivariate network modeling, foreign adaptation could be used to improve and increase model accuracy to ultimately reduce a pathogen's impact on a given community.

5.1.1 Data Analysis

As COVID-19 continues to spread and proliferate amongst the world's population, cases and deaths will continue to increase. Developing an accurate model predicated upon transparent data is paramount to ensuring a timely and accurate method to disrupt its spread.

The United States county network provides an opportunity to model SARS-CoV-2 as adjacencies are formed either through a binary relationship, great circle distance between county centroids, or the shortest path. As such, counties exert influence on one another as individuals personally or professionally commute between counties and in turn carry the virus.

5.1.2 GNAR Model

Although the daily case and survivor combinations diverged, the GNAR package demonstrated its capability to effectively model COVID-19 cases within the United States. By enabling a weekly count, we were able to mitigate this daily divergent behavior as GNAR Model 3 continued its precision and maintained its accuracy. Additionally, the accuracy of MASE and MAPE continued to increase for the models in subsections 4.2.3 and 4.2.4. To provide maximum accuracy, case and survivors should be modeled weekly and deaths

should be modeled daily. This combination would enable public health officials to make public policy recommendations to disrupt and ultimately halt SARS-CoV-2's progression.

5.2 Department of Defense Impact

Service members and their families are not insulated from infectious diseases and viral pandemics. They live amongst and interact with the various communities surrounding the installations and are, consequently, susceptible to infection. Once an individual is infected, a unit's medical readiness may decrease as the number of deployable personnel shrinks. As the number of deployable personnel declines, a Commander's ability to achieve and promote the tactical, operational, and strategic interests of the United States and its allies across the competition continuum and achieve unified action may be limited.

In March 2020, SARS-CoV-2 disseminated amongst the United States Ship (USS) *Theodore Roosevelt* crew at an exponential rate forcing their evacuation (LaGrone 2020a). As a result, the entire ship sat idle and unable to project air power, provide crisis response, and deter aggression (Thompson 2019). As great power competition returns, adversaries can threaten the international order, exploit these vulnerabilities, and achieve their objectives unopposed. Therefore, it is critical that officials from the Defense Health Agency, CDC, local public health administrations, and Commanders be agile, adaptable, and understand the operational and personnel risk in a given area and which public health prevention measures to enact.

5.3 Future Work

While the model's accuracy demonstrated GNAR's capability, additional work could be done to potentially increase its performance and ultimately its utility as an epidemiological model.

5.3.1 Supplementary Data Collection

While the GNAR model is univariate, additional data can be collected to potentially provide additional information. The network's adjacency structure can be modified with edge weights represented by the shortest path distance between two counties. While the great circle distance can be used as an acceptable, albeit temporary, data surrogate for the short-

est driving distance between counties, it often resulted in inaccurate models. Additionally, county information such as unemployment can be collected and collated.

Open Street Map Networkx

Open Street Map (OSM) provides a free and editable map of the world upon which the Python package Open Street Map Networkx (OSMnx) is built (Boeing 2017b). The OSMnx package enables users to download geospatial data from Open Street Map to model and analyze real-world street networks (Boeing 2017a).

Users may create a graph from a specified latitude-longitude coordinate pair, address, bounding box, polygon, extensible markup language (XML), or place. The function `graph_from_place` accepts a query string or a list of strings and creates a graph from OSM within the specified boundaries of some geocodable place (Boeing 2021). Users are also able to specify an additional buffer distance around the polygon. We can then locate the intersection nearest a county's centroid as the starting point using the `get_nearest_node` function and providing a latitude-longitude coordinate pair. We can then use the same function and provide a second pair of latitude-longitude coordinates for the ending node in the adjacent county. Knowing the maximum distance between adjacent counties using the great circle distance values, we can apply a multiplicative buffer to ensure the adjacent county's centroid is captured and finally create a new graph. With this new graph, we can finally determine the shortest path in meters using `shortest_path` function.

Despite OSMnx's capabilities, it certainly has limitations. Since Open Street Map is an open and collaborative data set, there may be unintended errors or unverified information that can cause irregularities when attempting to exercise various functions within the OSMnx package (Boeing 2017b). Erroneous data such as a nonexistent road or intersection can result in an imprecise shortest path calculation, which may lead to incorrect analysis and conclusions.

