Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Exploring the Effects of Trump Support and Vaccine Hesitancy on the SARS-CoV-2 Time-Varying Reproduction Number

A County-Level Study in the Context of the 2020 U.S. Presidential Election

Author:
Emma Valentine Marie Landry

Supervisor:
Dr. Seth Flaxman

Submitted in partial fulfillment of the requirements for the MSci degree in Mathematics of Imperial College London

June 2021

This is my own work except where otherwise stated.

Signed:  Date: 15/06/2021

Acknowledgments

I would first like to thank my supervisor, Dr. Seth Flaxman, for his continued assistance, guidance, and encouragement throughout the project, as well as for his enthusiasm for the subject. I would also like to express my gratitude to Dr. Oliver Ratmann, Andrea Brizzi, and Valerie Bradley, for devoting time to my research and for sharing their knowledge on the topics I was exploring. My thanks go to all the other members of the Imperial College COVID-19 Response Team for their reactivity. Last but not least, I wish to thank my family for their constant support all through my degree.

Abstract

In the midst of the COVID-19 pandemic, the time-varying reproduction number R_t has been an essential tool for measuring immunity in populations and the effects of non-pharmaceutical interventions. This thesis uses a semi-mechanistic Bayesian model to estimate R_t values at the county-level in the contiguous United States based on observed deaths data, where less populated counties are grouped together to ensure a sufficient number of observations. Bayesian linear regression models are then fit to those values using Trump electoral support in the 2020 U.S. presidential election and vaccine hesitancy rates as predictors. I find that the effect of these two covariates on the reproduction number varies with time, and I identify a change of sign in both regression coefficients in the weeks following the election. These findings appear consistent with the results obtained for a state-level model, which was constructed to take advantage of availability of more detailed data. A shortcoming of the modelling approach, both at the state and county levels, is that the reproduction number estimates are obtained by considering each region independently. Introduction of a hierarchical structure through partial pooling of parameters could help improve the models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims

The main aim of this thesis is to provide a better understanding of the SARS-CoV-2 dynamics in the United States, focusing on the county-level. The focus will be set particularly on modelling the reproduction number using the methods developed by the Imperial College COVID-19 response team, specifically those provided in the `R`[1] `epidemia`[2] package, that implements the semi-mechanistic Bayesian model introduced by Flaxman et al. [3].

Motivated by the evidence given by Messner and Payson [4], which shows there is a higher variation in outbreak rates at the county-level than at the state-level and indicates studies should be performed at the most granular level possible, I thus first aspire to obtain functional models that can provide accurate estimates of the SARS-CoV-2 reproduction number for these smaller geographical areas.

The subsequent goal this thesis aims to achieve is to understand the relationship between this reproduction number and quantities that would not yet have been explored in detail, partly due their novelty at the time of the project: Trump support in the 2020 Presidential election, and vaccine hesitancy. These two components differ greatly to those that tend to be considered in studies focusing on the reproduction number: they do not have a direct physical effect on the spread as will social distancing measures [5] or mobility [6].

Both are of particular interest as they pertain to unique events in history. On one hand, the 2020 US Presidential election was the first one taking place in the midst of a global pandemic, which had significant effects on the way it was run, as described in the following chapter. On the other, the COVID-19 vaccines have been developed faster than any other: the FDA approved the Pfizer-BioNTech vaccine on December 12^{th}, 2020 after less than a year of research, while the previously fastest was the mumps vaccine in the 1960s, which took four years from sampling to approval [7]. In this context, the objective hence is to establish whether vaccine hesitancy and Trump electoral support are related to the reproduction number, and if so, in what way.

1.2 Thesis Outline

The thesis starts with the background, providing an overview of the needed epidemiological definitions, introducing the mathematical approaches taken in modelling of infectious

diseases and how they build up to the methods used in the project to estimate the reproduction number R_t. The chapter also provides motivation for the study of this quantity along with vaccine hesitancy and Trump electoral support, by building on the literature covering these topics. Chapter 3 presents all the datasets used during the project, explaining the information they contain as well as how they are manipulated. Chapter 4 outlines the methods that were followed for obtaining results. First, the semi-mechanistic Bayesian model is defined, and the details of its implementation follow. The second part of the chapter covers the Bayesian linear regression models that explore the relationship between R_t, vaccine hesitancy and Trump support. Chapter 5 presents the findings of this project in the same order as the methods were presented. Finally, the discussion in Chapter 6 analyses those findings, providing explanations for the observations that followed from the results. Chapter 6 also provides the limitations to the methods and approaches taken in this report, proposing improvement suggestions as well as hinting at further topics of study this thesis motivates.

Chapter 2

Background

2.1 Modelling of Epidemics

2.1.1 The Time-Varying Reproduction Number

The basic reproductive number R_0, taken in the epidemiological context, is defined by Heffernan et al. [8, p. 281] as 'the mean number of individuals infected by a single infected individual during their entire infectious period, in a population which is entirely susceptible'. It was however initially developed in the context of demographics, before the applications to vector-borne diseases and directly transmitted human infections were developed [8].

Extending this definition, the time-varying reproduction number, denoted as R_t, is defined as the number of new cases of a disease provoked by a single infected individual while they are infectious with the disease of interest [9] at a specific time point, and as a necessary consequence, accounting for individuals becoming immune. In addition to measuring immunity (from vaccination or previous infection), it also used to measure the impact of non-pharmaceutical interventions (NPIs) such as social distancing [3]. It is intuitively clear that the benchmark of interest is 1: a value larger than that means that transmission is increasing, and therefore that the epidemic is worsening, whereas values less than 1 imply slowing down, ideally with convergence to 0 for the epidemic to die out.

This statistic, as indicated by its subscript t, is time specific, making it a useful tool to observe the evolution of an epidemic over time. It allows to track the effect of interventions, providing information on whether control needs to be intensified. It can be desirable to monitor the values of R_t through approaches based on the number of reported cases as it reflects variation in transmission intensity [10]. The issue with the value of reported cases is that there is a delay between the moment the disease is transmitted and the time the infected individual becomes infectious. This is firstly due to the incubation period of a disease. Although it can sometimes be quite short, such as with the case of influenza, which has a median incubation period of 1.4 days [11]; this period during which the disease is not detectable can be much longer, for example with infectious mononucleosis, for which the incubation period is estimated to be in the range of 33-49 days [12]. As a result it may never be identified if an individual is infectious: PCR or rapid lateral flow tests will only inform on whether the individual is infected, and antibody tests will inform regarding past infection and unlikely infectiousness. The delays between infection and reporting can also be a consequence of limitations with detection methods, including time for tests to return results or limited available personnel on weekends and holidays.

In theory, the time-varying reproduction number for a disease can be computed on the global scale, however it is sensible to restrict it to a limited geographical area. Indeed, the dynamics of a disease will vary greatly regionally, and the statistic is therefore not very informative. Limiting R_t spatially is also coherent with the implementation of targeted public health policy, which will tend to focus on the country level, or even possibly on a within-country regional level (such as the state or county level in the United States).

2.1.2 Estimating the Time- Varying Reproduction Number

A wide variety of approaches to the modelling of epidemics have been developed in the literature, that can generally be split into two broad categories: deterministic and statistical models [13]. The former consists mainly of the simple compartmental disease models [14], which include the well-known SIR model.

The model was developed by Kermack and McKendrick [15] and considers the number of susceptibles, infectious and recovered individuals at time t, respectively denoted as $S(t), I(t)$, and $R(t)$. Adopting the notation given by van den Driessche [14], the model can be described by the following system of ordinary differential equations (ODEs).

$$\frac{dS(t)}{dt} = -\beta S(t)I(t) \tag{2.1a}$$

$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t) \tag{2.1b}$$

$$\frac{dR(t)}{dt} = f\gamma I(t) \tag{2.1c}$$

In the above, β denotes the disease transmission rate, γ denotes the recovery rate and f denotes the fraction of infectious individuals who recover from the disease. In this context, the basic reproduction ratio is defined as [14]

$$R_0 = \frac{\beta S(0)}{\gamma}. \tag{2.2}$$

This definition is consistent with the comments made earlier: a value of $R_0 < 1$ will indeed lead to infections going to zero, whereas with $R_0 > 1$ the number of infected individuals will increase (before going back to zero as eventually there will remain no susceptibles in the population).

Many other compartmental disease models build on the concepts of the SIR model, for example by considering a latent period after infection during which an infected individual cannot yet transmit the infection, consequently introducing a fourth quantity corresponding to the number of exposed individuals, leading to the the SEIR model [14]. Another deterministic approach worth mentioning is the one Heffernan et al. [8] refer to as the 'survival function' approach. Letting $F(a)$ be the probability for a newly infected individual to remain infectious until time a, and $b(a)$ the average number of newly infected individuals that an infectious individual will produce per unit time for a total time a, Heesterbeek and Dietz [16] give R_0 as

$$R_0 = \int_0^\infty b(a)F(a)da. \tag{2.3}$$

Expressions for the functions $b(a)$ and $F(a)$ can be obtained directly from the SIR equations, but can also be extended to a wider range of models [14].

The second group of models presented by Myers et al. [13] are the statistical models. Examples of purely statistical models include Generalized Linear Models, time series approaches using Auto Regressive Integrated Moving Average (ARIMA), or other methods based on machine learning [2]. The approaches of most relevance to the methods adopted in this thesis are those on focusing the estimation of R_t using case data.

The first approach worth noting, that differs from most others in that it treats cases directly as observations, is the one taken by Wallinga and Teunis [17]. Derived with the aim of modelling the severe acute respiratory syndrome (SARS) that emerged in late 2002 in southern China, the model describes the outbreak of the infectious disease as a directed network. This method requires only case incidence data and the distribution of the time between the onset of symptoms of a primary case and the onset of symptoms of a secondary case (the serial interval)[9], a desirable property, however it has a few significant shortcomings. Firstly, as already touched upon in subsection 2.1.1, cases do not necessarily show consistency over time, to which is added the issue of under-reporting implying uncertainty around the incidence of the infectious disease[18]. Furthermore, the estimates are right-censored as the method requires data from later than time t to obtain R_t [9].

Other approaches such as those presented by Ferguson et al. [19], Fraser [20], Bettencourt and Ribeiro [21], Kelly et al. [22], Cori et al. [9] account for under-reporting in the definition of the models [23], however the issue remains of sole reliance on case data. These drawbacks therefore motivate the adoption of the model implemented by Scott et al. [24] in their `R` [1] package `epidemia`. It allows the use of different types of observations, and in particular enables the implementation of models based on observed deaths data. Such data will have some limitations: deaths could have been missed, there can exist discrepancies across geographical areas, delays in reporting are unavoidable however difficult to predict [3]. It is however still believed to be more reliable than cases. Furthermore, the model implemented in `epidemia` can be described as 'semi-mechanistic', meaning it isn't either completely deterministic nor completely statistical. This therefore allows it to depart from the imposed structure of compartmental models which make generalizing difficult [25], while preserving the inherent stochasticity of epidemic models [25] and keeping the infection dynamics plausible through interpretable mechanistic parameters which describe the propagation of the disease [2].

2.1.3 COVID-19 Pandemic Context

The coronavirus disease 2019 (COVID-19) is a highly infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [26]. Starting from its first case detected in late December 2019 in Wuhan, Hubei Province, China, SARS-CoV-2 very rapidly spread to entire world, leading to the declaration by the World Health Organization on March 11, 2020 of a global pandemic. The consequences have been catastrophic, with over 3.5 million deaths reported as of June 3, 2021 as given by the European Centre for Disease Prevention and Control (ECDC) [27]. In such a context, it is therefore of utmost importance to understand the virus and the dynamics of the epidemic in order to halt its spread; by informing policy decisions, motivating changes in individual behaviors, and developing the appropriate medical solutions. Due to its unobservable nature, developing models that

provide accurate estimates of the COVID-19 reproduction number R_t is thus a key task, as it provides good indication of spread.

The package `epidemia` [2] has been built based on the methods presented by Flaxman et al. [3], whose focus when estimating R_t in European countries was understanding the effects of NPIs, which included complete lockdown, public events banned, school closures, self-isolation, and social distancing introduction. Since this first presentation of the methods, many more results have been obtained implementing them. These include the work by Unwin et al. [6] which conducted a state-level analysis in the U.S. of R_t, using mobility data as a proxy for NPIs and accounting for the states' age structure. Other notable work using the Bayesian semi-mechanistic approach in the United States include Monod et al. [28], who use mobile-phone data to track age-specific mobility and reconstruct human contact patterns; Olney et al. [29], who look at the state-level effects of social distancing interventions; and Smith et al. [30], who investigate the association of temperature, humidity, ultraviolet radiation, and population density with R_t. Hawryluk et al. [31] and Faria et al. [32] apply these methods to data in Brazil. The former usies hospital data to fit a joint subnational model to determine the COVID-19 epidemiological distributions, and the latter incorporates both genomic and mortality data to track the P.1 lineage of the virus.

2.2 Relevance of Trump electoral Support and Vaccine Hesitancy

2.2.1 COVID-19 and the 2020 U.S. Presidential Election

The context of the COVID-19 pandemic undoubtedly set the 2020 U.S. Presidential Election apart from any previous one. As a result of the fact that the Constitution does not specify any mechanisms for changing the expiry of the presidential mandate (January 20^{th}, 2021), the election was maintained in the midst of the pandemic [33]. Some changes that can easily be observed are firstly the changes in campaigning. Due to COVID-19 restrictions, the typical campaigning including door-knocking, public events, and large rallies was impossible; and both candidates chose distinct strategies [33]. On one hand, Donald Trump maintained large rallies, including some held in indoor venues, where lack of social distancing and mask-wearing was often observed. On the other hand, Joe Biden replaced these events by 'drive-in' rallies, where supporters would remain in their own cars. The contrast between these two approaches already hints at distinct behaviors towards COVID-19 encapsulated by the Democratic and Republican parties. This election was also distinct from any other in terms of the voting methods: in contrast with 2016 where 60% of the ballots were cast in person on election day, this value was only 28% in 2020 [34].

The impacts of the pandemic can also be directly be witnessed in the outcome of the election. Baccini et al. [35] explored those effects and found significant evidence that COVID-19 cases decreased electoral support for Trump, and in particular, this impact was stronger in states without a stay-at-home order, states where Trump won in 2016, swing states, and urban counties. They propose two explanations for those findings. One is that voters sanctioned Trump for his handling of the pandemic, and the other is that the public health threat and economic recession may have lead to a shift in preferences to expansion of a social safety net, which are policies Democrats are more likely to focus on.

Another key element when considering the election in the context of the pandemic, touched

upon in the mention of the campaign choices of both candidates, are the behaviors of Democratic and Republican voters. These differences are well represented by two polls conducted by the Pew Research Center [36], which are represented in tables 2.1 and 2.2. These show a clear discrepancy between Republican and Democratic attitudes to policies aiming to slow the spread of COVID-19, as well as a discrepancy in how each are comfortable with going to indoor places which will naturally imply less social distancing. It is worth noting these polls data from March and June 2020 respectively, and may not represent behaviors at the time of elections, however they do provide motivation for a joint study of political opinions and COVID-19 dynamics. Given the results by Flaxman et al. [3] showing the effect of NPIs in reducing R_t in Europe, and the work of Olney et al. [29] looking at social distancing in the United States and also finding that social distancing measures such as lockdowns or schools closing have a non-trivial effect on the decrease of R_t, it appears reasonable to hypothesize on the existence of a link between Trump support in the election and the R_t value.

	Agreement on necessity (%)	
Policy	Republican	Democratic
Restricting international travel to the U.S.	96	94
Cancelling major sports and entertainment events	87	95
Closing K-12 schools	85	94
Asking people to avoid gathering in groups of more than 10	82	92
Limiting restaurants to carry-out only	78	91
Requiring most businesses other than grocery stores and pharmacies to close	61	81
Postponing upcoming state primary elections	66	73

Table 2.1: Bipartisan agreement with necessity of policies to address the pandemic. These results are obtained from a survey by the Pew Research Center [36], conducted on March 19-24, 2020. The terms Republican and Democratic include those leaning towards each party respectively. The margin of error for this survey is $\pm 1.5\%$.

	% comfortable with	
Event	Republican	Democratic
Going out the grocery store	87	73
Visiting with a family member or close friend inside their home	88	68
Going to a hair salon or barbershop	72	37
Eating out in a restaurant	75	28
Attending an indoor sporting event or concert	40	11
Attending a crowded party	31	8

Table 2.2: Bipartisan comfort with going out to different places. These results are obtained from a survey by the Pew Research Center [36], conducted on June 16-22, 2020. The terms Republican and Democratic include those leaning towards each party respectively. The margin of error for this survey is $\pm 1.8\%$

2.2.2 Vaccine Hesitancy

As a result of the high effectiveness of vaccines in prevention against infection and transmission [37], it naturally follows that vaccination rates should have a direct effect on the

reproduction number R_t. In fact, there exists a direct relationship between "herd immunity", term coined by Topley and Wilson [38] meaning the 'indirect protection from an infectious disease that happens when a population is immune either through vaccination or immunity developed through previous infection' [39], and the basic reproduction number R_0. Indeed, the level of herd immunity required to block transmission is given by $1 - 1/R_0$ [40].

Given that the principal focus of this thesis on the period surrounding the elections, proportions vaccinated are not an adequate quantity to focus on, as doses started being administered on December 14^{th}, 2020 in the United States. A related quantity which may however be explored is vaccination hesitancy, defined as 'the delay in acceptance or refusal of vaccination despite availability of vaccination service' [41]. It naturally cannot not have the same impact on R_t as vaccination numbers do, as the latter represent a physical modification to the characteristics of the population that impact directly the dynamics of the disease. Results from the Axios-Ipsos survey [42] give that, in the week leading up to March 23^{rd} 2021, 52% of unvaccinated Americans report have visited friends or relatives, as opposed to 41% of those vaccinated. Additionally, results from May 11^{th} 2021 show that vaccinated people are more likely than vaccinated people to wear a mask all the time (65 % against 46 %). The unvaccinated population can by no means taken to be equivalent to the hesitant population, however the above results could hint at the fact that vaccine hesitancy may be accompanied by attitudes that have been shown to facilitate the spread of the disease.

Moreover, the Axios-Ipsos findings from April 6, 2021 give that 31% of Republicans are resistant to the vaccine, compared to the value of 19% in the general population; a result that hints at a potential relationship between vaccine hesitancy and political orientation. Although those results may not necessarily be very accurate [43], they certainly provide motivation for an exploration of the relationships between R_t, Trump support, and vaccine hesitancy.

Chapter 3

Data

The entire analysis is performed on all the states of the USA, excluding the states of Alaska and Hawaii. This choice is made due to these being non-contiguous, the epidemic dynamics can be expected to be different from the rest of the territory. The District of Columbia, Puerto Rico, and the Island Areas are also not included in the analysis.

3.1 COVID-19 Data

3.1.1 Deaths and Cases

The first part of the analysis consists in obtaining estimates of R_t, as described in the next chapter. The aim of this study is to implement a model at the county-level in the United States. The deaths data (as well as cases, used in visualizations but not as part of the model) is obtained from USAFacts [44]. As per the detailed methodology [45], the datasets are obtained through aggregation of data from the Centers for Disease Control and Prevention (CDC), state-, and local-level public health agencies. The positive case counts and death counts follow the CDC reporting guidelines [45]. Therefore, presumptive positive cases are counted as confirmed cases. Furthermore, deaths are considered COVID-19 related if it is established that the disease clearly played a direct role in causing death. COVID-19 needs to be the unique cause of death, however if an individual dies while coincidentally being infected by the disease, it will not automatically be counted as a COVID-19 death.

The data is then pre-processed for use with functions from the `epidemia R` package. Although downloaded at a later date, the dates considered are from the start of reporting in the data set (January 22^{nd}, 2020) to January 31^{st}, 2021. The study is restricted to those dates as the main interest is data from 2020, due to the presidential elections taking place on November 3^{rd}, 2020; with January data necessary for predictions of the end of the year due to the model needing data from weeks forward to construct predictions, as a result of the average 3 weeks from infection to death.

It has been identified by collaborators at Imperial College London that the estimations may not be converging and accurate when the death numbers are too low. Therefore, to avoid such problems, certain counties are grouped to represent to have a joint higher population size (the population data is taken from the United States Census Bureau [46], and are county estimates for 2019, built on base data from the 2010 census). These counties are grouped geographically such that a group contains only neighboring counties (see figure A.1

in the supplementfor a map of these groupings). The county adjacency dataset was obtained through the United States Census Bureau website as well [47].

Although not the main focus of this project, an analysis is also performed at the state-level. The data required comes from the same sources as for the county level, that is the deaths data is obtained from USAFacts [44].

3.1.2 Vaccination Hesitancy

In the second part of the analysis, linear models are fit predicting R_t with vaccination hesitancy as well as with presidential election data.

State-level

At the state level, vaccination hesitancy is available directly through the Household Pulse survey [48] (motivated by Bradley et al. [43]'s findings, the limitations of such data is detailed in the discussion). The Census Bureau introduced the survey on April 23, 2020 as a way to measure how the COVID-19 pandemic impacts households. However, questions regarding vaccinations were only introduced in 2021, therefore the dataset used corresponds to the two-week period from January 6, 2021 to January 18, 2021. As per the methodology, the survey was conducted on 68,348 respondents, and the data then contains estimates for the entire population. The data provides information at the state-level regarding the likeliness of respondents to get vaccinated. Defining hesitancy to be all the categories but "Will definitely get a vaccine" (including respondents who received a first dose but are not sure about getting a second one), it also provides the reasons hesitant respondents give for not planning to receive the vaccine.

County-level

The Household Pulse survey only provides vaccination hesitancy data at the state level, therefore a different source is required for the county-level. It is obtained from the Office of the Assistant Secretary for Plannning and Education (ASPE) website [49]. The methodology for the dataset describes a two-step approach: first regresssion, then post-stratification.

First, the Household Pulse survey data I described in the state-level approach is used, where the ASPE excludes the respondents who gave "definitely" and "probably" as answers regarding their intent to receive the vaccine [49] (differing from the state-level where only "definitely" is excluded). The ASPE then obtains logistic regresssion coefficients for a variety of predictors of vaccine hesitancy available in the survey data.

The second step makes use of the 2019 American Community Survey (ACS) microdata [50], applying the regression coefficients to the ACS respondents, and averaging predicted values by using ACS survey county-level weights [49]. This "post-stratification" step allows to move from the estimates for each demographic group in the total population, to individual predictions for a representative sample of the U.S. population that can be used to approximate county-level proportions of hesitancy.

3.2 Presidential Election Data

3.2.1 Forecasts

In order to best predict R_t in the months leading up to the elections through the Trump support predictor, I use daily election forecasts given by FiveThirtyEight [51], covering the period from June 1^{st}, to election day, November 3^{rd}. As indicated in the methodology, these forecasts are constructed from a variety of publicly accessible polls, with weights assigned to them based on their sample size and how FiveThirtyEight rates the pollster. The obtained averages are then adjusted to account for a multitude of elements, such as the fact that certain polls will consistently be biased towards one party. The polls are also combined with additional demographic and economic data to enhance the forecasts. The dataset however only provides the information at the state-level, and for county-level analysis I therefore solely rely on the presidential election results.

3.2.2 Results

State-level

The state-level returns for elections to the U.S. presidency are obtained through the MIT Election Data + Science Lab [52]. The data gives results for the presidential elections since 1976, as collected from the House of Representatives website which provides official results. The dataset is then cleaned to contain only relevant entries, only keeping results from 2020 and 2016 (in order to consider to use the change in votes from one election to the other as a predictor).

