Bayesian Spatiotemporal, Sample Survey, and Forecasting Methods for Analyzing COVID-19 Infections and Mortality

by

Justin James Ian Slater

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Statistical Sciences
University of Toronto

Bayesian Spatiotemporal, Sample Survey, and Forecasting Methods for Analyzing COVID-19 Infections and Mortality

Justin James Ian Slater
Doctor of Philosophy

Department of Statistical Sciences
University of Toronto
2023

Abstract

For decades, mathematicians and statisticians have been modelling infectious diseases to forecast case/death counts, estimate important epidemiological quantities, and understand the dynamics of disease spread. This dissertation offers methodological insights into each of these three challenges using novel spatial, spatio-temporal, and Bayesian modelling methods, with applications to COVID-19 data. Alongside methodological contributions, this thesis also presents estimates of important epidemiological quantities which, subject to peer review, could be utilized by public health professionals and policy makers. There are four primary contributions of this work: 1) a subnational, single-wave COVID-19 mortality forecasting model that accounts for day-of-the-week effects, which was shown to outperform the most highly-cited model during the first viral wave; 2) a mobility-augmented spatial model for COVID-19 case counts, where cellphone-derived mobility data is shown to capture dependence between areal units better than physical proximity; 3) a novel, interpretable spatio-temporal infectious disease model where infectiousness is a function of mobility between areal units, resulting in estimates of the risk associated with travelling in two Spanish Communities; 4) a modular Bayesian framework based on mixture modelling of serological data and disaggregated deaths data to estimate COVID-19 incidence and infection fatality rates, resulting in estimates of these quantities across Canada for various strata. Although the applications in this thesis are to COVID-19 data, the proposed methodology can be applied to a wide spectrum of problems across infectious disease epidemiology.

Acknowledgements

This thesis would not have happened without my incredible support system of supervisors, collaborators, friends, and family. I would now like to thank them in no particular order.

My only hope in my academic career is that I can provide my future students with the support and guidance that was given to me by my supervisors, Drs. Patrick Brown and Jeff Rosenthal. The fact that two of the brightest statisticians on Earth would guide me through this experience is truly humbling, and has shaped the course of my career for the better.

In addition to the financial support provided by my supervisors, I am grateful for NSERC funding a portion of my studies.

This work was done with the help of key collaborators that I would like to thank. Jorge Mateu provided me with novel datasets and ideas that lay the foundation for the work done in Chapters 3 and 4 of this thesis. Aiyush Bansal, Harlan Campbell, and Paul Gustafson provided medical and methodological guidance for Chapter 5 of this work. I am hoping that this is only the beginning of our collaborations and friendships.

I would like to thank my committee member Monica Alexander for her questions/advice along the way, and for teaching the most useful/relevant course that I have ever taken. I would also like to thank my external examiner, Dr. Andrea Riebler for her insightful questions/comments on my thesis, and for writing the paper that inspired Chapter 3 of this work.

I would like to thank Dr. So-hee Kang for being the most enthusiastic teacher I have ever met, and for being my teaching mentor who has inspired me to explore innovative teaching methods in the future.

A big thank you to my fellow students and members of the Bayesian Reading Group. I know that the past few years were not ideal conditions for doing a PhD, but knowing that I made life-long friends and future collaborators along the way made it all worth it.

I would like to thank my parents, Fred and Anna, and brothers, Evan and Alex, for their unconditional support of my decision to pursue a career in academia. I wasn't always the best student, but my family always did everything that they could to help me succeed

in my academic pursuits, and ensured that I got into university in the first place. I owe everything to them.

Finally, I would like to thank my partner, best friend, and colleague, Emily. Making you proud has always been my number one motivator, and I will continue to do so throughout my academic career. You made this experience everything that I could ask for and more. Now it's your turn.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

For decades, mathematicians and statisticians have been modelling infectious diseases to forecast case/death counts, estimate important epidemiological quantities, and understand the dynamics of disease spread. This dissertation offers methodological insights into each of these three challenges using novel spatial, spatio-temporal, and Bayesian modelling methods, with applications to COVID-19 data. Alongside methodological contributions, this thesis also presents estimates of important epidemiological quantities which, subject to peer review, could be utilized by public health professionals and policy makers. There are four primary contributions of this work: 1) a subnational, single-wave COVID-19 mortality forecasting model that accounts for day-of-the-week effects, which was shown to outperform the most highly-cited model during the first viral wave; 2) a mobility-augmented spatial model for COVID-19 case counts, where cellphone-derived mobility data is shown to capture dependence between areal units better than physical proximity; 3) a novel, interpretable spatio-temporal infectious disease model where infectiousness is a function of mobility between areal units, resulting in estimates of the risk associated with travelling in two Spanish Communities; 4) a modular Bayesian framework based on mixture modelling of serological data and disaggregated deaths data to estimate COVID-19 incidence and infection fatality rates, resulting in estimates of these quantities across Canada for various strata. Although the applications in this thesis are to COVID-19 data, the proposed methodology can be applied to a wide spectrum of problems across infectious disease epidemiology.

1.1 Thesis Outline

This thesis is divided into four chapters, each corresponding to a published or submitted journal article that addresses a different statistical modelling challenge pertaining to COVID-19. Each chapter stands independently, with its own abstract, bibliography, and appendices.

In Chapter 2, we develop a parameter-driven model that accurately and consistently estimates COVID-19 mortality at the subnational level early in the pandemic, using only daily mortality counts as the input. We use a Bayesian hierarchical skew-normal model with day-of-the-week parameters to provide accurate projections of COVID-19 mortality. We validate our projections by comparing our model to the projections made by the Intitute for Health Metrics and Evaluation, and highlight the importance of hierarchicalization and day-of-the-week effect estimation when forecasting COVID-19 mortality.

In Chapter 3, we develop a mobility-augmented spatial model for COVID-19 case counts. We investigate the efficacy of using cellphone-derived mobility data to model dependence between areal units in spatial models for COVID-19. We do this by extending Besag York Mollié (BYM) (Besag et al., 1991) models to include both a physical adjacency effect and a mobility effect. The mobility effect is given a Gaussian Markov random field model, with the number of trips between regions used as edge weights. Using two Spanish Communities as examples, we leverage modern parametrizations of BYM models (Riebler et al., 2016) and find that the number of people moving between regions better explains variation in COVID-19 case counts than physical proximity data. We conclude that these data should be used in conjunction with physical proximity when developing spatial models for COVID-19 case counts.