An additional limitation is OSMnx's inability to find a unique nearest node given a pair of latitude and longitude coordinates. To calculate the shortest path between two counties, the user must input a graph containing the desired starting and ending nodes, defined as street intersections, into the `shortest_path` function. OSMnx provides a helper function, `get_nearest_node`, which accepts a pair of latitude and longitude coordinates. The

`get_nearest_node` function then seeks out the nearest intersection within the specified graph. However, if the graph does not contain ample unique nodes within the vicinity of a starting pair of latitude and longitude coordinates or an ending pairs of latitude and longitude coordinates, the function will return the identical node. As a result, the calculated shortest path will be zero.

County Data

In addition to county adjacencies, the United States Census Bureau provides population information for each county to include, but not limited to the following: population demographics, population density, and total size when it conducts its census every ten years (United States Census Bureau 2020b). The Census Bureau also provides population estimates every year, which may provide a more accurate representation of the population in a given county. Unemployment and poverty data could be used to determine counties with vulnerable populations as those individuals may be unable to seek medical care and lack the financial resources to isolate. Despite additional data, a predictive model will never be 100% since the human component is never fully predictable. As such, lurking variables will always exist and may falsely identify a strong relationship or may obscure the true relationship. As a result, they may continue working, which could increase their transmission rate amongst Susceptible individuals.

Through additional data collection and computation, we can determine the total number of intensive care beds per county as per the Department of Homeland Security (DHS) (Department of Homeland Security 2021). Since medical facilities provide varying capabilities, the hospital types and capacities provided by local and regional health officials to the DHS can be used to represent medical capacity within a given county (Department of Homeland Security 2021). While former President Trump and state governors ordered medical capacity be increased through methods such as the deployment of United States Naval Ship (USNS) *Comfort* or building field expedient hospitals to provide additional medical services and support to local hospitals, determining the exact capacity, duration, and county location of those hospitals may be infeasible (LaGrone 2020b).

Modeling ZIP Codes

Although the CDC monitors COVID-19 infections and deaths at the county level, the CDC could further segregate cases and deaths into individual respective ZIP codes. According to the United States Postal Service, there are nearly 42,000 ZIP codes in the United States (USPS 2021). Since ZIP codes comprise the entirety of homes within the United States, they also form a network with edges being modeled as binary, great circle distance, or shortest path distance. Although individuals are able to move freely amongst the counties, movement between ZIP codes may provide a more realistic representation of daily life and community interactions due to the relatively smaller distance between nodes.

Foreign Modeling

Although we utilized the United States' county adjacency structure as the underlying network for the GNAR model, any governance network can be applied. A network can be formed by using a nation's townships, counties, municipalities, regional municipalities, and districts as nodes. Edges can be similarly modeled as either binary, great circle distance, or shortest path distance. The data and subsequent network could be used as discussed in Chapter 3 to create a GNAR model for a given country, province, or region.

Countries within the European Union and the United Kingdom could provide a sufficient opportunity for modeling. Most countries within the European Union employ a similar governance hierarchy with provinces and municipalities representing states and counties within the United States. European nations also enforced strict lockdown measures to curb the spread of SARS-CoV-2 (Di Donato et al. 2020). Europe's climate, terrain, and population disposition are similar to the United States. The two continents share a similar weather cycle, ecosystems, and dense population centers surrounded by rural minorities. However, nations within the African Union may not be ideal as the aforementioned criteria are not comparable. As such, one can apply the GNAR package to compare and contrast public policy decisions and their impact on COVID-19 infections and deaths within the African Union and European Union.

5.3.2 Principal Component Analysis

Principal component analysis (PCA) is an unsupervised machine learning approach where a subset of features $X_1, X_2, \dots, X_n$ is used in lieu of a traditional response variable, Y (James

et al. 2013). Principal component analysis can be utilized to generate “a lower dimensional representation of a data set that contains as much as possible of the variation” (James et al. 2013). Since GNAR only accepts univariate data, PCA can be used to create a new, univariate response variable Y that contains information from county features.