County-level

Although ideally the same source would have been taken for the county-level results, the MIT Election Data + Science Lab only provide such data up until the 2016 election. Instead, the results are taken from McGovern's Github repository [53]. As described in the README, and also visible from the data processing source code, the dataset is obtained from scraping county level results data from the New York Times website.

Chapter 4

Methods

Below I describe all the methods adopted in the project. I first focus on describing the model used to derive estimates for the reproduction number at the county-level, before providing details pertaining to the linear regression task.

4.1 Estimating the Time-Varying Reproduction Number

The first part of the analysis consists in estimating the time-varying reproduction number using the approach taken by Bhatt et al. [23], that builds on the work of Flaxman et al. [3], using the implementation in the `R` package `epidemia` [24]. Each county (or group of counties when applicable) is modelled as a single homogeneous population, independently of others, rather than jointly through shared parameters. Such an approach has some limitations, which are expanded upon in the discussion.

4.1.1 Statistical Model

Notation

Before providing the specifics of the model, I first introduce the notation that will be used extensively throughout this thesis.

$\boldsymbol{Y} = (Y_1, ..., Y_n)$	Observed time series of data (daily deaths counts in the context of this thesis)
y_t	Expected value of the data distribution for the number of deaths, at time t
i_t	Number of new infections at time t (unobserved)
$i_{v:0}$	Seeded infections on days $v, v+1, ...0$
α_t	Ascertainment rate at time t (typically interpreted as the proportion events at time t recorded in the data, known as the Infection Fatality Rate (IFR) in the case of deaths data)
π_t	Infection to observation distribution
ϕ	Auxiliary parameter (applicable for certain families of GLMs)
R_t	Reproduction number at time t
g	Probability mass function for the time between infections, known as the generation time distribution

Definition of Model

As given by Scott et al. [54], the model for the data distribution can be expressed the following way, asssuming a seeding period $[v, 0]$ during which the process is initialized

$$Y_t \sim p(y_t, \phi) \tag{4.1a}$$

$$y_t = \alpha_t \sum_{s=v}^{t-1} i_s \pi_{t-s} \tag{4.1b}$$

The discrete π_s values in (4.1b) allow to form a linear combination of the past infections, weighting these by the probability of each leading to a death. The ascertainment rate α_t accounts for the discrepancies between actual death numbers, and recorded death numbers.

The motivation for a discrete time equation for new infections comes through a continuous time renewal process approach, as was taken by Bhatt et al. [23]. A counting process $\{N(t) : t \geq 0\}$ is called a renewal process if the interarrival times are independent and identically distributed with an arbitrary distribution [55]. The process is initialized by infection values in the seeding period $[v, 0]$, that constitute one of the parameters in the model. Letting $N^I(t)$ denote the number of infections occurring up to time t, the process intensity is given by

$$\lambda(t) = R(t) \int_{v<s<t} g(t-s) N^I(ds), \quad t > 0, \tag{4.2}$$

where g is the probability density function for the time between infections, and $R(t) > 0$ is the reproduction number at time t [23].

Now, to transition into a discrete model, I_t is introduced, the number of infections at time t. As the interest lies in discrete time points t, it follows that $I_t = N_t^I - N_{t-1}^I$. Then the discrete version of (4.2), as given in [23], is

$$\mathbb{E}[I_t | R_{1:t}, I_{v:t-1}] = R_t L_t, \qquad L_t \equiv \sum_{s=v}^{t-1} I_s g_{t-s}, \tag{4.3}$$

where L_t is known as the total infectiousness until time t, and g is a discretized version of the continuous density from (4.2). Letting $i_t \equiv \mathbb{E}(I_t | R_{1:t}, I_{v:0})$, and taking the conditional expectation given $R_{1:t}$ and $I_{v:0}$ on both sides of (4.3) as given by Bhatt et al. [23], one obtains

$$\text{LHS} = \mathbb{E}\{\mathbb{E}[I_t | R_{1:t}, I_{v:t-1}] | R_{1:t}, I_{v:0}\} = \mathbb{E}[I_t | R_{1:t}, I_{v:0}] = i_t, \tag{4.4}$$

where the second equality follows from the law of total expectations. For the right hand side,

$$\text{RHS} = \mathbb{E}[R_t L_t | R_{1:t}, I_{v:0}] \tag{4.5a}$$

$$= R_t \mathbb{E}[L_t | R_{1:t}, I_{v:0}] \tag{4.5b}$$

$$= R_t \mathbb{E}[\sum_{s=v}^{t-1} I_s g_{t-s} | R_{1:t}, I_{v:0}] \tag{4.5c}$$

$$= R_t \sum_{s=v}^{t-1} \mathbb{E}[I_s | R_{1:s}, I_{v:0}] g_{t-s} \tag{4.5d}$$

$$= R_t \sum_{s=v}^{t-1} i_s g_{t-s}. \tag{4.5e}$$

As LHS = RHS by (4.3), the equation describing new infections at times $t > 0$ is then given by

$$i_t = R_t \sum_{s=v}^{t-1} i_s g_{t-s}. \tag{4.6}$$

The unknown parameters of the full model, encapsulated by equations (4.1a), (4.1b) and (4.6), are therefore: $i_{v:0}, \boldsymbol{R}, \phi, \boldsymbol{\alpha}$. The remainder of the quantities can either be obtained directly from the equations presented, or are pre-specified and fixed, which is why the method is referred to as "semi-mechanistic". These unknowns parameters are then assigned priors and inferred via MCMC methods, as described in the next part.

4.1.2 Implementation

County Groupings

Although one could wish to work with each county separately to obtain estimates at the most granular level, from an epidemiological perspective this is simply unfeasible. Indeed, numerous US counties have very small populations, as low as 169 in Loving County, Texas. It naturally follows that deaths (and cases), which constitute our observations, will be close to zero during the entire period of analysis. This makes convergence and obtaining meaningful results impossible using the methods presented. A low population setting can firstly make it difficult for the model to identify the beginning of the pandemic. It can also lead to the model underestimating the value of R_t. Indeed, data that gives observed deaths or cases equal to zero during several days in a row could lead to an estimation of R_t being zero as well, implying the pandemic has died out, which is a stronger assumption than can reasonably be made.

The approach taken to remedy to this problem is to group counties with low populations together. The issue that naturally follows is the specification of what is considered a "low population". Any choice will to some extent be arbitrary as naturally there exists no formal method for determining such a threshold. However, a good choice would be one that is not too large so as to avoid making groups that lose out on too much of the county-level dynamics, while large enough to mitigate the problem of insufficient observations. The threshold value that tended to give satisfactory results as suggested by collaborators, including Andrea Brizzi, working with similar problems, with yet unpublished studies, is a population of 150,000.

Given the threshold value, certain assumptions then need to be made before grouping counties. Firstly, the groupings are only made within a state, i.e. a group can only contain counties belonging to the same state. The motivation for this choice is to ensure that these counties will be bound by similar restrictions and policies regarding the pandemic. Indeed, the Tenth Amendment to the US constitution has been interpreted to include police powers, which implies the right to take action during a public health crisis; including the implementation of lockdowns or quarantines [56].

Furthermore, I chose that a county can only be grouped with a neighbouring county. The main motivation is mainly from a travel perspective. I illustrate this taking a specific example from the the Bureau of Transportation statistics data. In the month of November, over 96% of the average daily trips made were of a distance less than 50 miles [57]. Due to the overwhelming majority of trips being short-distance, it follows that travel was mainly within-county, or to counties close geographically. Thus, the spread of COVID-19 by a given individual shouldn't in general extend to further distances. That is not to say that the groupings create a hermetic barrier stopping the spread, however it motivates modelling the reproduction number within a geographically compact area rather than in two counties across the state from each other.

Using the adjacency and population data sets described in the previous chapter, the method to obtain the groups of counties within each state is as follows. Let $P = (P_1, ..., P_N)$ denote the list containing the population of each grouping, where "grouping" can refer to either a single county (initially that will be the case for all groupings), or a group of counties. Also let $Q^i = (Q^i_1, ..., Q^i_{M^i})$ denote the list containing the populations of the groupings adjacent to grouping i. Then the pseudo-code for the algorithm is:

Algorithm 1: Grouping Counties Algorithm

while $\min_k P_k$ **do**
  Set $i \leftarrow \arg\min_k P_k$;
  Set $j \leftarrow \arg\min_l Q^i_l$;
  Replace grouping i by $i \cup j$, with population $\min_k P_k + \min_l Q^i_l$;
  Remove grouping j
end

The `R` implementation of this algorithm is given in the appendix as listing B.1. To visualize the output of this algorithm, the groupings are shown on a map for four different states in Figure 4.1. Additionally, Figure A.1 in the supplement gives a map of the entire United States, with groupings identified by color.

Modelling Choices and Assumptions

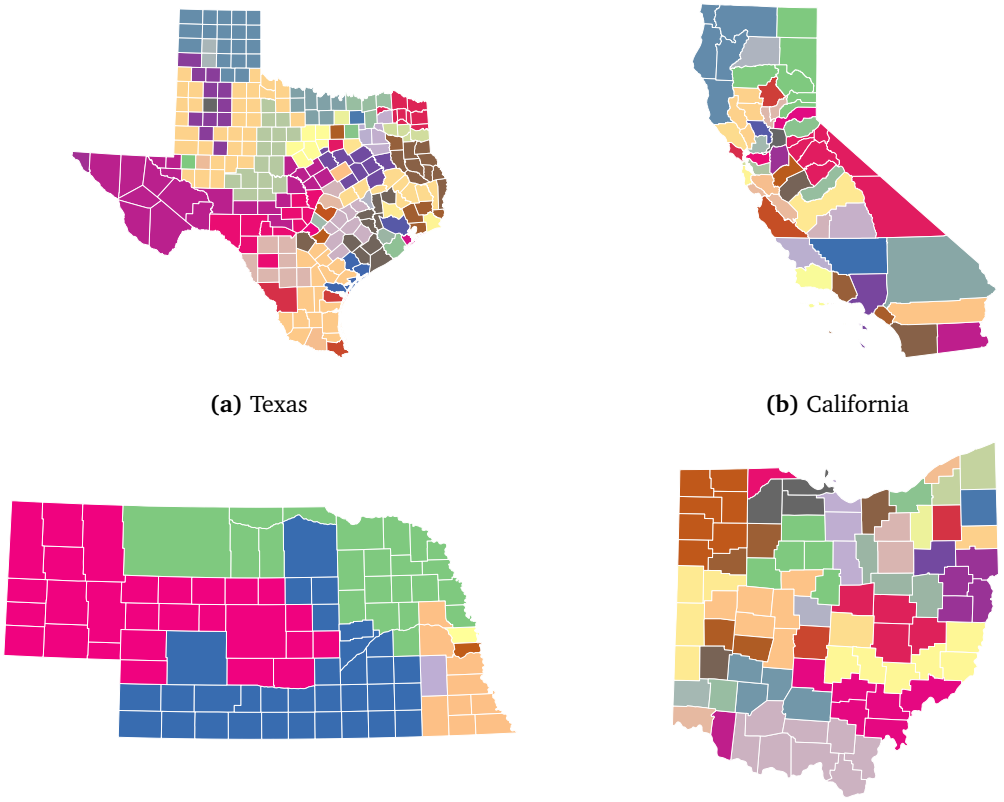
As given at the end of subsection 4.1.1, the unknowns parameters of the model are assigned priors [54], such that

$$\mathbb{P}(i_{v:0}, \boldsymbol{R}, \phi, \boldsymbol{\alpha}|\boldsymbol{Y}) \propto \mathbb{P}(i_{v:0})\mathbb{P}(\boldsymbol{R})\mathbb{P}(\phi)\mathbb{P}(\boldsymbol{\alpha}) \prod_{t=1}^{n} \mathbb{P}(Y_t|y_t, \phi). \tag{4.7}$$

Due to the independence of the parameters, $\mathbb{P}(i_{v:0})\mathbb{P}(\boldsymbol{R})\mathbb{P}(\phi)\mathbb{P}(\boldsymbol{\alpha})$ is the joint prior, while $\prod_{t=1}^{n} \mathbb{P}(Y_t|y_t, \phi)$ is the likelihood of the model. $\boldsymbol{R}$ and $\boldsymbol{\alpha}$ correspond to the vectors of values between times $t = 1$ and $t = n$, corresponding to the times at which observations Y_t are available.

In the most general case, `epidemia` models the reproduction number using a linear predictor η, then transformed with a link function g such that $R = g^{-1}(\eta)$ [54], where

$$\boldsymbol{\eta} = \boldsymbol{\beta}_0 + X\boldsymbol{\beta} + Z\boldsymbol{b} + Q\boldsymbol{\gamma}. \tag{4.8}$$



(a) Texas **(b)** California

(c) Nebraska **(d)** Ohio

Figure 4.1: Obtained county groupings for (a) Texas, (b) California, (c) Nebraska, and (d) Ohio. Notice that California consists mainly of single-county groupings, in contrast with Nebraska which consists of few groupings containing many counties. Plots for states closer to the East or West coast, where population density is high, will resemble California (b) the most, whereas other states in the Great Plains and Rocky Mountains regions will tend to display similar characteristics to Nebraska (c).

In (4.8), X is the design matrix for the fixed effect, Z is the model matrix for the random effects, and Q is the model matrix for the autocorrelation terms. As no covariates are used (which could for instance be non-pharmaceutical interventions or political events), and the model treats each county independently on its own rather than taking a multilevel approach, the model in fact reduces to

$$R_t = g^{-1}(\gamma_t) \quad [23]. \tag{4.9}$$

The stochastic process $\{\gamma_t\}$ is modelled as a random walk, that is $\gamma_s = \gamma_{s-1} + \epsilon_s$, where $\{\epsilon_t\}$ is a white-noise process. Specifically, the ϵ_t, referred to as the steps of the random walk, are taken to be Gaussian with zero mean, and unknown scale σ. This variance therefore needs to be assigned a prior to itself, which is taken to be half-normal prior with mean zero and variance 0.02^2 (the default value given to the `prior_scale` parameter of the `rw` function in `epidemia`). The index s is used rather than t in order to distinguish the time index taken in the random walk from the times of observations. The random walk is in fact modelled weekly (whereas observations are available daily). The link function is taken to be the default option in the `epirt` function, that is the the log link, with $g(R) = \log(R)$, or equivalently $g^{-1}(\eta) = \exp(\eta)$.

The next parameter of interest is the vector of seeded infections, $i_{v:0}$. In the package `epidemia` used for the implementation, the daily seeds are assumed to have constant value over the seeding period, i.e. $i_k \equiv i, k = v, ..., 0$ [54]. This means that only one parameter i needs its value to be inferred by the model, then giving a vector of repeated $v + 1$ repeated values that can be used in (4.6). The model for i is given as

$$i \sim \text{Exp}(\tau^{-1}) \tag{4.10a}$$
$$\tau \sim \text{Exp}(\lambda_0) \tag{4.10b}$$

The value of λ_0 is taken to be 0.03, the default value in the `epiinf` function. Once the seeded infections are obtained, the i_t for $t > 0$ are obtained deterministically, requiring the probability mass function of the time between infections. This is approximated by the serial interval distribution as presented in Bi et al. [58], that is the Gamma distribution with mean 6.3 days and standard deviation 4.2 days. Solving for the shape and rate parameter gives:

$$\begin{cases} \frac{\alpha}{\beta} = 6.3 \\ \frac{\alpha}{\beta^2} = 4.2^2 \end{cases} \implies \begin{cases} \alpha = \frac{9}{4} \\ \beta = \frac{5}{14} \end{cases}, \tag{4.11}$$

so $g \sim \text{Gamma}(\alpha, \beta)$. The values are encoded at discrete time points with non-zero values up to 100 days. These are displayed in comparison with the true density in Figure 4.2.

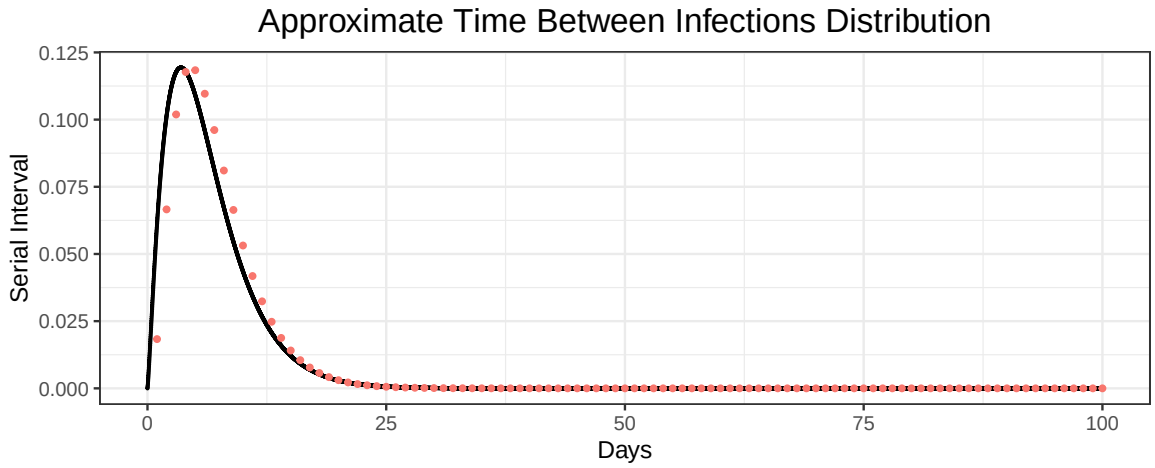


Figure 4.2: Approximated value of probability of infection at days 1 to 100. The solid black line corresponds to the probability density function of the Gamma distribution, and the dots correspond to the discrete probability values encoded in `EuropeCovid$si`. Lack of perfect correspondence can be explained by necessity to normalize the discrete data for summing to unity.

Finally, to conclude the specification of the model, the priors on α and ϕ need to be specified to complete the observations model. Taking Y_t to be death observations, the ascertainment rate α_t represents the infection fatality rate (IFR) [54]. It is modelled similarly to R_t as a transformed linear predictor, i.e. $\alpha = g^{-1}(\eta)$, where g is a link function. The linear predictor is taken to be an intercept only model. Following an example code, it the prior on the intercept parameter is taken to be $\mathcal{N}(1, 0.5)$.

The sampling distribution $\mathbb{P}(y_t, \phi)$ is taken to be from the negative binomial family, the default parameter in `epiobs`, which allows to account for overdispersion. This implies the auxiliary parameter ϕ is required. It is assigned the default prior $\mathcal{N}(10, 25)$. Furthermore,

the link function is also set to the default of a logit function, i.e. $g(\alpha) = \log\left(\frac{\alpha}{1-\alpha}\right)$, or equivalently $g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$.

The last quantity needing defining is the infection to death distribution. Similarly to the serial interval, it is obtained directly from `epidemia`, the exact distribution being provided in Flaxman et al. [3]. It is given as the sum of two Gamma distributions: the infection-to-onset distribution with mean 5.1 and coefficient of variation 0.86; and the onset-to-death distribution, with mean of 17.8 and coefficient of variation 0.45. The shape and rate of the first one are:

$$\begin{cases} \frac{\alpha_1}{\beta_1} = 5.1 \\ \frac{\alpha_1}{\beta_1^2} = (5.1 \times 0.86)^2 \end{cases} \implies \begin{cases} \alpha_1 \approx 1.4 \\ \beta_1 \approx 0.3 \end{cases}, \tag{4.12}$$

and for the second one are:

$$\begin{cases} \frac{\alpha_2}{\beta_2} = 17.8 \\ \frac{\alpha_2}{\beta_2^2} = (17.8 \times 0.45)^2 \end{cases} \implies \begin{cases} \alpha_2 \approx 4.9 \\ \beta_2 \approx 0.3 \end{cases}. \tag{4.13}$$

As the rate is approximately the same, the infection to death distribution is approximately Gamma($\alpha_1 + \alpha_2$,0.3). The discrete values at 101 data points are plotted in comparison with this density in Figure 4.3.

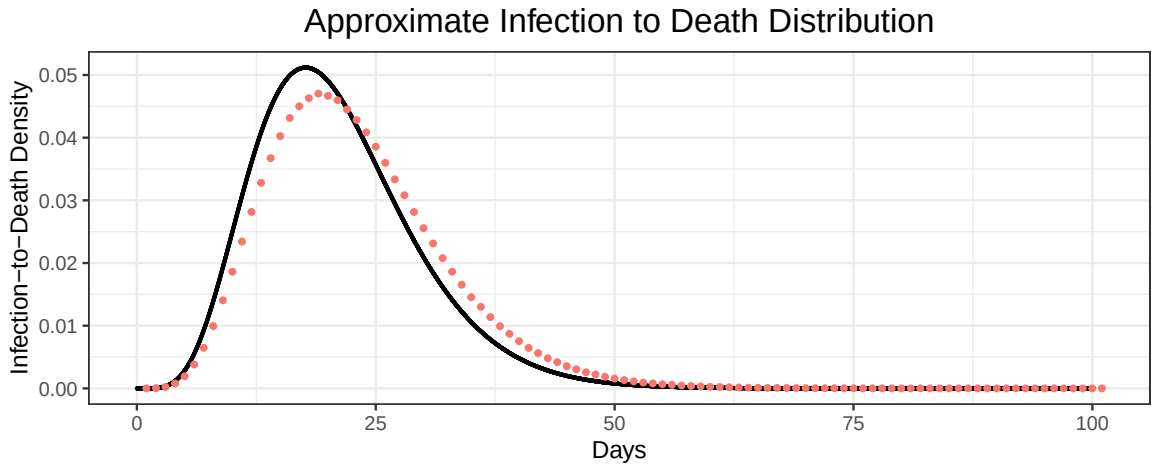


Figure 4.3: Approximated value of the infection-to-death distribution at days 1 to 101. The solid line represents the gamma probability density function, and the dots the discrete probability values. The match is not exact due to approximations in computations, but shows a peak at around 20 days, which would be the mean time from infection to death.

This concludes the specifications of the model, with all quantities defined and attributed prior distributions. The next step is inference.

Hamiltonian Monte Carlo

It can be noticed that combining the multiplying the priors in (4.7) will lead to an analytically intractable posterior distribution, and Markov Chain Monte Carlo (MCMC) methods are therefore required.

MCMC methods are based on the concept of Markov chains, a type of stochastic process

which satisfies the Markov property. That is, the sequence of random variables $X_1, X_2, ...$ is a Markov chain if it satisfies [59]

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, ..., x_1) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}). \tag{4.14}$$

The idea is to sample from the posterior distribution by drawing from an approximate distribution which depends on the previous value in the chain [59]. It belongs to the more general group of Monte Carlo methods, that is statistical techniques that rely on the collection of random samples for approximating numerical results [60].

There exist numerous MCMC algorithms and methods used in Bayesian inference of parameters. In this work, the models are fit using `epidemia` [24], which uses the Hamiltonian Monte Carlo (HMC) algorithm. Indeed, the model fitting function `epim` uses the probabilistic programming language Stan [61], which in fact runs the No-U-Turn Sampler, an extension to HMC [62].

Before introducing the more complex, however more computationally efficient, Hamiltonian Monte Carlo algorithm, it is important to present the Metropolis Algorithm as its building block, introduced by Metropolis et al. in 1953 [63].

Denoting the vector of parameters needing to be inferred as $\boldsymbol{\theta} = (i_{v:0}, \boldsymbol{R}, \phi, \boldsymbol{\alpha})$, and writing the posterior (referred to as the target distribution in the MCMC setting) as $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{Y})$, the Metropolis algorithm can be written as follows [59].

