Diagnosing Prevailing Trends and Disparate Impacts of COVID-19 at the County-level

Justin Kaashoek

A thesis presented to the Department of Applied Mathematics
in partial fulfillment of the requirements
for a Bachelor of Arts degree with Honors

Harvard College
Cambridge, Massachusetts
March 26, 2021

Abstract

Since the beginning of the COVID-19 pandemic, there has been a significant amount of research on the dynamics of the disease and its unique effects on different locations. There remains a lot to learn about the spatial and temporal trends of the disease, as well as the disparate impacts on different populations, especially at finer levels of resolution such as U.S. counties. This research attempts to understand the differences between how and when COVID-19 hit counties and provide a narrative for the social and political forces that helped shape those differences.

To develop a spatial and temporal understanding of pandemic spread, we first apply Dynamic Mode Decomposition (DMD), a dimensionality reduction technique that allows us to find a linear approximation to a non-linear dynamic system, to county-level COVID-19 deaths. We find that DMD in it most basic form largely fails as a method when applied to COVID-19 data.

To develop an understanding of the disparate impacts of COVID-19, we next applied a series of machine-learning models to predict county-level death rates across different waves of the disease using a combination of social vulnerability, demographic, political, and behavioral variables as input. We find that politics played an increasingly important role in determining disease spread as the pandemic matured. We also find that social vulnerabilities are consistently important predictors of disease impact.

Acknowledgements

This work could not have been completed without the guidance, feedback, and support of numerous individuals over the past four years. I am extremely grateful to have such brilliant and understanding people in my life.

I would first like to thank my advisor, Dr. Mauricio Santillana, for his invaluable guidance, his willingness to let me explore different areas of interest, and for constantly pushing me to produce my best work.

Dr. Nancy Krieger, Dr. Bill Hanage, Dr. Jarvis Chen, Christian Testa, Dr. David Lazer, Dr. Jon Green, Matthew Simonson, and Diya Sashidhar, thank you for your willingness to share your expertise and for your patience as I tried (and often failed) to implement it.

Dr. Lucas Stolerman, Leonardo Clemente, and Andre Nguyen, thank you for your feedback, data insights, and code snippets.

To Professor Salmaan Keshavjee and Dr. Cecile Viboud, thank you for fostering my budding interest in Global Health nearly three years ago.

To my parents, Frans and Mathilda, I am so grateful for your help and constant support. To Nick, thank you for being an ever-willing partner in having a laugh at our parents' expense.

Lastly, I am so lucky to have such truly wonderful friends in my life. Jess, Michelle, Lizzie, Belinda, and Joseph: I could not have asked for a better blocking group to help see me through thick and thin. PJ, thank you for being a constant source of love and laughs. Meaghan, thank you for being a rock who always seems to know when I need help before I do. Jeana, Emily, Maggie, Luke, and Rina: your brilliance and kindness constantly push me in the best of ways. Cam, Brinkley, YG, Lizzy Y., Christi, Jordan, Nick L., Jess Z., Fatima, and Sofi: I am so grateful to have shared these past four years with you.

Contents

Chapter 1

Introduction

This chapter explains the progression of the COVID-19 pandemic and the importance of studying the disease to further understand how and why it has spread the way it has. In Section 1.2, we examine previous works surrounding disease modeling, dynamic mode decomposition, and social determinants of health, exploring how each has been adapted to the current pandemic. Finally, we present our contributions to these areas in Section 1.3.

1.1 Problem Motivation

From February 2020 to the time of this writing, COVID-19 — the disease caused by the betacoronavirus SARS-CoV-2 — has caused over 2,400,000 deaths worldwide, over 500,000 of which are in the US [15]. After a large spike in cases over the winter months, both positive tests and COVID-19 deaths are trending downwards. Along with the introduction of vaccines developed by Moderna, Pfizer, and Johnson & Johnson, these trends are cause for hope. President Biden has also stated that he aims for the US to "declare independence" from the virus by July 4th [58].

These positive trends warrant cautious optimism as we are seemingly approaching what could be the end of a devastating public health emergency, but the lull in cases and deaths also calls for reflection. The fact remains that the United States comprises only 4% of the worldwide population but saw over 20% of worldwide deaths. The US witnessed over twice as many deaths as Brazil,

the country with the second largest number of deaths. Despite the volume of COVID-19 research that has emerged over the past year, much remains to be learned about the disease, how it spread, and why the United States was so heavily impacted.

Gaining insights in these areas is still of great importance. First, even with the decline in cases, the majority of the U.S. population remains susceptible to SARS-CoV-2 [37]. Knowing which subset of the population is most vulnerable could help to inform non-pharmaceutical interventions (NPIs) such as school and business closures, work-from-home policies, and travel bans, or vaccine strategies. At the same time, knowing which counties are less vulnerable could help inform business re-openings.

Second, understanding how the pandemic spread throughout the country, determining which populations were impacted when, and analyzing the populations that were most vulnerable could inform future policies that could help prevent a similar outbreak from occurring. Although COVID-19 is a unique disease from an epidemiological perspective, and has its own reproduction number and mortality rate, there is evidence to suggest that another infectious disease might spread in a similar way [69]. COVID-19 was first observed in populous cities that were internationally connected before spreading to more rural areas. It seems likely that another infectious disease will follow a similar pattern, assuming a disease origin in another country. Populous cities are also more likely to be overcrowded, which may increase infectious spread and severity [69]. Understanding the nuances and less obvious trends of how COVID-19 spread and which populations it impacted could be essential in mitigating future diseases.

Finally, there is a lack of robust understanding around the relationship between politics and demographics when analyzing the severity of COVID-19. Over the past year, COVID-19 has become a highly politicized disease [13, 40]. Mask wearing and social distancing, two common NPIs that have been shown to reduce transmission, have become political statements [3, 23, 27, 30, 41]. How important are these behaviors in relation to demographic variables such as income and how have the relationships between these variables changed over the past year? By only focusing on the political, it is easy to lose sight of at-risk populations that experienced intense outbreaks

because family members could not afford to work from home or worked in places that imposed few regulations [25]. By separating out these effects and attempting to determine relationships between variables, we can develop a better understanding of the evolving roles these variables played in how hard different areas in the country were hit. This understanding, in turn, can help inform policies to address existing inequities and inform metrics such as the CDC Social Vulnerability Index (SVI). The SVI takes into account factors such as income, household crowding, and education and is meant to serve as a quantitative measure for areas that need additional support in the event of a an external stressor on health, such as a natural distaster [1].

It is important to develop our understanding of the disease at the county-level for four reasons. First, county-level analyses allow for more specific insights into disease spread and impact. Second, state-level impacts have been more commonly studied, and county-level data are less well understood. Third, counties within the same states, even neighboring counties, experienced the pandemic differently. Finally, counties within the same state have different demographics, vulnerabilities, and political leanings. Analyses on the state-level are not fine-grained enough and wash out these differences.

This research seeks to address the questions of how COVID-19 spread and where the disease had disproportionate impacts through two county-level analyses. First, we apply dynamic mode decomposition (DMD), a matrix decomposition technique, to county-level deaths in an attempt to discern prevailing spatial and temporal trends of the disease. This analysis is largely exploratory, and in addition to providing insights on COVID-19, attempts to understand if DMD is a useful technique to apply to epidemiological modeling. Second, we apply a combination of clustering and regression methods to understand the relationship between different political, behavioral, demographic, and social vulnerability variables and the severity of COVID-19 outbreaks at the county-level over the past year.

1.2 Related Works

Dynamic Mode Decomposition is a technique that was not originally designed for epidemiological modeling. The technique was first applied to study fluid dynamics and was developed to extract relevant flow features from experimental data [56, 57]. The basic idea behind the technique, extracting linear dynamics from a complex non-linear system using experimental data, has been extended to study the spread of disease. Specifically, DMD has been used to study influenza, where it was mainly used as a diagnostic, retrospective tool, rather than a predictive method [49]. The technique has yet to be applied to COVID-19, and there remain questions about the effectiveness of the method in epidemiological settings, particularly in providing predictive accuracy.

There is an abundance of other models that are frequently used for disease forecasting. Two examples of common models are SIR models and autoregressions. The SIR model assigns a population to one of three groups: S (susceptible), I (infected), R (recovered). Certain parameters of the model, such as how contagious a disease is or how common interactions are between populations, control the predicted spread of the disease. Autoregressive models fit linear regressions on subsets of consecutive time series data to predict future disease prevalence.

The majority of COVID-19 epidemiological research over the past year has made use of these types of models to allow for real-time prediction of COVID-19 cases. This research has allowed governors and public health officials to implement and recommend NPIs in hopes of mitigating an outbreak before it occurs. The NPIs have imposed serious economic burdens on small business, and predictive models have also been essential in determining when it is safe to re-open businesses without putting people at risk. Predictive models have been implemented in countries around the world, such as India, South Africa, Italy, and China [17, 48, 52, 71]. Other research has looked at the relationship between environmental factors and the spread of disease, a common relationship to examine in infectious disease modeling due to the higher transmissibility of many diseases in more humid climates [47].

In the United States, there has been similar predictive work done at the state-level [72]. At the county-level, however, models have produced mixed results, especially in counties with smaller

populations [9]. There are a number of potential explanations for why such models have not performed well. First, county-level data is noisy, and especially in smaller counties with less robust data collection, there can be large day-to-day fluctuations. Second, a lack of standardized data reporting has made it difficult to create a general dataset of county-level COVID-19 data across the country. Third, trends, demographics, and behaviors vary even among neighboring counties [31]. This heterogeneity makes it difficult for models that take into account the dynamics of nearby counties, as many SIR models do.

Other works have examined the disease from a social epidemiology lens, instead of attempting to predict the number or spread of COVID-19 cases. These works focus on the relationship between socio-economics, demographics, vulnerabilities, and COVID-19. Specifically, it has been well-documented that the disease has placed a higher burden on minority groups, areas of high household crowding, and low income neighborhoods [7, 25, 32, 33, 59]. There have been similar observations made in other countries, such as Italy [8]. Other results have shown that population density has also impacted the spread of the disease, as areas of higher population density in China experienced more severe outbreaks [51].

Much of this work in the United States has been conducted at the state-level and examines regional differences or differences within a specific state [25, 33]. Studies on the county-level have shown correlations between certain county-level factors and transmission rates [6]. The main demographic focus in these works has been on gender, proportion of African Americans, and income inequality. It has been shown that men, a higher percentage of African-Americans, and higher income inequality are associated with increased transmission rates [6]. County-level analysis had either been focused solely on demographics variables and excluded political variables, been focused around earlier months in the pandemic, or failed to account for the dynamic nature of certain variables. For example, it has been shown that mask use within states has changed as the pandemic progressed over the past year, and this should be taken into account when studying county-level data [35].

Analyzing the disparate impacts of diseases and other health conditions in this way is not

unique to COVID-19, and in general, there has been a shift towards accounting for and including demographics and vulnerabilities in epidemiological analyses [5, 44, 66]. One example of such analyses that is relevant to COVID-19 is that it has been well-documented that minority populations are less likely to seek out health care or have less access to care [62, 67]. Such findings hold during a pandemic, when access to care is of great importance. Specifically, after the 2009 H1N1 pandemic, research showed that Blacks and Hispanics were at a higher risk of exposure to H1N1 and that 75% of Hispanics lacked a regular healthcare provider [50].

Just as the pandemic cannot be separated from social determinants of health, COVID-19 has also been inextricably tied to political ideologies. The disease has become highly politicized, and although attitudes towards the disease have shifted throughout the year, Democratic-leaning communities have generally tended to believe that the disease is more of a public health threat than predominantly Republican communities [42, 55]. With the lack of federal mandates, public health measures have largely fallen to state governors [25]. The stringency of interventions, as well as the media coverage of the disease, have fallen largely along party lines as Republican governors tended to have less strict measures than Democratic governors [2]. Adherence to public health measures such as mask wearing has seen a political dimension as well, with Democrats showing higher rates of mask usage [30]. These findings suggest both a top-down and bottom-up effect of politics. In the top-down effect, right-leaning governors are less likely to institute strict regulations, which have been proven to help mitigate the effects of COVID-19 [26]. In the bottom-up effect, right-leaning individuals are less-likely to abide by health guidelines or are less likely to engage in behaviors such as mask wearing or social distancing regardless of policies.

While difficult to separate the effects of social vulnerability, demographic, political, and behavioral (tied closely to political) variables, previous works have detailed the importance of studying the relationships between social vulnerability and politics [5, 44]. Related to these works, there have been efforts to develop conceptual frameworks that detail how politics, vulnerability, and disease might interact [68]. Specifically, the WHO had laid out a framework that, explains how, "...social, economic and political mechanisms give rise to a set of socioeconomic positions, whereby

populations are stratified according to income, education, occupation, gender, race/ethnicity and other factors; these socioeconomic positions in turn shape specific determinants of health status (intermediary determinants) reflective of people's place within social hierarchies" [68]. This research will draw upon this framework, as explained in more detail in Figure 3.6.

1.3 Contributions

This research presents the two main avenues through which we attempt to address gaps in the existing literature. In both of these avenues, we use county-level COVID-19 data from Johns Hopkins' publicly available data repository, which contains data on 3,143 counties [15]. Specifically, we use cumulative death counts for all available counties in the United States. These data contain counts from January 22, 2020 to February 12, 2021, the date of the most recent data pull. We choose to focus on death counts because, especially early on in the pandemic, daily case counts have been shown to be highly correlated with state testing capacity [29]. Deaths are believed to be more reliable for this time period in the pandemic [29].

In Chapter 2, we present a methodological contribution by applying dynamic mode decomposition (DMD) to county-level deaths in the United States. We first perform singular value decomposition (SVD) to examine the prevailing spatial and temporal trends of the system. We then perform DMD to (1) retrospectively examine the prevailing dynamics of COVID-19 in the United States and (2) attempt to predict future dynamics. Our results suggest that DMD in its most basic form is not an informative technique for disease modeling, and therefore, is unable to provide any new understandings in the spread of COVID-19.

In Chapter 3, we attempt to robustly and quantitatively analyze the effects of political, behavioral, demographic, and vulnerability factors on how intensely U.S. counties experienced COVID-19. We split the year into three time periods (1: Feb. 2020 - May 2020, 2: June 2020 - Sept. 2020, and 3: Oct. 2020 - Feb. 2021) to address the temporal nature of certain variables. Our main contribution is in our analysis of period 3, the period of the majority of deaths within the country.

We show that partisanship played an increasingly important role as the pandemic progressed and that variation in factors such as a county's political leaning and the stringency of governor interventions are two of the strongest causes of variation in pandemic impact during the winter months, after controlling for the effects of other variables. Throughout the pandemic, vulnerabilities and demographic factors played important roles in predicting pandemic severity.

Chapter 2

Uncovering Prevailing Trends using Dynamic Mode Decomposition

While DMD was not originally conceived in an epidemiological context, the method has been extended to study infectious diseases, particularly influenza. The benefits of the method are that it allows us to model the system based on simulated, experimental, or historical data [56]. These data are the only required input for the method to work, which removes the need for a set of governing equations. Other models, specifically SIR-type models, rely on these governing equations and require a basic reproductive number, R_0, that estimates the number of individuals a single infectious person will infect. Estimating this value, especially in the case of a new disease such as COVID-19 can be challenging. Add in the fact that behaviors such as mask wearing or social distancing reduce R_0, and these models quickly become reliant on assumptions in the absence of extensive research. DMD, on the other hand, can be performed solely on historical data, which can often be readily found for infectious diseases such as influenza. For COVID-19, this method would not work well early in the pandemic because we have only a few months of data. Now that we have approximately a year's worth of daily data, the method has promise. If our results were to lend insights into the spread of COVID-19 and corroborate previous findings, we would have evidence to suggest that DMD is a powerful tool that can be used to model disease dynamics without the

assumptions and equations other models rely on.

2.1 Background Knowledge

In this section, we introduce the math behind DMD. The method relies on principal component analysis (PCA), a dimensionality reduction technique, and time-series analysis that allows us to identify frequencies of a time-varying signal [49]. In Section 2.1.1, we introduce and explain singular value decomposition, a method that helps to conduct PCA. In Section 2.1.2, we explain how to extend PCA to DMD.

2.1.1 Singular Value Decomposition and Principal Component Analysis

Singular value decomposition (SVD) is a type of matrix decomposition for rectangular matrices [21]. Say we have a data matrix $D \in \mathbb{R}^{n \times m}$. The equation for the SVD of our matrix is

$$D = U\Sigma V^T$$

where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{m \times m}$. U and V contain the left and right singular values of D, respectively. The singular values themselves are stored as entries of the diagonal matrix $\Sigma \in \mathbb{R}^{n \times m}$. The magnitude of each singular value indicates the amount of energy stored in that mode. They are stored in descending order such that the first columns of U and V, which correspond to the first singular value, are the most prevailing spatial and temporal trends. The columns of U and V are the orthonormal eigenvectors of DD^T and D^TD, respectively.