In Chapter 4, we build a spatio-temporal mechanistic model using cellphone-derived mobility networks. One limitation of the model developed in Chapter 3, was that it does not capture the underlying dynamics of disease spread, which we aim to overcome in this chapter. We extend the Endemic-Epidemic modeling framework in a principled manner, incorporating temporally changing mobility network data. We do so by deriving our model from first principles as done in Bauer and Wakefield (2018), and quantify the risk associated

with travelling throughout the first year of the pandemic in two Spanish Communities.

In Chapter 5, we develop statistical methodology to estimate incidence and infection fatality rates (IFR) in Canada for various demographic groups. This is done using serological data from the Action to Beat Coronavirus serosurvey conducted by the Centre for Global Health Research. We develop a modular Bayesian framework where the number of infections (incidence) is estimated based on multivariate mixture models fit to antibody test results. We then combine these estimates into a model using disaggregated deaths data to estimate IFR. In doing so, we account for uncertainty from both the estimated number of infections and incomplete deaths data to provide estimates of IFRs, while not allowing the deaths data to influence incidence estimates.

In Chapters 4 and 5, we provide estimates of epidemiologically important quantities that are, at the time of writing this thesis, still subject to peer review and should not be cited or used by policy makers.

1.2 Attribution

My work is supported by the Natural Sciences and Engineering Research Council (PGSD3-559264–2021, Chapters 3-5), and the Centre for Global Health Research (Chapter 5). This work is the aggregation of four research papers, each corresponding to its own chapter. For each project, I was the primary contributing author.

Chapter 2 corresponds to Slater et al. (2021), which is based on joint work with Patrick E. Brown and Jeffrey S. Rosenthal. I was responsible for statistical analysis, methodological development, and writing the research paper. Brown proposed the high level idea of the project, and suggested that I work on it. Brown and Rosenthal were in supervisory roles, providing modelling suggestions, advice on framing results, and guidance regarding the general direction of the work.

Chapters 3 and 4 correspond to a published paper (Slater et al., 2022b) and a manuscript in review (Slater et al., 2022c), respectively. Both chapters are based on joint work with Patrick E. Brown, Jeffrey S. Rosenthal, and Jorge Mateu. I was responsible for model development and implementation, as well as writing the research papers. Mateu acquired the

data and provided continued advice regarding interpretation of the data. Brown provided the high level idea for Chapter 3, while the idea for Chapter 4 was my own. For both chapters, Brown, Mateu, and Rosenthal all provided modelling suggestions, advice on framing results, and guidance regarding the general direction of the work.

Chapter 5 corresponds to a manucript in review (Slater et al., 2022a), and was joint work with Aiyush Bansal, Harlan Campbell, Patrick E. Brown, Jeffrey S. Rosenthal, and Paul Gustafson. I was responsible for statistical analysis, modelling framework development and implementation, and writing the research paper. Data for this study was provided by the Centre for Global Health Research. Bansal provided medical expertise, acquired data pertaining to COVID-19 deaths, helped with interpretation of the serosurvey data, and contributed to the literature review. Brown suggested the initial modelling framework, which was iterated on and improved over the course of the project. Campbell provided key methodological critique which lead to great improvements in the work. Brown, Rosenthal, and Gustafson provided modelling suggestions, advice on framing results, and guidance regarding the general direction of the work.

1.3 Bibliography

Bauer, C. and Wakefield, J. (2018). Stratified space–time infectious disease modelling, with an application to hand, foot and mouth disease in China. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1379–1398.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165.

Slater, J. J., Bansal, A., Campbell, H., Rosenthal, J. S., Gustafson, P., and Brown, P. E. (2022a). A Bayesian approach to estimating COVID-19 incidence and infection fa-

tality rates. *Pre-print*. http://probability.ca/jeff/ftpdir/mixture_manuscript_submitted.pdf.

Slater, J. J., Brown, P. E., and Rosenthal, J. S. (2021). Forecasting subnational COVID-19 mortality using a day-of-the-week adjusted Bayesian hierarchical model. *Stat*, 10(1):e328.

Slater, J. J., Brown, P. E., Rosenthal, J. S., and Mateu, J. (2022b). Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. *Spatial Statistics*, 49:100540.

Slater, J. J., Brown, P. E., Rosenthal, J. S., and Mateu, J. (2022c). Leveraging cellphone-derived mobility networks to assess COVID-19 travel risk. *Pre-print*. http://probability.ca/jeff/ftpdir/mixture_manuscript_submitted.pdf.

Chapter 2

Forecasting subnational COVID-19 mortality using a day-of-the-week adjusted Bayesian Hierarchical model

Abstract

As of October 2020, the death toll from the COVID-19 pandemic had risen over 1.1 million deaths worldwide. Reliable estimates of mortality due to COVID-19 are important to guide intervention strategies such as lockdowns and social distancing measures. In this chapter, we develop a parameter-driven model that accurately and consistently estimates COVID-19 mortality at the regional level early in the epidemic, using only daily mortality counts as the input. We use a Bayesian hierarchical skew-normal model with day-of-the-week parameters to provide accurate projections of COVID-19 mortality. We validate our projections by comparing our model to the projections made by the Intitute for Health Metrics and Evaluation, and highlight the importance of hierarchicalization and day-of-the-week effect estimation.

2.1 Introduction

As of October 2020, the death toll from the COVID-19 pandemic had to risen over 1.1 million deaths worldwide, with deaths in many regions rising again following decreases in late spring and early summer. Although this number is likely an under estimate of the true number of deaths, it is more reliable than the reported number of cases, which is largely a function of the number of people tested. Reliable estimates of mortality due to COVID-19 are useful for guiding intervention strategies such as lockdowns and social distancing measures. Estimates are needed at the regional (e.g provincial or state) level, as the spread of the disease can vary greatly within a particular country.

There have been many attempts at forecasting COVID-19 cases and mortality. Extensions of Susceptible, Exposed, Infectious, or Recovered (SEIR) models have been considered (Anastassopoulou et al., 2020; Sarkar et al., 2020). Various time series models have also been considered (Perc et al. 2020; Petropoulos and Makridakis 2020; Chakraborty and Ghosh 2020). Perhaps most notably, the Institute for Health Metrics and Evaluation (IHME) has made their predictions available since March 25th 2020 (Friedman et al. 2020), and have been cited as the gold standard regional level projections. However, in all of these forecasting methods, there has been little attempt at accounting for differences in deaths by day-of-the-week, which if left unaccounted for, can drastically bias long-term forecasts depending on which day of the week the observed data ends. Additionally, making projections for regions can be difficult where there are a relatively small number of deaths. COVID-19 forecasting methodology needs to be able to handle low daily mortality counts, while providing reasonable mortality estimates for each region.