Algorithm 2: Metropolis Algorithm

Initialize: Sample $\boldsymbol{\theta}^0$ belonging to the support of $\mathbb{P}(\cdot|\boldsymbol{Y})$ from a starting distribution, which can be selected in different ways. Set n =1.

for $n = 1$ *to* M **do**

- Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(n-1)}$;
- Generate a candidate value $\boldsymbol{\theta}^*$ from a symmetric proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$;
- Compute the ratio $r \leftarrow \frac{\mathbb{P}(\boldsymbol{\theta}^*|\boldsymbol{Y})}{\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{Y})}$;
- Generate $U \leftarrow u \sim U(0, 1)$;
- **if** $U \leq \min(1, r)$ **then**
  - Set $\boldsymbol{\theta}^{(n)} \leftarrow \boldsymbol{\theta}^*$;
- **else**
  - Set $\boldsymbol{\theta}^{(n)} \leftarrow \boldsymbol{\theta}$;
- **end**

end

It can be noted that the proposal of candidate values is in fact a random walk. This therefore leads to the issue of an inefficient exploration of a complex multivariate posterior [59]. Hamiltonian Monte Carlo is a method that allows to avoid these behaviors.

HMC is based on the physical concept of Hamiltonian dynamics. To each of the d parameters of the model (which can be considered to be position variables in a physical setting) is associated what is called a momentum variable, which shall be denoted as r_i, for $i = 1, ..., d$ [64]. In the Stan implementation, the momentum variables are drawn independently from the standard normal distribution [62] (although it can be generalized to a multivariate normal distribution with covariance matrix different from identity). One can then obtain the

Hamiltonian function from the joint density $\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{r}|\boldsymbol{Y})$, defined as follows [65]

$$
\begin{aligned}
H(\boldsymbol{\theta}, \boldsymbol{r}) &= -\log[\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{r}|\boldsymbol{Y})] && \text{(4.15a)}\\
&= -\log[\mathbb{P}(\boldsymbol{r}|\boldsymbol{\theta})] - \log[\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{Y})] && \text{(4.15b)}\\
&= -\log[\mathbb{P}(\boldsymbol{r})] - \mathcal{L}(\boldsymbol{\theta}), \qquad \mathcal{L}(\boldsymbol{\theta}) \equiv \log[\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{Y})] && \text{(4.15c)}\\
&\propto -\frac{1}{2}\boldsymbol{r}^T\boldsymbol{r} + \mathcal{L}(\boldsymbol{\theta}) && \text{(4.15d)}
\end{aligned}
$$

The system consisting of the position and momentum variables can then be described by what are known as Hamilton's equations [64]:

$$
\begin{aligned}
\frac{d\boldsymbol{\theta}}{dt} &= \frac{\partial H}{\partial \boldsymbol{r}} && \text{(4.16a)}\\
\frac{d\boldsymbol{r}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} && \text{(4.16b)}
\end{aligned}
$$

To approximate the solution to Hamilton's equations, a discretization of time is needed, using small stepsizes ϵ. Then the method that is used is called the leapfrog method, which makes half steps for the momentum variables and full steps for the position variables [64]. The method is defined by the following equations:

$$
\begin{aligned}
\boldsymbol{r}^{t+\epsilon/2} &= \boldsymbol{r}^t + \frac{\epsilon}{2}\Delta_\theta \mathcal{L}(\boldsymbol{\theta}^t) && \text{(4.17a)}\\
\boldsymbol{\theta}^{t+\epsilon} &= \boldsymbol{\theta}^t + \frac{\epsilon}{2}\boldsymbol{r}^{t+\epsilon/2} && \text{(4.17b)}\\
\boldsymbol{r}^{t+\epsilon} &= \boldsymbol{r}^{t+\epsilon/2} + \frac{\epsilon}{2}\Delta_\theta \mathcal{L}(\boldsymbol{\theta}^{t+\epsilon}) && \text{(4.17c)}
\end{aligned}
$$

For a step size of ϵ, the algorithm for the method is as given below [62].

Algorithm 3: Leapfrog method

Input : $\boldsymbol{\theta}$, $\boldsymbol{r}$, ϵ
Output : $\boldsymbol{\theta}^*, \boldsymbol{r}^*$
Set $\boldsymbol{r}^* \leftarrow \boldsymbol{r} + (\epsilon/2)\Delta_\theta \mathcal{L}(\boldsymbol{\theta})$
Set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta} + \epsilon \boldsymbol{r}$
Set $\boldsymbol{r}^* \leftarrow \boldsymbol{r}^* + (\epsilon/2)\Delta_\theta \mathcal{L}(\boldsymbol{\theta}^*)$

The leapfrog method is applied L times, which corresponds to the number of steps. The final $\boldsymbol{\theta}^*$ and $\boldsymbol{r}^*$ are then equivalent to what was given as the candidate values resulting from the proposal distribution in the Metropolis Algorithm. Hamiltonian Monte Carlo can then be seen as a transformed Metropolis Algorithm using these candidate values and with ratio r involved in the acceptance probability given by [62]:

$$
r = \frac{\mathbb{P}(\boldsymbol{\theta}^*, \boldsymbol{r}^*|\boldsymbol{Y})}{\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{r} * |\boldsymbol{Y})} = \frac{\exp(\mathcal{L}(\boldsymbol{\theta}^*) - \frac{1}{2}\boldsymbol{r}^{*T}\boldsymbol{r}^*)}{\exp(\mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{r}^T\boldsymbol{r})}. \tag{4.18}
$$

The algorithm for Hamiltonian Monte Carlo is then as provided in Algorithm 4 [62]. It can be shown that HMC converges to the posterior distribution, the proof is given by Neal [64].

In comparison with the Metropolis Algorithm, the Hamiltonian Monte Carlo algorithm involves two parameters that very clearly need tuning, ϵ and L (note that in fact the Metropolis Algorithm also has a parameter needing tuning, the standard deviation of the proposal

Algorithm 4: Hamiltonian Monte Carlo Algorithm

Input : ϵ, L

Initialize: Sample $\boldsymbol{\theta}^0$ belonging to the support of $\mathbb{P}(\cdot|\boldsymbol{Y})$ from a starting distribution, which can be selected in different ways. Set n =1.

for $n = 1$ *to* M **do**

- Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(n-1)}$;
- Sample $\boldsymbol{r} \sim \mathcal{N}(0, I)$;
- Set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}, \boldsymbol{r}^* \leftarrow \boldsymbol{r}$;
- **for** $i = 1$ *to* L **do**
  - Set $\boldsymbol{\theta}^*, \boldsymbol{r}^* \leftarrow \text{Leapfrog}(\boldsymbol{\theta}^*, \boldsymbol{r}^*, \epsilon)$
- **end**
- Compute the ratio $r \leftarrow \frac{\exp(\mathcal{L}(\boldsymbol{\theta}^*) - \frac{1}{2}\boldsymbol{r}^{*T}\boldsymbol{r}^*)}{\exp(\mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{r}^T\boldsymbol{r})}$;
- Generate $U \leftarrow u \sim U(0, 1)$;
- **if** $U \leq \min(1, r)$ **then**
  - Set $\boldsymbol{\theta}^{(n)} \leftarrow \boldsymbol{\theta}^*$;
- **else**
  - Set $\boldsymbol{\theta}^{(n)} \leftarrow \boldsymbol{\theta}$;
- **end**

end

density which has a direct influence on the acceptance probability). This parameter selection is in fact key for satisfactory performance of the sampler. A very low ϵ leads to an acceptance probability of the proposed states that is also very low; and an ϵ that is too high can both waste computational time and lead to behavior ressembling the random walk from the Metropolis algorithm HMC aims to eliminate [64]. Regarding the second parameter, which can be seen as the trajectory length in the physical setting, a too low value of L gives a similar result to a too large ϵ, i.e. random walk behavior, whereas a too large value generates trajectories that loop back and repeat their steps, which is wasteful [62]. Fortunately, the HMC algorithm implemented in Stan makes uses of an adaptation called the No-U-Turn-Sampler (NUTS), that doesn't require manual tuning of these parameters [62].

The specifics concerning the working of the NUTS algorithm are not of primary focus to this thesis, the details are given by Hoffman and Gelman [62]. However, it is worth noting that ϵ is tuned during a warm-up phase, to match a target acceptance rate [66] (the Stan default value of 0.8 is used in the `epidemia` implementation).This fixed value is then used for all iterations. On the other hand, L varies at each transition, depending on the value of the momentum and position vectors, in order to satisfy certain conditions regarding the trajectory, chosen through a tree building algorithm [66].

For the implementation, I run 4 chains for each model fit. This choice is motivated by the work of Gelman and Rubin[67], which gives two main advantages to running several independent sequences. It allows to take advantage of the variability in the starting distribution, and improves monitoring of convergence through easier estimation of variability. Each chain runs for 2,500 iterations, a value obtained through trial and error, that is large enough for most chains to have time to converge while not requiring an exorbitant computation time and memory.

4.2 Regressing R_t on Trump Support and Vaccination Hesitancy Rates

The main aim of this project is to provide results and an analysis on the county level. However, due to limited availability of quality data, I started with performing analysis on the state level. I obtained R_t estimates using state COVID-19 deaths data and implementing the methods described in section 4.1. This then allowed me to use election results forecasts available daily starting from June 1^{st}, and use the Household Pulse Survey vaccination hesitancy data [48] (as opposed to the county level which doesn't have directly surveyed data available, it is only obtained from the state values).

4.2.1 Bias Correction in State-Level Forecasts

It has been found by Panagopoulos [68] that in 2020, most preelection polls have underestimated Trump support, leading to statistically significant biases. In fact, this is not a one-off occurence: in the same paper, the author shows that an overall pro-Democratic bias exists in polling data in the elections from 1996 to 2020, where only the 2000 and 2012 elections show a bias towards Republicans. Skelley [69] proposes possible explanations why the 2020 election was particularly affected by such bias. The pandemic context appears to play an important role in over-representation of Democrats in polls, who may have become more likely to answer pollsters as a result of staying at home and strong anti-Trump feelings. Furthermore, Skelley [69] advances that those who feel most alienated are less likely to respond to polls, as well being more likely to support a Republican candidate.

Now looking specifically at the state-level data from FiveThirtyEight [51] that I shall be using, these pro-Democrat biases can very clearly be noticed. Indeed, looking at the difference between each state's Trump support in the November 3^{rd} election and the mean of the forecasts taken daily between June 1^{st} and November 3^{rd}, the national average is $+1.229\%$ for Joe Biden. Perhaps even more surprisingly, the average difference between the official election results and the forecasted value on November 3^{rd} is $+1.899\%$ for Joe Biden. The values for each state can be found in table A.1.

As this election forecast data is required for regression at different time points, I perform bias correction before the model fitting step. This is in order to ensure consistency between the Trump electoral support values used at dates prior to the election, and dates after November 3^{rd}.

Several methods for correcting bias can be considered, and given the absence of a benchmark against which to compare the corrected values, the choice has to be made qualitatively. The first element of consideration is whether to treat the bias on the entire period of forecasts, or only at a specific date. It would appear most reasonable to treat the bias between the forecasts on November 3^{rd} and the election results. Indeed, since these correspond to the same date, it naturally follows that the Trump support values should match. Since the daily forecast values show that Trump support varies from day to day in each state, the alternative of trying to match the average value over the 5 month period from June 1^{st} to November 3^{rd} would be rather counter-intuitive. Therefore the bias correction will aim to match the forecast from November 3^{rd} to the election results.

The next choice that has to be made is whether the correction is applied to all states identi-

cally, or each state is treated separately. Seeing in table A.1 that the differences range from -1.481% in Maryland to +5.488% in North Dakota, the evidence would imply that the biases are in fact sufficiently different in each state to justify individual treatment. Finally, the method of correction needs to be selected. Although this correction is useful for consistency, it is not the key element of analysis and as long as relative Trump support across states is adequate, the final analysis will not be severely impacted. Therefore, I do not consider advanced methods and instead only consider either an additive or multiplicative approach. The former would consisting in simply adding the bias from November 3^{rd} to all the forecast values, and the latter would consist in multiplying all the forecasted values by

$$\frac{\text{Support on November } 4^{\text{th}}}{\text{Support on November } 3^{\text{rd}}}.$$

Both these approaches set the values at these two dates as equal. Given the earlier exposed possible explanations for the bias, the multiplicative method is better justified. Indeed, these biases imply an inherent disproportionality in the population sample, and thus the bias should scale multiplicatively with support.

Applying this multiplicative correction, the new average difference between election results and the mean forecasts over the 5 month period is -0.704%, that is after correction the average forecasts give slightly lower Biden support than was achieved in the elections. Treating these new forecasted values as estimates of Trump Support on a given day, this would mean that overall Trump support decreases in the months leading up to the election.

4.2.2 Bayesian Linear Regression

The final step is the analysis of the relationship between the three quantities of interest, that is to say the reproduction number R_t, the vaccination hesitancy rates, and support for Trump in the 2020 elections. More particularly, I focus on how good predictors are the latter two of R_t at a given time point.

I fit various models both on the state and on the county level, considering different time periods and exploring different predictors and combinations of the predictors. The exact specifications of the models that are most meaningful will be presented and described in the results section, however I first introduce the general method that I perform modelling with: Bayesian Linear Regression.

A linear model is defined as

$$y_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, ..., K; \tag{4.19}$$

where K is the number of data points, $\boldsymbol{Y}$ is the vector of outcomes (the values of R_t in the models I fit), α is an intercept term, $\boldsymbol{x}_i$ are the N predictors for observation i (which can be represented in the design matrix X), and ϵ_i are noise terms taken to be normally distributed with variance σ_ϵ^2. It can be noted that the intercept term can be, and is often, included in the coefficient vector $\boldsymbol{\beta}$, with a corresponding column of ones in the design matrix. However, since the Bayesian linear model will be implemented using the package `rstanarm` [70], distinguishing the intercept is important as it will be assigned a separate prior.

In the ordinary least squares setting, the equations that provide parameter values which minimize the sum-of-squared residuals are the following [71]

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \tag{4.20a}$$

$$\hat{\alpha} = \bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}^T \hat{\boldsymbol{\beta}} \tag{4.20b}$$

$$\hat{\sigma}_\epsilon^2 = \frac{(\boldsymbol{y} - \hat{\alpha} - X\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \hat{\alpha} - X\hat{\boldsymbol{\beta}})}{N} \tag{4.20c}$$

Both the `lm` function from `R`, and the `stan_lm` function from `rstanarm` that I use in fact perform a QR decomposition on the design matrix X [71]. In fact, Q is an orthogonal matrix constructed from the design matrix with added column of 1s for the intercept, and then the first column is removed to keep α separate (indeed this doesn't affect the intercept term as the first column of Q will consist of K entries with value $1/K$, which then add back up to α). This method allows to have column means of zero, as to ensure orthogonality with the first column of Q (then removed), the column sums need to be zero. This then gives [71]

$$\hat{\boldsymbol{\beta}} = R^{-1} Q^T \boldsymbol{y} = R^{-1} \boldsymbol{\theta}, \qquad \boldsymbol{\theta} \equiv Q^T \boldsymbol{y} \tag{4.21}$$

Now, since this is implemented in a Bayesian setting, these are not computed directly but instead, parameters are assigned priors. In fact the parameters that are assigned priors are ω,α, $\boldsymbol{u}$ and R^2 [71], therefore giving the posterior

$$\mathbb{P}(\omega, \alpha, \boldsymbol{u}, R^2 | \boldsymbol{y}, X) \propto \mathbb{P}(\boldsymbol{y} | \omega, \alpha, \boldsymbol{u}, R^2, X) \mathbb{P}(\omega) \mathbb{P}(\alpha) \mathbb{P}(\boldsymbol{u}) \mathbb{P}(R^2), \tag{4.22}$$

and the quantities of interest are then connected to these parameters by the equations [71]

$$\sigma_Y = \omega s_Y \quad (s_Y \text{ is the sample standard deviation of } \boldsymbol{y}) \tag{4.23a}$$

$$\sigma_\epsilon = \sigma_y \sqrt{1 - R^2} \tag{4.23b}$$

$$\boldsymbol{\beta} = R^{-1} \boldsymbol{u} \sigma_Y \sqrt{R^2 (N-1)} \tag{4.23c}$$

It is also worth noting that the likelihood is normally distributed with mean $\boldsymbol{\alpha} + \boldsymbol{x}^T \boldsymbol{\beta}$, which follows directly from the Gaussian noise.

I now provide explanations about the different parameters in the model. Firstly, it is given by Gabry and Goodrich [71] that the k^{th} element of $\boldsymbol{\theta}$ can be written as

$$\theta_k = \rho_k \sigma_Y \sqrt{N-1}, \tag{4.24}$$

where ρ_k is the correlation between the k^{th} column of Q and $\boldsymbol{y}$ and σ_Y is the variance of $\boldsymbol{Y}$. I start by proving this equation. Letting $q^{(k)}$ be the k^{th} column of Q (or alternatively the k^{th} of Q^T), the left hand side is:

$$\theta^k = \sum_{i=1}^{N} q_i^{(k)} Y_i. \tag{4.25}$$

and the right hand side,

$$\rho_k \sigma_Y \sqrt{N-1} = \frac{\sum_{i=1}^N (q_i^{(k)} - \bar{q}^{(k)})(Y_i - \bar{y})}{(N-1)\sigma_Y \sigma_{q^{(k)}}} \sigma_Y \sqrt{N-1} \tag{4.26a}$$

$$= \frac{\sum_{i=1}^N (q_i^{(k)} - \bar{q}^{(k)})(Y_i - \bar{y})}{\sqrt{N-1}\sigma_{q^{(k)}}} \tag{4.26b}$$

Noting that Q has zero mean columns, and that, due to orthogonality of the columns, the sum of squares must be 1, implying that $\sigma_{q^{(k)}} = 1/\sqrt{N-1}$, Equation 4.26b thus equals:

$$= \frac{\sum_{i=1}^{N}(q_i^{(k)})(Y_i - \bar{y})}{\sqrt{N-1}1/\sqrt{N-1}} \tag{4.27a}$$

$$= \sum_{i=1}^{N} q_i^{(k)} Y_i = \theta_k \tag{4.27b}$$

This proves that Equation 4.24 holds. Then Gabry and Goodrich [71] define $\boldsymbol{\rho} = \sqrt{R^2}\boldsymbol{u}$, where $\boldsymbol{u}$ is a unit vector uniformly distributed on the surface of a hypersphere and $R^2 = \boldsymbol{\rho}^T\boldsymbol{\rho}$ can be recognized as the coefficient of determination. Although in this setting R^2 will not be computed in practice as it is simply inferred, it is worth noting its definition, as it is used as a goodness-of-fit measure in the linear models. In a Bayesian setting, where S simulations have been performed, the proportion of variance explained at simulation s is given by [72]

$$R_s^2 = \frac{\text{Var}(\sum_{i=1}^{K}\hat{y}_i^s)}{\text{Var}(\sum_{i=1}^{K}\hat{y}_i^s) + \text{Var}(\sum_{i=1}^{K}\hat{e}_i)}, \tag{4.28}$$

where $\hat{\boldsymbol{y}}^s$ corresponds to the predicted values with the inferred parameters from iteration s and $\boldsymbol{e}^s = \boldsymbol{y} - \hat{\boldsymbol{y}}^s$. R^2 can then be summarized by the posterior median of the S R_s^2 values (or another summary).

Returning to the inference problem, the prior on R^2 is given by a Beta$(\frac{K}{2}, \eta)$ distribution [71]. Due to a lack of a prior information, I set the location parameters value to be $\eta = 0.5$, which is given to be a safe choice.

Finally, ω is defined via an improper uniform prior on $\log\omega$, as is α (note that this can be modified but I chose to adopt the default in `rstanarm`). The parameter estimates are then returned according to (4.23a), (4.23b) and (4.23c), once again using HMC, which is the algorithm `rstararm` relies on.

Chapter 5

Results

All the results presented in this section have been coded in `R`. The full data and code are available at `https://github.com/emma-landry/M4R`.

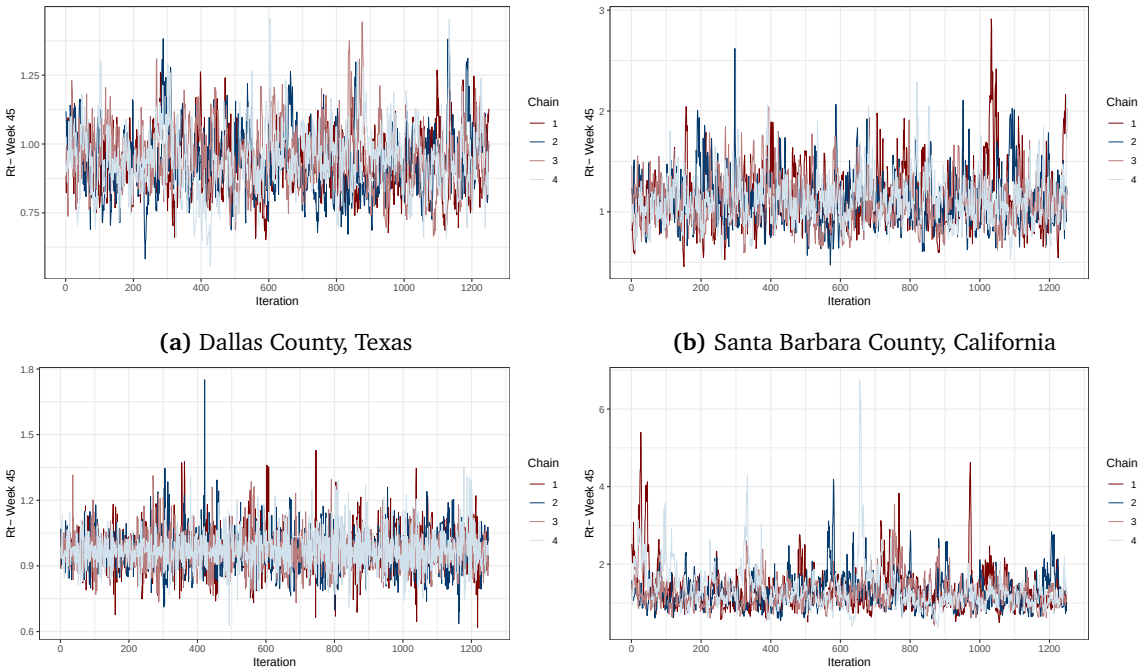
5.1 Reproduction Number and Infections

In this first section I present some summaries of the model fits obtained at the state and county levels with `epidemia` [24]. The code, which can be found as listing B.2 in the appendix, was run using Imperial College's High Performance Computers. 48 array jobs were submitted so that each state could be run separately. Furthermore, 4 CPUs were requested for each job, to allow simultaneous running of the 4 chains. The PBS file is presented in the appendix as listing B.3. Note that Texas is the state consisting of the highest number of county groupings (55), and since the counties are modelled one after the other, the walltime request was 72h.

Due to the large number of models fit (48 and 796 at the state and county levels respectively), and to the large number of parameters in the model (this can vary from model to model depending on the week modelling starts, but roughly 65), it is naturally impossible to show the results for all of them. Therefore, this section mainly focuses on week 45, which is the week during which the U.S. Presidential election took place. For county-specific visualizations, 4 are selected: Dallas County, Texas; Santa Barbara County, California; a 9-county grouping in Nebraska; and 3-county grouping in Ohio. The former grouping consists of Richardson, Nemaha, Pawnee, Johnson, Otoe, Gage, Dodge, Sauners, and Cass Counties; while the latter consists of Ottawa, Sandusky and Wood Counties. Note that although this choice of 2 counties and 2 groups of counties out of a total of 796 groups of counties may appear arbitrary, it allows to consider both single counties as well as groups, and provide an overview covering distinct geographical regions of the United States.