SVD is also useful because it allows us to reconstruct the matrix D using only the first r columns of U, Σ, and V. Instead of having an n by m matrix, then, we will have an n by r matrix. The proportion of the total energy of the system that is captured by using a rank r approximation is

$$\frac{\sum_{i=1}^{r} \sigma_i}{\sum_{j=1}^{m} \sigma_j}$$

where σ_i corresponds to the ith singular value. Generally, we can select r using an elbow plot. An elbow plot shows the value $\frac{\sigma_i}{\sum_{j=1}^{m} \sigma_j}$ for $i \in 0, 1, \ldots, m$. Because the singular values are stored in descending order, a "kink" in the elbow plot represents a singular value that does not significantly contribute to the total energy of the system. Subsequent values, then, will also not significantly contribute. The value r is chosen to be near a kink in an elbow plot.

PCA generally refers to this process of reducing a high-dimensional data structure to a lower dimension by capturing the components that most contribute to the variance of the data. It follows that SVD is a useful method that allows us to conduct PCA.

2.1.2 Dynamic Mode Decomposition

The DMD presented in this thesis is based off work done by Joshua Proctor and Nathan Kutz [34, 49]. DMD seeks to find a relationship between measurements at time t and time $t + 1$. Again, let $D \in \mathbb{R}^{n \times m}$ be our data matrix, where we have n observations (rows) at m time-steps (columns). For all pairs of data, then, we have

$$D_{t+1} = AD_t$$

where A is a propagator matrix that takes us from our observations at time t to our observations at time $t + 1$. Let X be a matrix containing the first $m - 1$ time steps ($D[, 0 : m - 1]$) and X' be a matrix of timesteps 1 through m ($D[, 1 : m]$). Both matrices are, then $\mathbb{R}^{n \times (m-1)}$. We can think of X as a matrix of previous measurements and X' as a matrix of future measurements. In our case, measurements are taken daily, so $\Delta t = 1$. We are looking for a propagator A that will transform our previous measurements one time step into the future:

$$X' = AX$$

We can find A using the Moore-Penrose pseudo-inverse [63]:

$$A = X'X^{\dagger}$$

This can be inefficient and time-expensive to calculate, so we can perform PCA on X to find a lower-rank approximation:

$$X \approx U_r \Sigma_r V_r^T$$

Here, $U_r \in \mathbb{R}^{n \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, $V_r^T \in \mathbb{R}^{r \times (m-1)}$. We then calculate $\tilde{A}$, the approximation to A:

$$\tilde{A} = U_r^T X' V_r \Sigma_r^{-1}$$

We then calculate the eigendecomposition of $\tilde{A}$:

$$\tilde{A}W = W\Lambda$$

where W is a matrix where each column is an eigenvector and Λ is a diagonal matrix with corresponding eigenvalues on the diagonal. The final dynamic modes are:

$$\phi = X' V_r \Sigma_r^{-1} W \tag{2.1}$$

We can think of ϕ as a projection of our data (X') onto the eigenvalues of $\tilde{A}$.

2.2 Methods

For our analysis, we pre-processed John's Hopkins county-level death data. First, we filter the dataset to include only counties within the 50 states, which excludes the Diamond Princess and Grand Princess cruise ships, Guam, Puerto Rico, the Virgin Islands, Northern Mariana Islands, and American Samoa. We then convert the cumulative death counts to new deaths per day.

In certain cases, this resulted in a negative number of new deaths due to changing county data collection procedures or reporting errors. In such cases, we assume that there were no new deaths that day. We also attempted to interpolate such cases without a significant change in results.

We then apply a seven day smoothing to the resulting dataset to remove some noise. We remove counties that have fewer than 5 deaths as these generally had particularly noisy pandemic curves, which excludes 131 counties. Finally, we convert the time series for each county to z-scores by removing the mean and dividing by the standard deviation. This standardization will allow us to compare the shape of the pandemic curves in counties with extreme differences in population.

The resulting data set is stored in a matrix, D, where each row of the matrix contains the standardized time series of death counts in a particular county, and each column represents one day's worth of counts. D is, then, a 3114 x 382 matrix because we have 3114 counties and 382 days of data. We will let n and m refer to the number of rows and columns of the matrix, respectively.

2.2.1 Dynamic Mode Decomposition

We first calculate the SVD of our D matrix. The left singular vectors, in this case, correspond to the prevailing spatial modes, while the right singular vectors correspond to the temporal modes. We want to perform a dimensionality reduction, so we examine an elbow plot, as shown in Figure 2.1. We see that the largest mode only contains about 5.5% of the total energy of the system, meaning that none of the modes contain an extremely high proportion of the total energy. We decide to use 180 modes to reconstruct the data.

We first perform a retrospective analysis by calculating DMD on the entire data matrix D. We divide D into X and X' by assigning the first $m-1$ columns of D to X and columns $1, 2, \ldots, m$ of D to X'. We then calculate ϕ, the dynamic modes as in Equation 2.1. We can think of each entry in a column of ϕ as corresponding to a county in the United States. The magnitude of that entry describes the locations involved in that dynamic of disease spread while the angle of the entry describes the phase of the given county's peak infection time relative to the other counties [49].

We now need a method for selecting dynamic modes. First, we can calculate the frequency of

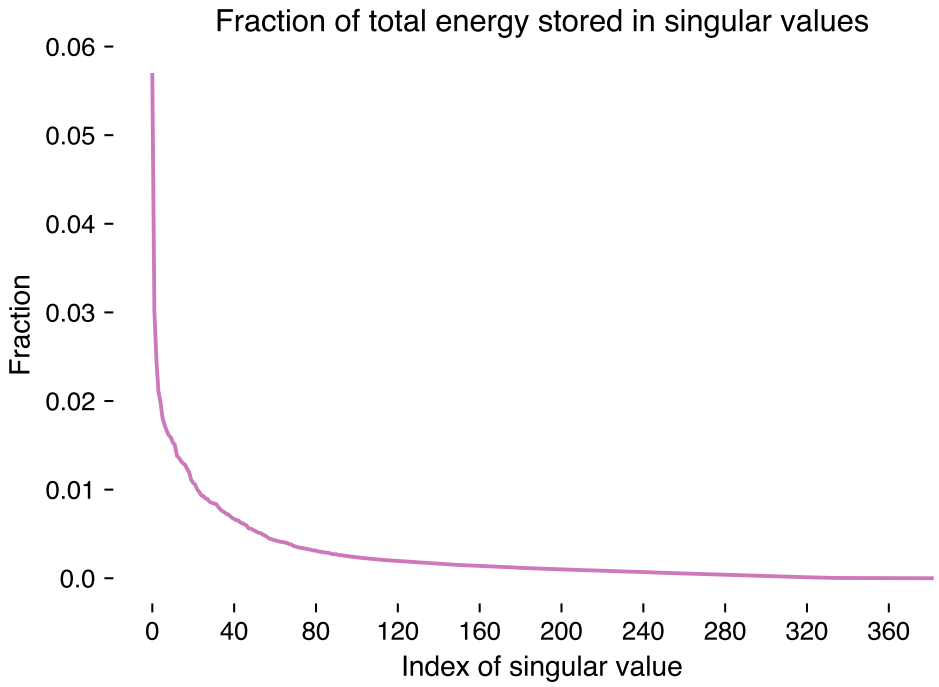

Figure 2.1: An elbow plot showing the fraction of the total energy stored in all singular values. We see that none of the singular values store an extremely high proportion of the energy. We therefore select 180 modes in our dimensionality reduction, as this accounts for 90% of the total energy of the system. By selecting 180 modes, we select more modes than indicated by the kink in this figure, which allows us to retain a higher percentage of the total energy of the system.

oscillation for eigenvalue j:

$$f_j = \frac{(\text{imag}(\log(\lambda_j))}{2\pi}$$

Note that imag means taking the imaginary part of the complex eigenvalue.

We can select the dynamic modes of interest according to the magnitude of $\lambda_j^p ||\phi_j||$, for some p, relative to the frequency [49]. This value, $\lambda_j^p ||\phi_j||$, represents how much each dynamic mode is impacting the system p steps in the future. We want to select modes that are still impacting the system p time steps into the future.

We can then visualize both the magnitude and angle of each dynamic mode on a map of the United States. We can also visualize the time dynamics of the jth eigenvalue by calculating $e^{\log(\lambda_j)t}$ for $t \in 1, 2, \ldots, m$. This calculation will provide a visualization of both the exponential growth or decay (real part of the eigenvector) and oscillation (imaginary part of the eigenvector).

Finally, we can attempt to reconstruct the system. The reconstructed system at time t is determined by the following:

$$D_{recon,t} = \phi e^{\Lambda t} b \tag{2.2}$$

where we can find b from our initial conditions:

$$D_0 = \phi b \implies b = \phi^\dagger D_0$$

We can also apply a similar calculation to attempt to predict future states of the system. In this prospective analysis, we look to predict one time step into the future. To evaluate how well this works, we start by calculating the dynamic modes using only the first 80% of days ($0.8 \cdot m = 305$ days) available to us. We can then predict the next day's deaths using Equation 2.2. More concretely, we start at time step $t = 305$ and perform the following at every subsequent time step:

1. Calculate the dynamic modes using the first t columns of D

2. Find b using D_t (the t th column of D) as the initial condition

3. Predict D_{t+1} by calculating $\phi e^{\Lambda(t+1)}b$

We perform this predictive analysis twice, once using only counties with a population greater than 100,000 ($n = 596$) and once using all counties. We evaluate the error by calculate the MSE of the predicted deaths and actual deaths at each time step.

2.3 Results

Figure 2.2 contains a plot of the first three right singular vectors of our data matrix D. These three vectors correspond to the three modes with the most energy of the system. It is important to note that singular vectors are only accurate up to a negative sign. The plotted vectors have been reflected across the line $y = 0$. The trends of the orange line are similar to those in counties that were largely unaffected by COVID-19 until experiencing an outbreak over the winter. The blue line could be representative of counties that experienced two outbreaks: a more mild one over the summer and a severe one over the winter. The green line also indicates two peaks: one around April 2020 and another over the winter months, similar to many Northeastern counties.

We can now move on to the DMD of the system. We reconstruct our data using 180 singular vectors and find the eigenvalue spectrum shown in Figure 2.3. The eigenvalues that fall within the unit are associated with modes that decay over time.

We can now select DMD modes based on the $\lambda_j^p||\phi_j||$ value of the mode relative to the mode's frequency, as discussed in Section 3.2.1. Figure 2.4 displays this plot using $p = 100$. We see that there is a cluster of points in the top left of the plot. We select 5 of these modes and plot a visualization of them in Figure 2.5. The maps on the left of this figure show the relative phase of each county's peak infection time [49]. The maps on the right show which counties were involved in the dynamic of disease spread associated with each DMD [49].

The map in the left column of the first row, showing the relative timing of peak infection, picks up on intuitive dynamics that counties in the Southeast had a similar peak infection times.

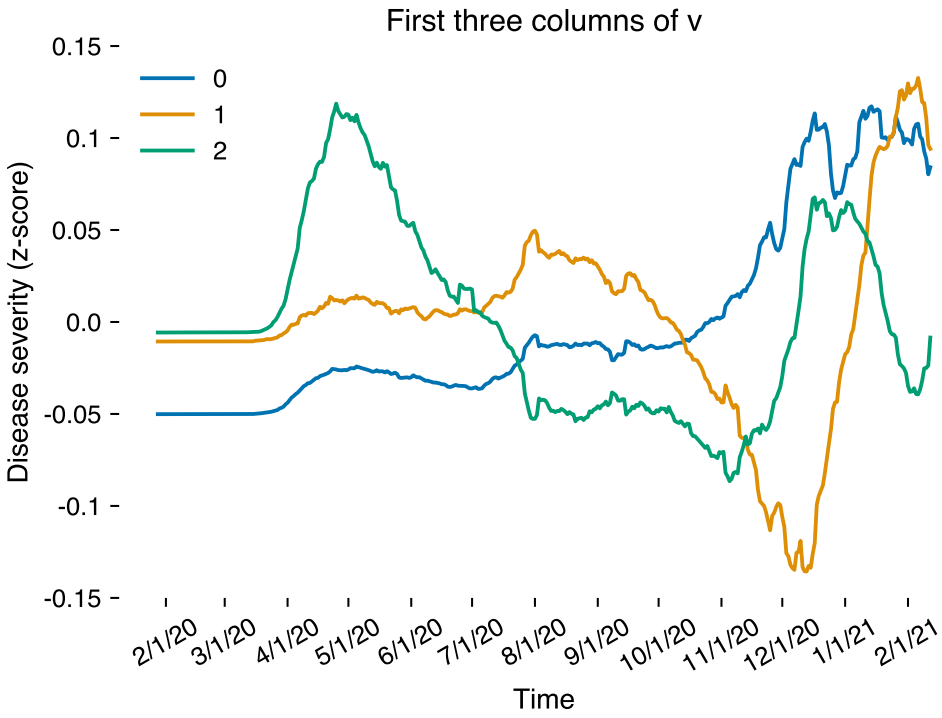

Figure 2.2: The first three right singular vectors of the entire system. These vectors represent the prevailing temporal trends of the system. We see dynamics that match the general progression of COVID-19 over the past year: an early peak between March 2020 and June 2020 followed by a universal spike in the winter months.

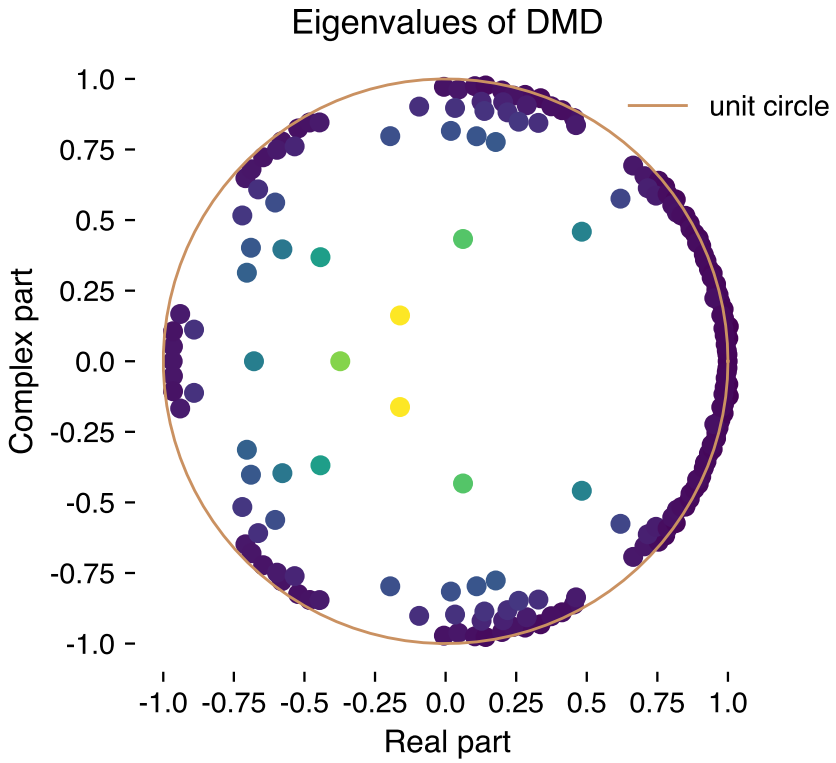

Figure 2.3: The eigenvalues of the DMD of the system. Points are colored based on the magnitude of the eigenvalues. We see that, as expected, the majority of points are near the unit circle (the tan line). All of the eigenvalues that fall within the unit circle are decaying modes. Those that fall outside of the unit circle are growth modes.

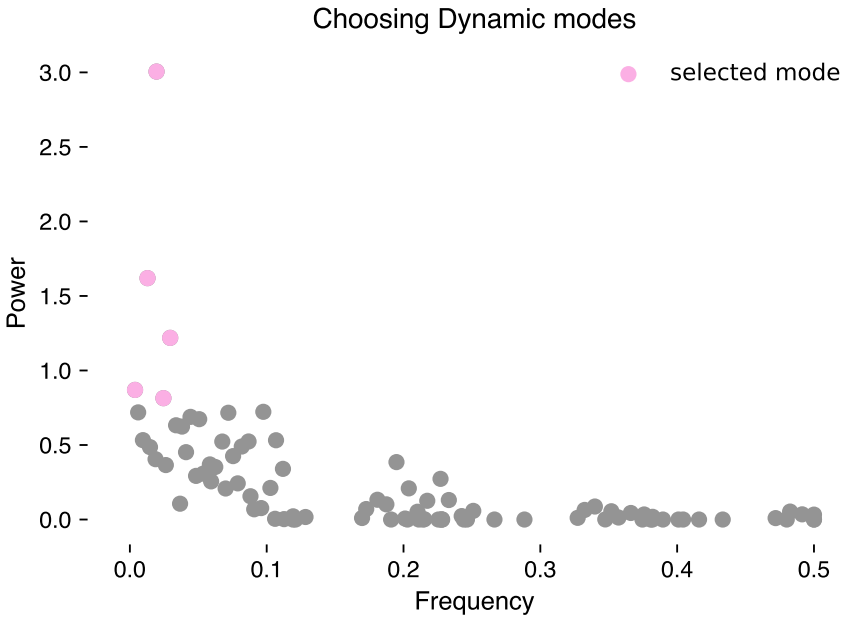

Figure 2.4: A plot of the frequency of a given DMD mode ($\frac{(\text{imag}(\log(\lambda_j))}{2\pi}$) on the x axis and $\lambda_j^{100}||\phi_j||$ on the y axis. We select modes that have a high $\lambda_j^{100}||\phi_j||$, which are the modes on the top left of the plot, as indicated by the pink color. Figure 2.5 shows a visualization of these modes.