Figure 2.1 shows the COVID-19 daily death counts in the United States from March 2nd to June 25th. There are several key features of these daily deaths that seem to be prevalent in every country or region's Coronavirus mortality counts. Firstly, note the rapid rise in daily deaths relative to the decline. Capturing this skewness in the daily death counts is essential for accurately forecasting COVID-19 mortality, and is not captured using the Normal density initially used by the IHME (note that the IHME has since switched to an SEIR model). Additionally, notice the weekly periodicity in daily death counts. It appears

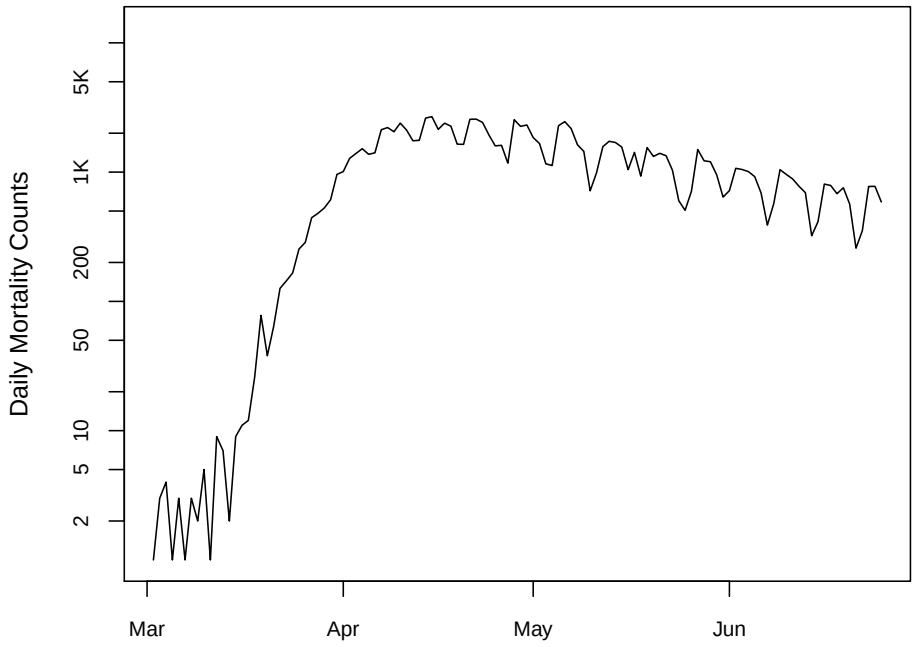


Figure 2.1: United States daily COVID-19 mortality from `www.coronavirus.app` (Scriby, Inc., 2020) from March 1, 2020 to June 25th 2020

that certain days of the week tend to have higher death counts than others, which is an important feature to capture, so that our cumulative death forecasts don't depend on what day of the week they are made. This weekly periodicity has been confirmed using spectral analysis (Bukhari et al. 2020), but the relative risks of mortality between certain days-of-the-week are still indeterminate. Studying day-of-the-week effects has proven useful in other fields such as actuarial science (Crevecoeur et al. 2019) and economics (Berument and Kiymaz 2001).

The goal of this chapter is to develop a data-driven model that accurately and reliably estimates COVID-19 mortality at the regional level early in the epidemic, using daily mortality counts as the input. We do so by developing a hierarchical Bayesian model where the daily death counts are assumed to follow a skew-normal density function, and vary by day-of-the-week. In doing so, we estimate the number of daily deaths at the peak of the epidemic, the date of the peak of the epidemic, and other epidemiologically significant features. The hierarchical nature of our model will allow for accurate estimation of cumu-

lative death counts in regions where the epidemic is still in the early stages by borrowing information from regions where the epidemic has matured. We compare the forecasting performance of our model to the projections made by the Institute for Health Metrics and Evaluation (IHME), as well as the observed mortality values and highlight the importance of hierarchicalization and accounting for day-of-the-week effects when projecting mortality counts during an epidemic.

This work is an extension of the model presented in Brown et al. 2020, which has shown promise as a national-level forecasting model. The two main contributions of our work are to allow for subnational level data (hierarchicalization), and estimation/modelling of day-of-the-week effects.

2.2 Methods

2.2.1 The Skew-Normal Model

We saw in Figure 2.1 that a key feature of daily coronavirus mortality counts is the rapid rise relative to the fall in daily death counts. A natural choice for the response distribution of the deaths per day is the negative binomial distribution with the trend in deaths per day in a region following a skew normal curve. The negative binomial is often used in infectious disease modelling, where events are positively correlated, causing larger variances than if the events were independent (Lloyd-Smith et al., 2005). The skew-normal density provides a good base to model the trend in death counts, but is far too simple to capture the full range of shapes of epidemics by itself. Firstly, although a majority of the deaths in a region occur during the main epidemic, a small number of deaths can occur outside of this epidemic. Additionally, in order to be able to compare various regions' mortality counts, we need to "standardize" our estimates based on how many deaths we would expect to see in that region from all causes. Lastly, we need to add a multiplicative term to our skew-normal that accounts for differences in daily mortality counts by day-of-the-week. By including all of these considerations, we arrive at the full model.

The statistical model (referred to as the "DOW model") used to estimate daily mortality,

	Parameter	Prior	Description
	$A_{i,j}$	$N(\text{Mar } 29, 100^2)$	Location Parameter
	$A_{i,Brazil}$	$N(\text{Jun } 17, 45^2)$	Location Parameter
	$\frac{1}{\sqrt{\tau_j}}$	Exp(1/10)	Overdispersion Parameter
	D_{ij}	Exp(1/10000)	Spark Term
C_{ij}	α_j	$N_+(50, 40^2)$	Mean of C_{ij}, the severity parameters
	θ_C	$N_+(2, 0.66^2)$	Scale of the severity parameters
B_{ij}	η_j	$N_+(60, 30^2)$	Mean of B_{ij}, the duration parameter
	θ_B	$N_+(9, 3^2)$	Scale of the duration parameters
K_{ij}	ζ_j	$N_+(3, 2^2)$	Mean of the K_{ij}, the skewness parameters
	θ_K	$N_+(3, 2^2)$	Scale of the skewness parameters
	$R_{j,m}$	$N_+(1, 2^2)$	Day-of-the-week parameters

Table 2.1: Prior distributions for the DOW model in (1)