Since COVID-19 infections are reported at the county level, we can generate a data object that contains a set of county features such as population demographics and density along with a response variable, or infections, for each county. We can then apply PCA to reduce the dimensionality in an attempt to incorporate these features.

We can apply PCA to generate the first principal component, which is a normalized linear combination of selected features $X_1, X_2, \dots, X_n$ that contains the largest sample variance (James et al. 2013). The second principal component is the linear combination of $X_1, X_2, \dots, X_n$ features that have maximal variance out of all linear combinations that are uncorrelated with the first principal component (James et al. 2013). As a result, the first and second principal components are orthogonal. Consequently, we can generate a new variable using a linear combination of these two principal components to elucidate the maximum amount of variance in two dimensions.

5.3.3 Model Parameters

While we utilized the model parameters as depicted in Table 3.4, a SIR compartment model may be adapted. Additionally, modeling the total infections and deaths may be modeled for daily and weekly time intervals.

SIR Implementation

As discussed in Section 2.3.2, COVID-19 infections are theorized to follow a SEIR compartment model (Hoehl et al. 2020) and (Patil and Kotwal 2020). However, public health officials, epidemiologists, and researchers can integrate the simpler SIR model into the GNAR model. Currently, the CDC provides daily information for the total number of COVID-19 cases and deaths. However, a more accurate representation of infections can be maintained with minimal resources.

Section 5.3.1 stated that the United States Census Bureau maintains an approximation of each county’s population. This approximation serves as the initial starting point for the

Susceptible compartment. Public health officials can then subtract the total number of active Infected and Recovered to approximate the remaining number of Susceptible individuals within their given counties.

While every person is certainly unique and cannot be universally modeled, public health officials could divide COVID-19 infections into mild or severe symptoms. For those with mild symptoms, health officials could approximate the duration an individual is actively infected by applying the recovery rate, γ . If symptoms worsen, individuals would presumably seek local medical care where hospitals would transition an individual's accountability from mild to severe. If an individual with mild symptoms does not seek medical care and the expected duration is complete, they could be considered Recovered and can be aggregated with death counts from local and state officials as discussed in Section 3.1.3.

Certainly, this adaptation requires multiple assumptions and is not without its limitations. Medical testing and screening would need to identify every infection. Individuals with mild symptoms recover at the recovery rate, γ . Individuals are able to seek medical assistance as symptoms worsen from mild to severe. County populations are isolated and do not interact with one another.

However, local health officials can maintain a more accurate count of active COVID-19 infection in their respective counties. Therefore, officials can then compartmentalize a given county's population into either Susceptible, Infected, or Recovered, which can create additional univariate data for GNAR to model.

Rates of Change

Since the data set provides a time series with cumulative infections and deaths, one can also model the change in deaths and cases at either daily or weekly time intervals. This rate of change can be used to describe the percentage change in a given response variable over a period of time. Consequently, public health officials can use this rate of change to determine trends and identify momentum within the data set. As a result, local and state officials can propose increased or decreased public health measures.

GNAR Parameters

As discussed in Section 3.2.5 and available in the documentation, GNAR parameters such as latency and neighbor weight can be modified (Leeming et al. 2020). For example, we applied a one-day latency period in `alphaOrder` but a longer duration such as five or ten may increase model accuracy. Additional emphasis could be applied to the county adjacencies by modifying `betaOrder`.

Monthly Time Intervval

The SARS-CoV-2 vaccine was the fastest vaccine ever produced (Cohen 2020). However, had the vaccine not been implemented at an exceptional rate and the pandemic continued indefinitely, a monthly time interval could be applied, which could potentially increase the model's accuracy. As demonstrated in Chapter 2, pandemics are not novel and will continue to emerge and disappear for the foreseeable future (Naguib et al. 2020). Therefore, a monthly time interval may apply to the next pandemic.

5.4 Summary

The very nature of the global landscape in the 21st century can be characterized as one of uncertainty and unpredictability. Through the rise of globalization, a process compelled by investments and international trade facilitated through the increase use of information technology, societies are becoming better integrated, more interconnected, and expanding and subsequently condensing natural environments. As a result, the natural barrier between animals and humans will continue to erode placing humans at higher risk for zoonotic disease transmission and the next pandemic (Senthilingam 2017).