Similarly, counties in the Northeast are shaded the same color. The magnitude plot for this row suggests that Northeastern counties were involved in this dynamic of disease spread, which could mean that the mode is picking up on the arrival of COVID-19 in the US during Febrary 2020 - March 2020. Other maps in this plot, however, do not pick up on any clear trends. The middle column shows DMD modes with decaying, oscillatory dynamics (rows 1 and 4) and modes with growing, oscillatory dynamics (rows 2, 3, and 5). The periods of the DMD modes are 265 days, 51 days, 77 days, 40 days, and 33 days, respectively. This suggests that COVID-19 has an oscillatory pattern: an outbreak occurs, lockdown measures are tightened causing a decrease in deaths, and when measures are relaxed or individual behavior changes, deaths spike again. The longer oscillatory frequencies of 265 and 77 days could correspond more to this pattern. An oscillation period of 33, 40, or 51 days is extremely frequent, however. Perhaps these modes reflect how quickly COVID-19 spread between adjacent counties or simply show the noisiness of the data.

As the final step to our retrospective analysis, we attempt to reconstruct the system using Equation 2.2. The results are shown in Figure 2.6. We see that the DMD reconstructions are quite

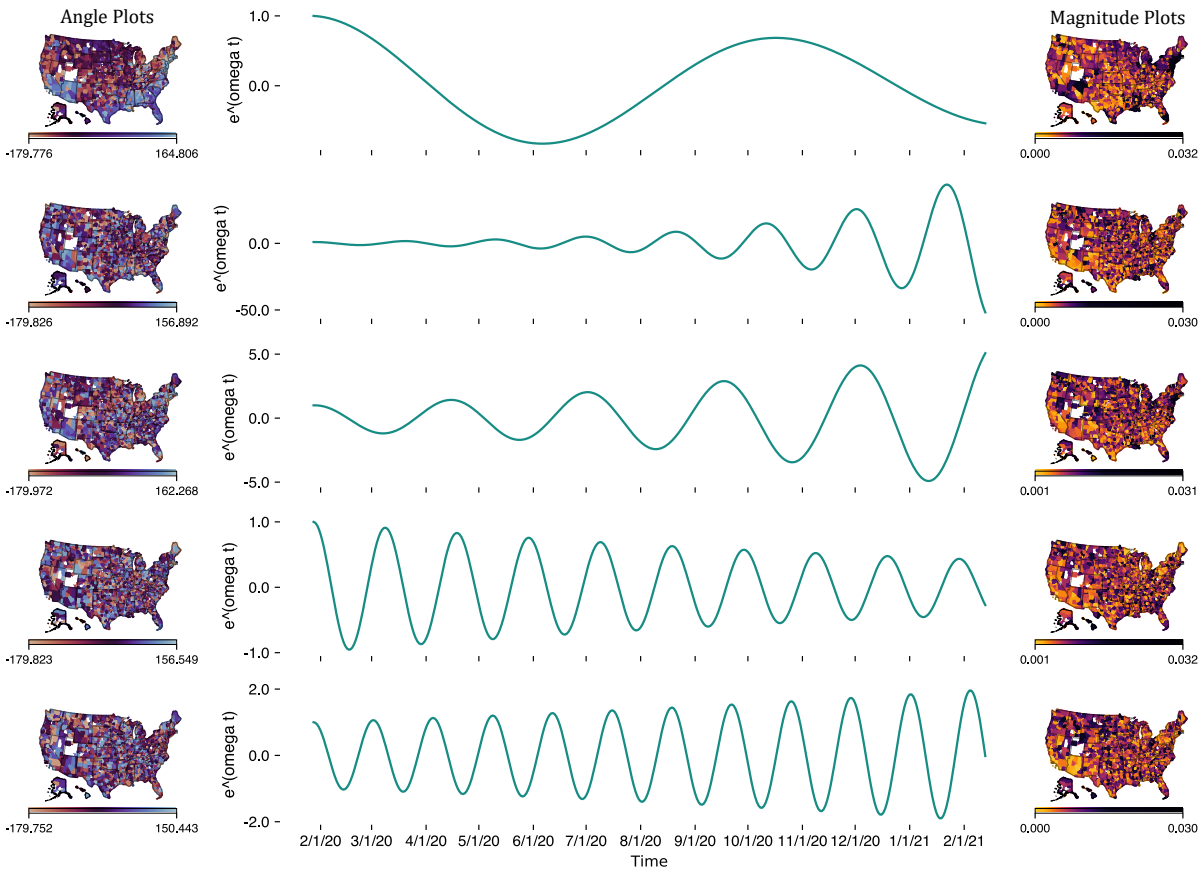

Figure 2.5: A visualization of 5 selected DMD modes. Maps in the left column contain a plot of the relative phase of each county represented by the given DMD mode, while maps in the right column contain a plot of the counties that are involved in the given dynamic of disease spread. The modes plotted are all small in magnitude. The center column contains time series plots with time on the x axis and $e^{\log(\lambda)t}$ on the y axis, where λ is the eigenvalue associated with the mapped DMD mode. The first mode, indicated by the first row, appears to pick up on some intuitive dynamics. The angle plots show a clustering of counties that have similar peak infection times in the Southeast and Northeast. The magnitude plot similarly shows a clustering of relatively high-magnitude areas in the Northeast. Other plots are less intuitive. The center column shows that we have two decaying modes (rows 1 and 4) and three growth modes (rows 2, 3, and 5).

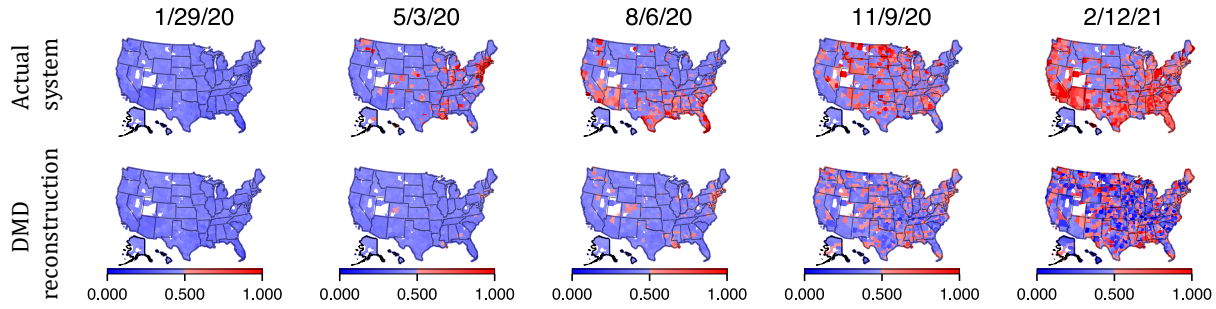

Figure 2.6: A visualization of the actual system (top row) compared to the DMD reconstruction system (bottom row), as determined using Equation 2.2. The plot shows the system state at 5 evenly spaced time points between January 28, 2020 and February 12, 2021. We see that DMD reconstruction performs poorly. We do not have data for counties shown in white.

different from the actual state of the system.[1] All counties start off with no COVID-19 activity and slowly become more active in both the actual system and the DMD reconstructed system. The areas of activity are largely different, however.

While these results suggest that DMD does not perform well in this COVID-19 setting, we still attempt to apply the method to perform prospective analysis. Figure 2.7 shows a plot of predicted z-score vs. actual z-score in 10 randomly-selected counties in the United States. The mean squared error of these estimates is 11.73. Figure 2.8 shows the same plot but filtering for counties with a population greater than 100,000. The mean squared error of these estimates is 31.35. We see that filtering for only large counties yields a higher MSE, likely due to the fact that larger counties saw more deaths which leaves more room for predictions to be incorrect.

2.4 Discussion

In the first few steps of our analysis, we observed intuitive results that suggested that the techniques presented in this thesis could be useful tools in studying the dynamics of COVID-19. For example, the three right singular values that resulted from SVD displayed intuitive temporal trends. The

[1] In a conversation with the author of [49] about his work, Joshua Proctor said that DMD often perform poorly in reconstructing original data. Because our eigenspectrum contains values inside the unit circle, over time these values will decay to 0. We also have eigenvalues outside of the unit circle, which will go to infinity over time

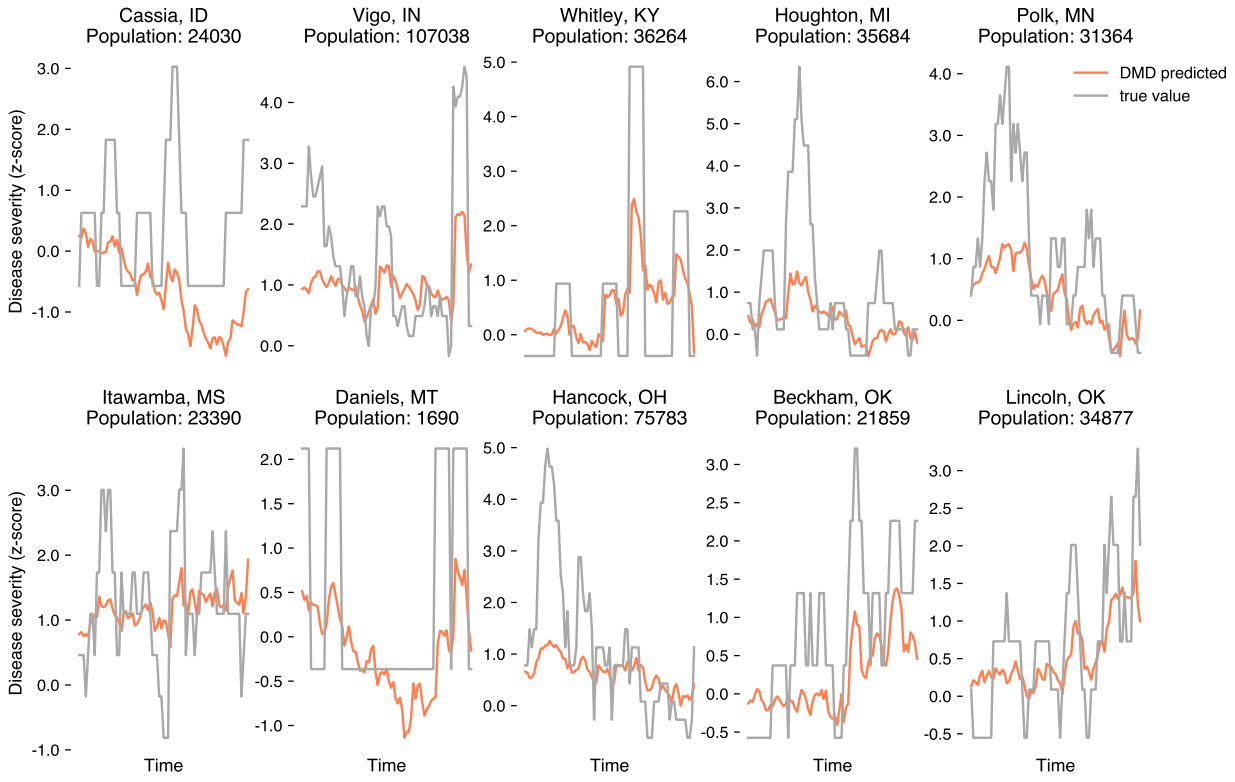

Figure 2.7: A plot of the DMD predictions (orange line) and the actual pandemic severity (gray line) for 10 random counties in the United States, without any filter for population. DMD predictions largely follow the overall shape of the true values, but do not pick up on the magnitudes of changes well. We also note the noisy pandemic curves in some small counties that make predictions challenging.

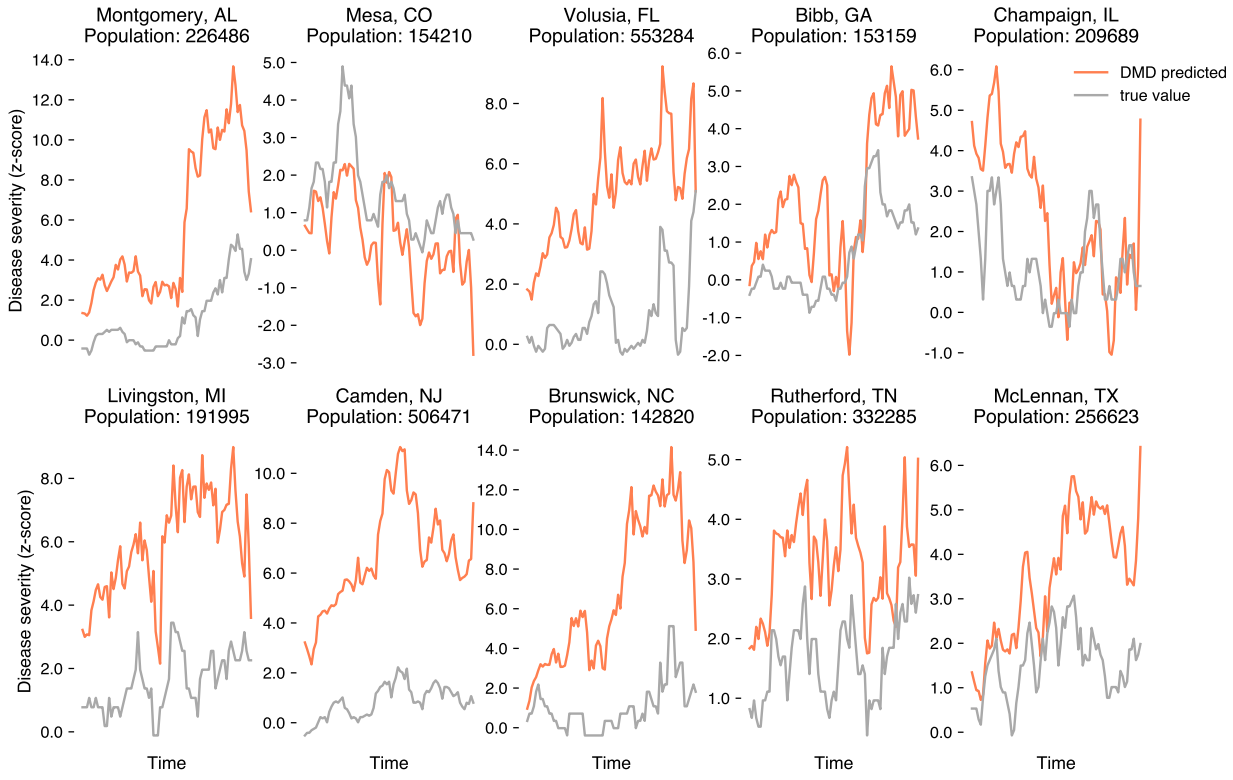

Figure 2.8: A plot of the DMD predictions (orange line) and the actual pandemic severity (gray line) for 10 random counties in the United States, filtering for population greater than 100,000. We see that DMD picks up on fluctuations better than in Figure 2.7.

dynamic modes that resulted from DMD yielded phase lengths trends and oscillation frequencies that passed sanity checks and one of our modes indicated intuitive tends.

Unfortunately, many of the following results were not clear. Some of the dynamic modes were counter-intuitive and we observed poor reconstructions. The failings of this approach are likely due to the relative simplicity of this analysis. We observed that many dynamic modes are raised to an exponent less than 1, which will cause these modes to decay over time. Other modes that are raised to an exponent greater than 1 dominate the system over time. These shortcomings call for more sophisticated variations of DMD. One possibility, which is saved for future analysis, is optimized DMD [4].

Another possible limitation of using DMD on COVID-19 data, which is difficult to overcome, is the lack of long-term data on the disease. In the past, DMD has been applied to study flu dynamics, which has years of available data and much clearer oscillatory behavior [49]. There has also been extensive research on flu dynamics, so it is easier to determine if DMD results match previous work. For this project, we have only a year of available data. It might be the case that not enough time has elapsed for us to fully understand the behavior of the system. COVID-19 data from counties is also noisy, especially relative to flu data. This might make it more difficult to discern the prevailing dynamics of the system.

Chapter 3

The Evolving Roles of Partisanship and Vulnerability in the COVID-19 Pandemic

In this chapter, we use multiple data sources to comprehensively analyze the relationship between political variables, measures of vulnerability, demographic differences, health-related behaviors, and COVID-19-related deaths at the county-level. We focus specifically on understanding the heterogeneous impact of COVID-19 during the winter months because, of the nearly 500,000 deaths, over half of them have occurred since October. By October, COVID-19 had been pervasive throughout society for nearly 8 months and the benefits of non-pharmaceutical interventions (NPIs) had already been firmly established in the literature [26, 38, 41, 70].