5.1.1 Traceplots

The first key step in the analysis is examining convergence of the chains. I start with a qualitative representation using traceplots. These provide the iteration on the x-axis and the parameter value at the corresponding iteration on the y-axis. The traceplots are obtained using the `R` `bayesplot` package [73], which I then further personalize using functionalities from the `ggplot2` package [74].



(a) Dallas County, Texas

(b) Santa Barbara County, California

(c) Nebraska counties

(d) Ohio counties

Figure 5.1: Week 45 R_t traceplots for (a) Dallas, Texas, (b) Santa Barbara, California, (c) Nebraska counties, and (d) Ohio counties. Figure is obtained using `bayesplot` [73], and represents the 1,250 iterations after the burn-in has been discarded (1,250 iterations).

Figure 5.1 represents the traceplots for the previously defined counties/ groups considered for county-level visualizations. All 4 of them appear to have explored the state space well, and appear to have converged. Traceplot (c) corresponding to the Nebraska counties is however the one exhibiting the best behavior, and the chains would need to be run for a large number of iterations to consistently obtain such results with all counties.

Traceplots were also obtained for the 50 states, and are shown in Figure A.2 in the Supplement. A large number of those show chains that do not appear to have converged, with chains that do not appear to have explored the state space and seem to have dynamics still dependent on the initial values. These poor results are particularly flagrant for states such as Indiana, New York, Kansas or Virginia. These results are discussed in more detail in the next chapter.

5.1.2 Convergence analysis

After a qualitative evaluation of convergence of the chains, I move on to a quantitative assessment, using two tools: the Effective Sample Size (ESS) and the $\hat{R}$ statistic. Both of these values are returned for Stan [61] objects, and are parameter-specific. They are defined as follows.

Effective Sample Size

Letting N bet the length of the sample (i.e. the length of the chain after burn-in is discarded), and denoting the lag τ autocorrelation by ρ_τ, the effective sample size is given by [75]

$$N_{\text{eff}} = \frac{N}{\sum_{\tau=-\infty}^{\infty} \rho_\tau} = \frac{N}{1 + 2\sum_{\tau=1}^{\infty} \rho_\tau}. \tag{5.1}$$

The autocorrelation sequence elements can however not be evaluated exactly, and therefore need to be estimated. Supposing that M chains have been run, each for N iterations; and denoting

$$\begin{aligned}
\hat{\rho}_{\tau,m} &\quad \text{: lag } \tau \text{ correlation estimate in chain } m \\
s_m^2 = \frac{1}{N}\sum_{n=1}^{M}\left(\theta_m^{(n)} - \frac{1}{N}\sum_{n'=1}^{M}\theta_m^{(n')}\right)^2 &\quad \text{: the chain } m \text{ sample variance} \\
W = \frac{1}{M}\sum_{m=1}^{M} s_m^2 &\quad \text{: the average within-sample variance} \\
B = \frac{N}{M-1}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{n=1}^{N}\theta_m^{(n)} - \frac{1}{MN}\sum_{m'=1}^{M}\sum_{n=1}^{N}\theta_{m'}^{(n)}\right)^2 &\quad \text{: the between-chain varaince} \\
\widehat{\text{var}}^{+} = \frac{N-1}{N}W + \frac{1}{N}B &\quad \text{: the multi-chain variance estimate;}
\end{aligned}$$

the combined autocorrelation is defined to be [75]

$$\hat{\rho}_\tau = 1 - \frac{W - \frac{1}{M}\sum_{m=1}^{M} s_m^2 \hat{\rho}_{\tau,m}}{\widehat{\text{var}}^{+}}. \tag{5.2}$$

The effective sample size estimated and returned for a Stan object is then [75]

$$\hat{N}_{\text{eff}} = \frac{M \cdot N}{1 + 2\sum_{\tau=1}^{\infty} \hat{\rho}_\tau} \tag{5.3}$$

This value is significant in that in can be interpreted as the number of independent samples in the chain, and lower values will lead to an increase in uncertainty of estimation of the posterior quantities (e.g. the medians or the quantiles) [75].

$\hat{R}$ Statistic

The potential scale reduction statistic, denoted by $\hat{R}$, is defined using the notation introduced above as [75]

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}}{W}}. \tag{5.4}$$

This measure gives the average ratio of the pooled variance across all chains to the average variance of samples from each chain, which implies that if the chains have converged to a common distribution, these two values should be the same and thus the value of $\hat{R}$ should be 1 [75].

The ESS and $\hat{R}$ values are thus returned for the counties whose traceplots were observed in the previous section, and reported in table 5.1. The obtained vales are consistent with the comments made regarding the traceplots. Indeed, per the guidance of Guo et al. [76], samples are suggested to be used for $\hat{R}$ values less than 1.05, and ESS values greater than 100, to ensure sufficient convergence and more reliable estimates (note that those values are simply rough approximations). All the values in the table show good convergence, and in particular the Nebraska with the $\hat{R}$ value of 1.000, which was previously noticed to have the highest quality traceplot.

County	ESS	Rhat
Santa Barbara County, CA	669	1.004
Dallas County, TX	541	1.005
Nebraska counties	5359	1.000
Ohio counties	264	1.013

Table 5.1: Summary of convergence statistics for 4 counties in California, Texas, Nebraska, and Ohio.

It is also worth analysing convergence of the chains corresponding to the remaining R_t values, not just the ones corresponding to the week of election day. Figure 5.2 represents the weekly time series of the corresponding $\hat{R}$ statistic, for the same four counties studied until now. It shows consistent excellent convergence for the Nebraska grouping, and overall no concerning results for any of the other counties as peaks reaching values greater than 1.02 are rare, and at no time do the values approach the approximate threshold of 1.05.

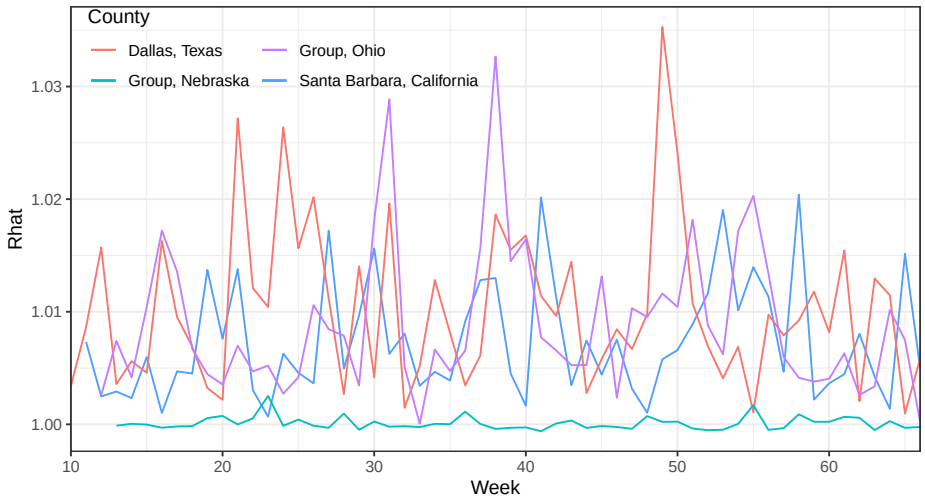
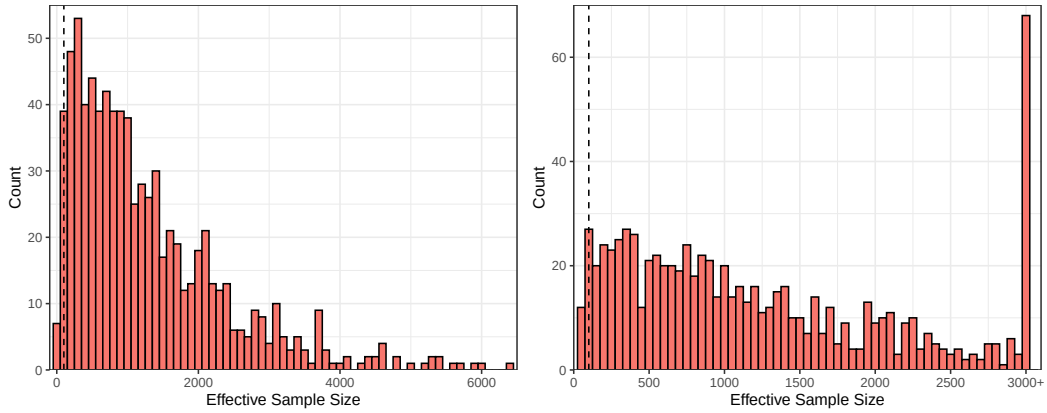


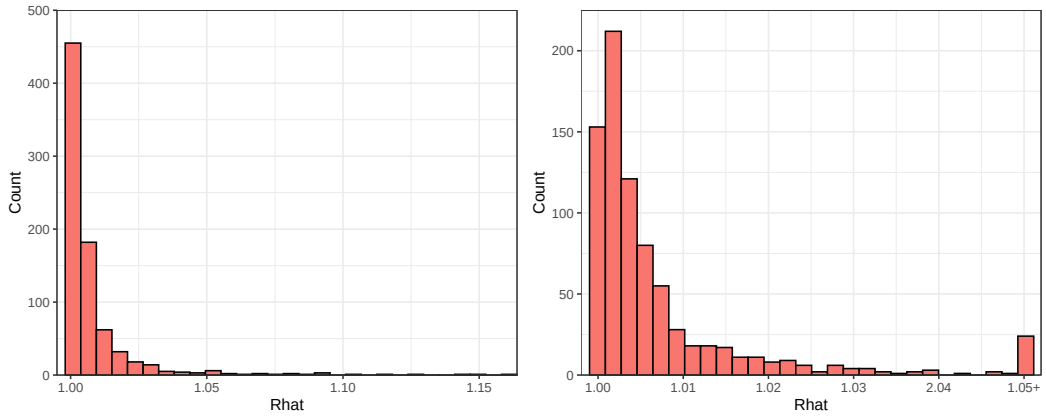
Figure 5.2: $\hat{R}$ values for weekly R_t chains. Each of the 4 counties is represented by a separate line, with varying start date and ending on week 66.

Before moving on to an overview of the state-level convergence, I also illustre the convergence diagnostic statistics for all counties, on the week of November 3^{rd}. Figure 5.3 displays these values as a histogram. Noting that the right-hand side ESS histogram has binwidth of 50, the total number of counties having an ESS less than 100 is approximately 40. Furthermore, roughly 25 out of the 796 county groupings have an $\hat{R}$ greater or equal to 1.05. This

therefore shows that not all counties have converged, and results later obtained using R_t posterior medians may therefore not be truly accurate, however it is reassuring to see that only a small proportion of the counties is affected by such convergence issues.



(a) Effective Sample Size



(b) $\hat{R}$

Figure 5.3: County-level histograms representing convergence on week 45 . (a) represents the effective sample size distribution for the 796 county groupings and (b) the $\hat{R}$ distribution. The left panels represent the distribution exactly, whereas the right groups the smaller bins at greater x values.

Now moving on to the state-level, table A.2 in the supplement provides the values of $\hat{R}$ and of the ESS. Unsurprisingly, the states which showed poor convergence in their traceplots some of the lowest ESS values and highest $\hat{R}$. I briefly explore the effects of an increased number of iterations for the states already mentioned previously that stood out by their poor convergence, that is New York, Indiana, Kansas, and Virginia. A new model is run on HPC for each of these 4 states, which generates chains of 20,000 thousand iterations rather than 2,500 as done initially.

Table 5.2 shows the values of the convergence statistics for the two runs, and the improvement in convergence is very apparent. Indeed, all the ESS values values go from being in the range of 20 to getting over the 100 approximate threshold, and similarly all $\hat{R}$ values drop below the value of 1.05. It is however worth noting that although running the chains for 20,000 iterations leads to significant improvement, it may be worth to choose an even larger value for guaranteed convergence of all states at all dates. As the main issue with lack

State	First run		Second run	
	ESS	$\hat{R}$	ESS	$\hat{R}$
New York	20	1.147	304	1.014
Indiana	18	1.277	251	1.007
Kansas	16	1.364	198	1.022
Virginia	14	1.267	136	1.048

Table 5.2: Comparison of convergence for 2,500 and 20,000 iterations. The states of interest are New York, Indiana, Kansas, and Virginia, which previously showed poorest convergence, and the results are given regarding the chains for R_t on week 45.

of convergence is inaccurate posterior estimates, I also provide in table 5.3 the 2.5 %, 50 % and 97.5% posterior quantiles for the week 45 R_t chains.

State	First run			Second run		
	2.5%	50%	97.5%	2.5%	50%	97.5%
New York	0.490	0.993	2.156	0.433	0.888	1.827
Indiana	0.492	0.968	1.972	0.426	1.019	2.534
Kansas	0.273	1.089	2.543	0.324	0.975	2.892
Virginia	0.151	0.445	1.361	0.178	0.596	1.689

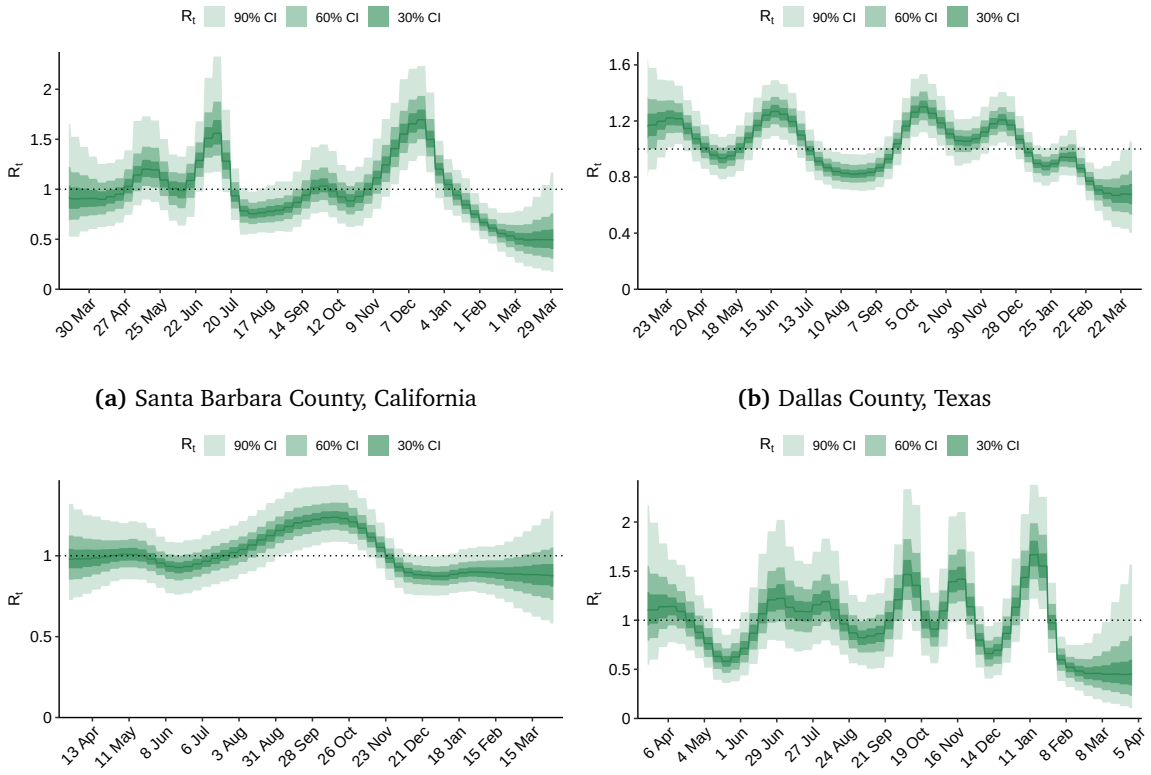
Table 5.3: Week 45 R_t posterior quantiles with 2,500 and 20,000 iterations. The values are obtained for the same four states as above, which displayed poorest convergence with 2,500 iterations.

Table 5.3 shows that the 97.5% percentile is the most affected by poor convergence, with values that vary greatly between the run of 2,500 iterations and the run of of 20,000 iterations. Reassuringly this difference is less pronounced for the median, which is the quantity of interest in later analysis. However, it is worth noting the issue when R_t is close to 1. The lack of convergence may lead a value that should be slightly greater than 1 to be given as less than 1 (or vice-versa), which may numerically not appear of great importance, but epidemiologically speaking, is very meaningful. Following these results, it would therefore be recommended to run all the models with this higher number of iterations (both on the state-level and the county-level), however due to the time-constraints of this thesis this is not feasible. Later results must as a consequence be considered (particularly) critically, with the drawbacks of lower numbers of iterations kept in mind.

5.1.3 R_t, infections and deaths

Before moving on to the regression task, I present certain results that allow to illustrate the workings of the model in the broader epidemiological context.

Figure 5.4 gives a visualization of the R_t values from the start of the pandemic, showing the 90%, 60% and 30% credible intervals. The weekly random walk modelling of R_t is visible in the discontinuities, however the plots give rather smooth trends. Although the range of values attained differs, it is interesting to notice that for Dallas County, Santa Barbara County, and the Ohio counties, all displays peaks and troughs at similar dates. Furthermore, these three counties display values of R_t that can go up quite high (perhaps slightly less so for Dallas County). The Nebraska counties displays a different trend, with values of R_t that oscillate less, and that remain closer to 1 during the whole period of study. There is however



(a) Santa Barbara County, California

(b) Dallas County, Texas

(c) 9-county grouping, Nebraska

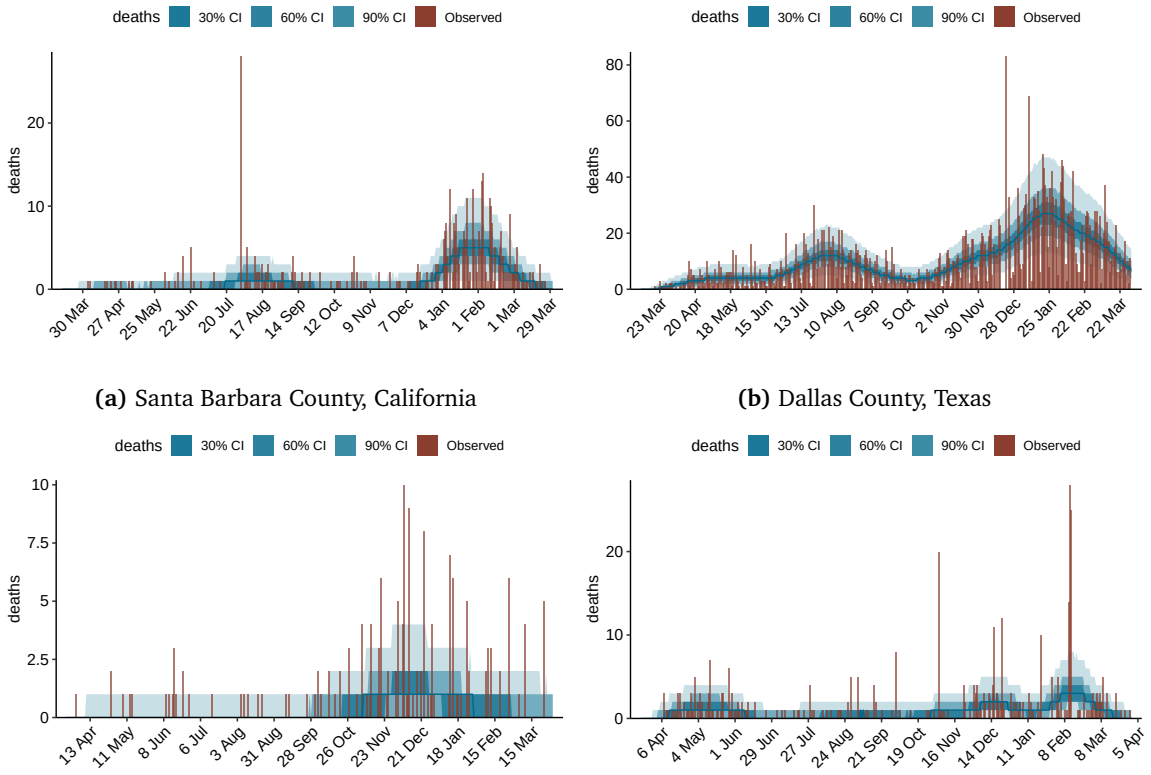
(d) 3-county grouping, Ohio

Figure 5.4: R_t times series with CIs for (a) Santa Barbara County, California (b) Dallas County, Texas, (c) Nebraska counties and (d) Ohio counties. The dotted line represents the critical value of 1. The plots are obtained using the inbuilt plotting function for R_t offered in `epidemia` [2].

still an apparent distinction between the autumn 2020 months where R_t is clearly greater than 1, and say, the beginning of 2021 where the 30% and 60% CIs give values lower than 1.

Figure 5.5 represents a time series of deaths, with both the reported values and the estimated values. Overall, the estimated model captures the data well: the overall shapes match. However, a significant improvement in fit is apparent when there is more deaths. Indeed, Dallas County and Santa Barbara county, where the bars corresponding to observed deaths are compactly represented and outline a clear shape, the estimation is smooth and matches the observations both in overall shape and in the range of values attained. On the other hand, for the Nebraska and Ohio counties, there are less reported deaths daily, and more days without any reported deaths at all, which gives estimates that do not outline such a clear trend, and that do not reach the higher values, which are attained on certain days.

Finally, figure 5.6 shows the infections estimated by the model. As infections are not a value measured or reported at the county-level in the US, the estimates are compared with the reported cases. Although these quantities are entirely distinct, as cases will not be reported on the day of infections, and many infections will simply not be reported as a case, comparing the infections estimate to reported cases is coherent in that they should follow approximately the same trend. Ignoring the difference in actual values, this hypothesis holds for Santa



(a) Santa Barbara County, California

(b) Dallas County, Texas

(c) 9-county grouping, Nebraska

(d) 3-county grouping, Ohio

Figure 5.5: Deaths times series with CIs for (a) Santa Barbara County, California, (b) Dallas, Texas, (c) Nebraska counties and (d) Ohio counties. The red bars correspond to the observed deaths, and the blue shaded area gives 90%, 60 % and 30% intervals for the estimated number of deaths by the model.These plots are obtained using the inbuilt plotting function for observations offered in `epidemia` [2].