Y_{ij}, in region i of country j is given by:

$$\begin{aligned} Y_{ij} &\sim \text{NegBinom}[\lambda_{ij}(t), \tau_j] \\ \lambda_{ij}(t) &= R_{j,m[t]} E_{ij} [C_{ij} f(t; A_{ij}, B_{ij}, K_{ij}) + D_{ij}] \\ C_{ij} &\sim \text{Gamma}(\alpha_j/\theta_C, \theta_C) \\ B_{ij} &\sim \text{Gamma}(\eta_j/\theta_B, \theta_B) \\ K_{ij} &\sim \text{Gamma}(\zeta_j/\theta_K, \theta_K) \end{aligned} \tag{2.1}$$

Prior distributions for model parameters can be found in Table 2.1. The function f is a skew-Normal density function with three parameters: the "location" parameter, A_{ij}, indicate the date at which the daily deaths reaches its peak and is analogous to the mean of a normal distribution; B_{ij} represents the duration of the epidemic, and is analogous to the standard deviation in the Normal density; and K_{ij} is the "skewness" parameter, which describes the ratio of the initial incline relative to the decline. The skewness is a key parameter for capturing the shape of daily mortality trends. The E_{ij}'s are the age standardized death counts in each region of the included countries. This was computed by obtaining age distribution information from census data of each country and comparing it to the deaths-by-age breakdown in Italy on March 29, 2020. Note that E_{ij} is a constant, as it is calculated apriori for each region. Inclusion of the E_{ij} allows for comparable heights of peaks between regions, which is captured in the parameter C_{ij}. A high C_{ij} means that there

were a large number of coronavirus related deaths, relative to the expected number of deaths in that region. The parameter D_{ij}, known as the "spark" term, captures the few deaths that were outside of the main epidemic. $R_{j,m[t]}$ captures the day-of-the-week effect for country j on day-of-the-week $m[t]$, with $R_{j,Sunday}$ fixed at 1. The overdispersion parameter, τ_j, allows the variance of the daily mean deaths to vary by country by a multiplicative factor.

The number of parameters in this model can grow very quickly depending on the number of countries/regions included in analysis, and can be difficult to implement without carefully chosen priors. One of the advantages to Bayesian analysis is the incorporation of prior knowledge to guide parameter estimation. That is, we can use information about duration and severity from epidemics that are already over (e.g Spain) to set priors for regions where the epidemic has yet to peak (e.g Brazil).

Note that C_{ij}, B_{ij} and K_{ij} are all modelled hierarchically. The advantage of this is that for regions low death counts can "borrow" information from other regions in the same country to estimate the severity, duration, and skewness of the epidemic. For example, the estimate of C_{ij} will be a weighted average between the country's mean and the region's mean, but will tend more toward the country's mean when the number of events is small (Gelman and Hill, 2006). Modelling A_{ij} hierarchically was considered, however it was deemed inappropriate because in regions with small death counts, the location parameter would tend toward the country average. This is problematic because the reason that the death counts are low in that region is likely because the epidemic has yet to run its course. For this reason, we decided to estimate the location parameter separately for each region. Modelling the day-of-the-week effects, $R_{j,m[t]}$, hierarchically was also considered because it would provide a day-of-the-week effect estimate for each region. However, this was deemed computationally too cumbersome, as this would add over 300 parameters to our model.

2.2.2 Daily mortality projections for four countries

We applied our model to daily COVID-19 death counts from 95 regions from four countries: U.S states; Canadian provinces; Spanish Autonomous Communities; and Brazilian states. These four countries were chosen based on demographic data availability at the regional level. Data are from `www.coronavirus.app` (Scriby, Inc., 2020), where any region with 50

or more deaths as of June 25th was included in the analysis. Parameters for our models were estimated using No-U-turn sampling (Hoffman and Gelman, 2014) within the Stan software (Carpenter et al., 2017). Four chains were used with 3000 iterations of warm-up and 1000 iterations of sampling, which were then thinned by a factor of 10 (leaving 400 posterior samples for each parameter). Convergence of Markov Chains was assessed using trace plots alongside the Gelman-Rubin Statistic ($\hat{R} < 1.05$) (Gelman et al., 1992).

Forecasts were created from the posterior samples of $\lambda_{ij}(t)$ up until October 1st 2020. Given that λ_{ij} has a day-of-the-week effect, the posterior samples of $\lambda_{ij}(t)$ will be oscillatory. Forecasts were also made at the country level by computing

$$\bar{\lambda}_j(t) = \sum_i \lambda_{ij}(t)$$

for each posterior sample.

2.2.3 Day-of-the-week effect estimates

The model produces estimates for 24 day of the week effect parameters: one for each day of the week (except for Sunday which is fixed at 1) for each of the four countries. In order to gain some insight as to whether the estimated day-of-the-week effects are simply due to differences in reporting, we pulled proportions of historical deaths by day-of-the-week in Canada from various non-COVID related causes: circulatory, pulmonary, circulatory and pulmonary, and non-circulatory pulmonary. We will compare the estimated mortality rates from these causes to the estimated day-of-the-week effects to see if there is a similar trend. If not, then this suggests that coronavirus may be more likely to cause death on certain days of the week, or simply that coronavirus mortality has its own unique reporting artifacts.

2.2.4 Validating Forecasts

Institute for Health Metrics and Evaluation (2020) has made projections for the United States available since March 25th 2020, and has since expanded the number of regions they include in their model. In order to validate our model projections, we ran our model using the same daily mortality data as the IHME, and compared both model's cumulative death