To conduct this analysis, we separate the past year of the pandemic into three time periods and develop a comprehensive narrative of the relationships between explanatory variables and county-level deaths in each period. In the first period, which spans from February to May, the pandemic caught the country by surprise. There was minimal existing research on how to mitigate disease spread and the pandemic took hold in large, urban areas, particularly in the Northeast (see Figure 3.1). By the second period, which spans the summer months from June to September, research had firmly established the importance of mask wearing, social distancing, and other NPIs. The disease spread to new areas, however, and specifically more southern or rural counties experienced

their first wave. By period three, which spans the winter months, the disease had been present for nearly 8 months and there was both a robust literature and precedent on how to successfully mitigate disease spread. As shown in Figure 3.1, however, the country experienced over half of its total deaths in these months. The impacts of the disease were not uniform, however, as counties in certain areas were able to better prevent deaths better than others.

In our analysis, we explore why the disease had an intense, heterogeneous impact in this third period. By dividing our analysis into distinct periods, we are able to to account for the dynamic nature of certain variables such as mask wearing and provide a narrative as relationships change [35]. We show that partisanship played an increasingly important role as the pandemic progressed, and that variation in factors such as a county's political leaning and the stringency of governor interventions are two of the strongest predictors of variation in pandemic impact during the winter months. Throughout the pandemic, vulnerabilities and demographic factors played important roles in predicting pandemic severity.

3.1 Methods

This section presents the data sources we used, an initial descriptive analysis of interactions between a subset of variables, and models that we will use to rigorously test a set of hypotheses on the driving forces behind COVID-19 outcomes.

3.1.1 Data Sources and Caveats

This section presents the data sources and caveats for the variables we study. All the variables are shown in Figure 3.6.

Response Variable

Our response variable of interest is, again, county-level death as provided by Johns Hopkins's publicly available data repository [15]. We perform the same pre-processing and smoothing as

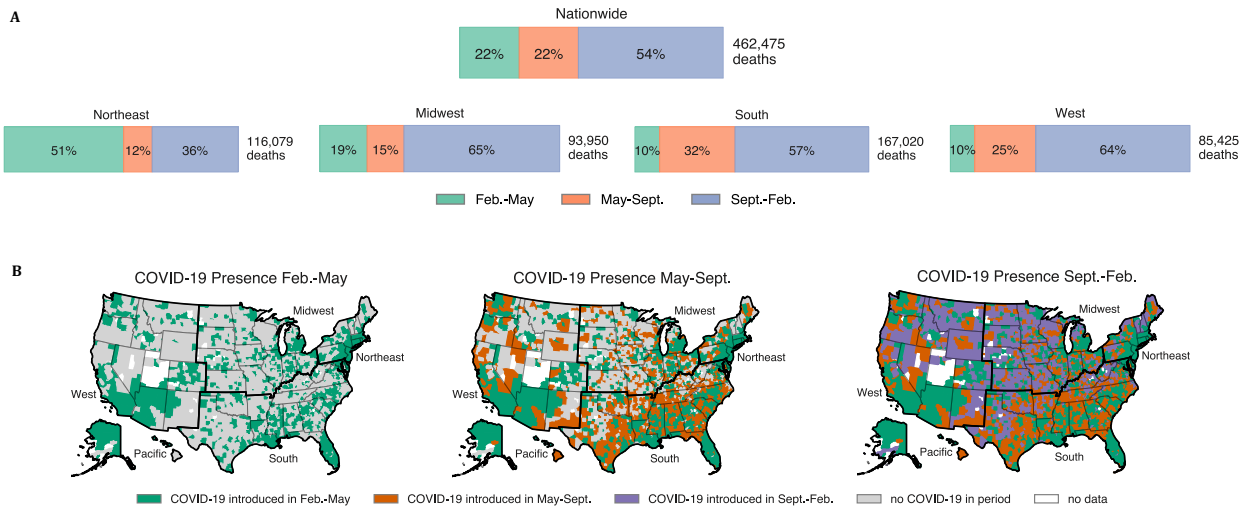

Figure 3.1: A breakdown of COVID presence across the time periods of interest, where the start and end dates of time periods have been adjusted such that each period is the same length. In (**A**), the top half of the figure, we present the percentage of deaths in the nation by time period, both nationwide and by census region. This part of the figure is inspired by [19]. Other than the Northeast, which was hit hard in the first period, the nation was hit hardest in period 3. This provides justification for a closer lens on this period. In (**B**), the bottom half of the figure, we visualize COVID-19 presence across the country at the county-level. COVID-19 is considered present in a county if it experiences 5 COVID-related deaths. We see the movement of COVID from the cities along the coasts to the center of the county throughout the year.

indicated in Section 2.2.

American Community Survey

We obtain race, education, age, income, population density, and household crowding from the U.S. Census Bureau American Community Survey's 5-year estimates [64]. We break down race by percent White not Hispanic, Black, Asian, Hispanic, and other. Each of the White, Black, Asian, and Hispanic percentages indicate estimates for the number of individuals who are exactly one race. We include percent People of Color (POC) only in our descriptive analysis (detailed in Section 3.1.3).

We calculate education as the percent of the population with at least a high school diploma. A high "education" value, then, indicates a highly educated county.

We break down age by percent of the population under the age of 19 and the percent of the population over the age of 65.

Income represents the median household income of a county.

Population density is calculated by dividing a county's population by its available land area, as provided by the Census Bureau.

Household crowding is also provided by the ACS and is the estimation of the number of households with more people than rooms.

Note that the values of all variables from the ACS are considered constant across the pandemic. This means, for example, that our income variable does not account for income shocks that occurred as a result of the pandemic. These are all pre-pandemic values.

Center for Medicare and Medicaid Services

We obtain nursing home data from the Centers for Medicare & Medicaid Services, which provide data for the average number of nursing home residents per day, organized by provider [61]. We divide this number by a county's population to determine the percent of the population that is in a nursing home.

It is important to note that these data were most recently updated in February 2021. As a result, they might inherently reflect the impact that COVID-19 has had on the elderly population.

VERA Institute

We obtain incarceration data from the Vera Institute of Justice, an institute that has worked to determine jail and prison populations at the county-level [65]. We elect to use only the jailed population as a proxy for the incarcerated population because prison data was only available in all states in 2014, whereas jail data is available since 2018. We believe that using incarceration data that is more than five years old might not reflect the current incarcerated population. Delaware, Rhode Island, Vermont, and Connecticut do not have local jails, and therefore, do not participate in the Bureau of Justice Statistics jails data collection [65]. We say that these states do not have any jailed population.

PLACES

We obtain obesity data from the CDC's PLACES dataset [11]. Obesity is only calculated using respondents who are aged 18 years or older.

U.S. Department of Transportation

Aviation data is obtained from the U.S. Department of Transportation and contains the scheduled international passenger traffic every year [14]. We average the provided quarterly numbers from 2019. We then classify an airport as a major international airport if it is within the top third of airports with the most international passenger traffic. We calculate the distance to a major airport as the straight-line distance from the center of a county, as provided by Johns Hopkins, to the closest major airport. If a county contains a major airport, this distance is zero.

Oxford COVID-19 Government Response Tracker

We determine governor stringency using the Oxford COVID-19 Government Response Tracker (OxCGRT) [24]. OxCGRT tracks all government COVID-19 responses and ranks each regulation on a scale from 0 to 100 based on the stringency of the response, where 100 is the most stringent. We calculate the average stringency for all state-wide responses in each period, excluding the last four weeks of the period to account for the lag between governor interventions and deaths [18, 22].

Election Results

Political leaning is obtained from The New York Times' 2020 election data. We calculate political leaning as the number of individuals who voted for Joe Biden subtracted from the number of individuals who voted for Donald Trump divided by the total number of voters. This calculation results in a value between -1 and 1, where a value of -1 represents a county where all voters voted for Joe Biden and a value of 1 represents a county where all voters voted for Donald Trump.

Delphi Epidata

Public mask usage data is obtained from Facebook's COVID-19 symptom survey and is an estimate of the percentage of people who wore a mask for most or all of the time while in public in the 5 days before filling out the survey [16]. This survey contains only data beginning on September 8, 2020. Because mask usage is known to be a dynamic variable that has changed over the past year, we choose to include only mask usage data for period 3 [35].

Unfortunately, these data are sparse at the county-level in comparison to our other data sources. Figure 3.2 suggests that counties with these data might not be representative of all counties. Of the counties without mask data, approximately 2000 are Republican and approximate 200 are Democratic. The excluded counties, however, tend to have higher death rates. While the counties with mask data might not be representative, because excluded counties are majority Republican with higher death rates, a result that suggests that Republican counties have lower mask use and higher death rates would likely not be impacted if we had more mask data.

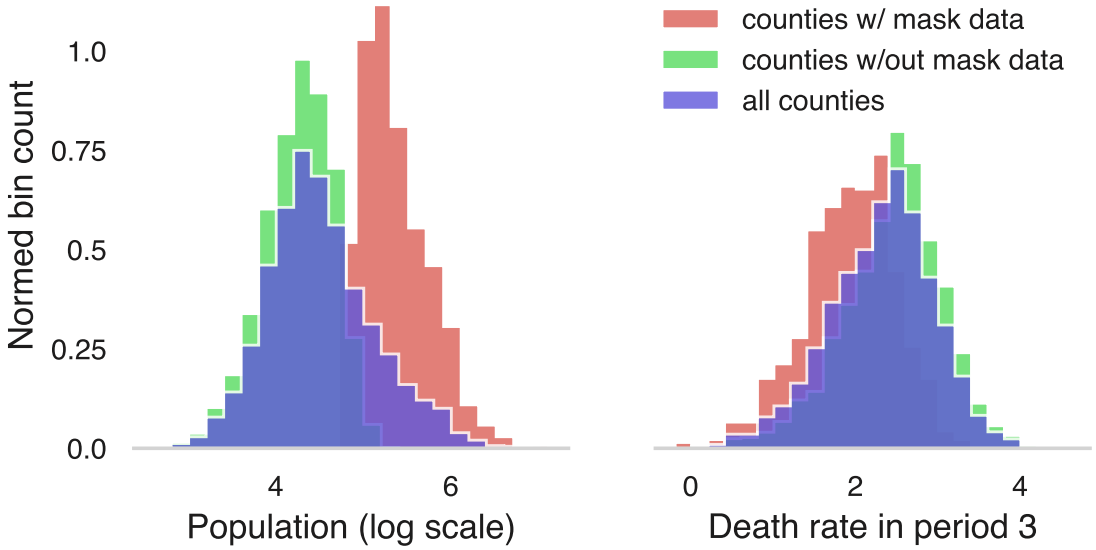

Figure 3.2: A comparison of counties with and without mask use data. We see that the counties with mask use data are generally larger and have lower death rates

Because of the limitation of these data, however, we present results both including and excluding counties with mask use data.

3.1.2 Clustering

Before further analyzing these variables, we perform a clustering analysis to provide a more empirical justification for analyzing waves of COVID-19 separately. We calculate z-scores for the smoothed daily death data to allow for cross-county comparisons. These z-scores are stored in a matrix, A, where each row of the matrix contains the standardized time series of death counts in a particular county and each column represents one day's worth of data. A is a 3119 x 382 matrix because we have 3119 counties and 382 days of data. We will let n and m refer to the number of rows and columns of the matrix, respectively.

We perform k-means clustering, a common unsupervised learning clustering technique, on our data [36, 39]. We also explored hierarchical clustering, but this method yielded less clear results. In clustering the data, we can think of the time series for each county as a point in an m dimensional space. With k-means clustering, k counties are randomly chosen as the initial clusters. All other

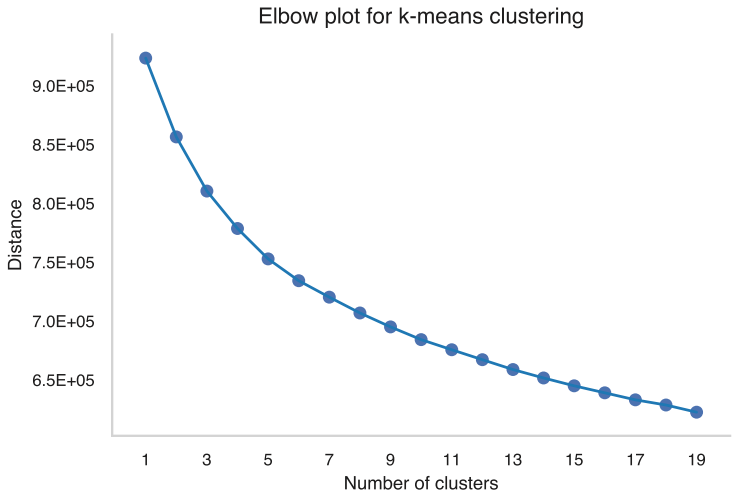

Figure 3.3: Elbow plot from performing k-means clustering on county-level time series. We first normalize each county's time series of COVID-19 deaths using z-scores. We then cluster the normalized time series using k-means clustering with a Euclidean distance metric. This elbow plot tells us that a logical choice for k is anywhere between $k = 2$ and $k = 5$.

counties are assigned to one of the k clusters based on their distance to the cluster center. For this analysis, we use simple Euclidean distance, although we also tried other distance metrics such as dynamic time warping. We select the number of clusters, k, using an elbow plot. An elbow plot for k-means clustering shows the total distance from each county to its assigned cluster center. Because increasing the number of clusters necessarily decreases the total distance, a "kink" in the elbow plot represents a new cluster that does not significantly reduce the total distance. A kink in the elbow plot, then, represents a reasonable choice for k.

Figure 3.3 shows the elbow plot that results from performing k-means clustering on A. We do not observe an obvious kink, but note that after $k = 3$, additional clusters do not result in as great a decrease in total distance to cluster center. To select the optimal clusters, we run 3-means clustering with 10 different combinations of counties as the initial 3 clusters and select the configuration that has the smallest total distance from each county to its assigned cluster's center.

3.1.3 Descriptive Analysis

We can begin to understand the relationship between a subset of our variables and the impacts of COVID 19 by conducting a descriptive analysis. We separate the relationship each of these variables could have on how a county experiences COVID-19 into two possible effects. First, a variable could be related to "seeding:" how likely a county is to first experience COVID-19. Second, a variable could be related to "spreading:" how severely a county experiences COVID-19. We say that a county is seeded in a given period if that county first experiences five deaths in that period. Once a county is seeded, it is seeded for the remaining periods. To determine severity for a county in a given period, we calculate the number of deaths per 10,000 people in that county during that time period (i.e., total number of deaths in the county during the time period multiplied by 10,000 then divided by the population of the county). We only calculate severity for counties that have been seeded.

To develop a descriptive understanding of the relationship between six choice variables (political leaning, income, percent POC, population density, mask use, and household crowding) and seeding and spreading, we separate counties into quantiles based on each variable. We first separate counties into five buckets based on quantiles for these six variables. Note that for the percent POC variable, we chose pre-determined bucket cutoffs.

We then further separate counties in each bucket based on governor stringency in each period. We consider "loose" stringency to be the third of states that have the least strict state-wide interventions, "middle" stringency to represent the middle third, and "strict" states to be the top third. We calculate two values to address the two possible effects. First, we calculate the proportion of counties in each bucket that are seeded. Second, among seeded counties, we calculate median death rate of counties in a given period. Table 3.1 shows the number of counties in each bucket that have been seeded. By period 3, all buckets have at least 9 counties seeded and the majority of buckets have over 30 counties seeded.

		Loose Governor					Middle Governor					Strict Governor				
Variable	**Period**	**Strong Dem.**	**Weak Dem.**	**Weak Rep.**	**Rep.**	**Strong. Rep**	**Strong Dem.**	**Weak Dem.**	**Weak Rep.**	**Rep.**	**Strong. Rep**	**Strong Dem.**	**Weak Dem.**	**Weak Rep.**	**Rep.**	**Strong. Rep**
Political Leaning	1	64	24	23	13	5	50	21	15	4	3	62	40	28	11	4
Political Leaning	2	103	61	75	59	37	112	118	104	79	53	81	64	41	25	20
Political Leaning	3	138	141	173	163	116	192	198	201	192	190	124	112	77	60	34
		Low	**Middle-low**	**Middle**	**Middle-high**	**High**	**Low**	**Middle-low**	**Middle**	**Middle-high**	**High**	**Low**	**Middle-low**	**Middle**	**Middle-high**	**High**
Median Household Income	1	14	21	35	48	94	18	16	24	41	92	42	43	40	43	82
Median Household Income	2	76	80	79	89	76	163	131	101	95	174	54	49	66	72	129
Median Household Income	3	218	175	160	162	135	229	238	247	216	245	86	96	106	135	152
Percent POC	1	74	51	43	27	17	73	43	31	22	22	25	59	79	62	25
Percent POC	2	190	96	62	34	18	127	157	190	114	76	141	83	59	47	41
Percent POC	3	418	162	137	77	56	525	224	219	128	79	327	93	64	50	42
Mask Usage	3	62	40	36	19	9	61	61	60	64	51	16	38	43	56	76
Household Crowding	1	2	6	30	44	130	2	10	20	50	109	5	13	31	49	152
Household Crowding	2	13	42	78	137	130	32	80	125	184	243	6	22	53	88	202
Household Crowding	3	147	188	175	191	149	140	241	284	255	255	32	71	117	152	204
Population Density	1	4	8	15	46	139	1	9	14	48	119	0	21	26	45	158
Population Density	2	13	45	89	122	131	21	116	135	158	234	5	16	45	100	205
Population Density	3	129	245	193	153	130	142	212	271	249	301	39	64	106	191	176

Table 3.1: The number of counties in each bucket shown in Figure 3.5B. In period 1, certain buckets have low counts (i.e. population density buckets for loose and middle governors). By period 3, all buckets have at least 9 counties (the minimum bucket counts in period 3 are 9, 16, 19, 32, and 34).