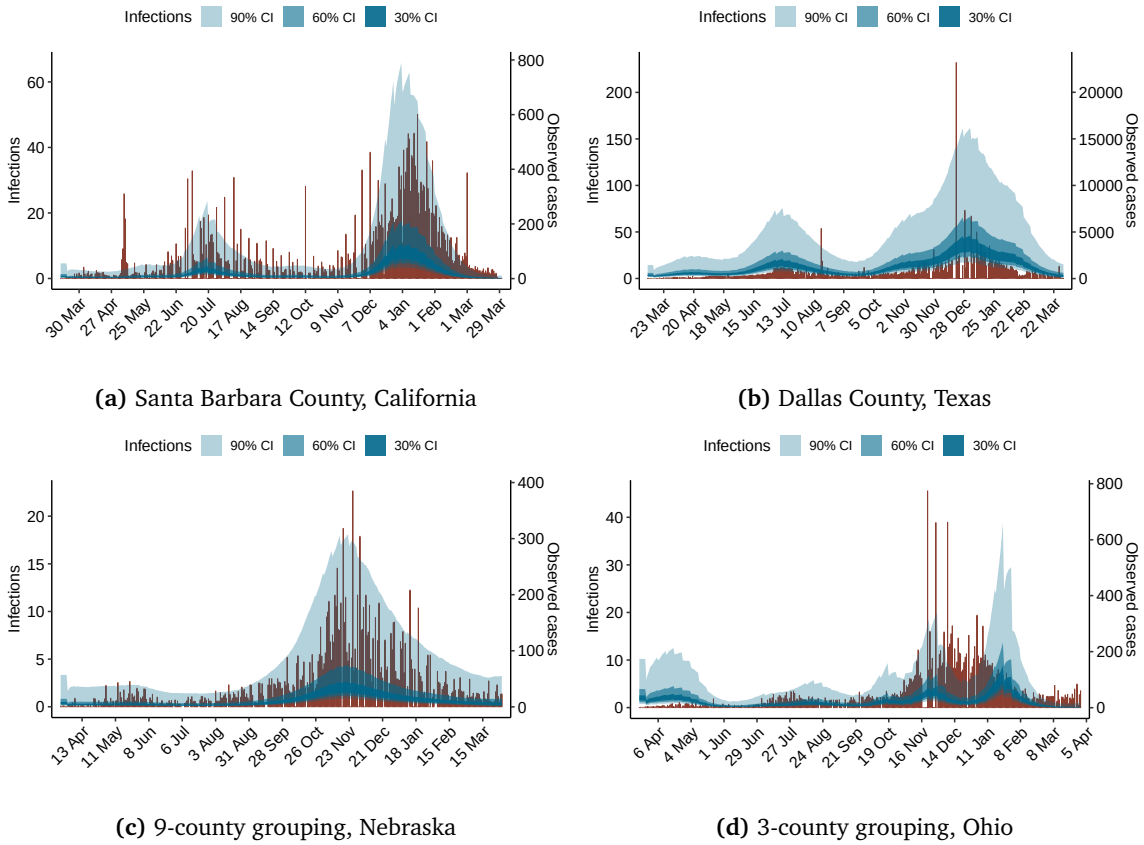
Barbara County, Dallas County (ignoring the value of 23,000 cases on December 21st, which most likely is a result of mishandling data rather than a true value), and the Nebraska counties. This is less the case for the Ohio counties, where the peaks in cases do not coincide with the model estimate. Potential explanations of this phenomenon are explored in the discussion chapter.

5.2 R_t, Trump Support and Vaccination Hesitancy

After obtaining the R_t values, which have now been closely investigated, I move on to the regression task.

5.2.1 Exploratory Data Analysis

Before fitting the Bayesian Linear Regression models, it is important to perform a prior inspection of the data, which consists of the election results (and polls on the state-level), and the vaccination hesitancy rates.



(a) Santa Barbara County, California

(b) Dallas County, Texas

(c) 9-county grouping, Nebraska

(d) 3-county grouping, Ohio

Figure 5.6: Infections times series with CIs for (a) Santa Barbara County, California, (b) Dallas County, Texas, (c) Nebraska counties and, (d) Ohio counties. Estimated values for infections are represented by the blue shaded area corresponding to 90%, 60% and 30% credible intervals and are indicated by left axis. The right axis corresponds to the number of reported cases, given by the red bars. These plots are obtained using the inbuilt plotting function for infections offered in `epidemia` [2] and personalized using `ggplot2`[74].

Map Visualizations

Figure 5.7 shows the different variables on a map. Before moving on to analysis, it is important to remember that the vaccine hesitancy data defined 'Hesitant' differently in the datasets used on the county and states-levels. Therefore, a perfect correspondence between the corresponding figures cannot be established as they do not measure the exact same quantity.

The first striking result is the fact that when looking at the county-level representations, the shades can vary greatly within a single state's borders, for all three quantities of interest. For instance, Arizona and New Mexico have very split support for Trump depending on the county, vaccine hesitancy varies greatly in the South Dakota counties, Wisconsin and Georgia have a significant split with counties whose R_t is less and greater than 1. Such results provide further motivation for the county-level approach, as the state-level appears to hide dynamics that appear more locally.

Comparing the county-level to the state-level map for R_t provides a good sanity check for the derived estimates. Indeed, it is reassuring to find that states containing many counties with lower R_t will also exhibit an overall lower R_t value on election day, and vice-versa.

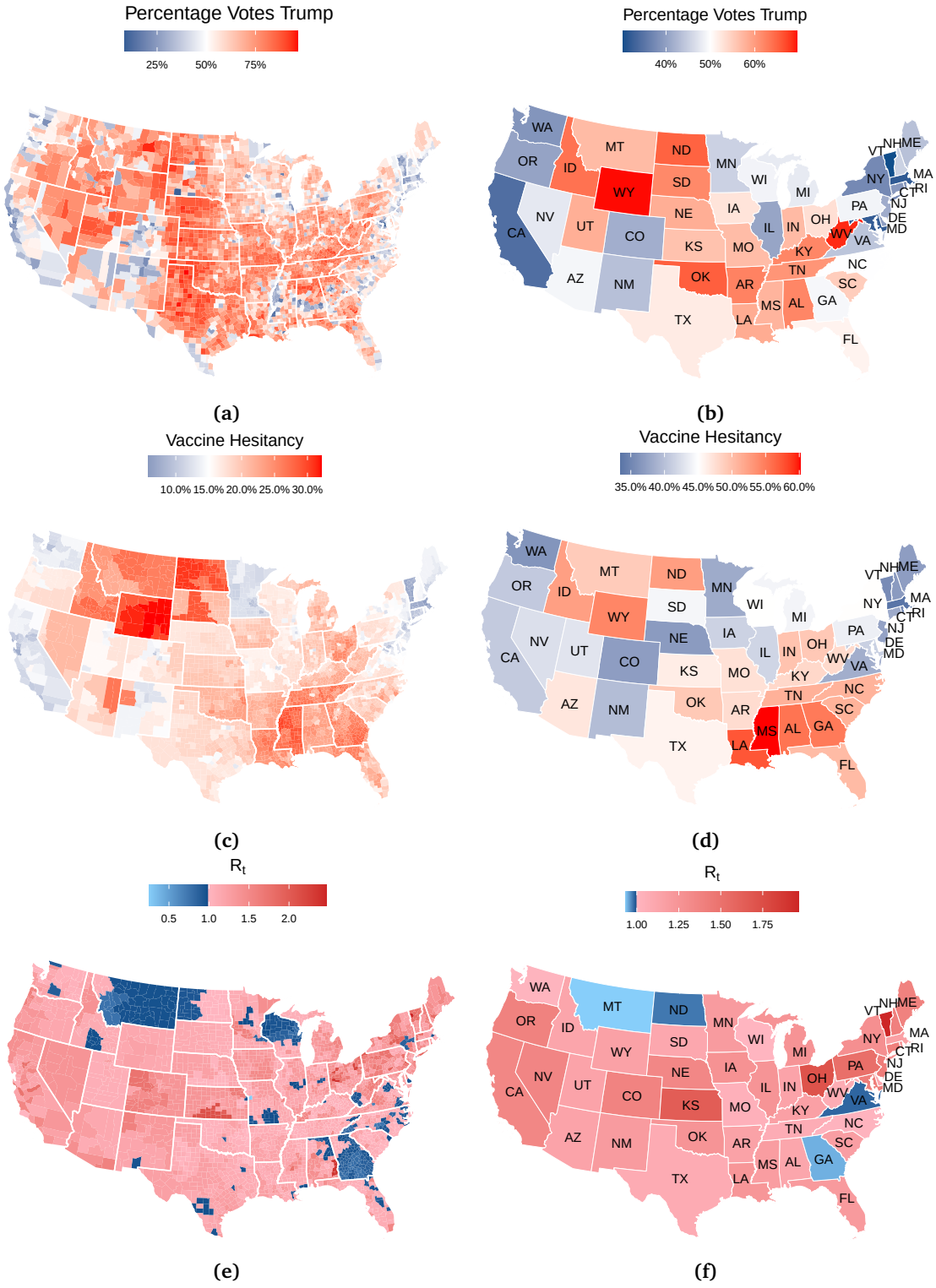


Figure 5.7: Map representation of Trump support, vaccine hesitancy, and R_t. The left hand-side provides visualizations at the county-level while the right-hand side consists of state-level representations. Percentage votes Trump corresponds to the November 3^{rd} election results, and R_t is taken from that day as well.

Finally, interesting links can be noticed between the three different represented quantities. Comparing the plots for the percentage votes for Trump to vaccine hesitancy, it appears that there is a correspondence in the shades. Indeed, counties or states that are more blue in subfigures (a) and (b), that is have lower Trump support, tend to also be blue in the vaccine hesitancy subfigures (c) and (d). The opposite relationship can be witnessed with regards to the R_t values. Indeed, counties and states that appear to have darker red colors, corresponding to high R_t values, tend to correspond to states and counties with lower vaccine hesitancy and Trump support.

Pearson Correlation

A more rigorous approach to the exploration of relationships between the various predictors later used in the regression models is one looking at the Pearson correlation coefficient. For two samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ is given by

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{5.5}$$

where $\bar{x}, \bar{y}$ denote the respective sample means.

Figure 5.8: Pearson correlation matrix for covariates at the state-level.

The covariates considered on the state-level are the following :

- `HESIT`: the percentage of the population that is hesitant (i.e. not certain to get the vaccine)
- `NO_LIKE`: the percentage of the population that is hesitant because they do not like vaccines

- NO_VAX_TRUST: the percentage of the population that is hesitant because they do not trust the COVID-19 vaccines
- NO_GOV_TRUST: the percentage of the population that is hesitant because they do not trust the government
- OTHERS_NEED: the percentage of the population that is hesitant because they believe others may need the vaccine more
- VOTES_2020: the percentage of the population that voted for Trump in the election
- CHANGE_2020: the percentage difference between votes for Trump in 2020 and in 2016

Figure 5.8 represents the Pearson coefficient values between all the pairs of covariates. Red tones correspond to positive correlation while blue tones correspond to negative correlation. It is interesting to notice that most covariates are positively correlated with each other (albeit not always strongly), apart from OTHERS_NEED which displays negative Pearson coefficient values when considered with HESIT, NO_VAX_TRUST and NO_GOV_TRUST, and CHANGE_2020 which also has values negatively correlated with NO_LIKE, NO_VAX_TRUST and NO_GOV_TRUST. These results can then allow to fit models that avoid multicollinearity in regression. Additionally, figure A.3 provided in the supplement provides scatter plots of each pair of variables as well as density estimates of the data for each covariate.

As such detailed data is not available at the county-level, the only covariates considered in the Pearson correlation analysis are the votes in the election, percentage hesitant, and percentage strongly hesitant.

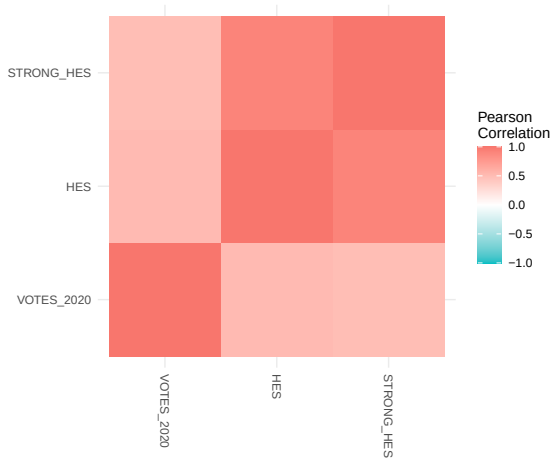


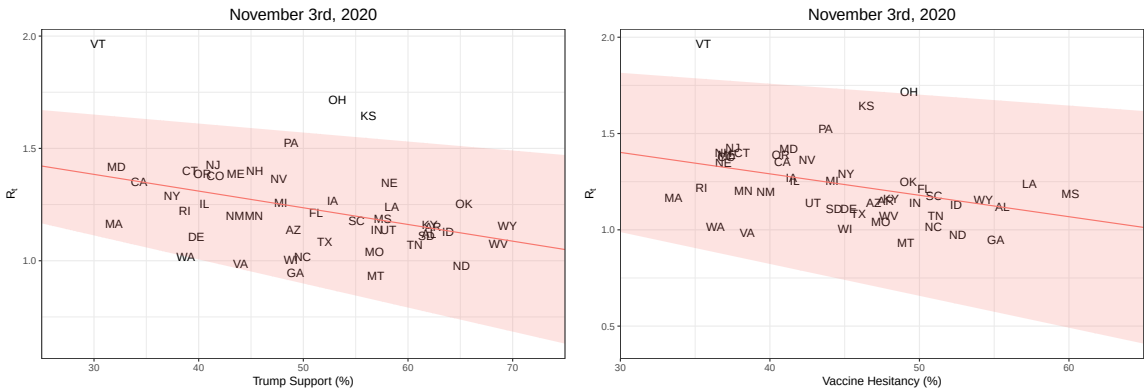
Figure 5.9: Pearson correlation matrix for covariates at the county-level.

Figure 5.9 shows that these three covariates are positively correlated with each other, however the relationship is particularly strong between hesitancy and strong hesitancy, thus providing evidence that using both these as predictors in the same model may be unnecessary.

5.2.2 Regression

On Election Day

First focusing on election day, (i.e. using R_t and results from this data), effects of the different covariates are explored by fitting 6 different regression models which include varied combinations of these. The 10%, 50% and 90% posterior percentiles for the parameter values are provided in table A.3 from the supplement. It is worth noting that none of the 6 models reach particularly high values of R^2: the highest posterior median is obtained with model 6 (0.2353), which includes all accessible covariates. However, this statistic is not necessarily very meaningful: indeed the Pearson correlation analysis has shown that some of the covariates are highly correlated with one another, which would hint to the model being poorer, even though it has this higher value. Moreover, as the aim of this regression task is to analyse and understand the relationship between the chosen quantities and R_t, and not to obtain predictions, the value of R^2 is not worth fixating on. Indeed, if one's aim was to simply obtain an R^2 closer to one, one could simply work with phenomena that are known to have a direct physical effect on the spread of the disease or on the calculation of R_t.

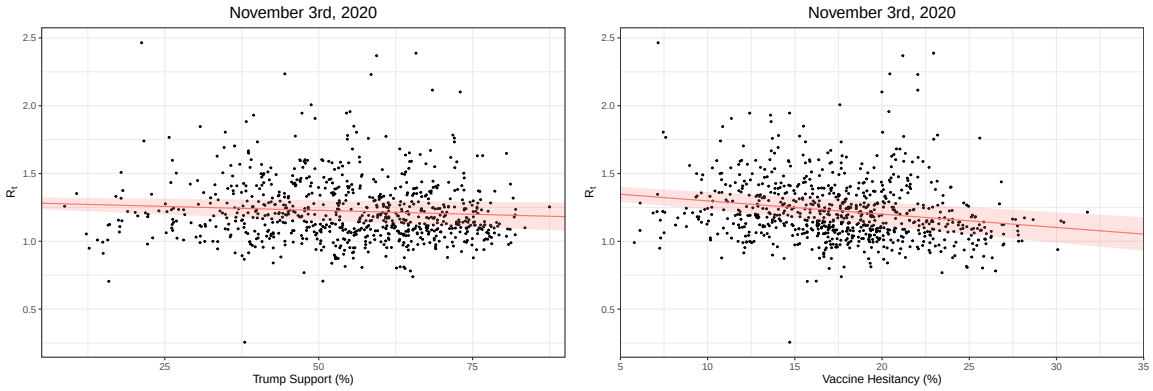


(a) **(b)**

Figure 5.10: Election day state-level scatter plots with regression line and CI. (a) represents R_t against Trump support, and (b) represents R_t against vaccine hesitancy. The shaded area corresponds to the 80% credible interval.

Figure 5.10 represents scatter plots of the 48 states, along with a regression line obtained from the simple intercept and single parameter Bayesian models, taking the election results and vaccine hesitancy covariates respectively. A downward sloping line can be noticed in both plots, implying a higher R_t for lower Trump support and vaccine hesitancy. It is however worth noting that the vaccine hesitancy values are the ones obtained in the period between January 6^{th} and January 18^{th}, and the results may have been different had the vaccination hesitancy been taken in the period of November 3^{rd}.

The same analysis is conducted at the county level, the coefficients of the various models fit are also provided in the supplement, in table A.5. Figure 5.11 represents the scatter plots equivalent to those presented in figure 5.10. There are two important elements worth noting. Firstly, with the increased number of data points (796 instead of 48), the credible intervals have become much narrower. Furthermore, the regression lines for both Trump support and vaccine hesitancy have become less steep, especially in the Trump support plot



(a) **(b)**

Figure 5.11: Election day county-level scatter plots with regression line and CI. (a) represents R_t against Trump support, and (b) represents R_t against vaccine hesitancy. The shaded area corresponds to the 80% credible interval.

which appears to have a slope close to zero. This is understandable as the points appear more scattered than they did in the state-level figures.

Time- varying coefficients

Given the time-varying nature of R_t, I also explore the value of the regression coefficients as a biweekly time series. The R_t and poll forecast values are averaged over two week periods, and Bayesian regression models are then fit using both Trump Support and vaccine hesitancy as predictors at each time point. The coefficient posterior medians, as well as 10% and 90% quantiles are reported, and represented in Figure 5.12. Since week 11 corresponds to the time elections happened, the Trump support values used from that moment onward are simply the election results. It is interesting to see that the values of the coefficients vary significantly through time, and in particular that they switch from being positive to negative (i.e. the correlation structure changes). It is however worth noting that the values of the coefficients for Trump support, represented on the right axis, are much closer to zero than those for hesitancy, so the varying is of much smaller scale. Another noteworthy element is that both curves appear behave in an 'anti-symmetric' manner: while the Trump support coefficients are positive, the hesitancy coefficients tend to be negative (apart from the first 2 months). Furthermore, both curves appear to reach a local maximum and minimum respectively for Trump support and hesitancy at period 9. Remembering that the vaccine hesitancy values are taken from early January, it is interesting to see that the dependence of R_t on them can vary greatly at different times.

A similar analysis is conducted at the county-level, where not only vaccine hesitancy, but also Trump support values are fixed as the forecast data is not available at the county-level, with the time series of coefficients represented in Figure 5.13 along with the credible intervals. The trend followed by the hesitancy parameter is very similar to the one observed at the state-level and additionally, are close in value, but the second switch in curve slope happens at period 13 rather than 9. For the Trump support, the general trend is still resembling the state-level, but the coefficient values are now much greater, reaching up to 0.3 (while the maximum value at the state-level was not much greater than 0.01). It is also worth noting that the credible intervals are narrower (relative to the values), which was already

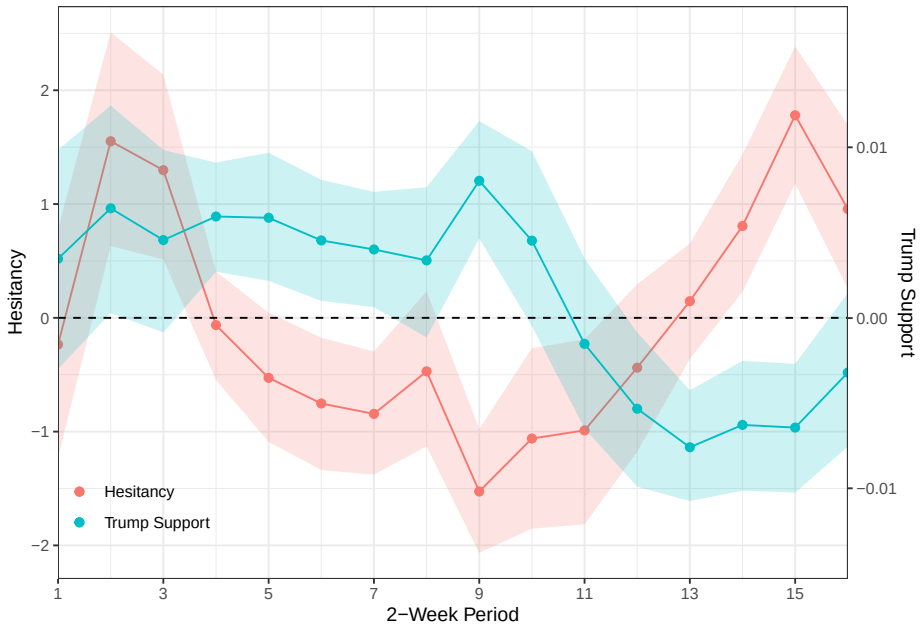


Figure 5.12: State-level time-varying hesitancy and Trump support coefficients. The 2-week periods considered start from the period 06/01/2020- 06/14/2020 to the period 01/11/2021-01/24/2021. Note that election day is part of period 11. The dashed line corresponds to a coefficient value of zero, the boundary between positive and negative correlation. A second y-axis is used for the Trump support coefficient values as the values are much lower than those for vaccine hesitancy.

commented on previously and attributed to the larger dataset size.

As quite a significant switch in behaviors is noticed between the earlier periods and the later ones, I finish by representing the data accompanied by their regression line at two separate dates: September 28^{th} (period 8) and December 16^{th} (period 14). Similarly to the approach taken to obtain the figures for election day, separate single parameter Bayesian linear models are obtained for the Trump support and vaccine hesitancy figures.

As the Trump hesitancy coefficients are always quite close to zero, the differences in the regression line for the state level plots in 5.14 based on that covariate are not particularly noticeable. The scattered points representing the states also appeared quite spread out without a clear pattern appearing. On the other hand, a stark difference can be noted between the September 28^{th} and December 16^{th} plots for vaccine hesitancy. Indeed, the regression line moves from a negative slope to a positive slope. It is interesting to note that the R_t values display quite different characteristics: on December 16^{th} many more states have R_t less than, or very close to, 1. The effect of vaccine hesitancy would thus appear to be rather strongly dependent on the dynamics of the pandemic.

At the county-level, represented in figure 5.15, visualizing the changes in slope of regression lines is slightly more difficult as the range of R_t value is greater (from 0.5 to 2.5 approximately). However, the switch in slopes from positive to negative is more apparent for the Trump support covariate than it was previously, and the change on the vaccine hesitancy plots remains observable as well. Once again, it can be observed that the credible intervals

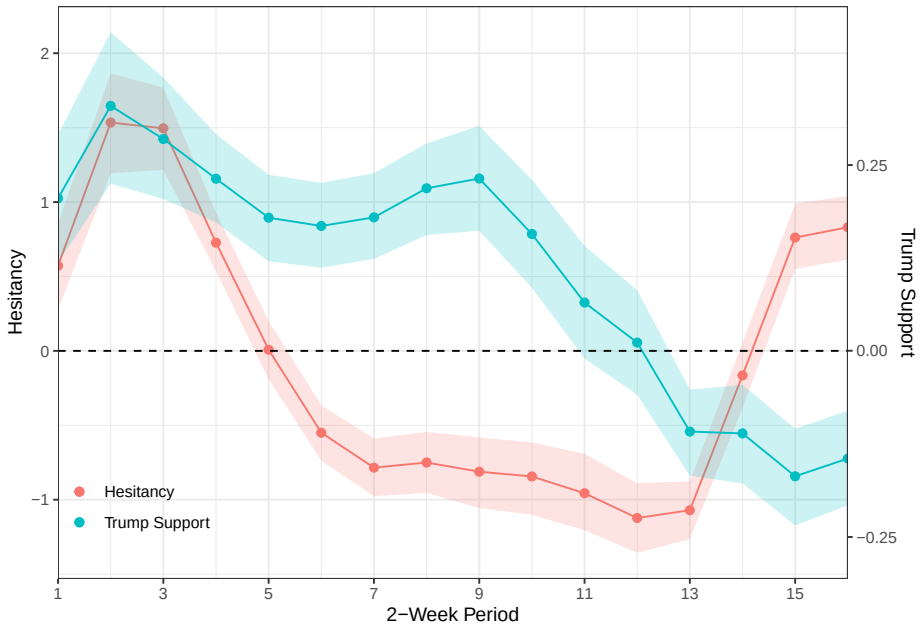


Figure 5.13: County-level time-varying hesitancy and Trump support coefficients. The 2-week periods are the same as those considered on the state-level. A second y-axis is once again used for the Trump support coefficient values.

are narrower in the county-level setting.