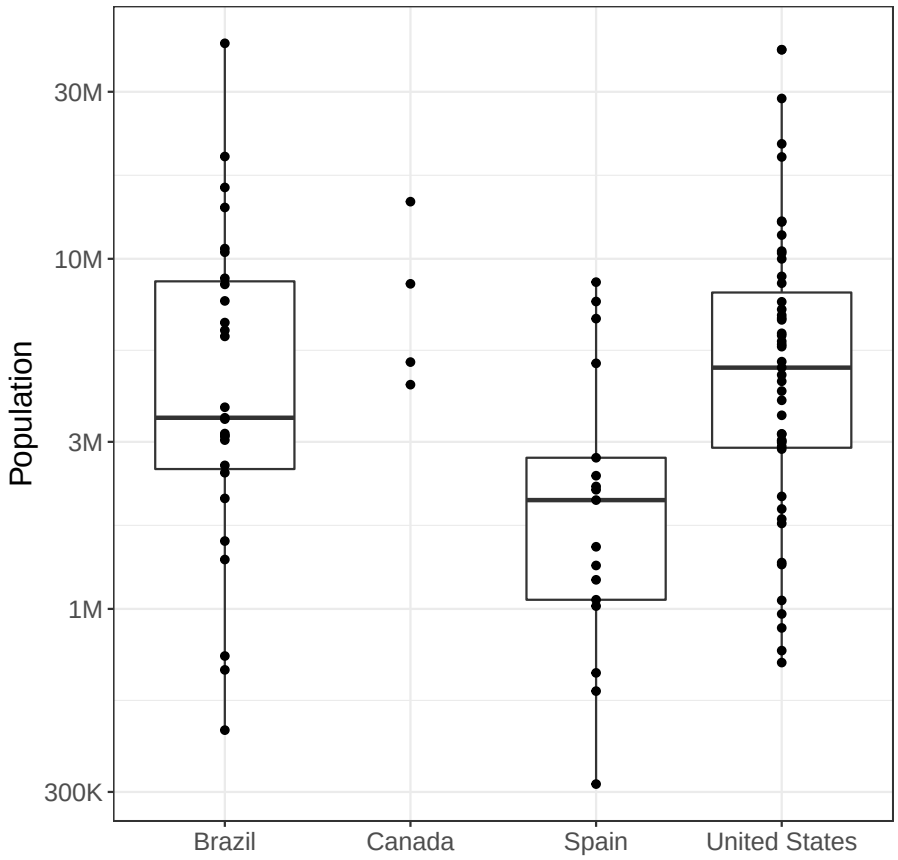


Figure 2.2: Populations of regions (black dots) by country. Note that the boxplot was omitted for Canada, as only 4 Canadian provinces were included

counts to the observed death count on June 30th 2020 for 14 different time points ranging from April 1st to June 25th. Additionally, we ran our model without a day-of-the-week effect to see whether the forecasting performance of our model relies on the day-of-the-week, and is not only outperforming the IHME model for other reasons. This model will be referred to as the "non-DOW model".

Models were compared by assessing consistency and accuracy of projections throughout time. Consistency was assessed by examining the amount of overlap between successive intervals, with subsequent intervals hopefully being narrower and mostly contained in previous intervals. Successive interval overlap is important to ensure that a model's results are consistent throughout time. If two successive intervals do not overlap, then at least one of those intervals must not contain the true value. Overlap was measured at 13 time points, since the first time point does not have a previous interval. At each of these 13 time points, the proportion of the interval that is contained in the previous interval is calculated for each region individually, resulting in 13 proportions per region. These proportions are then averaged across all regions to determine which model, on average, had the most consistent predictions.

The first measure used to assess the accuracy of the models was the standardized Root Mean Squared Log Error (sRMSLE) at time t, which was computed as:

$$sRMSLE(t) = \sqrt{\underset{ij}{\text{mean}}\left\{ \frac{\log(P_{ij}(t)/O_{ij}(t))^2}{\log(O_{ij}(t))} \right\}}$$

where $P_{ij}(t)$ and $O_{ij}(t)$ are the predicted/observed number of deaths from region i of country j. Note that each time point has a different number of regions with available data. The mean squared error is often used as a metric to assess accuracy of point estimates. We chose to use the log error because of the skewness in the reported deaths. Additionally, we chose to standardize the result to ensure all included regions have the same weight.

Accuracy of model forecasts was also assessed by computing the proportion of time points that the prediction intervals contain the observed cumulative death counts on June 30th 2020. To avoid the issue of excessively wide intervals appearing the best, we also plot the mean log-length of the intervals for each model, at each time point. The model which

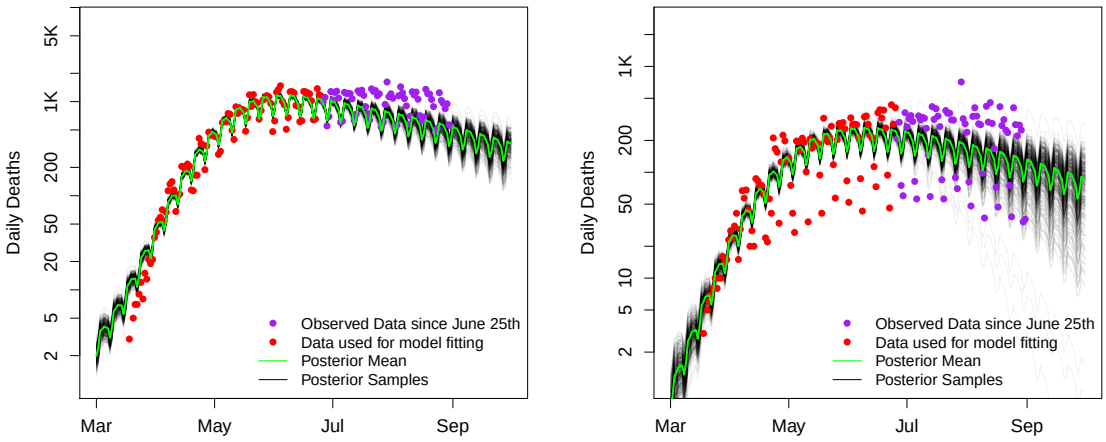
contains the true value the most often, relative to the mean log-length of the intervals, was considered the favourable model by this metric.

2.3 Results

2.3.1 Forecasts of Daily Mortality in Four Countries

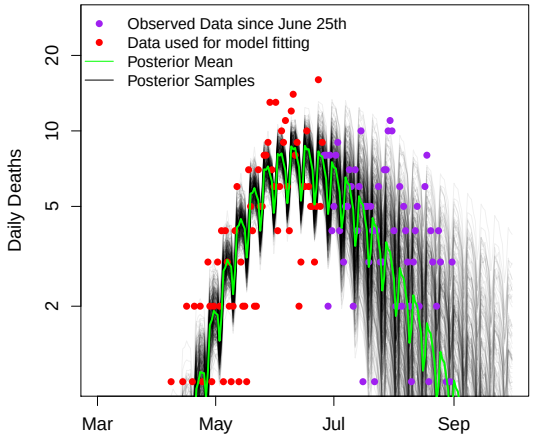
Figure 2.2 visualizes the variation in the population sizes of the regions considered. Note that the regions in Spain tend to be smaller than those of the other three countries and for Canada only the four largest provinces were included. Figures 2.3-2.5 show forecasts for countries and a subset of the regions studied, a full set of results is available in the online supplement to this chapter. The forecasts for all of Brazil are shown in Figure 2.3a, with the red points representing the data that was used to fit the model, and the purple representing the observed values from June 26th to August 31st. Up until July, our model fits the data quite well, indicated by the red points clustered around the posterior samples, and the day-of-the-week effect is well captured. However, starting in July, our model slightly underestimates the mean daily death counts. Figures 2.3b-c show the daily mortality forecasts in Sao Paulo, the region in Brazil with the most deaths, and Acre, a region with few deaths. In Acre, our model captures the trend of daily deaths reasonably well, suggesting that the hierarchical nature of our model is helping provide good forecasts for small regions. Starting in July, we are underestimating daily deaths in Sao Paulo. Although we estimate the mean daily deaths in Sao Paulo before July, it appears that we are slightly underestimating the day-of-the-week effect in Sao Paulo, indicated by the dispersion of the red and purple points having higher variance than the day-of-the-week effect allows for. This is due to the fact that we only allowed for one day-of-the-week effect for each country.