3.1.4 Regressions

The descriptive analysis described in the previous section helps to inform our hypotheses on what variables might be important in seeding and spreading within a county, but to develop a more robust understanding, we must be rigorous in our analysis. To provide a clear narrative, we separate our analysis into the three periods and test a set of hypotheses in each period. These hypotheses could be related to either seeding or spread. To test the seeding hypotheses, we run a logistic regression.

To test the spreading hypothesis, we run two different regressions for robustness: a LASSO regression and a random forest regression. The LASSO regression can help select only the most important variables by utilizing regularization, and the random forest regression better captures non-linearities in the data. To account for possible non-linear relationships within the data in our linear models, we test four monotonic transformations of our input variables: squared, square root, log, and exponential. If any of these transformations yield a higher Pearson's correlation with either death rate or the log of death rate, we apply the transformation to the given variable. We find the optimal transformation for each variable in each time period. In all three time periods, the majority of explanatory variables yield higher correlations to the log of our response variables, so we take the log of our response variable in our regressions. For both models, we train on all data and predict purely in-sample. Out-of-sample predictions for the random forest model are included in our robustness checks.

We run the LASSO regression using Python sklearn's LassoCV regression. This regression

seeks to minimize the following equation:

$$\frac{1}{2 \cdot N} \cdot ||y - Xw||_2^2 + \alpha \cdot ||w||_1 \tag{3.1}$$

where N is the number of data points, y is the observed death rate, X is a matrix of the independent variables, w is a vector of coefficients, and α determines how much to penalize coefficients. A large α will force more coefficients towards zero. We run the LassoCV model with values of α ranging from 0 to 1 with a step size of 0.001 and 5-fold cross-validation. We weight the regression by the log of a county's population.

In our random forest regressions, we first fit the model using all variables and 70% of counties as training data to ensure that we find the optimal model that does not over-fit when trained on all of our data. We perform a grid search using 5-fold cross validation on our training data and pick the hyper-parameters that yield the lowest MAE. We repeat this process three times, which yields the random forest configuration with optimal hyper-parameters for each period. To calculate feature importance in the random forest regression, we calculate permuted feature importance. This metric performs a random permutation on a given explanatory variable and calculates the resulting loss in accuracy of the model.

To address spatial autocorrelation in residuals, we include latitude and longitude as predictors in all models. We choose to exclude latitude and longitude in our visualization of feature importance and model coefficients because latitude and longitude are not inherent measures of risk or disease spread. We report Moran's *I* for our regressions, a measure of spatial autocorrelation [43]. We calculate this value using a 5-nearest neighbors weight matrix (see robustness checks in section 3.2.2 for more details). A low Moran's *I* (near 0) suggests that there is no spatial autocorrelation, which is desired to ensure a spatially normal distribution of residuals. Generally, we achieve low Moran's *I*, but we also run a spatial lag model that accounts for the spatial distribution of our data in our robustness checks.

Finally, it is important to make a distinction between counties that are included in seeding vs. spreading hypothesis testing. In spreading hypotheses, a county that was either seeded in the

period of interest or a previous period is considered. In seeding hypotheses, however, we consider all counties but only predict counties that were first seeded in the period of interest. For example, for spreading hypotheses in period 2, we only include counties that were seeded in period 1 or 2. In testing period 2 seeding hypotheses, we only attempt to predict counties that were first seeded in period 2 but include all counties (i.e., counties that were seeded in period 1, period 3, or were never seeded all have a response value of 0).

3.2 Results

Figure 3.4A shows the average of z-scored deaths across counties in each cluster and Figure 3.4B shows the counties that belong in each cluster. We note three periods (January-May, June-October, November-February) and push the end dates to ensure that each period is the same length:

- Period 1: January 26, 2020 - June 2, 2020
- Period 2: June 3, 2020 - October 7, 2020
- Period 3: October 8, 2020 - February 12, 2021

The growth in all three cluster centers in period 3, as shown in Figure 3.4A, as well as the breakdown of deaths presented in Figure 3.1, provide justification for our focus on period 3. Even within these cluster centers, we begin to see that COVID-19 has a more severe impact in certain areas of the country than others. Figure 3.4C-I plot a subset of the variables that we will study to motivate just how heterogeneous the United States is along a number of different metrics.

3.2.1 Descriptive Analysis

We begin to understand the relationship between governor stringency, COVID-19, and the variables presented in Figure 3.4 by considering each variable separately, as shown in Figure 3.5. Recall that counties are bucketed both based on the variable indicated by the title of the subplot and the

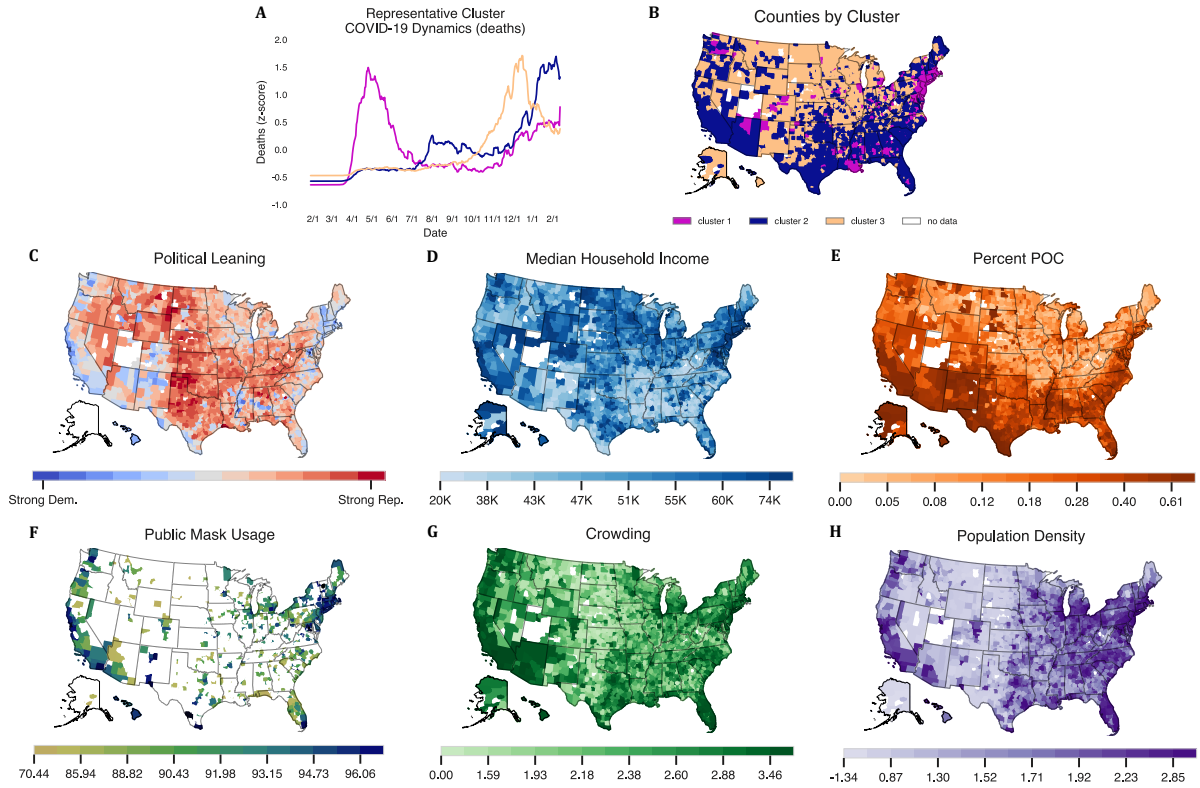

Figure 3.4: A summary of 3-means clustering (**A, B**) and maps of counties included in the analysis shaded by the six different population variables of interest (**C-H**). Counties without any color are missing data. (**A**) The cluster centers that result from k-means clustering. The center of a given cluster is determined by averaging the z-scored deaths of all counties assigned to that cluster. (**B**) A map of cluster assignments for all U.S. counties with data and at least 1 death. Northeastern counties and counties surrounding New Orleans largely comprise the first cluster, Southern counties generally comprise the second cluster, and remaining counties comprise the third cluster. (**C-H**) show counties shaded according to the population variable suggested by the title. Figures (**D**) and (**G**) show the log of population density and crowding. Figure (**F**) visualizes mask use, a dynamic variable, and shows period 3 averages.

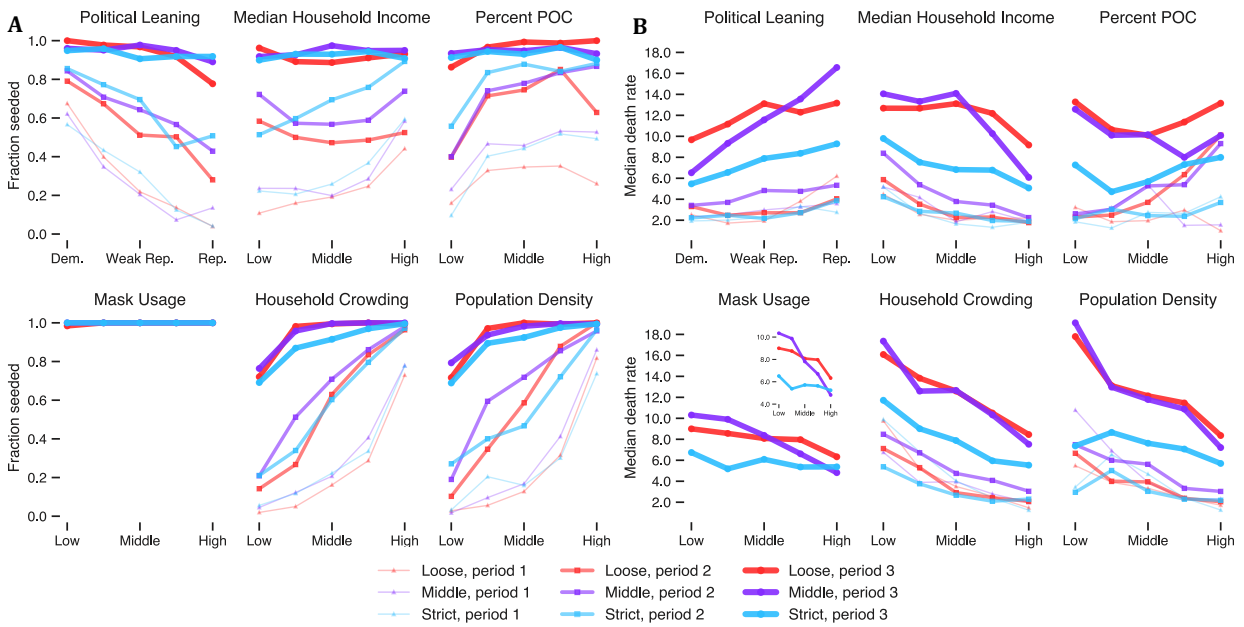

Figure 3.5: A plot of proportion of counties in each bucket that have been seeded with COVID-19 **(A)** and the median death rate (deaths per 10,000) for each of the six population variables broken down by period and the stringency of state-level mandates **(B)**. **(A)** In period 1, the majority of counties that are seeded with COVID-19 are Democratic, highly population dense counties. By period 3, most counties are seeded with COVID-19. **(B)** Across all six variables, the death rate in period 3 is the highest while the death rate in period 1 is the lowest, suggesting that outbreaks worsened throughout the year. All variables show clear differences depending on the stringency of governor interventions. The inset mask usage plot in the second row of **(B)** shows the same as the main mask use plot just with a different y-axis scale to more clearly show the downward trend of the lines.

stringency of governor interventions in a given period. Figure 3.5A shows the fraction of counties within each bucket that are seeded in a given period (have experienced more than 5 deaths by the end of the period) and Figure 3.5B shows the median death rate of seeded counties in a bucket, where death rate is the number of deaths per 10,000 individuals in a county during a period.

In Figure 3.5A we observe that Democratic counties are more likely to be seeded in period 1. Similarly, areas of high crowding and high income are also more likely to be seeded. Low percent POC areas are more likely to avoid the pandemic in period 1. In period 2, we observe similar trends as more counties become seeded. By period 3, almost all counties are seeded. There is no noticeable difference between counties with differing levels of governor stringency in the seeding

plots.

The prevailing trends in Figure 3.5(B) are that, in periods 1 and 2, counties experienced less severe outbreaks than in period 3. The main difference between period 1 and 2 is that there is no discernible difference in governor stringency in period 1, whereas in period 2, loose and middle governors largely experienced worse outbreaks. The trends in period 3, the period of greatest interest, are more stark than the previous periods. Across all six variables, we see a clear difference between strict governors and governors in the loose and middle categories, with strict governors experiencing less severe outbreaks. Republican counties experienced more severe outbreaks than Democratic counties. Intuitively, high income and high mask usage counties also all experienced less severe outbreaks. High crowding areas experienced less severe outbreaks as well, which could be an indication of the disease moving to more rural areas. Percent POC trends are less linear. Low POC areas experienced slightly worse outbreaks across all three level of governor stringency than middle-low, middle, and middle-high POC areas. High POC areas experienced the worst outbreaks.

While these results suggest strong trends, we have not yet been able to isolate the effects of a variable. Isolating these effects is our goal in the following section.

3.2.2 Regressions

We now add all the variables discussed in Section 3.1.1. We filter our analysis to counties that have data available for all data sources (resulting in $n = 2926$ counties). Figure 3.6 shows a conceptual framework that presents these variables and the means by which these variables might impact COVID-19 outcomes. As suggested by this figure, the relationships are complicated. By studying the variables together, we hope to begin to isolate individual impacts, even if the mechanism by which that impact comes about is not always clear.

One of the main challenges in this analysis is that many of the variables are conceptually related and correlated. Figure 3.7 visualizes the correlations in variables among certain subsets of counties, showing variables that have a magnitude of Spearman's Rank-Order correlation greater

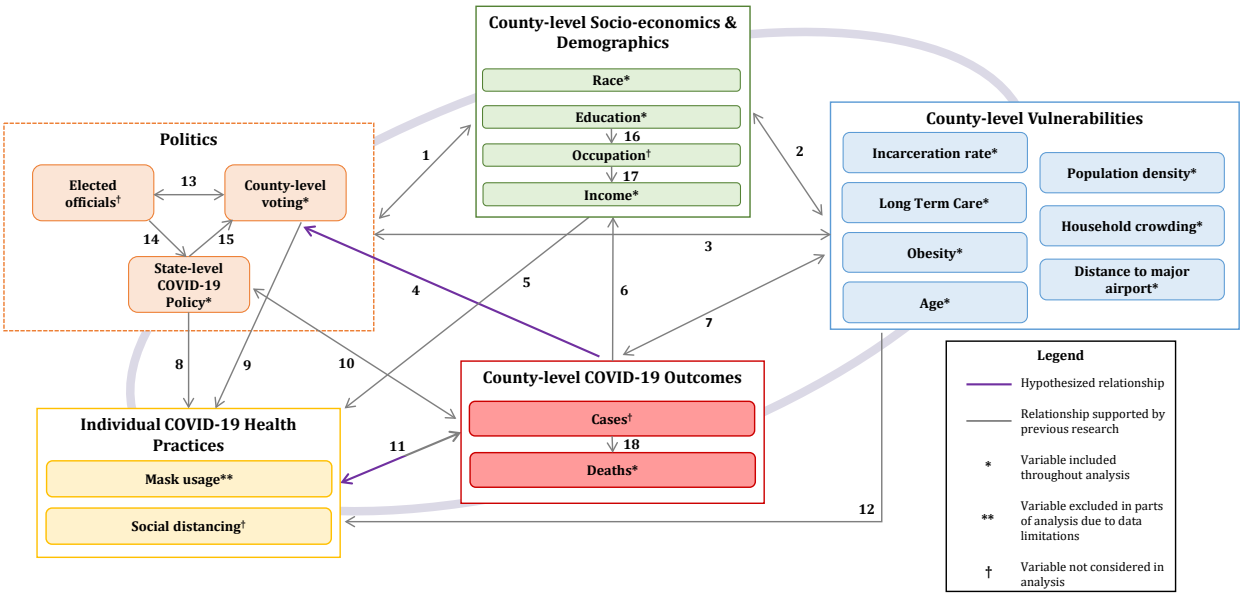

Figure 3.6: A summary of the ways in which variables impact or might be impacted by COVID-19. (1-3, 16-17) are derived from [68]. (4) We hypothesize because of the differing political coverage surrounding the pandemic [20]. (5, 12) are supported by [45]. (6) is supported by [28]. (7) is supported by [10, 12]. (8, 9) are supported by [3, 30]. (10) is supported by [26] and the well-documented re-opening of the economy when case counts are low. (11) is supported by [23, 41, 70]. (13-15) follow from our democratic system of government and (18) follows from the many COVID-related deaths across the world.

than 0.5 [60]. Note that these correlations change depending on what analyses we run. Our seeding analyses include all counties, regardless of period (correlations shown in Figure 3.7A), but our spreading analyses only include seeded counties, which change from period-to-period. Therefore, correlations change depending on the type of hypothesis we are testing and in what period we are testing the hypothesis. We also include a correlation plot of counties with mask use data (Figure 3.7D) because these data are sparse, so correlations change again.