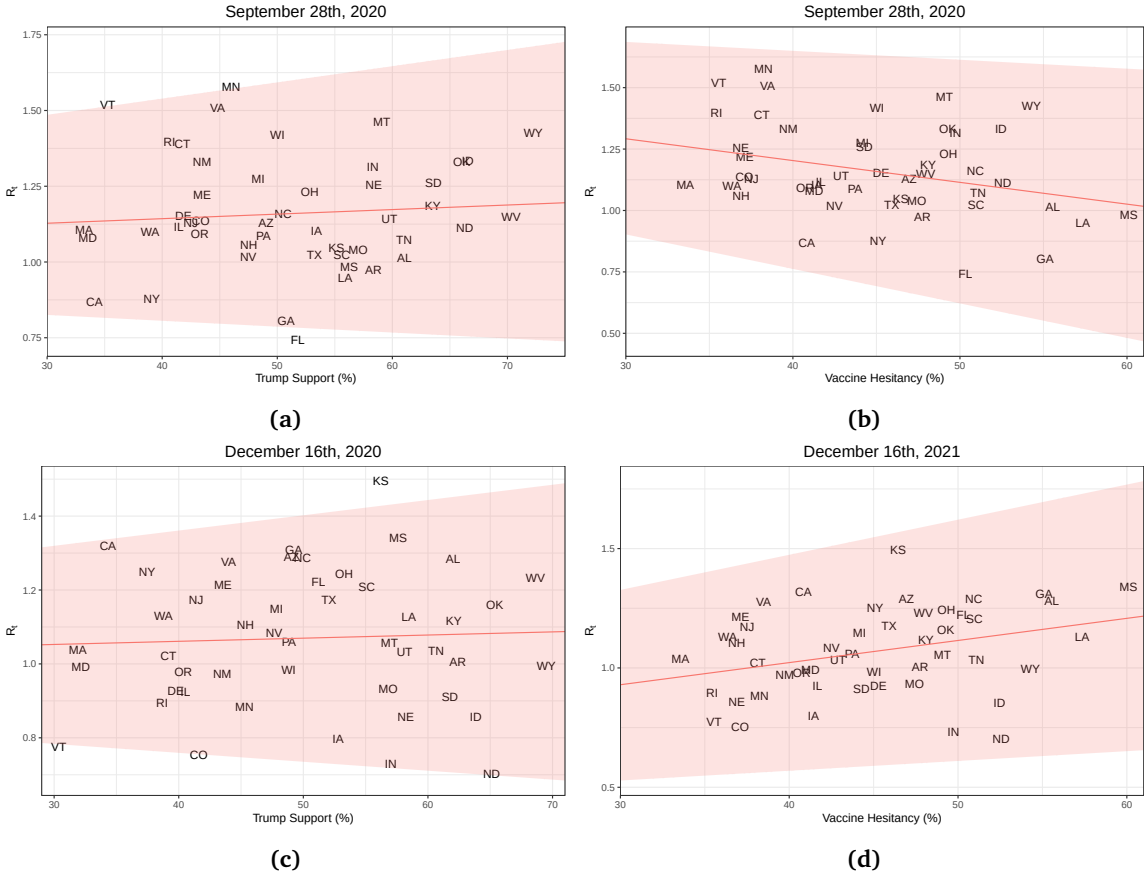


Figure 5.14: State-level scatter plots for September 28^{th} and December 16^{th}. Regression lines and 80 % credible intervals are obtained by fitting intercept + single parameter Bayesian linear models.

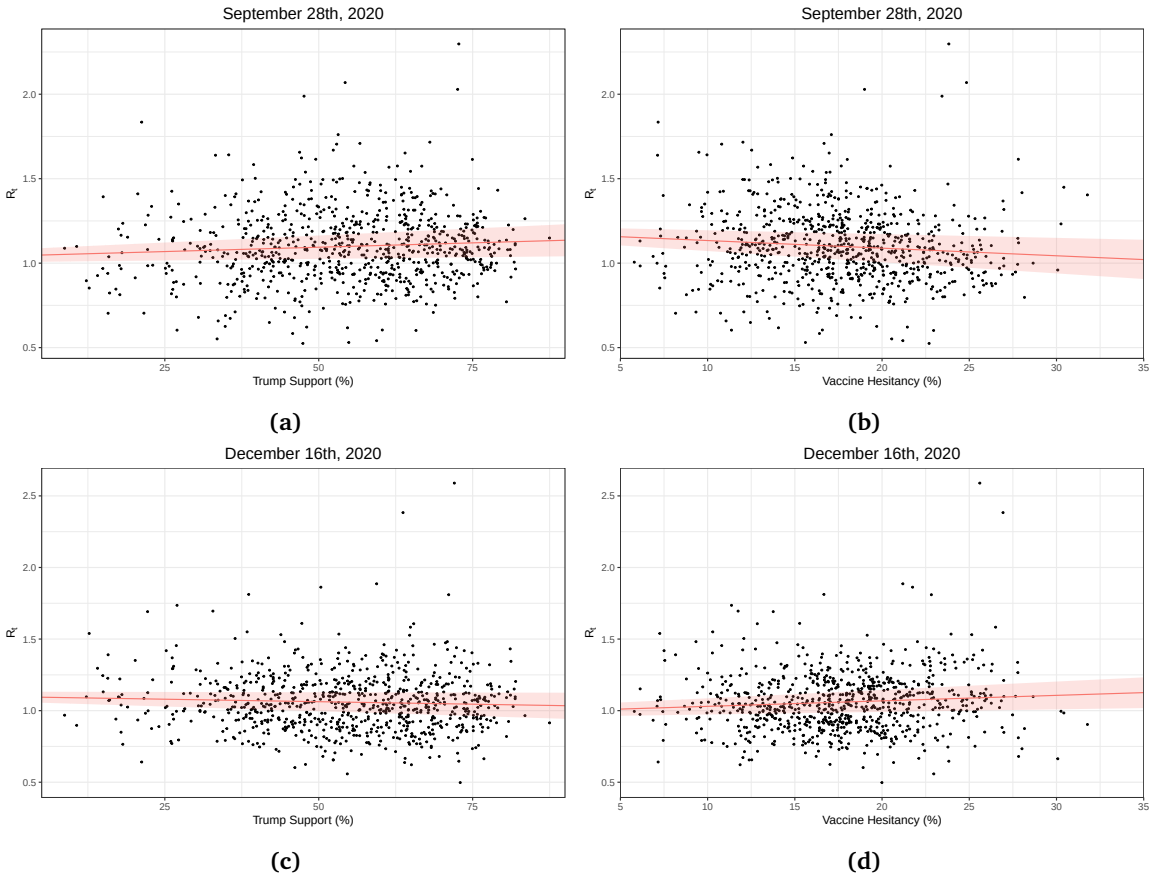


Figure 5.15: County-level scatter plots for September 28^{th} and December 16^{th}. Regression lines and 80 % credible intervals are obtained by fitting intercept + single parameter Bayesian linear models.

Chapter 6

Discussion

6.1 Analysis of Findings

6.1.1 Understanding convergence issues

In the previous section, convergence issues were frequently noticed when fitting the models using 2,500 iterations MCMC iterations. Interestingly, not all states suffered from those issues equally, which implies that specific characteristics for the states of interest lead to poor convergence, rather than the chosen number of iterations being simply too small for the parameter space to be explored.

These issues, which have for instance touched states like New York, Kansas, Indiana, or Virginia, could perhaps be explained by bimodality in the target distribution. Indeed, as given by Scott et al. [2], periods with lower death counts can be explained by policies suppressing the virus, and by the herd immunity built by the virus spreading through the susceptible population. The model may then remain stuck in the local mode induced by herd immunity, a problem MCMC methods often face when working with multimodal target distributions. As the model for R_t is built as a random walk, the consequences of being trapped in the herd immunity local mode at times where death counts are particularly low can then expand to later dates, even if deaths may start increasing again.

Although the four county-level models that were explored in depth appeared to have converged, it was later noticed that the estimated number of infections appeared inconsistent with the number of observed cases, for the Ohio counties (which would suggest such problems might have occurred with many of the 796 models). Although the infections are modelled daily (as is allowed by the generation time distribution), its deterministic expression from (4.6) can be seen to also depend on R_t, which is modelled weekly. This could thus lead to some discrepancy with the observations given daily. However, as this issue does not appear for all models, a more likely explanation of the issue witnessed with the Ohio counties model would be linked directly with the data available. It appears that the estimated infections follow a similar trend to the deaths estimations, where peaks appear to coincide with very high numbers of observed deaths. These could perhaps be the consequence of reporting delays and errors. Additionally, the observation data also displays a few peak values that stand out from the rest, possibly suggesting case reporting inaccuracies other counties did not experience as strongly. Joining these two sources of error together could thus explain the clear discrepancy between estimated infections and observed cases.

6.1.2 Interpreting regression results

The fit of various regression models at the state-level and at the county-level did not show a very conclusive result regarding the relationship between R_t and the two predictors considered: vaccine hesitancy and Trump support. Indeed, taking the time-varying approach showed coefficient values that alternated from positive to negative, making it difficult to conclude on the effect these two quantities had on the reproduction number.

Following some of the arguments presented in the background section, these results could appear surprising. Indeed, first focusing on the Trump support, it can be noted that the coefficient values are very close to zero. However, Sy et al. [77] conclude that counties with greater population density have greater rates of transmission of SARS-CoV-2, as a result of increased contact rates in such areas. Additionally, it is worth noting that rural counties vote overwhelmingly for Trump: it is estimated that in the 2020 elections only around 10% of the rural counties went to Biden [78]. Given this information, the lack of clear results may therefore not necessarily be an indicator of lack of correlation (and even potential indirect causation), but rather of a flaw in the study setting. A perhaps more informative approach would be consist of a study restricted to counties with similar characteristics to obtain a more controlled experiment, and thus where Trump support could be more representative.

Similar arguments can also be presented regarding vaccine hesitancy: the components driving the variation in the reproduction number may be blurring the effects of the covariate. Additionally, there is the problem that the vaccine hesitancy data is taken at a fixed date (also the case for Trump support data at the county-level). Not only has it been given by Fridman et al. [79] that in fact attitudes regarding vaccines have varied through time and thus taking the values from January to represent hesitancy at earlier times will be inaccurate, it also follows that since R_t is changing, the effect of the vaccine numbers will also have to vary at the different dates to fit models that best fit the given data.

6.2 Limitations

Following from the analysis above, and considering the various observations made throughout the thesis, I now outline and summarize the principal shortcomings of the work presented.

Firstly, it is important to underline the flaws represented by the various datasets. The Bayesian models used to estimate R_t were based solely on reported deaths data, which as has already previous been presented, may not always be perfectly accurate. Although in the months at the start of the pandemic reported case data was inaccurate as testing was not yet developed, this became less of a problem at later dates and thus a model included observed cases such as the one presented by Mishra et al. [80] may have helped with convergence issues and the inconsistency between estimated and observed quantities noted previously. The issues with the polling data have been previously mentioned, and one could perhaps implement a more refined bias correction technique, however following from arguments previously presented, the impact of the inaccuracies on final results is likely not significant. Finally, the vaccination hesitancy data obtained from the Census' Household Pulse survey has been shown to display discrepancies between vaccine willingness and hesitancy [43].

Other limitations that are important to underline are those inherent to the semi-mechanistic Bayesian model implemented in `epidemia`. Mishra et al. [80, p. 7] present a comprehensive list of those limitations in the context of the model they fit for the United Kingdom, most of which are also applicable to the work in this thesis. These include, but are not limited to, the model assuming a fixed probability distribution with fixed mean and standard deviation for the delays from infection to death distribution as well as for the generation time distribution; assuming the generation time and serial interval distributions are the same; implementation of new measures are not explicitly part of the model and thus their effect may only appear at a delayed date.

Finally, an important issue to remember is that not all models converged, and the length of the MCMC chains must have been much greater than 2,500 iterations. It was seen that 20,000 already seemed to help, however once everything is set-up, and without the time constraints of this project, the models could have run on HPC for much longer than that. It is worth noting that other than the time required for obtaining the models, memory also becomes an issue for those larger models. Indeed, when running the four chains for 2,500 iterations, the state fits have a size of approximately 6.5MB, and the posterior R_t draws approximately 15.5MB. For 20,000 iteration, those become approximately 44.5MB and 122MB respectively. Although the data itself is stored in the Research Data Store (RDS), running analyses using RStudio requires loading those into RAM, which can be very time consuming as well a quickly approach the laptop memory limits. Ideally, the code for analysis would therefore have been developed and tweaked on the smaller files as I have done, before then submitting jobs to HPC that generate the necessary plot and outputs using the better quality fits.

6.3 Potential Extensions

Before concluding this thesis, I present some extensions to be explored that could improve certain results, as well as further research on the topic that this project motivates.

At various points in the project the idea of an "ideal" model at the county-level was presented. The choice made was to simply group models by population, and treat them independently. This approach has the drawback of ignoring the effects that other populations may have on a given county. To remedy to this, a hierarchical model could be fit instead. This approach, which be implemented directly using `epidemia` [2] would then allow to share signal across some larger groups. Further work could thus focus on an exploration of the county grouping size, jointly with the definition of the larger groups within which parameters are shared, in the hopes of improving the county-level fits. Other improvements to the model could include a more careful selection of the initial parameters for the sampler, as suggested by Bhatt et al. [23], or by making use of the population adjustment feature in `epidemia` that removes already infected individuals from the population, but also allows the input of proportion of population that is removed from the susceptible group through other means (such as vaccination).

As the county-level data for vaccination hesitancy only had direct estimates available at the county-level, I relied on the data from the ASPE [49], which gives the estimates on the county-level using a two stage approach, starting from the state-level poll results. The methodology is not very detailed, thus to have a better grasp of the specifics of the model

used, multilevel regression with post-stratification (MRP) [81] could be performed using the ACS publicly available data. This would allow to fit a model with a hierarchical component, i.e. where state-level effects can be accounted for, as well as give greater control over the population characteristics included.

A notable issue that was faced during this thesis was the lack of time-varying data to work with in order to obtain regression models for R_t around election time. However, the availability increases at later dates: for example, vaccine hesitancy bi-weekly data can be taken from the Household Pulse Survey. It could thus be interesting to continue the work done in this research at later dates, fitting models with time-varying vaccine hesitancy values. Additionally, as vaccinations become readily available in the United States, one could also incorporate those numbers in the models. Once the elections have passed, Trump support is no longer a quantity that will be measured, however one could for example look at public trust in the government, maintaining the link with politics. Finally, the availability of Household Pulse Survey data regarding a vast range of domains, the methods of this thesis could be extended to also cover variables other than vaccine hesitancy or Trump support, such as proportions of population teleworking, number of trips taken, or changes to household spending (and reasons), to name a few.

Chapter 7

Conclusion

Using the methods first introduced in Flaxman et al. [3], this project explores the SARS-CoV-2 reproduction number R_t at the county level. Even though the areas considered were significantly smaller than other work working with the same implementation, and hence the observed death numbers were lower for each county or grouping considered, a very large majority of the models had $\hat{R}$ values lower 1.05, implying convergence. The R_t values were then taken to be the response variable of linear regression models, where vaccine hesitancy and Trump support in the 2020 U.S. election were the predictors. The results obtained did not show a very distinctive trend between the covariates and R_t over time, but the analysis and discussion aided with the understanding of the various dynamics driving the spread of the disease. Consistency between the state-level and county-level did however show that the models picked up on the changing dependence of R_t on the two covariates, where a switch in the effect of Trump support and vaccine hesitancy on R_t can be observed around election time.

This thesis; started, researched, coded, drafted and submitted; in the midst of the COVID-19 pandemic ties itself into a wide range of research and innovation in the field of mathematical modelling of infectious diseases. The high impact of the scientific work developed in decision-making truly shows how important these studies are, and illustrates the constant need for progress, to which this project has hopefully contributed. Although its main focus has been the reproduction number, it is however worth noting that R_t should not be considered in isolation for a thorough understanding of the dynamics, in particular in the realm of policy-making, as any estimate will inevitably not account for all subtleties [82].

References

[1] R Development Core Team. R: A Language and Environment for Statistical Computing, 2021. URL https://www.R-project.org/.

[2] James A. Scott, Axel Gandy, Swapnil Mishra, Samir Bhatt, Seth Flaxman, et al. epidemia : An R package for Bayesian , semi-mechanistic modelling of infectious diseases. *[preprint]*, 2021.

[3] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, -08 2020. doi: 10.1038/s41586-020-2405-7. URL https://www.nature.com/articles/s41586-020-2405-7.

[4] Wolfgang Messner and Sarah E. Payson. Variation in COVID-19 outbreaks at the US state and county levels. *Public Health*, 187:15–18, October 1, 2020. doi: 10.1016/j.puhe.2020.07.035. URL https://www.sciencedirect.com/science/article/pii/S0033350620303309.

[5] Charles Courtemanche, Joseph Garuccio, Anh Le, Joshua Pinkston, and Aaron Yelowitz. Strong social distancing measures in the United States reduced the COVID-19 growth rate. *Health Affairs*, 39(7):1237–1246, July 1, 2020. doi: 10.1377/hlthaff.2020.00608. URL https://www.healthaffairs.org/doi/10.1377/hlthaff.2020.00608.

[6] H. Juliette T. Unwin, Swapnil Mishra, Valerie C. Bradley, Axel Gandy, Thomas A. Mellan, Helen Coupland, et al. State-level tracking of COVID-19 in the United States. *Nature Communications*, 11(6189):1–9, 2020. doi: 10.1038/s41467-020-19652-6.

[7] Philip Ball. The lightning-fast quest for COVID vaccines — and what it means for other diseases. *Nature*, 589(7840):16–18, December, 2020. doi: 10.1038/d41586-020-03626-1. URL https://www.nature.com/articles/d41586-020-03626-1.

[8] Jane M. Heffernan, Robert J. Smith, and Lindi M. Wahl. Perspectives on the basic reproductive ratio. *Journal of the Royal Society, Interface*, 2(4):281–293, -09-22 2005. doi: 10.1098/rsif.2005.0042.

[9] Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, November 1, 2013. doi: 10.1093/aje/kwt133. URL https://doi.org/10.1093/aje/kwt133.

[10] Sam Abbott, Sam Hellewell, Robin N. Thompson, Katharine Sherratt, Hamish P. Gibbs, Nikos I. Bosse, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved with

reservations]. *Wellcome Open Res.*, (5):112, 2020. URL https://doi.org/10.12688/wellcomeopenres.16006.2.

[11] Anne Cori, Alain-Jacques Valleron, Fabrice Carrat, Gianpaolo Scalia Tomba, Gaetan Thomas, and Pierre-Yves Boëlle. Estimating influenza latency and infectious period durations using viral excretion data. *Epidemics*, 4(3):132–138, 2012. doi: 10.1016/j.epidem.2012.06.001. URL https://www.clinicalkey.es/playcontent/1-s2.0-S175543651200031X.

[12] Robert J. Hoagland. The incubation period of infectious mononucleosis. *American Journal of Public Health and the Nations Health*, 54(10):1699–1705, -10 1964. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1255045/.

[13] Monica F. Myers, David J. Rogers, J. Cox, Antoine Flahault, and Simon I. Hay. Forecasting disease risk for increased epidemic preparedness in public health. *Advances in Parasitology*, 47:309–330, 2000. doi: 10.1016/s0065-308x(00)47013-2.

[14] Pauline van den Driessche. Reproduction numbers of infectious disease models. *Infectious Disease Modelling*, 2(3):288–303, June 29, 2017. doi: 0.1016/j.idm.2017.06.002.

[15] William Ogilvy Kermack and Anderson G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, August 1, 1927. doi: 10.1098/rspa.1927.0118. URL https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118.

[16] Hans J. a. P. Heesterbeek and Klaus Dietz. The concept of r0 in epidemic theory. *Statistica Neerlandica*, 50(1):89–110, 1996. doi: https://doi.org/10.1111/j.1467-9574.1996.tb01482.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1996.tb01482.x.

[17] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516, September 15, 2004. doi: 10.1093/aje/kwh255.

[18] Cheryl L. Gibbons, Marie-Josée J. Mangen, Dietrich Plass, Arie H. Havelaar, Russell Jogn Brooke, Piotr Kramarz, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*, 14(147), 2014. doi: https://doi.org/10.1186/1471-2458-14-147.

[19] Neil M. Ferguson, Christl A. Donnelly, Mark E. J. Woolhouse, and Roy M. Anderson. The epidemiology of BSE in cattle herds in Great Britain. II. Model construction and analysis of transmission dynamics. *Philosophical transactions. Biological sciences*, 352 (1355):803–838, Jul 29, 1997. doi: 10.1098/rstb.1997.0063. URL http://rstb.royalsocietypublishing.org/content/352/1355/803.abstract.

[20] Christophe Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE*, 2(8):e758, August 22, 2007. doi: 10.1371/journal.pone.0000758. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000758.

[21] Luís M. A. Bettencourt and Ruy M. Ribeiro. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE*, 3(5):e2185, May14, 2008.

doi: 10.1371/journal.pone.0002185. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002185.

[22] Heath A. Kelly, Geoff N. Mercer, James E. Fielding, Gary K. Dowse, Kathryn Glass, Dale Carcione, et al. Pandemic (h1n1) 2009 influenza community transmission was established in one australian state when the virus was first identified in north america. *PLoS ONE*, 5(6), June 28, 2010. doi: 10.1371/journal.pone.0011341. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893203/.

[23] Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A. Scott. Semi-mechanistic Bayesian modeling of COVID-19 with renewal processes. *[Preprint]*, December, 2020. URL https://arxiv.org/abs/2012.00394v2.

[24] James A. Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. Modeling of Epidemics using Hierarchical Bayesian Models, 2020. URL https://imperialcollegelondon.github.io/epidemia/. R package version 1.0.0.

[25] Mick Roberts, Viggo Andreasen, Alun Lloyd, and Lorenzo Pellis. Nine challenges for deterministic epidemic models. *Epidemics*, 10:49–53, March 1, 2015. doi: 10.1016/j.epidem.2014.09.006. URL https://www.sciencedirect.com/science/article/pii/S1755436514000553.

[26] Marco Cascella, Michael Rajnik, Abdul Aleem, Scott C. Dulebohn, and Raffaela Di Napoli. *Features, Evaluation, and Treatment of Coronavirus (COVID-19)*. StatPearls. StatPearls Publishing, Treasure Island (FL), April 20, 2021. URL http://www.ncbi.nlm.nih.gov/books/NBK554776/.

[27] European Centre for Disease Prevention and Control. COVID-19 situation update worldwide, as of week 21, updated 3 June 2021, June 3, 2021. URL https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases.

[28] Mélodie Monod, Alexandra Blenkinsop, Xiaoyue Xi, Daniel Hebert, Sivan Bershan, Simon Tietze, et al. Age groups that sustain resurging COVID-19 epidemics in the United States. *Science (American Association for the Advancement of Science)*, 371(6536):1336, Mar 26, 2021. doi: 10.1126/science.abe8372. URL https://www.ncbi.nlm.nih.gov/pubmed/33531384.