History has shown that epidemics are undoubtedly unpredictable. At present, no public health expert is able to exactly predict the duration, severity, and scope despite their best effort to disrupt or even prevent pandemics. Data analysts, developers, and practitioners of predictive modeling tools must remain cognizant of the capabilities and deficiencies and not over promise. Since the first known virus infected and spread amongst humans, shaman, healers, scientists, researchers, physicians, and epidemiologists alike have sought to explain and ultimately contain the deadly pathogen of their time. By employing science, mathematics, and now modeling techniques aided by computers, epidemiologists and scientists

are able to better understand and potentially forecast a pandemic's impact on society and provide public health recommendations to policy officials. As we have shown, the GNAR model can be incorporated as one of these varying methodologies and modeling techniques to prevent future human suffering and loss of life.

THIS PAGE INTENTIONALLY LEFT BLANK

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>.
- USAFacts (2021a) US COVID-19 cases by state. Accessed February 2, 2021, https://static.usafacts.org/public/data/covid-19/covid_confirmed_usafacts.csv.
- USAFacts (2021b) US COVID-19 deaths by state. Accessed February 2, 2021, https://static.usafacts.org/public/data/covid-19/covid_deaths_usafacts.csv.
- Ahuja RK, Magnanti TL, Orlin JB (1993) *Network Flows: Theory, Algorithms, and Applications* (Prentice hall, New Jersey).
- Ban C (2014) Shannon County, S.D. to be renamed Oglala Lakota County. Shannon County, S.D. Accessed March 2, 2021, <https://www.naco.org/articles/shannon-county-sd-be-renamed-oglala-lakota-county>.
- Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.
- Barrett JS, Jayaraman B, Patel D, Skolnik JM (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.
- Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M (2020) Application of the ARIMA model on the COVID-19 epidemic dataset. *Data in Brief* 29, <https://doi.org/10.1016/j.dib.2020.105340>.
- Bergs A (2011) *Social Networks and Historical Sociolinguistics: Studies in Morphosyntactic Variation in the Paston Letters (1421-1503)*, volume 51 (Walter de Gruyter, Berlin, Germany), <https://doi.org/10.1515/9783110923223>.
- Boeing G (2017a) OSMnx 1.0.1 — OSMnx 1.0.1 documentation. Boeing, Geoff. Accessed February 15, 2021, <https://osmnx.readthedocs.io/en/stable/>.
- Boeing G (2017b) OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers Environment and Urban Systems* 65:126–139.
- Boeing G (2021) User reference — OSMnx 1.0.1 documentation. Boeing, Geoff. Accessed February 15, 2021, https://osmnx.readthedocs.io/en/stable/osmnx.html?highlight=shortest_path#osmnx.distance.shortest_path.

- Broomfield County (2021) History of Broomfield. City and county of Broomfield. Accessed March 2, 2021, <https://www.broomfield.org/386/History-of-Broomfield#:~:text=Becoming%20the%2064th%20County&text=The%20amendment%20passed%20on%20November,effect%20on%20November%2015%2C%202001>.
- Brownlee J (2017) Autoregression models for time series forecasting with Python. Machine Learning Mastery. Accessed February 16, 2021, <https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/#:~:text=Autoregression%20is%20a%20time%20series,range%20of%20time%20series%20problems>.
- Capitanelli A (2020) Modeling the spread of diseases. Towards Data Science. Accessed February 6, 2021, <https://towardsdatascience.com/modeling-the-spread-of-diseases-821fc728990f>.
- Cavallo JJ, Donoho DA, Forman HP (2020) Hospital capacity and operations in the coronavirus disease 2019 (COVID-19) pandemic—planning for the nth patient. *Journal of American Medical Association Health Forum* 1(3), <https://doi.org/10.1001/jamahealthforum.2020.0345>.
- Center for Disease Control and Prevention (2020) Emerging SARS-CoV-2 variants. CDC. Accessed February 6, 2021, <https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html>.
- Centers for Disease Control and Prevention (2020) Faq: COVID-19 data and surveillance. CDC. Accessed February 15, 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).
- Chaurasia V, Pal S (2020) Application of machine learning time series analysis for prediction COVID-19 pandemic. *Research on Biomedical Engineering* <https://doi.org/10.1007/s42600-020-00105-4>.
- Chen D, Yang Y, Zhang Y, Yu W (2020a) Prediction of COVID-19 spread by sliding MSEIR observer. *Science China Information Sciences* 63(12), <https://doi.org/10.1007/s11432-020-3034-y>.
- Chen TM, Rui J, Wang QP, Zhao ZY, Cui JA, Yin L (2020b) A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious diseases of poverty* 9(1):1–8.