Highly correlated and related variables pose a challenge because they might result in an ill-conditioned matrix of explanatory variables. The condition number is measure of how much an output can change given a small perturbation in the input [54]. A small condition number, as close to 0 as possible, is desired. The matrix that includes all variables and all counties (i.e., the matrix of explanatory variables in seeding tests) has a condition number of 133.99, which is relatively high. Ideally, we want condition numbers under 1, and will try to achieve this by removing highly correlated variables that are measures of a similar idea. We remove density (crowding serves as a proxy for this), percent under 19 (percent over 65 serves as a proxy for this), and certain race variables (percent White not Hispanic, Asian, and other, which percent Black and Hispanic proxy for). Note we do not include percent POC in any of our regressions. We keep all of obesity, education, and income despite their correlations because these variables address different concepts and vulnerabilities. Removing these variables reduces the condition number to 107.41. In our period 2 spreading analysis, which only includes counties that were seeded in period 1 or 2, removing these variables reduces the condition number from 150.98 to 119.35. Finally, in our period 3 spreading analysis, removing these variables reduces the condition number from 93.61 to 58.20 when using counties regardless of whether they have available mask use data or not and from 154.52 to 137.02 when using only counties with mask use data. These condition numbers are still higher than one, but removing any more variables would result in a less robust understanding of the pandemic. We proceed with our analysis keeping this in mind.

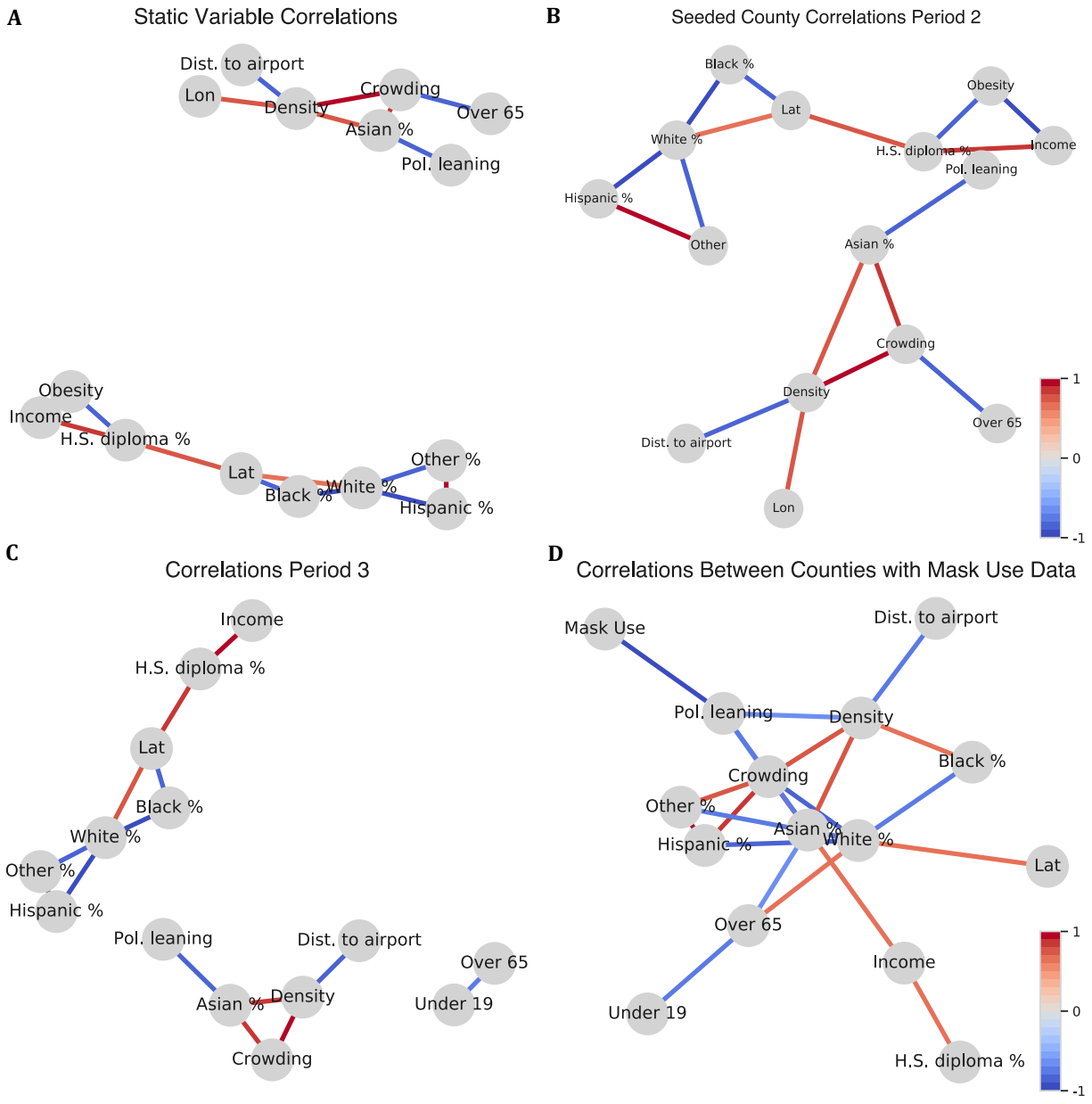

Figure 3.7: A summary of the Spearman's Rank-Order correlation (r_s) between variables. Only variables that have a $|r_s|$ greater than 0.5 appear on these plots. **(A)** Correlations when including all counties, as is the case with seeding analyses. **(B)** Correlations when including only counties that have been seeded in period 1 or 2, as is the case with period 2 spreading analyses. **(C)** Correlations when excluding only counties that were not seeded at any point in the year, as is the case with period 3 spreading analyses. **(D)** Correlations when including only counties with available period 3 mask use data.

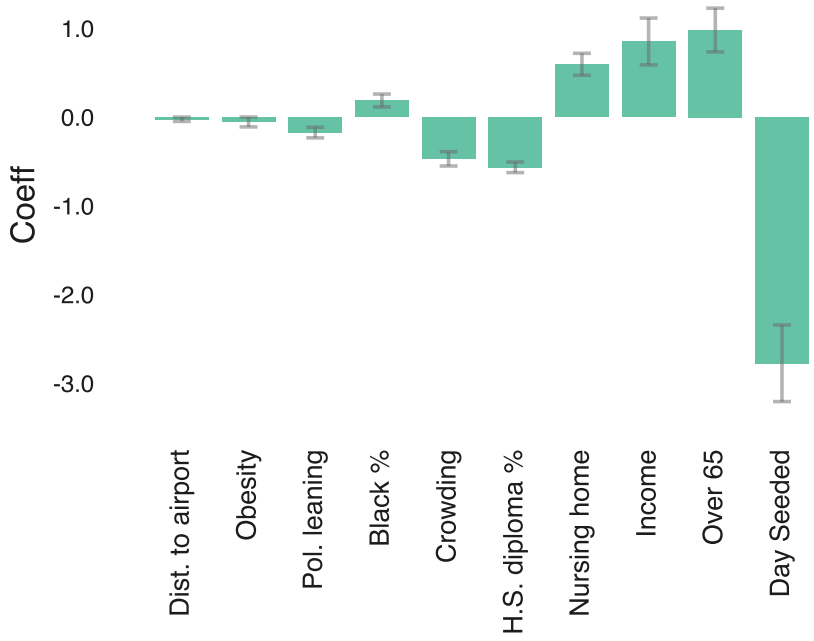

Figure 3.8: Coefficients of a LASSO model that predict spread in period 1. Error bars indicate one standard error. The mean absolute error between observed and predicted outcomes is 0.11. The day a county is seeded emerges as the clear strongest predictor. As a result, we focus mainly on seeding hypotheses in period 1.

Period 1

The narrative in period 1 is largely driven by seeding. From Figures 3.1, 3.4, and 3.8 (Moran's I of residuals: 0.17), we see that, while many counties were seeded with COVID, the outbreak was contained largely to the Northeast and counties that were seeded earlier had the larger outbreaks. As a result, we focus mainly on what might cause a county to be seeded earlier.

Seeding Hypothesis 1: Counties near large, international airports are more likely to be seeded. From the two models presented in Figure 3.9, we see that distance to a major airport is one of the strongest predictors of becoming seeded in period 1. The coefficient on this variable is negative, suggesting that counties farther from a major airport are less likely to be seeded. We conclude that this hypothesis is confirmed.

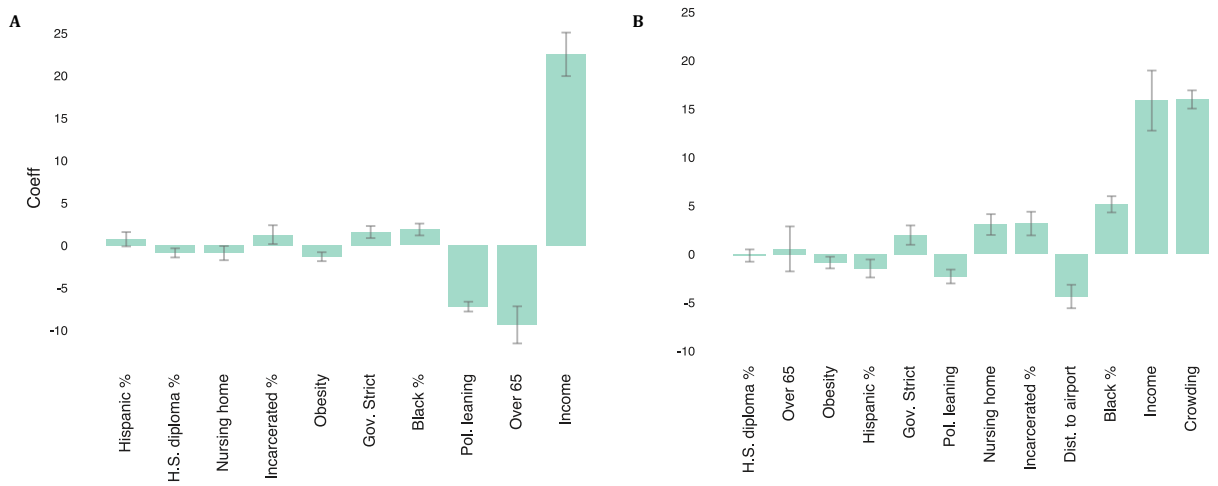

Figure 3.9: A summary of our seeding hypothesis testing in period 1. **(A)** visualizes the coefficients of variables after running a logistic regression, excluding both distance to a major airport and crowding. **(B)** shows the results of the same regression, this time including distance to a major airport and crowding. The coefficient on political leaning changes from -7.20 to -2.30. Error bars indicate one standard error.

Seeding Hypothesis 2: Democratic counties are more likely to be seeded, and this is largely due to proximity to international airports. Figure 3.9 shows the tests we run to address this hypothesis. We begin by removing distance to an airport and crowding from the analysis (Figure 3.9A; accuracy: 0.79; sensitivity: 0.81; specificity: 0.69). This results in a coefficient on political leaning of -7.20. Adding distance to an airport back into the model (accuracy: 0.80; sensitivity: 0.82; specificity: 0.73), the coefficient on political leaning reduces in magnitude to -6.26, a 13% change. Finally, we add crowding back into the model (Figure 3.9B; accuracy: 0.84; sensitivity: 0.86; specificity: 0.79), further reducing the coefficient on political leaning to -2.30, an additional 63% decrease. Between distance to an airport and crowding, then, the coefficient on political leaning has reduced by 68%.

We conclude that the result that Democratic counties are more likely to be seeded (as indicated by the negative coefficient) is largely due to both the effects of proximity to international airports and household crowding.

Period 2

In period 2, we still do not have accurate mask wearing data, so we cannot conduct any tests with behavioral variables. As a result, we focus on vulnerabilities, demographics, and political variables.

Seeding Hypothesis: Counties far from airports are more likely to be seeded as the disease spread to more rural Republican areas. This hypothesis is supported by Figure 3.10A, which shows the results from running the same logistic regression as Figure 3.9B, but using period 2 seeding as the response variable (accuracy: 0.69; sensitivity: 0.71; specificity: 0.56). We conclude that this hypothesis is true, as both the coefficients on political leaning and airport are positive. Other variables are more important in determining seeding in period 2, however.

Spreading Hypothesis 1: Among seeded counties, more severe COVID-19 spread is correlated with indicators of higher social vulnerability. We conclude that this hypothesis is true based on the feature importance and coefficients indicated in Figures 3.10B (Moran's *I* of residuals: 0.069) and 3.10D (Moran's *I* of residuals: 0.14). Some of the top predictors in both models are education, age, and race. We also note that these models predict the data well. The mean absolute error (MAE) between observed and predicted death rates is 0.065 and the R^2 is 0.42 for the LASSO model and 0.0094 (MAE) and 0.97 (R^2) for the random forest model.

Spreading Hypothesis 2: Seeded Republican areas will more likely have large waves, and this will be driven by variation in vulnerabilities. We also conclude that this hypothesis is true, with certain caveats. There may be a significant impact of behavioral variables such as mask use and social distancing, but we are unable to test these impacts in period 2 due to data limitations. Instead, we see that the importance and coefficient on political leaning is small relative to the coefficients on vulnerability variables. When we run the models without vulnerability variables (leaving just political leaning and governor stringency), we observe that the linear models readily pick out governor stringency over political leaning, while the random forest assigns relatively

equal importance to these variables. These tests provides evidence that governor stringency might explain some of the effect of political leaning. The story in this period, then, is that vulnerable populations were significantly impacted by COVID-19, and that Republican counties and counties with less strict regulations were hit hardest.

Period 3

In period 3, nearly all counties have been seeded (only 6% of the 2926 included in the regression analyses are unseeded), so we focus on spreading hypotheses. Because mask use data is sparse and not necessary representative of U.S. counties, we run our models first without any mask use data (Figure 3.10E-G). Separately, we run our models using only counties with mask use data (Figure 3.11), keeping in mind the limitations of our data.

Spreading Hypothesis 1: Spread is driven by vulnerability and preventative behaviors. This hypothesis is clearly confirmed by all models in Figures 3.10 and 3.11. The top predictors in all models are some combination of nursing home population (consistently the best predictor), income, education, crowding, political leaning, and governor stringency. The random forest model more readily picks out mask use, suggesting important non-linear interactions between terms.

Note that the LASSO model shown in Figure 3.10G results in reasonably accurate predictions (Moran's I of residuals: 0.17), with an MAE of 0.047 and R^2 of 0.35. The random forest model performs significantly better, with an MAE of 0.0061 and an R^2 of 0.98 (Moran's I of residuals: 0.054).