[29] Andrew M. Olney, Jesse Smith, Saunak Sen, Fridtjof Thomas, and H. Juliette T. Unwin. Estimating the effect of social distancing interventions on COVID-19 in the United States. *American Journal of Epidemiology*, (kwaa293), January 7, 2021. doi: 10.1093/aje/kwaa293. URL https://doi.org/10.1093/aje/kwaa293.

[30] Thomas P. Smith, Seth Flaxman, Amanda S. Gallinat, Sylvia P. Kinosian, Michael Stemkovski, H. Juliette T. Unwin, et al. Temperature and population density influence SARS-CoV-2 transmission in the absence of nonpharmaceutical interventions. *Proceedings of the National Academy of Sciences*, 118(225), June, 2021. doi: 10.1073/pnas.2019284118.

[31] Iwona Hawryluk, Thomas A. Mellan, Henrique Hoeltgebaum, Swapnil Mishra, Ricardo P. Schnekenberg, Charles Whittaker, et al. Inference of COVID-19 epidemiological distributions from Brazilian hospital data. *Journal of The Royal Society Interface*, 17(172):20200596, November 25, 2020. doi: 10.1098/rsif.2020.0596. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2020.0596.

[32] Nuno R. Faria, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372(6544):815–821, May, 2021. doi: 10.1126/science.abh2644.

[33] Kate Sullivan. Impact of COVID-19 on the 2020 US presidential election. *International IDEA*, November, 2020. URL https://www.idea.int/sites/default/files/impact-of-covid-19-on-the-2020-us-presidential-election.pdf. Case Study.

[34] Nathaniel Rakich. What Absentee Voting Looked Like In All 50 States, February 2, 2021. URL https://fivethirtyeight.com/features/what-absentee-voting-looked-like-in-all-50-states/.

[35] Leonardo Baccini, Abel Brodeur, and Stephen Weymouth. The COVID-19 pandemic and the 2020 US presidential election. *Journal of Population Economics*, 34:739–767, 2021. doi: 10.1007/s00148-020-00820-3.

[36] Claudia Deane, Kin Parker, and John Gramlich. A Year of U.S. Public Opinion on the Coronavirus Pandemic, March 5, 2021. URL https://www.pewresearch.org/2021/03/05/a-year-of-u-s-public-opinion-on-the-coronavirus-pandemic/.

[37] World Health Organization. Getting the COVID-19 Vaccine, 2021. URL https://www.who.int/news-room/feature-stories/detail/getting-the-covid-19-vaccine.

[38] William W. C. Topley and Graham S. Wilson. The spread of bacterial infection. the problem of herd-immunity. *The Journal of Hygiene*, 21(3):243–249, -05 1923. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2167341/.

[39] World Health Organization. Coronavirus disease (COVID-19): Herd immunity, lockdowns and COVID-19, December, 2020. URL https://www.who.int/news-room/q-a-detail/herd-immunity-lockdowns-and-covid-19?gclid=CjwKCAjwtpGGBhBJEiwAyRZX2tlnU468JtFmXDOZ4ha3ddesoLvGwKeGHeiqCYtH4dJ_el51tKCVHhoCk6UQAvD_BwE#.

[40] Roy M. Anderson, Carolin Vegvari, James Truscott, and Benjamin S. Collyer. Challenges in creating herd immunity to SARS-CoV-2 infection by mass vaccination. *The Lancet*, 396(10263):1614–1616, 2020. doi: 10.1016/S0140-6736(20)32318-7. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32318-7/abstract.

[41] Noni E. MacDonald and the SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164, August 14, 2015. doi: 10.1016/j.vaccine.2015.04.036. URL https://www.sciencedirect.com/science/article/pii/S0264410X15005009.

[42] Ipsos. America's pandemic state of mind coming to an end, June 8, 2021. URL https://www.ipsos.com/en-us/news-polls/axios-ipsos-coronavirus-index.

[43] Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Are we there yet? Big Data significantly overestimates COVID-19 vaccination in the US. *[preprint]*, June, 20201 June, 20201.

[44] US COVID-19 cases and deaths by state, -04-05T14:05:46.622Z 2021. URL https://usafacts.org/visualizations/coronavirus-covid-19-spread-map.

[45] Detailed Methodology and Sources: COVID-19 Data, April 24, 2020. URL https://usafacts.org/articles/detailed-methodology-covid-19-data/.

[46] County Population Totals: 2010-2019, June 22, 2020. URL https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.

[47] County Adjacency File, Aug 30, 2015. URL https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html.

[48] Household Pulse Survey Data Tables. URL https://www.census.gov/programs-surveys/household-pulse-survey/data.html.

[49] Office of the Assistant Secretary for Planning and Evaluation. VACCINE HESITANCY FOR COVID-19: STATE, COUNTY, AND LOCAL ESTIMATES. URL https://aspe.hhs.gov/pdf-report/vaccine-hesitancy.

[50] Public Use Microdata Sample (PUMS), February 23, 2021. URL https://www.census.gov/programs-surveys/acs/microdata.html.

[51] Nate Silver. 2020 Election Forecast, December 8, 2020. URL https://projects.fivethirtyeight.com/2020-election-forecast/.

[52] MIT Election Data Lab and Science. U.S. President 1976–2020, 2017. URL https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX.

[53] Tony McGovern. US_County_Level_Election_Results_08-20, December 17, 2020. URL https://github.com/tonmcg/US_County_Level_Election_Results_08-20. commit :nbsp;81a52aa8bb671952894c5a8badb8c3a3545df6f2.

[54] James Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. Model description. URL https://imperialcollegelondon.github.io/epidemia/articles/model-description.html.

[55] Shunji Osaki. *Renewal Processes*. Applied Stochastic System Modeling. Springer, Berlin u.a, 1992. ISBN 978-3-642-84681-6. doi: 10.1007/978-3-642-84681-6. URL https://doi.org/10.1007/978-3-642-84681-6_4.

[56] Anna Price and Louis Myers. Federal, State, and Local Government Responses to COVID-19, -11 2020. URL https://www.loc.gov/law/help/covid-19-responses/us.php#IV.

[57] Daily Travel during the COVID-19 Public Health Emergency, September 2, 2020. URL https://www.bts.gov/daily-travel.

[58] Qifang Bi, Yongsheng Wu, Shujiang Mei, Chenfei Ye, Xuan Zou, Zhen Zhang, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet infectious diseases*, 20(8):911–919, Aug 2020. doi: 10.1016/S1473-3099(20)30287-5. URL http://dx.doi.org/10.1016/S1473-3099(20)30287-5.

[59] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Basics of Markov chain simulation*, pages 275–292. Bayesian Data Analysis. Chapman and Hall/ CRC, New York, 3rd edition, July 6, 2015. ISBN 9780429113079. doi: https://doi.org/10.1201/b16018.

[60] A. M. Johansen. *Monte Carlo Methods*, pages 296–303. International Encyclopedia of Education (Third Edition). Elsevier, Oxford, January 1, 2010. ISBN 9780080448947. URL https://www.sciencedirect.com/science/article/pii/B9780080448947015438.

[61] Stan Development Team. The Stan Core Library, 2018. URL https://mc-stan.org/.

[62] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15 (47):1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

[63] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.

[64] Radford M. Neal. *MCMC using Hamiltonian Dynamics*, pages 116–162. Handbook of Markov Chain Monte Carlo. Chapman; Hall/ CRC, 2011.

[65] Stan Development Team. 15. MCMC Sampling, 2020. URL https://mc-stan.org/docs/2_26/reference-manual/hmc-chapter.html.

[66] Cole C. Monnahan, James T. Thorson, and Trevor A. Branch. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3):339–348, 2017. doi: 10.1111/2041-210X.12681.

[67] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, /11 1992. doi: 10.1214/ss/1177011136.

[68] Costas Panagopoulos. Polls and elections: Accuracy and bias in the 2020 U.S. general election polls . *Presidential Studies Quarterly*, 51(1):214–227, 2021. doi: 10.1111/psq.12710.

[69] Geoffrey Skelley. Why Was The National Polling Environment So Off In 2020?, February 23, 2021. URL https://fivethirtyeight.com/features/why-was-the-national-polling-environment-so-off-in-2020/.

[70] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2020. URL https://mc-stan.org/rstanarm.

[71] Jonah Gabry and Ben Goodrich. Estimating Regularized Linear Models with rstanarm, July 20, 2020. URL https://mc-stan.org/rstanarm/articles/lm.html.

[72] Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for Bayesian regression models. *The American Statistician*, 73(3):307–309, May 13, 2019. doi: 10.1080/00031305.2018.1549100.

[73] Jonah Gabry and Tristan Mahr. bayesplot: Plotting for Bayesian Models, 2021. URL https://mc-stan.org/bayesplot/. R package version 1.8.0.

[74] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[75] Stan Development Team. 16. Posterior Analysis, 2020. URL https://mc-stan.org/docs/2_27/reference-manual/effective-sample-size-section.html.

[76] Jiqiang Guo, Jonah Gabry, Ben Goodrich, and Sebastian Weber. Convergence and efficiency diagnostics for Markov Chains, 2020. URL https://mc-stan.org/rstan/reference/Rhat.html.

[77] Karla Therese L. Sy, Laura F. White, and Brooke E. Nichols. Population density and basic reproductive number of COVID-19 across United States counties. *PLOS ONE*, 16(4):e0249271, 21 avr. 2021. doi: 10.1371/journal.pone.0249271. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249271.

[78] August Benzow. Rural America is not all Trump country: A closer look at the rural counties that Biden won, November, 2020. URL https://eig.org/news/rural-america-is-not-all-trump-country.

[79] Ariel Fridman, Rachel Gershon, and Ayelet Gneezy. COVID-19 and vaccine hesitancy: A longitudinal study. *PLOS ONE*, 16(4):e0250123, 16 avr. 2021. doi: 10.1371/journal.pone.0250123. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250123.

[80] Swapnil Mishra, Jamie Scott, Harrison Zhu, Neil M. Ferguson, Samir Bhatt, Seth Flaxman, et al. A COVID-19 model for local authorities of the United Kingdom. *[preprint]*, 2020. doi: 10.1101/2020.11.24.20236661. URL https://www.medrxiv.org/content/10.1101/2020.11.24.20236661v1.full.pdf.

[81] Juan Lopez-Martin, Justin H. Phillips, and Andrew Gelman. Multilevel Regression and Poststratification Case Studies, February, 2021. URL https://bookdown.org/jl5522/MRP-case-studies/introduction-to-mrp.html.

[82] David Adam. A guide to R — the pandemic's misunderstood metric. *Nature*, 583 (7816):346–348, -07-03 2020. doi: 10.1038/d41586-020-02009-w. URL https://www.nature.com/articles/d41586-020-02009-w.

Appendix A

Supplement

A.1 Related Tables

State	Election Results	Mean Value	Difference	November 3rd	Difference
Alabama	62.032	59.512	2.520	59.616	2.416
Arizona	49.056	48.302	0.754	48.086	0.970
Arkansas	62.396	57.276	5.120	60.456	1.940
California	34.321	34.391	-0.070	34.112	0.209
Colorado	41.604	44.064	-2.460	42.822	-1.218
Connecticut	39.187	39.641	-0.454	37.542	1.645
Delaware	39.775	37.742	2.032	36.646	3.129
Florida	51.220	48.086	3.134	48.416	2.803
Georgia	49.237	50.515	-1.277	49.152	0.086
Idaho	63.838	62.037	1.801	59.476	4.362
Illinois	40.553	39.064	1.489	40.388	0.165
Indiana	57.021	55.417	1.604	54.761	2.260
Iowa	52.800	50.870	1.930	49.962	2.838
Kansas	56.213	55.247	0.966	55.718	0.495
Kentucky	62.087	59.504	2.583	58.400	3.686
Louisiana	58.461	57.211	1.250	58.775	-0.314
Maine	43.551	44.333	-0.782	43.368	0.183
Maryland	32.150	35.186	-3.036	33.631	-1.481
Massachusetts	31.908	33.405	-1.497	31.834	0.074
Michigan	47.837	45.444	2.393	45.518	2.319
Minnesota	45.285	45.744	-0.459	44.633	0.652
Mississippi	57.603	55.591	2.012	56.787	0.816
Missouri	56.800	54.119	2.680	54.063	2.737
Montana	56.918	54.810	2.109	52.212	4.706
Nebraska	58.224	58.524	-0.299	58.180	0.045
Nevada	47.666	45.788	1.878	46.214	1.452
New Hampshire	45.356	45.979	-0.623	44.210	1.146
New Jersey	41.397	40.242	1.155	39.435	1.962
New Mexico	43.497	42.734	0.763	42.574	0.923
New York	37.461	36.284	1.177	34.955	2.507
North Carolina	49.934	49.134	0.801	48.769	1.165
North Dakota	65.114	60.664	4.450	59.626	5.488

Ohio	53.271	49.737	3.534	49.792	3.480
Oklahoma	65.373	62.823	2.550	62.083	3.290
Oregon	40.367	41.281	-0.914	39.267	1.101
Pennsylvania	48.844	47.041	1.803	47.326	1.517
Rhode Island	38.670	37.182	1.489	35.669	3.002
South Carolina	55.093	53.878	1.215	53.375	1.718
South Dakota	61.769	59.227	2.543	57.800	3.970
Tennessee	60.660	57.487	3.173	57.259	3.401
Texas	52.058	51.482	0.575	50.305	1.753
Utah	58.130	56.450	1.680	55.455	2.675
Vermont	30.381	31.408	-1.027	29.070	1.311
Virginia	43.996	43.754	0.242	43.289	0.706
Washington	38.767	36.996	1.771	37.004	1.763
West Virginia	68.632	65.413	3.219	64.486	4.146
Wisconsin	48.822	46.611	2.211	45.378	3.444
Wyoming	69.500	68.192	1.307	65.778	3.722

Table A.1: Summary of state-level Trump support forecasts and results in the 2020 Presidential election. The "Mean Value" column corresponds to the mean values of the daily forecasts obtained between June 1^{st} and November 3^{rd}, and "November 3^{rd} to the forecasts on that date. The "Difference" columns correspond respectively to the values from these columns subtracted from the election results.

State	ESS	Rhat
Alabama	250	1.015
Arizona	175	1.012
Arkansas	443	1.009
California	161	1.017
Colorado	189	1.020
Connecticut	222	1.028
Delaware	204	1.019
Florida	75	1.046
Georgia	102	1.015
Idaho	435	1.003
Illinois	87	1.068
Indiana	18	1.277
Iowa	101	1.059
Kansas	16	1.364
Kentucky	101	1.036
Louisiana	63	1.057
Maine	219	1.025
Maryland	237	1.016
Massachusetts	101	1.030
Michigan	51	1.020
Minnesota	138	1.033
Mississippi	158	1.032
Missouri	84	1.046
Montana	456	1.015

State	ESS	Rhat
Nebraska	157	1.023
Nevada	286	1.016
New Hampshire	664	1.004
New Jersey	29	1.130
New Mexico	408	1.009
New York	20	1.147
North Carolina	138	1.045
North Dakota	681	1.006
Ohio	35	1.133
Oklahoma	24	1.180
Oregon	290	1.010
Pennsylvania	104	1.042
Rhode Island	51	1.073
South Carolina	343	1.014
South Dakota	665	1.009
Tennessee	138	1.037
Texas	85	1.030
Utah	398	1.008
Vermont	1080	1.003
Virginia	14	1.267
Washington	228	1.034
West Virginia	80	1.060
Wisconsin	226	1.002
Wyoming	892	1.007

Table A.2: Effective Sample Size and $\hat{R}$ for the state-level week 45 R_t draws. The values are obtained using Stan's [61] `summary` function for MCMC draws.

	Model 1			Model 2		
Parameter	10%	50%	90%	10%	50%	90%
Intercept	1.4151	1.5970	1.7712	1.4090	1.5861	1.7673
HESIT						
NO_LIKE						
NO_VAX_TRUST						
NO_GOV_TRUST						
OTHERS_NEED						
VOTES_2020	-1.0693	-0.7238	-0.3716	-1.0545	-0.6988	-0.3554
CHANGE_2020				-1.5843	-0.1311	1.4170
R^2	0.0309	0.1067	0.1946	0.0464	0.1433	0.2682

	Model 3			Model 4		
Parameter	10%	50%	90%	10%	50%	90%
Intercept	1.4781	1.7473	2.0053	1.4855	1.7239	1.9770
HESIT	-1.7208	-1.1423	-0.5552	-1.3532	-0.5890	0.1714
NO_LIKE						
NO_VAX_TRUST						
NO_GOV_TRUST						
OTHERS_NEED						
VOTES_2020				-0.9301	-0.4490	0.0152
CHANGE_2020						
R^2	0.0286	0.1026	0.1948	0.0657	0.1648	0.2868

	Model 5			Model 6		
Parameter	10%	50%	90%	10%	50%	90%
Intercept	1.5217	1.7979	2.0499	1.6554	2.1023	2.5611
HESIT	-1.5323	-0.9886	-0.4197	-1.5034	-0.6601	0.1730
NO_LIKE				-2.6686	-1.2026	0.2458
NO_VAX_TRUST				-2.2304	-0.9706	0.3166
NO_GOV_TRUST	-1.6740	-0.6204	0.3850	-1.0945	0.4300	1.8549
OTHERS_NEED				-1.5047	-0.6827	0.1317
VOTES_2020				-0.7775	-0.2654	0;2498
CHANGE_2020				-1.5645	-0.1578	1.3504
R^2	0.0530	0.1515	0.2661	0.1360	0.2353	0.3447

Table A.3: Regression coefficients, state-level. Values obtained from 6 different Bayesian linear regression models implemented using `rstanarm` [70].

2-Week Period	R^2
1	0.0434
2	0.2865
3	0.2488
4	0.1837
5	0.1099
6	0.0917
7	0.1019
8	0.0546

2-Week Period	R^2
9	0.2299
10	0.0828
11	0.1560
12	0.1759
13	0.2476
14	0.1108
15	0.2464
16	0.0896

Table A.4: R^2 at each date for state-level time-varying regression

	Model 1			Model 2		
Parameter	10%	50%	90%	10%	50%	90%
Intercept	1.2471	1.2866	1.3250	1.3412	1.2864	1.4304
VOTES_2020	-0.1823	-0.1161	-0.0475	-0.0317	0.0446	0.1175
HES				-1.3158	-1.0615	-0.7888
STRONG_HES						
R2	0.0010	0.0054	0.0127	0.0234	0.0384	0.0556

	Model 3			Model 4		
Parameter	10%	50%	90%	10%	50%	90%
Intercept	1.2838	1.3276	1.3702	1.3557	1.3959	1.4390
VOTES_2020	-0.0849	-0.0051	0.0771			
HES				0.1787	-0.9785	-0.7564
STRONG_HES	-1.6908	-1.2463	-0.8070			
R2	0.0118	0.0229	0.0385	0.0191	0.0310	0.0460

	Model 5			Model 6		
Parameter	10%	50%	90%	10%	50%	90%
Intercept	1.2925	1.3259	1.3585	1.3542	1.4021	1.4492
VOTES_2020				-0.0448	0.0377	0.1192
HES				-2.1292	-1.5933	-1.0567
STRONG_HES	-1.6371	-1.2537	-0.8680	0.1251	1.0015	1.8688
R2	0.0095	0.0190	0.0312	0.0267	0.0417	

Table A.5: Regression coefficients, county-level. Values obtained from 6 different Bayesian linear regression models implemented using `rstanarm` [70].