The forecasts for all of the United States are shown in Figure 2.4a. Notice that our model under predicts daily mortality counts throughout July and August. This is likely due to the fact that many states in the U.S have loosened their lockdown restrictions, which our model is unable to account for. Figures 2.4b-c show plots of the projections for Illinois and California. Our projections for Illinois seem to be quite accurate, as this is



(a) Brazil

(b) Sao Paulo



(c) Acre

Figure 2.3: Forecasting daily and cumulative deaths in Brazil as a whole, and the states of Sau Paulo and Acre.

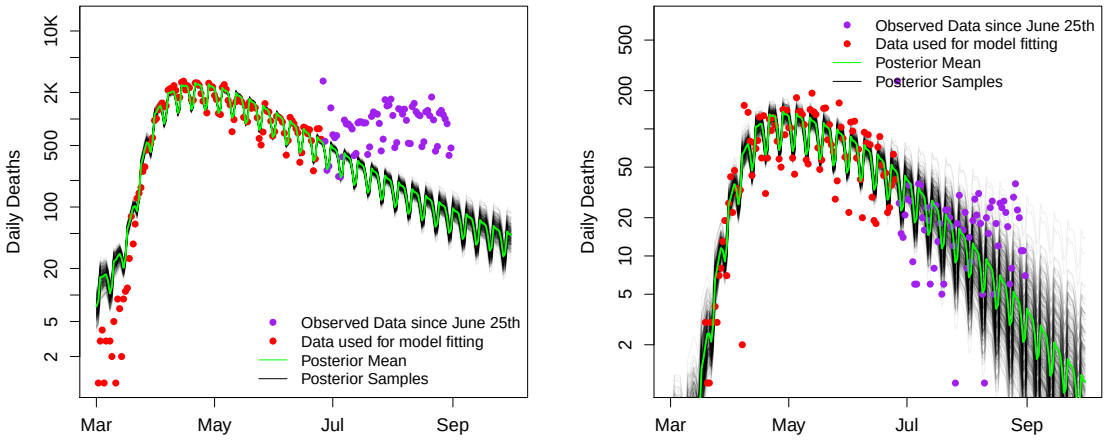
one of the states that is yet to experience any repercussions from reopening. However, our model is underestimating deaths in California likely because of loosening COVID-19 related restrictions.

Results for Spain and Canada are presented in Figure 2.5, and are less interesting due to the fact that the epidemics are largely over in these countries. When looking at the daily death plot for Spain, we see a small second wave not captured by the model. But the fact that this plot is on the log scale amplifies the apparent size of the second wave. Our model seems to project Canada's COVID-19 mortality reasonable well, likely due to the relatively firm COVID-19 restrictions in Canada. Ultimately, our model seems to predict daily COVID-19 mortality well in regions with firm COVID-19 restrictions. In 2.3.3, we will validate our projections under the assumption that COVID-19 related restrictions are held constant.

2.3.2 Day-of-the-week effects

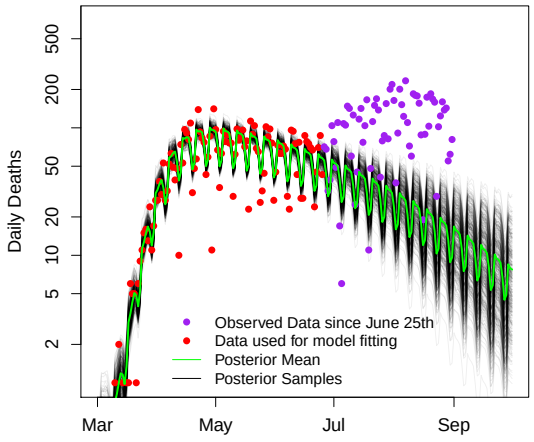
The 2.5th, 50th and 97.5th percentiles of the posterior distributions of the day-of-the-week effects are presented in Figure 2.6. With the possible exception of Spain, Sunday appears to report the lowest death counts, followed by Monday. In all countries, death counts seem to rise on Tuesday, and remain high until Friday or Saturday, and are still elevated relative to Sunday. The day-of-the-week effect is most pronounced in the United States, where Tuesday - Friday are all very similar, but are vastly different than the other days of the week. Brazil also shows a strong day-of-the-week effect, indicated by Monday's credible interval having almost no overlap to any other day-of-the-week. As expected, Canada's credible intervals are the widest, due to the fewest deaths overall.

The proportion of deaths by day-of-the-week relative to Sunday for historical (non-COVID) data from Canada are shown in Table 2.2. Note that there is very little, if any, similarities between our model's day of the week estimates and this data. These numbers only fluctuate by a few percentage points, and do not generally show a large spike on Tuesday as we saw in our COVID-19 day-of-the-week estimates. This could indicate that people are more likely to die due to COVID-19 on particular days, but is more likely explainable by differences in reporting between data sources.