Spreading Hypothesis 2: Seeded Republican areas are more likely to have large waves, and this is driven by variation in preventive behaviors. Political lean has a positive coefficient in all models, suggesting larger waves in Republican areas. It is also an important predictor, particularly in the random forest model. Preventative behaviors are important, but political lean is picking up on more than just mask use. When we add mask use, the importance of political lean in the random forest model decreases from 0.18 to 0.13, a 38% reduction. While it is difficult to interpret

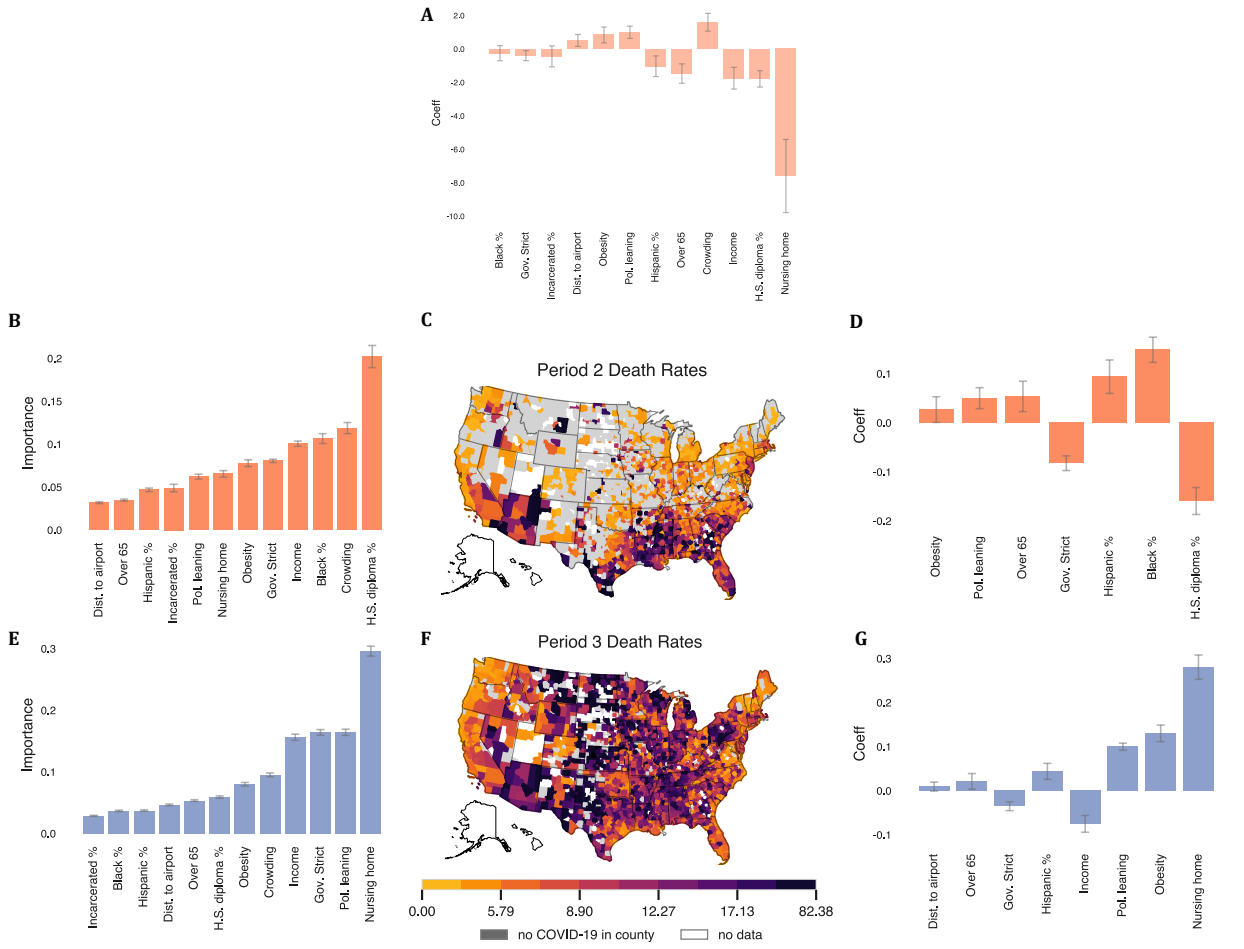

Figure 3.10: A summary of hypothesis testing in periods 2 and 3. **(A)** shows the results of a logistic regression that predicts whether or not a county is seeded in period 2. **(B,D)** shows coefficients and feature importance, for the random forest and LASSO model, respectively, when predicting period 2 death rates (visualized in **(C)**. **(E-G)** shows the same plots for period 3 death rates. Any variables that have a coefficient of zero in the LASSO model are excluded from the plot. We note the emergence of political leaning and governor stringency in period 3. Unsurprisingly due to the disproportionate impact of COVID-19 on the elderly, nursing home population is one of the best predictors of deaths. Error bars indicate one standard deviation (random forest models) and one standard error (LASSO models).

reductions such as this one when adding in correlated variables (Figure 3.7D shows mask use and political lean are correlated, with $r_s = -0.81$), this reduction provides evidence that the effect of political differences is partially driven by behaviors. It is clear, however, that political leaning is an indication of more than just mask use. It also important to note that, more than any other period, governor stringency is picked as an important predictor.

Robustness Checks

We additionally test all seeding hypotheses with L1 regularization, selecting a value of α, the penalization term on model coefficients, based on the model that results in the lowest AIC. We try 100 values for α evenly spaced in the interval from 0.01 to 1. We found that model convergence was slow, and for certain α values, some variables did not converge. The optimal α was also almost always 0.01, suggesting that regularization is not particularly insightful in this case.

We also test the spreading hypotheses with a spatial lag model that better accounts for spatial autocorrelations. We run the spatial lag regression using pysal's ml_lag model [53]. This model attempts to find the maximum likelihood estimate of the following:

$$y = \rho W y + \beta X + \epsilon \tag{3.2}$$

where y is the death rate, X is a matrix of independent variables, ρ and β are the spatial autoregressive coefficients and independent variable coefficients, respectively. Wy is a measure of the spatial lag. W is a weight matrix that determines how heavily nearby counties impact a given county. For our analysis, we construct W using k nearest-neighbors, where all k nearest-neighbors are weighted equally. To find the optimal k values, we run the spatial lag model using all variables with values of k ranging from 1 to 20. We find that $k = 5$ yields the lowest mean absolute error (MAE). We also constructed W using a kernel density function, but this weight matrix yielded less accurate results.

The results of the spatial model using the 5 nearest-neighbors weight matrix are shown in Figure 3.12. We observe that the narrative told by these spatial models is the same as the LASSO

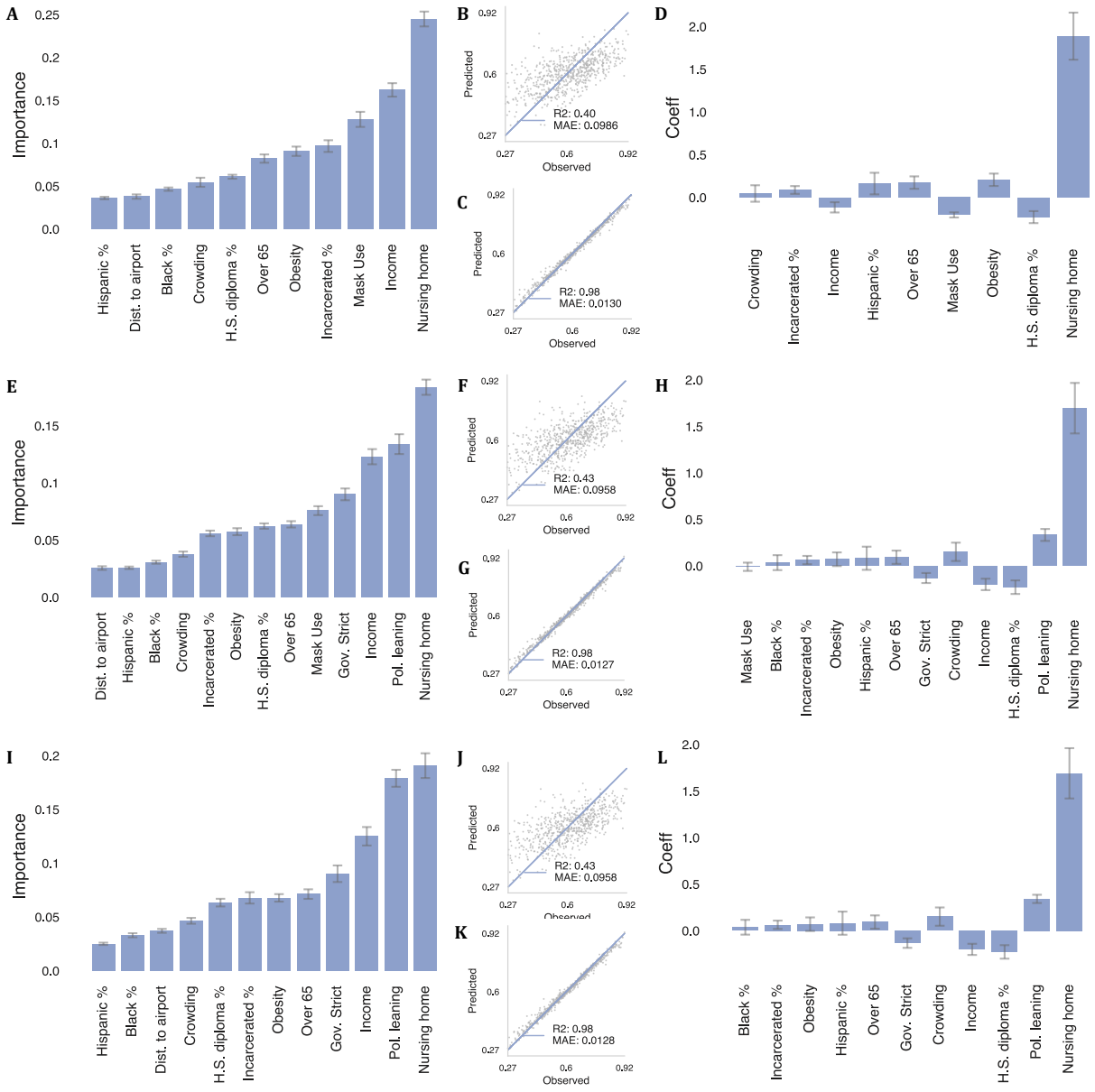

Figure 3.11: A summary of our hypothesis testing in period 3 for counties with mask use. **(A-D)** shows the result of the random forest and LASSO models when we exclude political leaning and governor stringency from the analysis (Moran's I of residuals: 0.11 and 0.21 for the random forest and LASSO, respectively). **(E-H)** shows the same results, this time including political leaning and governor stringency (Moran's I of residuals: 0.063 and 0.22 for the random forest and LASSO, respectively). Finally, **(I-L)** remove mask use and only contain political and vulnerability variables (Moran's I of residuals: 0.068 and 0.21 for the random forest and LASSO, respectively. **(B, F, J)** show the fits for the LASSO model, while **(C, G, K)** show fits for the random forest model. The random forest model suggests that mask use drives some of the variation in political leaning but not all. Generally, political leaning and governor stringency emerge as strong predictors in this period. Error bars indicate one standard deviation (random forest models) and one standard error (LASSO models).

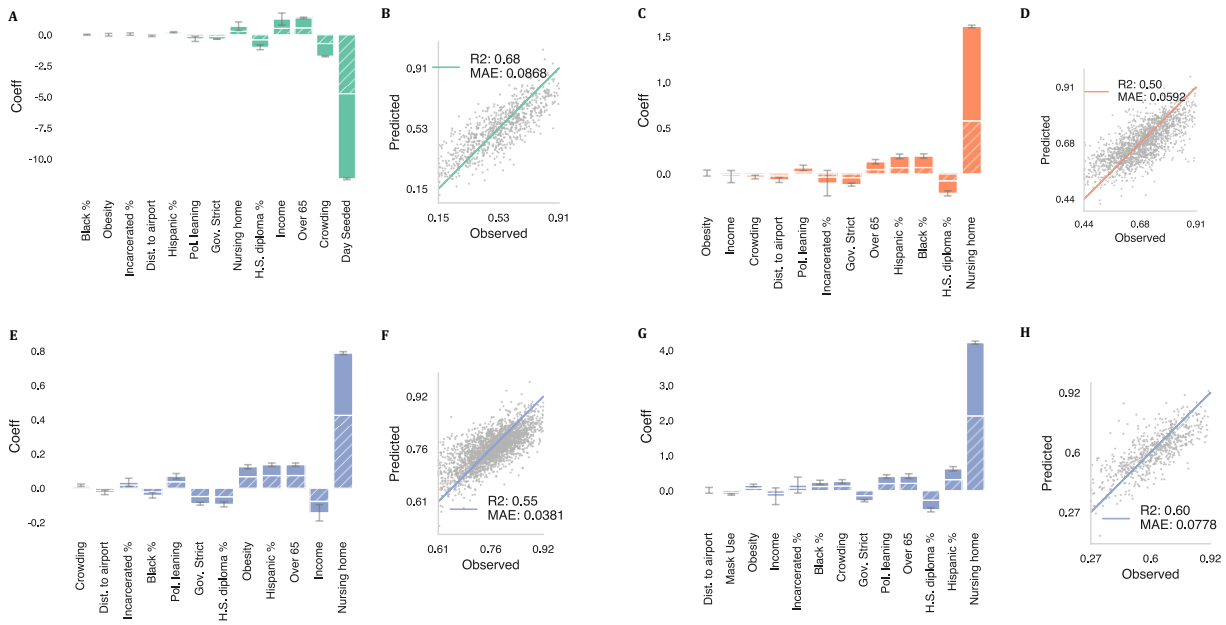

Figure 3.12: A summary of our spreading hypothesis testing across all periods for the spatial lag model. Hatched bars indicate the indirect effects of a given variable (i.e., if county 1 is near county 2, the performance in county 1 affects county 2 and vice versa). **(A-B)** show the coefficients and fit in period 1. **(C-D)** show the same same for period 2. **(E-F)** show the same for period 3, including counties without mask data. Finally, **(G-H)** are the results when including only counties with mask data. We observe a similar narrative as laid out in the previous section. Error bars indicate one standard error of the direct coefficients.

and random forest regressions.

We also vary the cutoff at which we consider a county seeded. We ran the same set of analyses using 1, 3, and 10 as the death cutoff (instead of 5). Additionally, we removed the weights on the LASSO regression to see if the narrative would change. The results were robust to all of these changes.

Finally, for completeness, we show out-of-sample predictions for the random forest models using 5 random 70/30 train-test splits in Figure 3.13. The average MAE for predictions in period 1 across these 5 runs is 0.053, the average MAE for period 2 predictions is 0.044, and the average MAE for period 3 predictions is 0.049. These are all low MAE's.

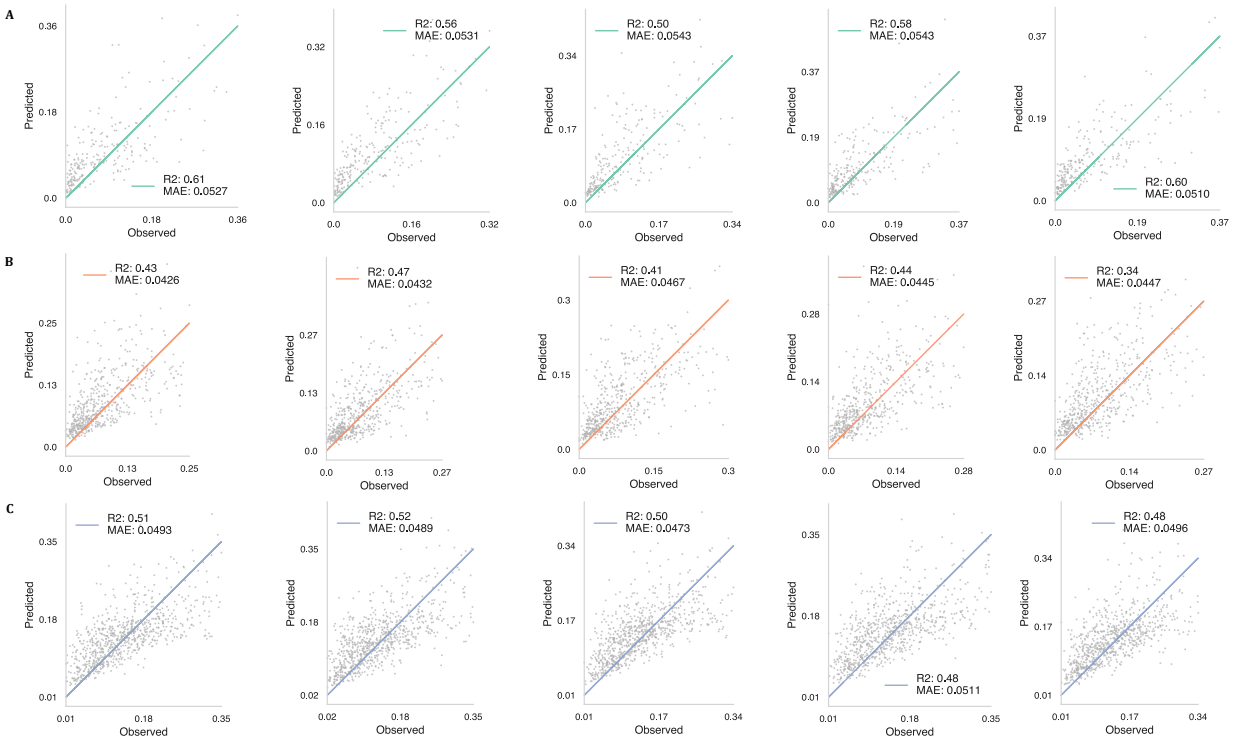

Figure 3.13: Out of sample predictions for the random forest model for five random train-test splits in each period. Row **(A)** shows period 1 results and rows **(B)** and **(C)** show period 2 and 3 results, respectively. The random forest models perform well in all three periods, with low mean absolute error.

3.3 Discussion

We were able to study the evolving relationships between political, behavioral, socio-economic, and vulnerability variables and COVID-19 outcomes. We showed that, as the pandemic progressed, partisanship played a bigger role in how severely COVID-19 impacted certain counties. Both governor stringency and political leaning emerged as two of the most important predictors in period 3. While mask use was able to explain some of the impacts of these variables, we observe that political leaning indicates other effects than just mask use. It is possible that social distancing, another variable that has been shown to be strongly associated with partisanship, could explain a large portion of the effects of political leaning. Our analysis also shows that socio-economics and vulnerabilities are essential pieces to the COVID-19 narrative. Throughout the pandemic, long term care, income, and race are important predictors of both seeding and spread. COVID-19 is clearly no exception to the historical trend that individual within at-risk populations experience worse health outcomes.

The political narrative presented in this research is in line with the increased politicization of the disease that was observed throughout the pandemic. While our understanding of the pandemic and how to sufficiently control disease spread developed, the partisan divide on pandemic views grew both at the federal and local levels. As the political divide grew, so too did the disparities in COVID-19 outcomes. Instead of putting the health and safety of the country's citizens first, individuals took actions (or inactions) that put Americans at risk.