- Chen YC, Lu PE, Chang CS, Liu TH (2020c) A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering* 7(4):3279–3294.
- Cohen S (2020) The fastest vaccine in history. UCLA Health Connect. Accessed March 5, 2021, <https://connect.uclahealth.org/2020/12/10/the-fastest-vaccine-in-history/>.
- Collins J, Abdelal N (2018) Spread of disease. Accessed April 16, 2021, <https://calculate.org.au/wp-content/uploads/sites/15/2018/10/spread-of-disease.pdf>.
- Cooper I, Mondal A, Antonopoulos CG (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>.
- Craig B, Phelan T, Jan-Peter S, Steinberg J (2020) Improving epidemic modeling with networks. *Federal Reserve Bank of Cleveland* 2020(23):1–8.
- Cryer JD (2008) *Time Series Analysis: With Applications in R* (Springer, New York), second edition.
- Csermely P, London A, Wu LY, Uzzi B (2013) Structure and dynamics of core/periphery networks. *Journal of Complex Networks* 1(2):93–123.
- Department of Homeland Security (2021) Homeland infrastructure foundation-level data. Department of Homeland Security. Accessed February 15, 2021, https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0.
- Di Donato V, Ruotolo N, Smith-Spark L (2020) Italy coronavirus: Military called to enforce lockdown as 627 people die in 24 hours. CNN. Accessed March 28, 2021, <https://www.cnn.com/2020/03/20/europe/italy-military-coronavirus-intl/index.html>.
- Easley D (2010) *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, Cambridge, England).
- Edunov S, Bhagat S, Burke M, Diuk C, Onur Filiz I (2016) Three and a half degrees of separation. Facebook research. Accessed February 26, 2021, <https://research.fb.com/blog/2016/02/three-and-a-half-degrees-of-separation/#:~:text=The%20majority%20of%20the%20people,and%204.2%20degrees%20of%20separation.&text=Estimated%20average%20degrees%20of%20separation,between%203%20and%204%20steps>.
- Erdem K (2021) Accessed February 14, 2021, <https://erdem.pl/2020/06/finding-correlations-in-time-series-data>.

- Erdős P, Rényi A (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5(1):17–60.
- Firth JA, Hellewell J, Klepac P, Kissler S, Jit M, Atkins KE, Clifford S, Villabona-Arenas CJ, Meakin SR, Diamond C, Bosse NI, Munday JD, Prem K, Foss AM, Nightingale ES, Zandvoort Kv, Davies NG, Gibbs HP, Medley G, Gimma A, Flasche S, Simons D, Auzenberg M, Russell TW, Quilty BJ, Rees QJ Eleanor Mand Leclerc, Edmunds WJ, Funk S, Houben RMGJ, Knight GM, Abbott S, Sun FY, Lowe R, Tully DC, Procter CI Simon Rand Jarvis, Endo A, O'Reilly K, Emery JC, Jombart T, Rosello A, Deol AK, Quaife M, Hué S, Liu RM Yangand Eggo, Pearson CAB, Kucharski AJ, Spurgin LG, Group CCW (2020) Using a real-world network to model localized COVID-19 control strategies. *Nature Medicine* 26(10):1616–1622.
- Fortunato S (2010) Community detection in graphs. *Physics Reports* 486(3):75–174.
- Frost J (2020) Guidelines for removing and handling outliers in data. Statistics By Jim. Accessed February 15, 2021, <https://statisticsbyjim.com/basics/remove-outliers/#:~:text=Outliers%20are%20data%20points%20that,findings%20or%20distort%20real%20results>.
- Gallagher J (2021) Covid immunity: Can you catch it twice? BBC News. Accessed March 14, 2021, <https://www.bbc.com/news/health-52446965>.
- Gera R (2021a) Clustering coefficients. Lecture, Complex Networks, February 23, Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA.
- Gera R (2021b) Groups of vertices and core-periphery structure. Lecture, Complex Networks, February 2, Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA.
- Gilbert EN (1961) Random plane networks. *Journal of the Society for Industrial and Applied Mathematics* 9(4):533–543.
- Glen S (2016) Power law and power law distribution. Statisticshowto. Accessed April 16, 2021, <https://www.statisticshowto.com/power-law/>.
- Golbeck J (2013) *Analyzing the Social Web* (Elsevier, Amsterdam).
- Guare J (1990) *Six Degrees of Separation: A Play* (Vintage Books, New York).
- 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>.

- Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using networkx. Technical Report LA-UR-08-05495; LA-UR-08-5495; TRN: US201006%%1254, Los Alamos National Lab, Los Alamos, New Mexico, <https://www.osti.gov/biblio/960616>.
- Harvard University (2020) If you’ve been exposed to the coronavirus. Harvard University. Accessed February 15, 2021, <https://www.health.harvard.edu/diseases-and-conditions/if-youve-been-exposed-to-the-coronavirus#:~:text=Some%20tests%20may%20be%20done,t%20begin%20producing%20antibodies%20immediately>.
- Hoehl S, Rabenau H, Berger A, Kortenbusch M, Cinatl J, Bojkova D, Behrens P, Bödinghaus B, Götsch U, Naujoks F, et al. (2020) Evidence of SARS-CoV-2 infection in returning travelers from Wuhan, China. *New England Journal of Medicine* 382(13):1278–1280.
- Hoffmann T, Peel L, Lambiotte R, Jones NS (2020) Community detection in networks without observing edges. *Science Advances* 6(4), <https://doi.org/10.1126/sciadv.aav1478>.
- 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.
- Humphries MD, Gurney K (2008) Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence. *PloS One* 3(4), <https://doi.org/10.1371/journal.pone.0002051>.
- Humphries MD, Gurney K, Prescott TJ (2006) The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society B: Biological Sciences* 273(1585):503–511.
- Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (OTexts, Online).
- Hyndman RJ, et al. (2006) Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting* 4(4):43–46.
- Institute for Disease Modeling (2021) SEIR and susceptible-exposed-infected-recovered-susceptible (SEIRS) models. Accessed February 6, 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).

- Johns Hopkins University (2021) Mortality analyses. Johns Hopkins Coronavirus Resource Center. Accessed March 16, 2021, <https://coronavirus.jhu.edu/data/mortality>.
- Kano T, Yasui K, Mikami T, Asally M, Ishiguro A (2021) An agent-based model of the interrelation between the COVID-19 outbreak and economic activities. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 477(2245), <https://doi.org/10.1098/rspa.2020.0604>.
- Kemper A (2010) *Valuation of Network Effects in Software Markets: A Complex Networks Approach* (Springer-Verlag, Berlin Heidelberg).
- Kermack WO, McKendrick AG (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* Cold Spring Harbor Laboratory Press. Accessed February 14, 2021, <https://www.medrxiv.org/content/early/2020/05/15/2020.05.10.20097469>.
- Knight M, Leeming K, Nason G, Nunes M (2020) Generalized network autoregressive processes and the GNAR package. *Journal of Statistical Software, Articles* 96(5):1–36.
- Kotu V, Deshpande B (2014) *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner* (Elsevier Science & Technology, San Francisco, California).
- LaGrone S (2020a) Navy, CDC to study COVID-19 outbreak on carrier theodore roosevelt. USNI News. Accessed April, 28 2021, <https://news.usni.org/2020/04/17/navy-cdc-to-study-covid-19-outbreak-on-carrier-theodore-roosevelt>.
- LaGrone S (2020b) Trump: Hospital ship Comfort to go to New York City, Mercy set for west coast support. USNI News. Accessed March 2, 2021, <https://news.usni.org/2020/03/18/trump-hospital-ship-comfort-to-go-to-new-york-city-mercy-set-for-west-coast-support>.
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature London* 521(7553):436–444.
- 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 CD (1982) *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting* (Butterworth-Heinemann, Oxford, United Kingdom).