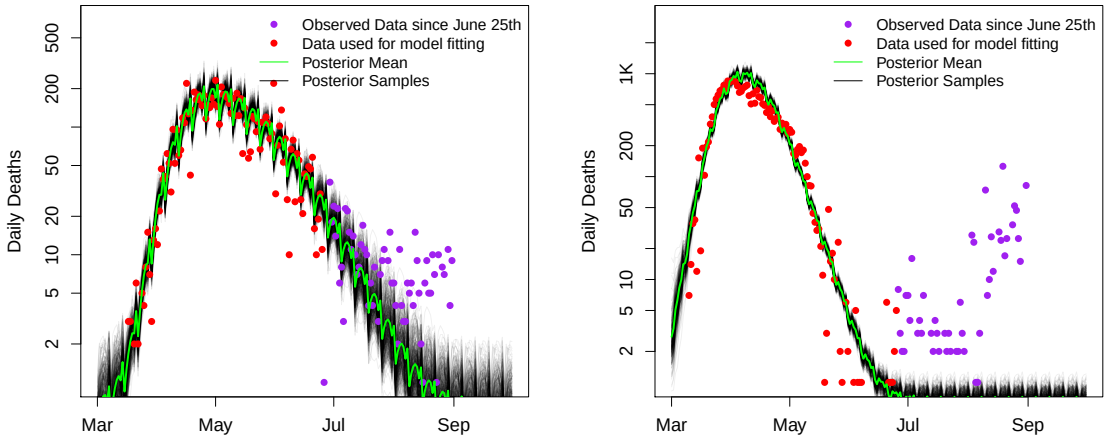
(a) United States

(b) Illinois



(c) California

Figure 2.4: Forecasting daily and cumulative deaths in the United States



(a) Canada (b) Spain

Figure 2.5: Forecasting daily deaths in the Canada and Spain

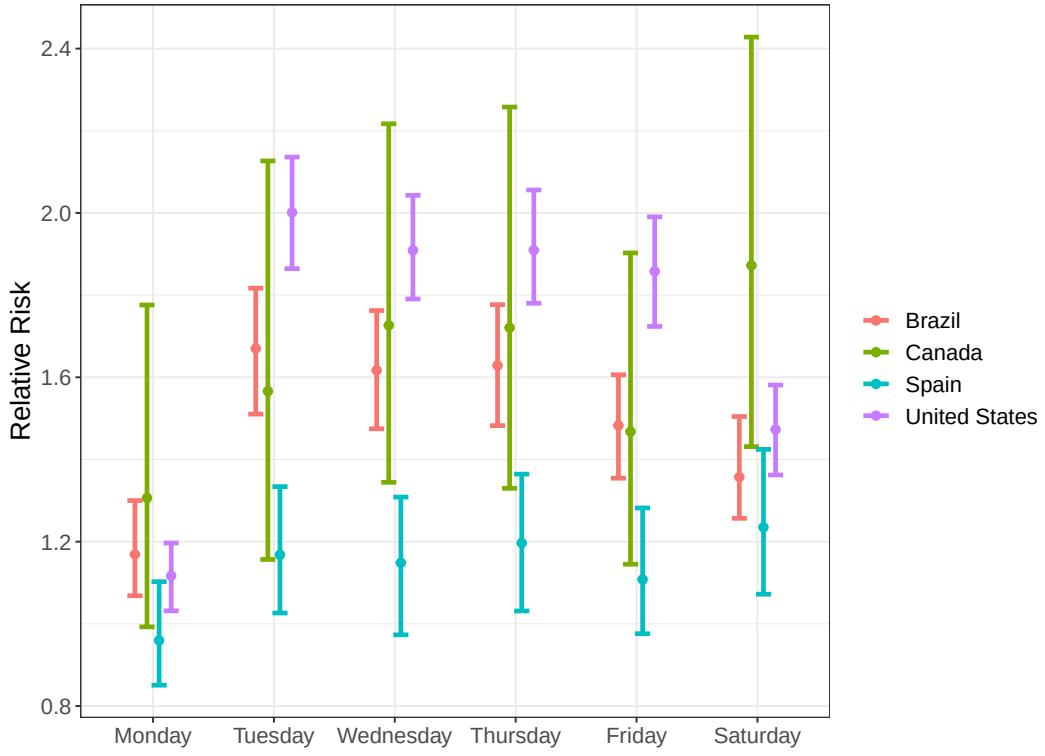


Figure 2.6: 2.5th, 50th, and 97.5th percentiles of posterior distributions for day of the week parameters relative to Sunday (which is fixed at 1.)

Cause	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Circulatory	1.016	1.003	1.005	0.995	1.003	0.997
Pulmonary	0.994	0.996	0.975	0.985	0.996	0.988
Circulatory and Pulmonary	1.011	1.002	0.999	0.993	1.002	0.995
Non-Circulatory Pulmonary	0.988	0.998	1.002	1.007	0.998	1.005

Table 2.2: Relative risks of days-of-the-week (relative to Sunday) for non COVID-19 related causes in Canada

2.3.3 Model Validation

Projections made at each of the 14 dates in all regions for the 3 models are shown in the online supplement. Figures 2.7-2.9 show projections for the four regions with the most deaths in each nation. In any plot where the Institute for Health Metrics and Evaluation (2020) results are missing, it is because they were not produced for that region at that time. In the plots where the DOW or non-DOW model were missing, it is because they had not yet achieved the minimum 50 deaths required to be included in our analysis. For regions with a large number of deaths (such as New York), the day-of-the-week model was very consistent, where 95% credible intervals have a large amount of overlap from date-to-date. The IHME projections are somewhat inconsistent for this region, indicated by non-overlapping intervals. However, in regions like Florida, the IHME model seems to outperform the DOW and non-DOW models, requiring a more formal investigation into model projection assessments. The mean proportion of overlap between successive intervals for all regions is shown in Figure 2.10b. The DOW model tends to show the highest mean overlap at 8/13 time points, the non-DOW model had the most overlap at 3/13 time points, follow by the IHME which had the most overlap at 2/13 time points. This suggests that the DOW model produces the most consistent projections of the 3 models, and indicates that in the long-run our projections are likely to be better suited than the IHME's.

The sRMSLE for each of the three models at each time point is shown in Figure 2.10a. This figure shows that the IHME is drastically underperforming when compared to the other two models between April 7th and May 4th. The sRMSLE's are otherwise comparable. Figure 2.10c shows the proportion of regions that each model's interval contained the observed June 30th cumulative death count. Somewhat surprisingly, we see a downward trend as we get closer to June 30th, as shorter term predictions should become easier. This