The politicization of the pandemic is especially concerning considering the growing political divide over the past decade [46]. It is known that population health and politics are closely tied, and this study has laid out that the growing political divide can lead directly to increased disparities in health outcomes. The mechanisms through which these pieces are tied are complicated, but the link is clear. Through this research, we hope to lay bare and call attention to the disparities. If research such as this is ignored based on perceived political motives or mistrust in science, we can expect continued negative health outcomes, as has been the case with COVID-19.

Chapter 4

Conclusions

This thesis has presented two different ways in which we tried to develop an understanding of COVID-19 at the county-level. In Chapter 2, we applied dynamic mode decomposition to county-level deaths both to test if the method could be extended to COVID-19 and to develop a more robust understanding of the spatial and temporal dynamics of the disease. We were unable to achieve the second of these goals, as our results suggest that DMD in its most primitive form is inaccurate in the context of COVID-19. In Chapter 3, we were able to examine the disparate impacts of the disease at the county-level, focusing on the winter months when the country was hit the hardest. We showed the strong correlation of both partisanship and vulnerabilities, particularly during this last period of the pandemic.

One of the biggest takeaways from these findings is that individual behavior matters. From the messages that world and community leaders send about the disease through their mandates or examples to our personal decisions to wear a mask or keep six feet apart, our actions have consequences. We may never feel those consequences ourselves, but they will almost certainly impact specific groups of individuals who have been shown to be disproportionately affected by COVID-19. How many of the 543,793 deaths over the past year were preventable had we been more conscious of this finding [15]? Thousands? Tens of thousands? Hundreds of thousands? The relationships between these variables and deaths as observed in our analysis clearly indicates

a strong connection.

Bibliography

1. Agency for Toxic Substances and Disease Registry. *CDC's Social Vulnerability Index* https://www.atsdr.cdc.gov/placeandhealth/svi/index.html.
2. Akovali, U. & Yilmaz, K. Polarised pandemic response and Covid-19 connectedness across US states. *VoxEU*. https://voxeu.org/article/polarised-pandemic-response-and-covid-19-connectedness-across-us-states (Dec. 10, 2020).
3. Allcott, H. *et al.* Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics* **191,** 104254. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7409721/ (Nov. 2020).
4. Askham, T. & Kutz, J. N. Variable projection methods for an optimized dynamic mode decomposition. *arXiv:1704.02343*. http://arxiv.org/abs/1704.02343 (Apr. 1, 2017).
5. Beckfield, J. & Krieger, N. Epi + demos + cracy: linking political systems and priorities to the magnitude of health inequities–evidence, gaps, and a research agenda. *Epidemiologic Reviews* **31,** 152–177. https://pubmed.ncbi.nlm.nih.gov/19474091/ (2009).
6. Bhowmik, T., Tirtha, S. D., Iraganaboina, N. C. & Eluru, N. A Comprehensive Analysis of COVID-19 Transmission and Fatality Rates at the County level in the United States considering Socio-Demographics, Health Indicators, Mobility Trends and Health Care Infrastructure Attributes. http://medrxiv.org/lookup/doi/10.1101/2020.08.03.20164137 (Aug. 4, 2020).
7. Blow, C. M. The Racial Time Bomb in the Covid-19 Crisis. *The New York Times*. https://www.nytimes.com/2020/04/01/opinion/coronavirus-black-people.html (Apr. 2, 2020).
8. Bonaccorsi, G. *et al.* Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences* **117,** 15530–15535. http://www.pnas.org/lookup/doi/10.1073/pnas.2007658117 (July 7, 2020).
9. Brooks, L. C. *et al. Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S.* International Institute of Forecasters. https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/.
10. Centers for Disease Control and Prevention. *Long-Term Effects of COVID-19* https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html.

11. Centers for Disease Control and Prevention. *PLACES: Local Data for Better Health* 2020. https://nccd.cdc.gov/PLACES/rdPage.aspx?rdReport=DPH_500_Cities.ComparisonReport.

12. Centers for Disease Control and Prevention. *People with Certain Medical Conditions* https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

13. Coppins, M. Trump's Dangerously Effective Coronavirus Propaganda. *The Atlantic*. https://www.theatlantic.com/politics/archive/2020/03/trump-coronavirus-threat/607825/ (Mar. 11, 2020).

14. Department of Transportation Office of the Assistant Secretary for Aviation and International Affairs. *International Report Passengers* 2021. https://data.transportation.gov/Aviation/International_Report_Passengers/xgub-n9bw.

15. Dong, E., Du, H. & Gardner, L. An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *The Lancet Infectious Diseases* **20,** 533–534. https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext (2020).

16. Farrow, D. C., Brooks, L. C., Rumack, A., Tibishirani, R. J. & Rosenfeld, R. Delphi Epidata API. *The Lancet Infectious Diseases*. https://github.com/cmu-delphi/delphi-epidata (2015).

17. Fenga, L. Forecasting the COVID-19 Diffusion in Italy and the Related Occupancy of Intensive Care Units. *Journal of Probability and Statistics* **2021,** e5982784. https://www.hindawi.com/journals/jps/2021/5982784/ (Jan. 16, 2021).

18. Ferguson, N. *et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand* (Imperial College London, Mar. 16, 2020). http://spiral.imperial.ac.uk/handle/10044/1/77482.

19. Gamio, L. Half of U.S. Coronavirus Deaths Have Come Since Nov. 1. *The New York Times*. https://www.nytimes.com/interactive/2021/02/10/us/coronavirus-winter-deaths.html (Feb. 10, 2021).

20. Gollwitzer, A. *et al.* Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour* **4,** 1186–1197. https://www.nature.com/articles/s41562-020-00977-7 (Nov. 2020).

21. Golub, G. H. & Reinsch, C. Singular value decomposition and least squares solutions. *Numerische Mathematik* **14,** 403–420. https://link.springer.com/content/pdf/10.1007/BF01436084.pdf (1970).

22. Grech, V. & Scherb, H. COVID-19: Mathematical estimation of delay to deaths in relation to upsurges in positive rates. *Early Human Development*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7528970/ (Oct. 1, 2020).

23. Guy, G. P. *et al.* Association of State-Issued Mask Mandates and Allowing On-Premises Restaurant Dining with County-Level COVID-19 Case and Death Growth Rates — United States, March 1–December 31, 2020. *MMWR. Morbidity and Mortality Weekly Report* **70,** 350–354. http://www.cdc.gov/mmwr/volumes/70/wr/mm7010e3.htm?s_cid=mm7010e3_w (Mar. 12, 2021).

24. Hale, T., Webster, S., Petherick, A., Phillips, T. & Kira, B. Oxford COVID-19 Government Response Tracker. https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19 (2020).

25. Hanage, W. P. *et al.* COVID-19: US federal accountability for entry, spread, and inequities—lessons for the future. *European Journal of Epidemiology* **35,** 995–1006. https://doi.org/10.1007/s10654-020-00689-2 (Nov. 1, 2020).

26. Haug, N. *et al.* Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour* **4,** 1303–1312. https://www.nature.com/articles/s41562-020-01009-0 (Dec. 2020).

27. Howard, J. *et al.* An evidence review of face masks against COVID-19. *Proceedings of the National Academy of Sciences* **118.** https://www.pnas.org/content/118/4/e2014564118 (Jan. 26, 2021).

28. Jenco, M. Study: COVID-19 pandemic exacerbated hardships for low-income, minority families. *AAP News.* https://www.aappublications.org/news/2020/06/03/covid19hardships060320 (June 3, 2020).

29. Kaashoek, J. & Santillana, M. COVID-19 positive cases, evidence on the time evolution of the epidemic or an indicator of local testing capabilities? A case study in the United States. *arXiv:2004.3128874.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3574849 (2020).

30. Kahane, L. H. Politicizing the Mask: Political, Economic and Demographic Factors Affecting Mask Wearing Behavior in the USA. *Eastern Economic Journal,* 1–21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7783295/ (Jan. 5, 2021).

31. Katz, J., Sanger-Katz, M. & Quealy, K. A Detailed Map of Who Is Wearing Masks in the U.S. *New York Times.* https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html.

32. Kendi, I. X. What the Racial Data Show. *The Atlantic.* https://www.theatlantic.com/ideas/archive/2020/04/coronavirus-exposing-our-racial-divides/609526/ (Apr. 6, 2020).

33. Krieger, N., Waterman, P. D. & Chen, J. T. COVID-19 and Overall Mortality Inequities in the Surge in Death Rates by Zip Code Characteristics: Massachusetts, January 1 to May 19, 2020. *American Journal of Public Health* **110,** 1850–1852. https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2020.305913 (Oct. 15, 2020).

34. Kutz, N. *Dynamic Mode Decomposition (Theory)* Apr. 14, 2018. https://www.youtube.com/watch?v=bYfGVQ1Sg98&ab_channel=NathanKutz.

35. Lazer, D. *et al. The COVID States Project #26: Trajectory of COVID-19-related behaviors* (OSF Preprints, Feb. 10, 2021). https://covidstates.org/reports.

36. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28,** 129–137. https://ieeexplore.ieee.org/document/1056489?reload=true (Mar. 1982).

37. Lu, F. S., Nguyen, A. T., Link, N. B., Lipsitch, M. & Santillana, M. Estimating the Early Outbreak Cumulative Incidence of COVID-19 in the United States: Three Complementary Approaches. *medRxiv 2020.04.18.20070821*. https://pubmed.ncbi.nlm.nih.gov/32587997/ (2020).

38. MacIntyre, C. R. & Chughtai, A. A. A rapid systematic review of the efficacy of face masks and respirators against coronaviruses and other respiratory transmissible viruses for the community, healthcare workers and sick patients. *International Journal of Nursing Studies* **108,** 103629. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7191274/ (Aug. 2020).

39. MacQueen, J. *Some methods for classification and analysis of multivariate observations* in. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (University of California Press, 1967). https://projecteuclid.org/euclid.bsmsp/1200512992.

40. McCarthy, T. Disunited states of America: responses to coronavirus shaped by hyper-partisan politics. *The Guardian*. https://www.theguardian.com/us-news/2020/mar/29/america-states-coronavirus-red-blue-different-approaches (Mar. 29, 2020).

41. McGrail, D. J., Dai, J., McAndrews, K. M. & Kalluri, R. Enacting national social distancing policies corresponds with dramatic reduction in COVID19 infection rates. *PLOS ONE* **15,** e0236619. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236619 (July 30, 2020).

42. Milligan, S. The Political Divide Over the Coronavirus. *US News & World Report*. //www.usnews.com/news/politics/articles/2020-03-18/the-political-divide-over-the-coronavirus (2021).

43. Moran, P. A. P. Notes On Continuous Stochastic Phenomena. *Biometrika* **37,** 17–23. https://doi.org/10.1093/biomet/37.1-2.17 (June 1, 1950).

44. Ng, E. & Muntaner, C. A Critical Approach to Macrosocial Determinants of Population Health: Engaging Scientific Realism and Incorporating Social Conflict. *Current Epidemiology Reports* **1,** 27–37. https://doi.org/10.1007/s40471-013-0002-0 (Mar. 1, 2014).

45. Papageorge, N. W. *et al. Socio-Demographic Factors Associated with Self-Protecting Behavior during the Covid-19 Pandemic* Working Paper w27378 (National Bureau of Economic Research, June 15, 2020). https://www.nber.org/papers/w27378.

46. Pew Research Center. *The Partisan Divide on Political Values Grows Even Wider* (Oct. 2017). https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/.

47. Poirier, C. *et al.* The role of environmental factors on transmission rates of the COVID-19 outbreak: an initial assessment in two spatial scales. *Scientific Reports* **10,** 17002. https://www.nature.com/articles/s41598-020-74089-7 (Oct. 12, 2020).

48. Pradeep, M. *et al.* Modelling and Forecasting of COVID-19 in India. *Journal of Infectious Diseases and Epidemiology* **6.** https://clinmedjournals.org/articles/jide/journal-of-infectious-diseases-and-epidemiology-jide-6-162.php?jid=jide (Sept. 18, 2020).

49. Proctor, J. L. & Eckhoff, P. A. Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International Health* **7,** 139–145. https://academic.oup.com/inthealth/article-lookup/doi/10.1093/inthealth/ihv009 (Mar. 2015).

50. Quinn, S. C. *et al.* Racial Disparities in Exposure, Susceptibility, and Access to Health Care in the US H1N1 Influenza Pandemic. *American Journal of Public Health* **101,** 285–293. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3020202/ (Feb. 2011).

51. Rader, B. *et al.* Crowding and the shape of COVID-19 epidemics. *Nature Medicine* **26,** 1829–1834. https://www.nature.com/articles/s41591-020-1104-0 (Dec. 2020).

52. Reddy, T. *et al.* Short-term real-time prediction of total number of reported COVID-19 cases and deaths in South Africa: a data driven approach. *BMC Medical Research Methodology* **21,** 15. https://doi.org/10.1186/s12874-020-01165-x (Jan. 11, 2021).

53. Rey, S. J. & Anselin, L. PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies* **37,** 5–27 (2007).

54. Rice, J. R. A Theory of Condition. *SIAM Journal on Numerical Analysis* **3,** 287–310. https://www.jstor.org/stable/2949623 (1966).

55. Roberts, D. Partisanship is the strongest predictor of coronavirus response. *Vox.* https://www.vox.com/science-and-health/2020/3/31/21199271/coronavirus-in-us-trump-republicans-democrats-survey-epistemic-crisis (Mar. 31, 2020).

56. Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics* **656,** 5–28. https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/abs/dynamic-mode-decomposition-of-numerical-and-experimental-data/AA4C763B525515AD4521A6CC5E10DBD4 (Aug. 2010).

57. Schmid, P. J. & Sesterhenn, J. Dynamic Mode Decomposition of numerical and experimental data. Conference Name: APS Division of Fluid Dynamics, MR.007. http://adsabs.harvard.edu/abs/2008APS..DFD.MR007S (Nov. 1, 2008).

58. Shear, M. D. Presidential Speech Highlights: Biden Calls For U.S. to 'Mark Our Independence From This Virus' by 4th of July. *The New York Times.* https://www.nytimes.com/live/2021/03/11/us/joe-biden-news (Mar. 11, 2021).

59. Solis, J., Franco-Paredes, C., Henao-Martínez, A. F., Krsak, M. & Zimmer, S. M. Structural Vulnerability in the U.S. Revealed in Three Waves of COVID-19. *The American Journal of Tropical Medicine and Hygiene* **103,** 25–27. http://www.ajtmh.org/content/journals/10.4269/ajtmh.20-0391 (July 8, 2020).

60. Spearman, C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* **15,** 72–101. https://www.jstor.org/stable/1412159 (1904).

61. The Centers for Medicare & Medicaid Services. *Provider Information* 2021. https://data.cms.gov/provider-data/dataset/4pq5-n9py.

62. Treviño, F. M., Moyer, M. E., Valdez, R. B. & Stroup-Benham, C. A. Health insurance coverage and utilization of health services by Mexican Americans, mainland Puerto Ricans, and Cuban Americans. *JAMA* **265,** 233–237. https://pubmed.ncbi.nlm.nih.gov/1984153/ (Jan. 9, 1991).

63. Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L. & Kutz, J. N. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics* **1,** 391. https://www.aimsciences.org/article/doi/10.3934/jcd.2014.1.391 (2014).

64. U.S. Census Bureau. 2014-2018. https://www.census.gov/programs-surveys/acs/data.html.

65. Vera Institute of Justice. *Incarceration Trends Dataset* 2020. https://github.com/vera-institute/incarceration-trends.

66. Wemrell, M., Merlo, J., Mulinari, S. & Hornborg, A.-C. Contemporary Epidemiology: A Review of Critical Discussions Within the Discipline and A Call for Further Dialogue with Social Theory. *Sociology Compass* **10,** 153–171. https://onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12345 (2016).

67. Williams, D. R. & Rucker, T. D. Understanding and Addressing Racial Disparities in Health Care. *Health Care Financing Review* **21,** 75–90. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194634/ (2000).

68. World Health Organization. *A conceptual framework for action on the social determinants of health: debates, policy & practice, case studies.* http://apps.who.int/iris/bitstream/10665/44489/1/9789241500852_eng.pdf (2010).

69. World Health Organization. *What are the health risks related to overcrowding?* http://www.who.int/water_sanitation_health/emergencies/qa/emergencies_qa9/en/.

70. Zhang, R., Li, Y., Zhang, A. L., Wang, Y. & Molina, M. J. Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 14857–14863. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7334447/ (June 30, 2020).

71. Zhang, Y. *et al.* Prediction of the COVID-19 outbreak in China based on a new stochastic dynamic model. *Scientific Reports* **10,** 21522. https://www.nature.com/articles/s41598-020-76630-0 (Dec. 9, 2020).

72. Zhu, S. *et al.* High-resolution Spatio-temporal Model for County-level COVID-19 Activity in the U.S. *arXiv:2009.07356*. http://arxiv.org/abs/2009.07356 (Sept. 16, 2020).

Code Base

All analysis is conducted with Python (version 3.6.9) using NumPy and Pandas for data analysis. All figures are generated using Matplotlib, Cartopy, and Powerpoint. Models were constructed using sklearn and statsmodel. The entire code base can be found here.