- Machhi J, Herskovitz J, Senan AM, Dutta D, Nath B, Oleynikov MD, Blomberg WR, Meigs DD, Hasan M, Patel M, Kline P, Chang RCC, Chang L, Gendelman HE, Kevadiya BD (2020) The natural history, pathobiology, and clinical manifestations of SARS-CoV-2 infections. *Journal of Neuroimmune Pharmacology: The Official Journal of the Society on NeuroImmune Pharmacology* 15:359–386.
- Milgram S (1967) The small world problem. *Psychology Today* 2(1):60–67.
- Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms* 6(2-3):161–180.
- Moore GE, et al. (1965) *Cramming More Components Onto Integrated Circuits* (McGraw-Hill, New York).
- Naguib MM, Ellström P, Järhult JD, Lundkvist Å, Olsen B (2020) Towards pandemic preparedness beyond COVID-19. *The Lancet Microbe* 1(5), [https://doi.org/10.1016/S2666-5247\(20\)30088-4](https://doi.org/10.1016/S2666-5247(20)30088-4).
- National Bureau of Economic Research (2010a) County distance database. National Bureau of Economic Research. Accessed February 15, 2021, <https://www.nber.org/research/data/county-distance-database>.
- National Bureau of Economic Research (2010b) County distance database. National Bureau of Economic Research. Accessed December 18, 2020, <https://nber.org/distance/2010/sf1/county/sf12010countydistancemiles.csv.zip>.
- Newman M (2008a) The physics of networks. *Physics Today* 61(11):33–38.
- Newman M, Barabasi AL, Watts DJ (2006) *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)* (Princeton University Press, Princeton, New Jersey).
- Newman ME (2008b) The mathematics of networks. *The New Palgrave Encyclopedia of Economics* 2(2008):1–12.
- Newman MEJ (2003) The structure and function of complex networks. *The Society for Industrial and Applied Mathematics review* 45(2):167–256.
- Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74(3), <http://dx.doi.org/10.1103/PhysRevE.74.036104>.
- Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Physical review. E, Statistical, Nonlinear, and Soft Matter Physics* 69, <https://doi.org/10.1103/PhysRevE.69.026113>.
- Newman MEJMEJ (2010) *Networks: an Introduction* (Oxford University Press, New York).

- 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 February 6, 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 February 7, 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 JM, Shay DK (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 HS (2016) Application of SIR epidemiological model: new trends. *International Journal of Applied Mathematics and Informatics* 10:92–97.
- Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R (2020) High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2 SARS-CoV-2. *Emerging Infectious Diseases* 26(7):1470–1477.
- Sanchez P, Sanchez S (TO APPEAR) *Simulation Analysis: AD Data Farming Framework* (TO APPEAR).
- Sanju'a R, Nebot MR, Chirico N, Mansky LM, Belshaw R (2010) Viral mutation rates. *Journal of Virology* 84(19):9733–9748.
- Senthilingam M (2017) Global pandemics: 7 reasons we're more at risk. CNN. Accessed March 5, 2021, <https://www.cnn.com/2017/04/03/health/pandemic-risk-virus-bacteria>.
- Silva PC, Batista PV, Lima HS, Alves MA, Guimarães FG, Silva RC (2020) Covid-abs: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals* 139, <https://doi.org/10.1016/j.chaos.2020.110088>.
- Singh D, Alanazi SA, Kamruzzaman MM, Alruwaili M, Alshammari N, Alqahtani SA, Karime A (2020) Measuring and preventing COVID-19 using the SIR model