A.2 Related Figures

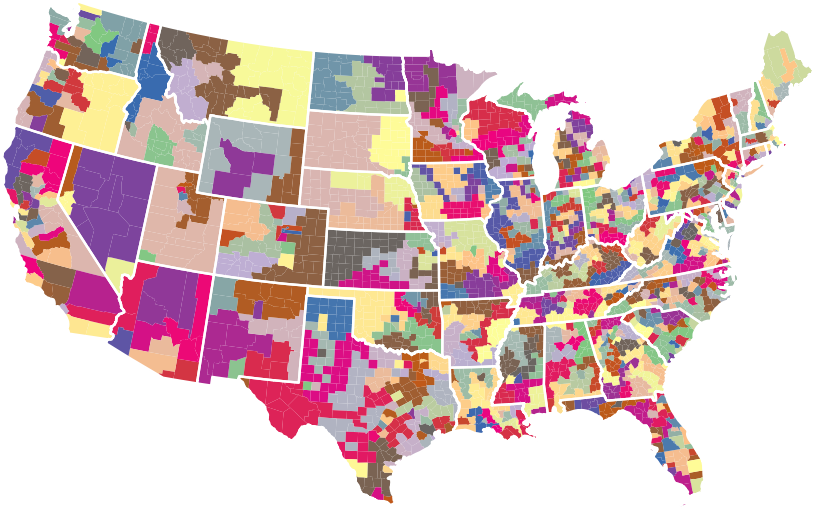
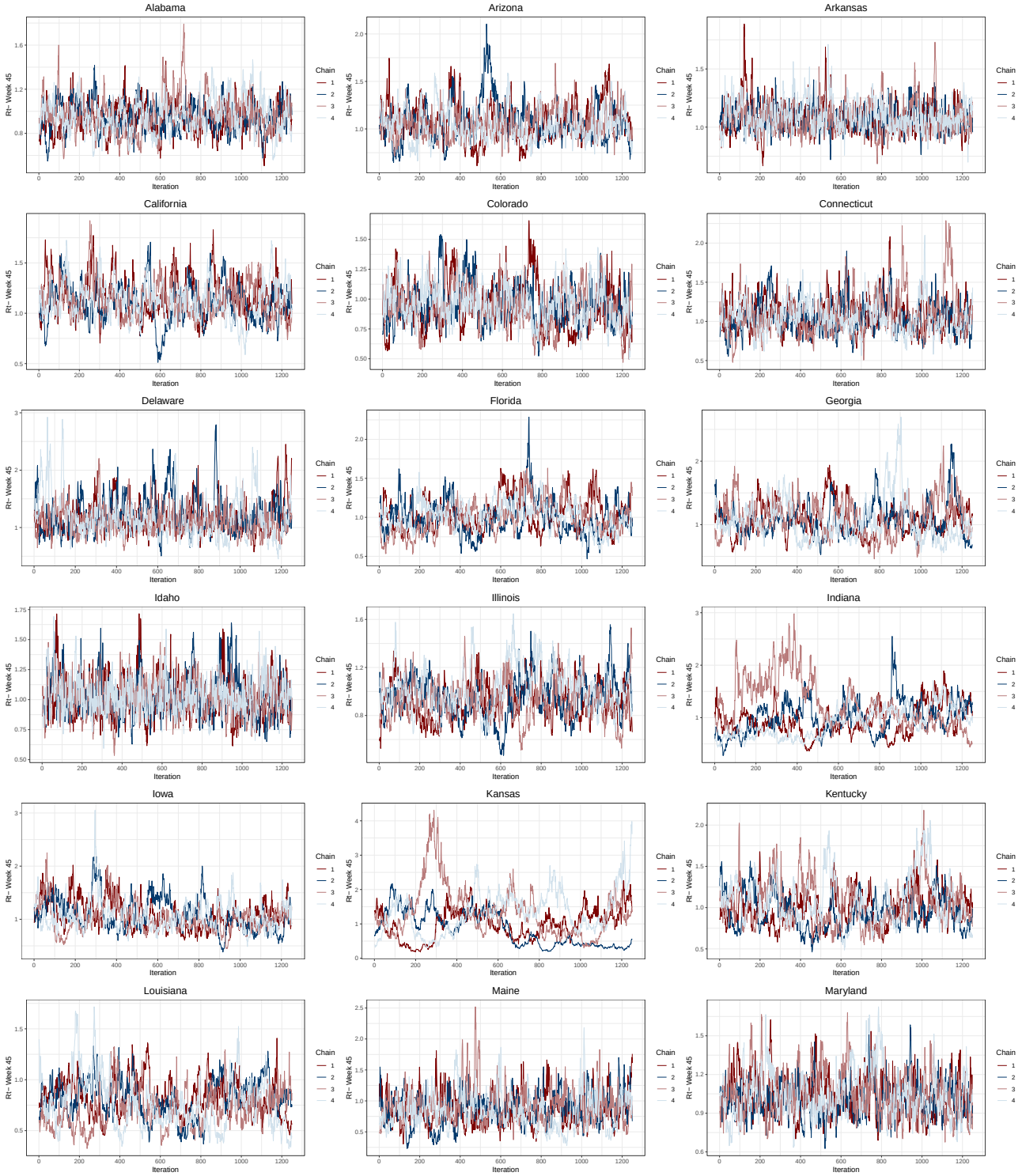
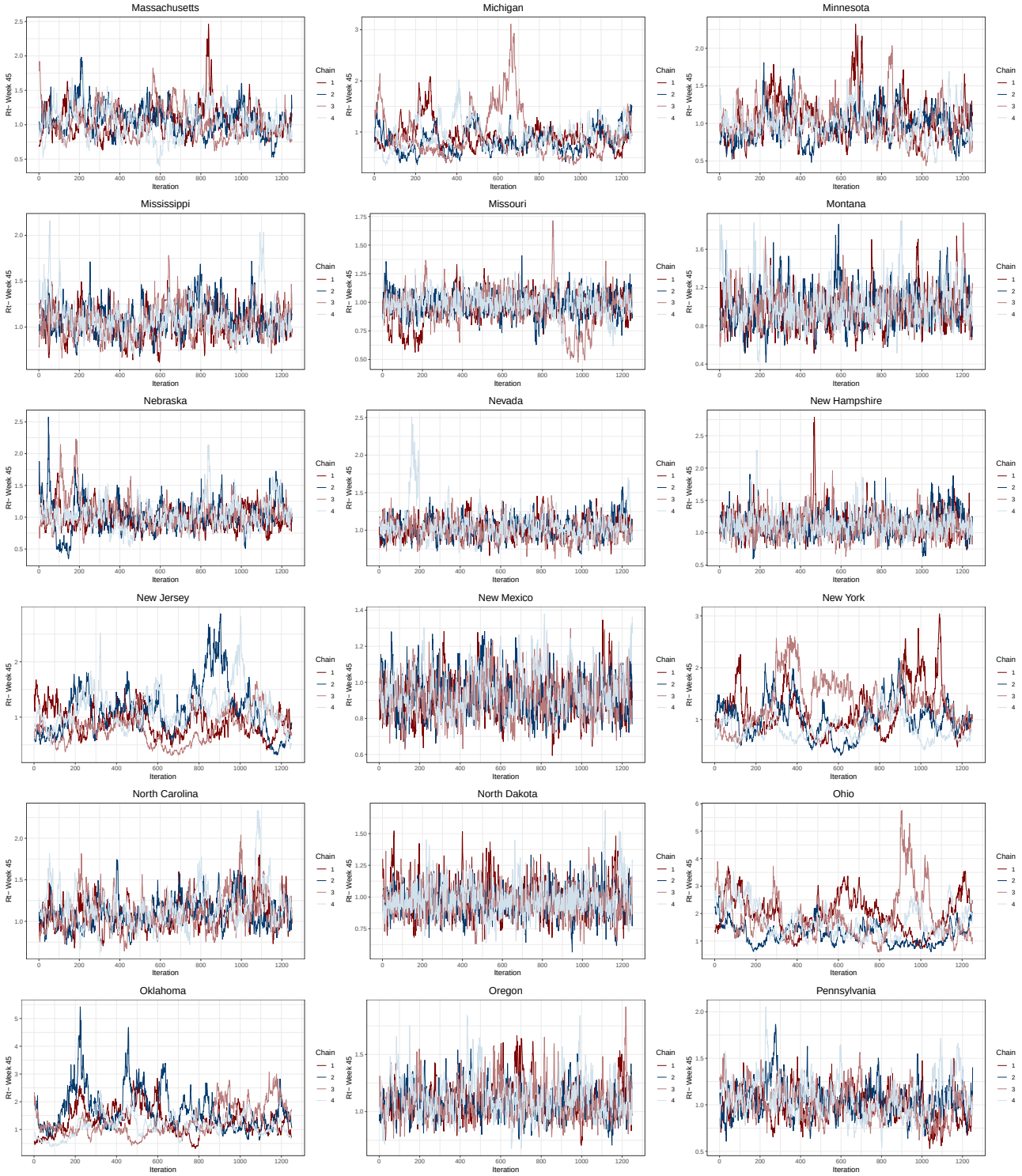


Figure A.1: Map of the USA representing the county grouping obtained for each state.


Alabama
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Rt~ Week 45
Iteration
Chain
1
2
3
4


Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rt~ Week 45
Iteration
Chain
1
2
3
4

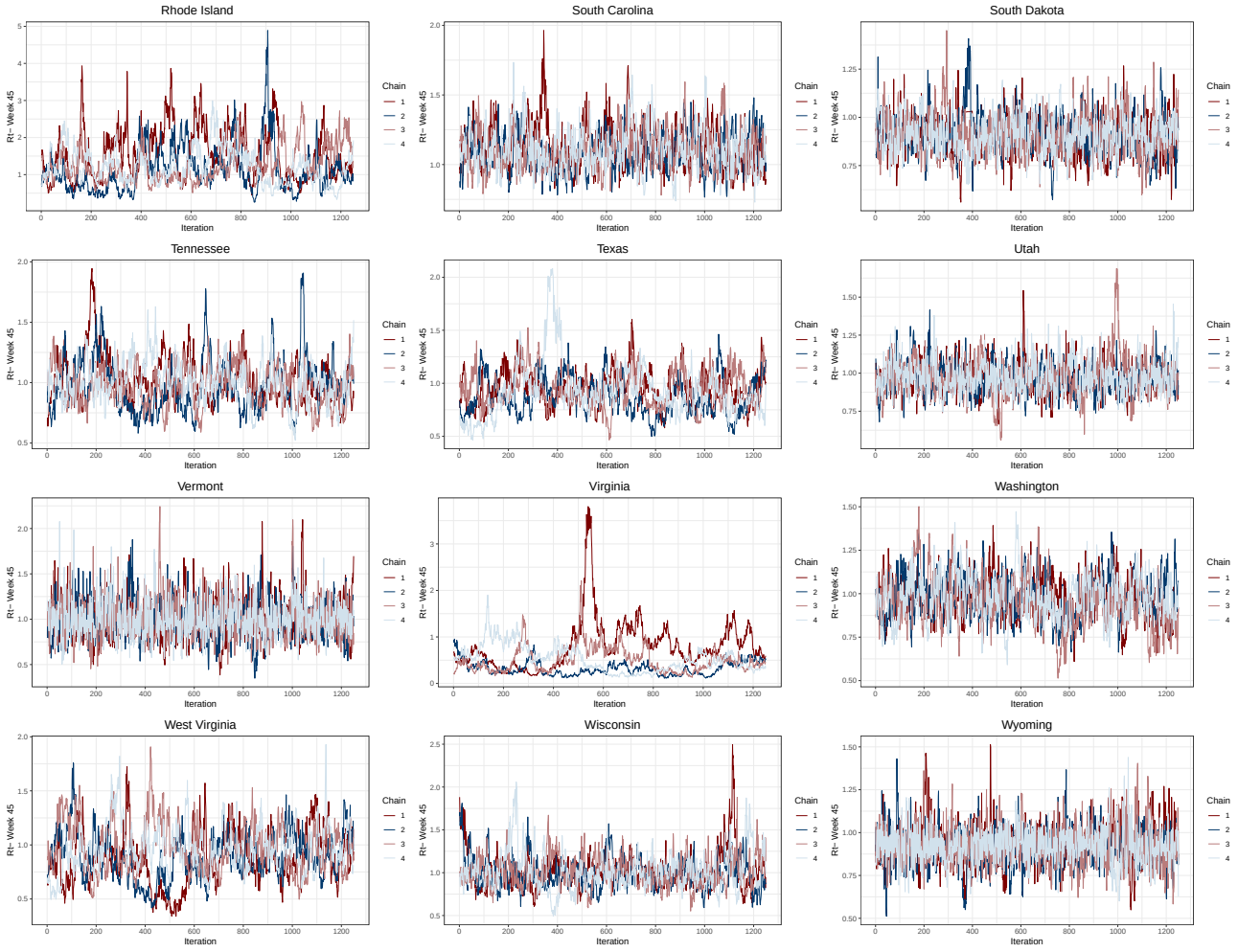


Figure A.2: Week 45 R_t traceplots for the contiguous states. The figure has been obtained using `bayesplot`[73], and corresponds to 1,250 iterations after burn-in has been discarded (1,250 iterations).

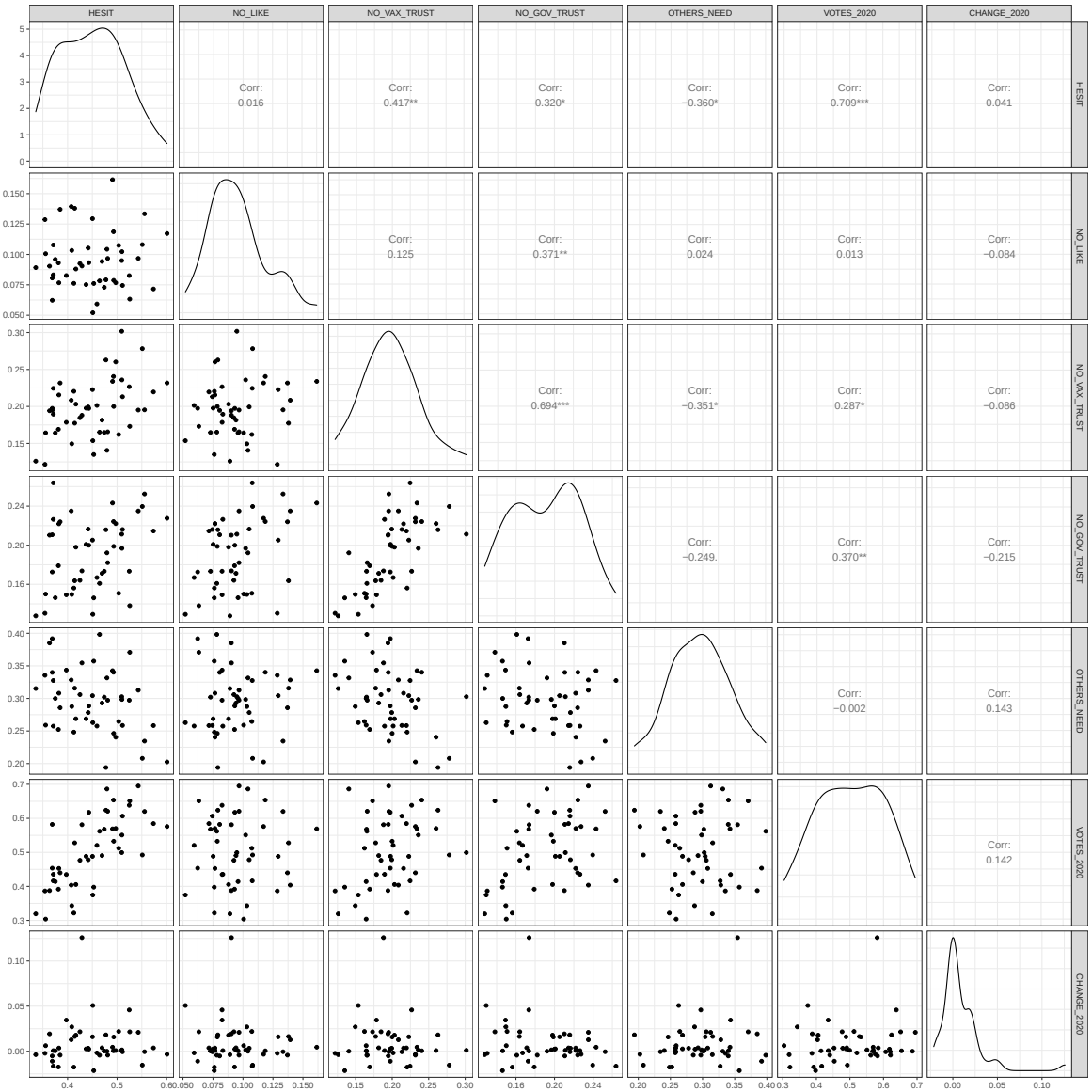


Figure A.3: Pairplot of the 8 covariates considered at the state-level. The diagonal figures represent the density estimate for each variable, and the scatter plots are constructed for each pair of variables. The figure is obtained using inbuilt features of `ggplot2` [74].

Appendix B

Code

As outlined at the beginning of the results section, the whole R code and data can be found in the following Github repository: https://github.com/emma-landry/M4R. In this appendix, I provide the implementations that aid with understanding the methods.

B.1 County Grouping Algorithm

```
#First data-preprocessing step
grouping_state <- function(cap,short){

  state_pop <- subset(pop_2019, state == cap) #select population data using
    capitalized full state name
  adj_state <- subset (adj, state == short) #select adjency using two letter
    state code

  adj_state <- subset(adj_state, county_fips != neighbor_fips) #remove each
    county from its own neighbors
  adj_state <- subset(adj_state, neighbor_state == short) #remove adjacent
    counties from diff state

  return(list(state_pop, adj_state))
}

#Combines the data into one data frame
grouping_step2 <- function(cap, short){
  #Pre-process data using grouping_state
  state_pop <- grouping_state(cap, short)[[1]]
  adj_state <- grouping_state(cap, short)[[2]]

  #join the two data frames and clean it
  state_df <-left_join(adj_state, state_pop, by = "county")
  state_df <- data.frame(state_df$county, state_df$county_fips, state_df$
    population, state_df$neighbor_county, state_df$neighbor_fips)
  colnames(state_df) <- c("county", "county_fips", "population", "neighbor_
    county", "neighbor_fips")

  state_df <- left_join(state_df, state_pop, by = c("neighbor_county"="county
    "))
  cnames <- colnames(state_df)
  cnames[3] <- "population"
  cnames[7] <- "neighbor_population"
  colnames(state_df) <- cnames
```

```
  state_df$state <- NULL

  return(state_df)
}

#Obtain the groupings
do_groups <- function(cap,short, tol){
  #Obtain fully prepared data frame
  state_df <- grouping_step2(cap, short)

  state_df$population <- as.numeric(state_df$population)
  min_pop <- min(state_df$population) #lowest population
  min_c <- state_df[which.min(state_df$population),]$county #index of county
    with lowest population
  i =1

  #iterate as long as some groupings have population less than tol
  while (min_pop <= tol){
    i = i+1
    min_c <- state_df[which.min(state_df$population),]$county #less populated
     county
    min_fips <-state_df[which.min(state_df$population),]$county_fips #fips of
     less populated county

    #Choose the neighbour with lowest population to merge
    tmp <- subset(state_df, county == min_c)
    min_neighbor <- tmp[which.min(tmp$neighbor_population),]$neighbor_county
    min_neighbor_fips <- tmp[which.min(tmp$neighbor_population),]$neighbor_
    fips
    group <- c(min_fips, min_neighbor_fips)

    tmp <- subset (state_df, county== min_c | county == min_neighbor)
    tmp <- subset (tmp , neighbor_county != min_c & neighbor_county != min_
    neighbor)
    string <- paste("Group", i)
    tmp$county <- rep(string, nrow(tmp))
    tmp$population <- rep ((tmp$population[1]+ tmp$population[nrow(tmp)]),
    nrow(tmp))

    tmp$county_fips <- rep(list(group), nrow(tmp))
    tmp <- unique(tmp)

    state_df<- subset(state_df, county != min_c & county != min_neighbor)
    state_df <- rbind(state_df, tmp)

    #We need to remove the grouped county as neighbor from others
    nrowmin<- length(state_df$neighbor_fips[state_df$neighbor_county == min_c
    ] )
    nrowminneigh <- length(state_df$neighbor_fips[state_df$neighbor_county ==
     min_neighbor])

    state_df$neighbor_fips[state_df$neighbor_county == min_c] <- rep(list(
    group),nrowmin)
    state_df$neighbor_fips[state_df$neighbor_county == min_neighbor] <- rep(
    list(group), nrowminneigh)

    state_df$neighbor_population[state_df$neighbor_county == min_c] <- rep(
    tmp$population[1], nrowmin)
    state_df$neighbor_population[state_df$neighbor_county == min_neighbor] <-
     rep(tmp$population[1], nrowminneigh)
```

```
77
78      state_df$neighbor_county[state_df$neighbor_county == min_c] <- rep(string
        , nrowmin)
79      state_df$neighbor_county[state_df$neighbor_county == min_neighbor] <- rep
        (string, nrowminneigh)
80
81      state_df= unique(state_df)
82
83      min_pop <- min(state_df$population)
84    }
85    return(state_df)
86  }
```

B.2 County level `epidemia` model

```
1  library(epidemia)
2  library(dplyr)
3  library(here)
4  library(readr)
5
6  options(mc.cores = parallel::detectCores())
7  options(mc.cores = 4)
8
9  #Obtain job details
10 job.index <- as.numeric(Sys.getenv("PBS_ARRAY_INDEX"))
11 print(job.index)
12 job.id <- Sys.getenv("PBS_JOBID")
13
14
15 short_state <- c("AL", "AK","AZ", "AR", "CA","CO","CT","DE","FL","GA","ID","
       IL","IN","IA","KS","KY",
16                   "LA","ME","MD","MA","MI","MN","MS","MO","MT","NE","NV","NH",
       "NJ","NM","NY","NC","ND",
17                   "OH","OK","OR","PA","RI","SC","SD","TN","TX","UT","VT","VA",
       "WA","WV","WI","WY")
18
19 long_state <- c("alabama", "alaska", "arizona", "arkansas", "california", "
       colorado", "connecticut",
20                   "delaware", "florida", "georgia", "idaho", "illinois","
       indiana","iowa", "kansas",
21                   "kentucky", "louisiana", "maine", "maryland", "massachusetts"
       , "michigan",
22                   "minnesota", "mississippi", "missouri","montana", "nebraska",
        "nevada", "new_hampshire",
23                   "new_jersey", "new_mexico", "new_york", "north_carolina", "
       north_dakota", "ohio",
24                   "oklahoma", "oregon", "pennsylvania", "rhode_island", "south_
       carolina",
25                   "south_dakota", "tennessee", "texas", "utah", "vermont", "
       virginia", "washington",
26                   "west_virginia","wisconsin", "wyoming")
27
28 #Identify the state being worked with
29 this.short <- short_state[job.index]
30 this.long <- long_state[job.index]
31
32 #Load grouping and deaths data for the given state
33 group.file.name<- paste0(this.long, "_groups.rds")
```

```
34
35 groups <- readRDS(here::here("data", "grouped_counties", group.file.name))
36 data <- readRDS(here::here("data", "covid", "prepped_data_Mar21.rds"))
37
38 state_data <- filter(data, state == this.short)
39
40 N <- nrow(groups)
41
42 #Loop over the N counties, to obtain a model for each
43 for (i in (seq (1:N))){
44
45   #Obtaining the data
46   fips <- unlist(groups$county_fips[i])
47
48   if (length(fips)==1){
49     data_sub <- filter(state_data, countyFIPS == fips)
50   }else{ #aggregate the data for all counties in the grouping
51     data_sub <- filter(state_data, countyFIPS %in% fips)
52     tmp <- aggregate(cbind(cases, deaths) ~ date, data=data_sub, FUN=sum, na.
       action = NULL)
53     data_sub <- data_sub[1:435,]
54     data_sub <- subset(data_sub, select = -c(countyFIPS, date,cases, deaths,
       state))
55     data_sub <- cbind(data_sub, tmp)
56     n <- nrow (data_sub)
57     data_sub$county <- rep(groups$county[i], n)
58   }
59   #Checking number of cases is sufficient
60   cases_no_NA = data_sub$cases
61   cases_no_NA[is.na(cases_no_NA)]=0
62   idx <- which(cumsum(cases_no_NA) >= 10)[1]
63
64   if (length(idx) == 0) {
65     stop(paste0("Fewer than 10 cumulative cases in entire epidemic. Not
       modeling."))
66   }
67
68   #Number the weeks for the weekly random walk
69   start_date <- data_sub$date[idx] - 7
70   data_sub <- filter(data_sub, date >= start_date)
71
72   data_sub <- mutate(data_sub, week = as.integer(format(date, "%V")))
73   new_year <- min(which(data_sub$week ==1))
74   data_sub[new_year:nrow(data_sub),]$week <- data_sub[new_year:nrow(data_sub)
       ,]$week +53
75
76   pops <- data.frame("county" = groups$county[i], "population"= groups$
       population[i])
77   data_sub$pop<- pops$population
78
79   args <- list(data=data_sub)
80
81   #Model for latent infections
82   inf <- epiinf( gen = EuropeCovid$si,
83                  pop_adjust = FALSE,
84                  susceptibles = pop) #check prior_tau
85
86   #Model for observations (deaths) vector
87   deaths <- epiobs(
88     #formula = deaths(county, date) ~ 1,
```

```
    formula = deaths ~ 1,
    family = "neg_binom", # overdispersion for daily counts
    i2o = EuropeCovid$inf2death,
    prior_intercept = rstanarm::normal(1, 0.5),
    link = "identity"
  )


  args$obs <- deaths

  args$rt <- epirt(
    formula = R(county, date) ~ rw(time=week)
  )

  #Define the HMC parameters
  args$algorithm <- "sampling"
  args$init_run <- TRUE
  args$inf <- inf
  args$iter <- 2.5e3
  args$chains <- 4
  args$seed <- 12345

  filename <- paste0(gsub("\\s", "_", groups$county[i]), "_", this.short,"_",
    job.id,".rds")

  fit <- do.call("epim", args) #fit the epidemiological model

  res <- list(
    fit = fit,
    county = groups$county[i],
    state = this.short
  )

  wd <- getwd()
  setwd("..")
  setwd("..")
  parent <- getwd()
  setwd(wd)

  #Saving the model to the Research Data Store
  saveRDS(res, file =  paste0(parent,"/Outputs/epidemia_fits/run7/", filename
    ))

  #SAving posterior medians to the Research Data Store
  rt <- posterior_rt(res$fit)
  draws <- rt$draws
  rt_medians <- apply(draws,2, median)
  rt_df <- data.frame("Date"= rt$time, "Rt_medians"= rt_medians)


  filename2 <- paste0(gsub("\\s", "_", groups$county[i]), "_", this.short,"_"
    ,job.id,"_medians.rds")
  write.csv(rt_medians, file =  paste0(parent,"/Outputs/rt_medians/run7/",
    filename2))
}
```

B.3 PBS file for Submitting Jobs to HPC

```
#!/bin/sh
#PBS -l walltime=72:00:00
#PBS -lselect=1:ncpus=4:mem=16gb
#PBS -J1-49

module load anaconda3/personal
source activate epidemia_env

cd $HOME/Msc/M4R/code
Rscript county_fit_deaths_only.R
```