- and machine learning in smart health care. *Journal of Healthcare Engineering* 2020, <https://doi.org/10.1155/2020/8857346>.
- So MK, Tiwari A, Chu AM, Tsang JT, Chan JN (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 February 15, 2021, <https://www.statisticssolutions.com/time-series-analysis/>.
- Sulasikin A, Nugraha Y, Kanggrawan J, Suherman AL (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: IEEE), <https://doi.org/10.1109/ISC251055.2020.9239072>.
- Svetunkov I (2021) Rolling origin. CRAN. Accessed March 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 February 7, 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 February 6, 2021, <https://arxiv.org/abs/2004.04735>.
- TheFactFile (2021) List of U.S. states and number of counties in each. Accessed January 15, 2021, <https://thefactfile.org/us-states-counties/>.
- Thompson L (2019) Ten capabilities that make the united states' navy's nuclear-powered aircraft carriers indispensable. Forbes. Accessed April, 28 2021, <https://www.forbes.com/sites/lorenthompson/2019/06/28/ten-capabilities-that-make-the-u-s-navys-nuclear-powered-aircraft-carriers-indispensable/?sh=7ff00dc395c>.
- Tolles J, Luong T (2020) Modeling epidemics with compartmental models. *Journal of American Medical Association* 323(24):2515–2516.
- Tomochi M, Kono M (2021) A mathematical model for COVID-19 pandemic—SIIR model: effects of asymptomatic individuals. *Journal of General and Family Medicine* 22(1):5–14.
- Traag V, Waltman L, van Eck N (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* 9(5233), <https://doi.org/10.1038/s41598-019-41695-z>.

- United States Census Bureau (2010) County adjacency file. United States Census Bureau. Accessed February 15, 2021, <https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html>.
- United States Census Bureau (2020a) 2019 FIPS codes. United States Census Bureau. Accessed February 15, 2021, <https://www.census.gov/geographies/reference-files/2019/demo/popest/2019-fips.html>.
- United States Census Bureau (2020b) 2020 census data. United States Census Bureau. Accessed February 1, 2021, <https://www.census.gov/programs-surveys/popest/technical-documentation/file-layouts.html>.
- USAFacts (2020) About USAFacts. USAFacts. Accessed February, 15, 2021, <https://usafacts.org/about-usafacts/>.
- USPS (2021) 42,000 ZIP codes. U.S. Postal Facts. Accessed March 5, 2021, <https://facts.usps.com/42000-zip-codes/#:~:text=There%20are%20nearly%2042%2C000%20ZIP%20Codes%20in%20the%20country.>
- Veremyev A, Prokopyev OA, Pasilião EL (2019) Finding critical links for closeness centrality. *Institute for Operations Research and the Management Sciences Journal on Computing* 31(2):367–389.
- Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393(6684):440–442.
- Wiemken TL, Kelley RR (2020) Machine learning in epidemiology and health outcomes research. *Annual Review of Public Health* 41(1):21–36.
- World Health Organization (2020) Impact of COVID-19 on people’s livelihoods, their health and our food systems. World Health Organization. Accessed February 19, 2021, <https://www.who.int/news/item/13-10-2020-impact-of-COVID-19-on-people’s-livelihoods-their-health-and-our-food-systems>.
- World Health Organization (2021) WHO coronavirus disease (COVID-19) dashboard. World Health Organization. Accessed February 13, 2021, https://covid19.who.int/?gclid=EAIaIQobChMI7M_PqZjo7gIV0ACtBh35Fw36EAAAYASAAEgLGePD_BwE.
- Yang J, Leskovec J (2012) Community-affiliation graph model for overlapping network community detection. Association for Computing Machinery. Accessed February 6, 2021, <https://dl.acm.org/doi/10.1109/ICDM.2012.139>.

Yoshida R (2020) Time series analysis. Lecture, Advanced Data Analysis, July 21, Department of Operations Research, Naval Postgraduate School, Monterey, CA.

Yum S (2020) Social network analysis for coronavirus (COVID-19) in the United States. *Social Science Quarterly* 101(4):1642–1647.

Zhang GQ, Cheng SQ, Zhang GQ (2012) A universal assortativity measure for network analysis. Cornell University. Accessed February 15, 2021, <https://arxiv.org/abs/1212.6456>.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

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