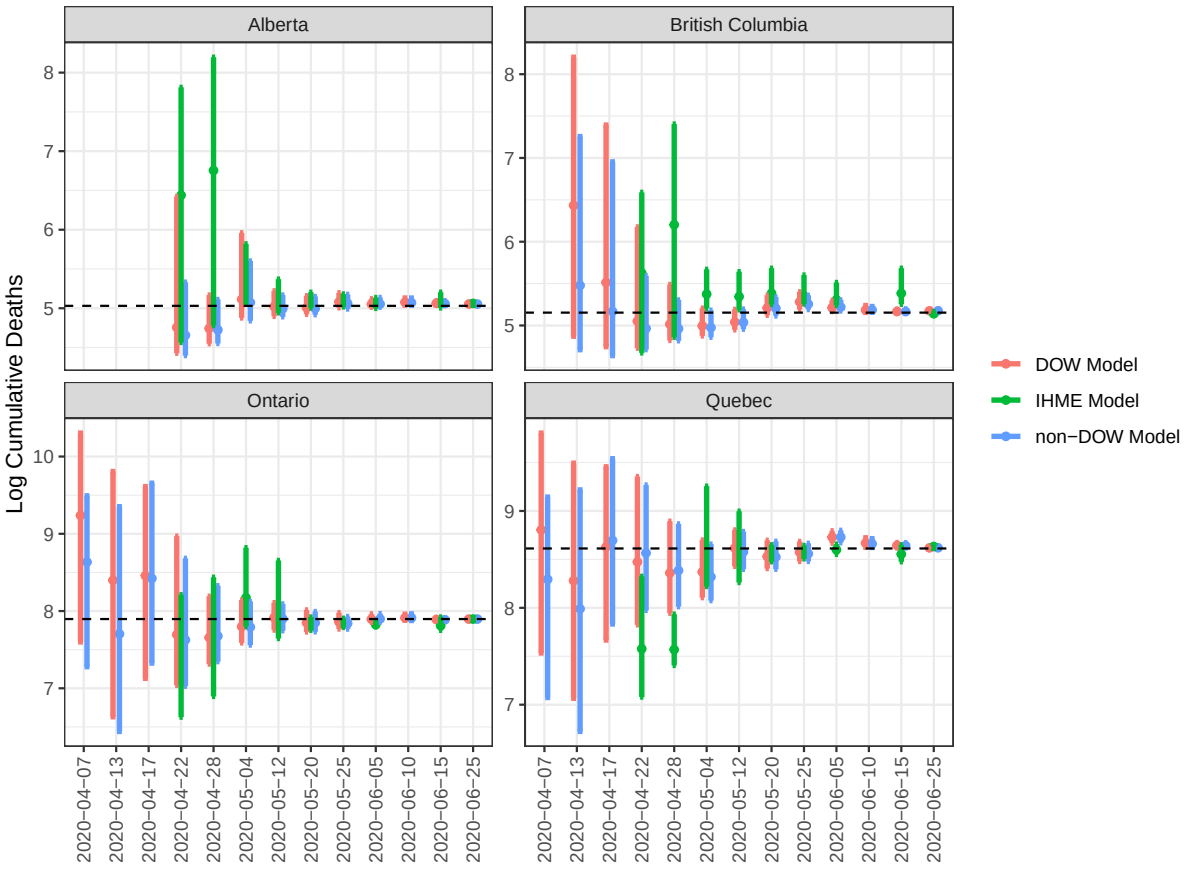


Figure 2.7: Cumulative mortality projections for June 30th made at 14 different dates starting April 1st 2020. The log of the cumulative mortality counts for June 30th are represented by the dashed line. Results are shown for the four Canadian provinces with the most COVID-19 deaths.

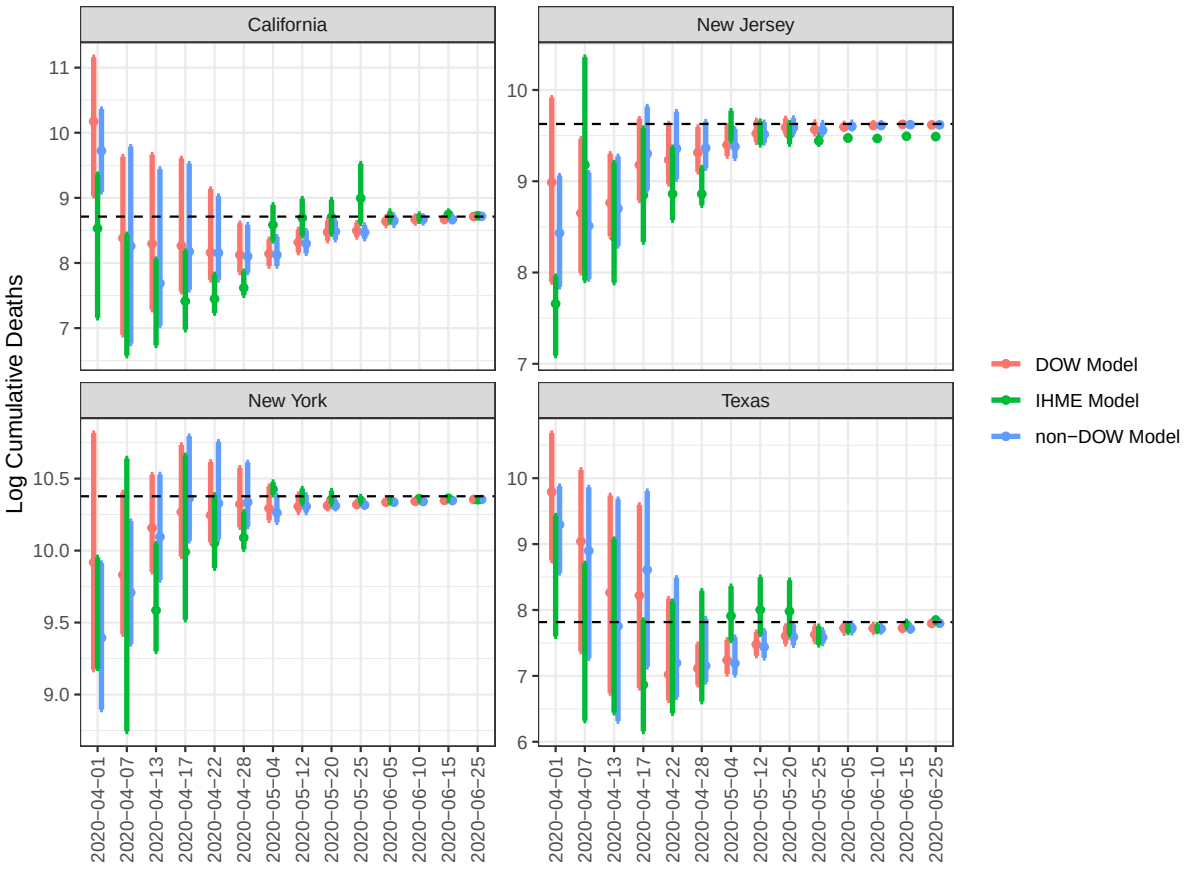


Figure 2.8: Cumulative mortality projections for June 30th made at 14 different dates starting April 1st 2020. The log of the cumulative mortality counts for June 30th are represented by the dashed line. Results are shown for the four U.S states with the most COVID-19 deaths.

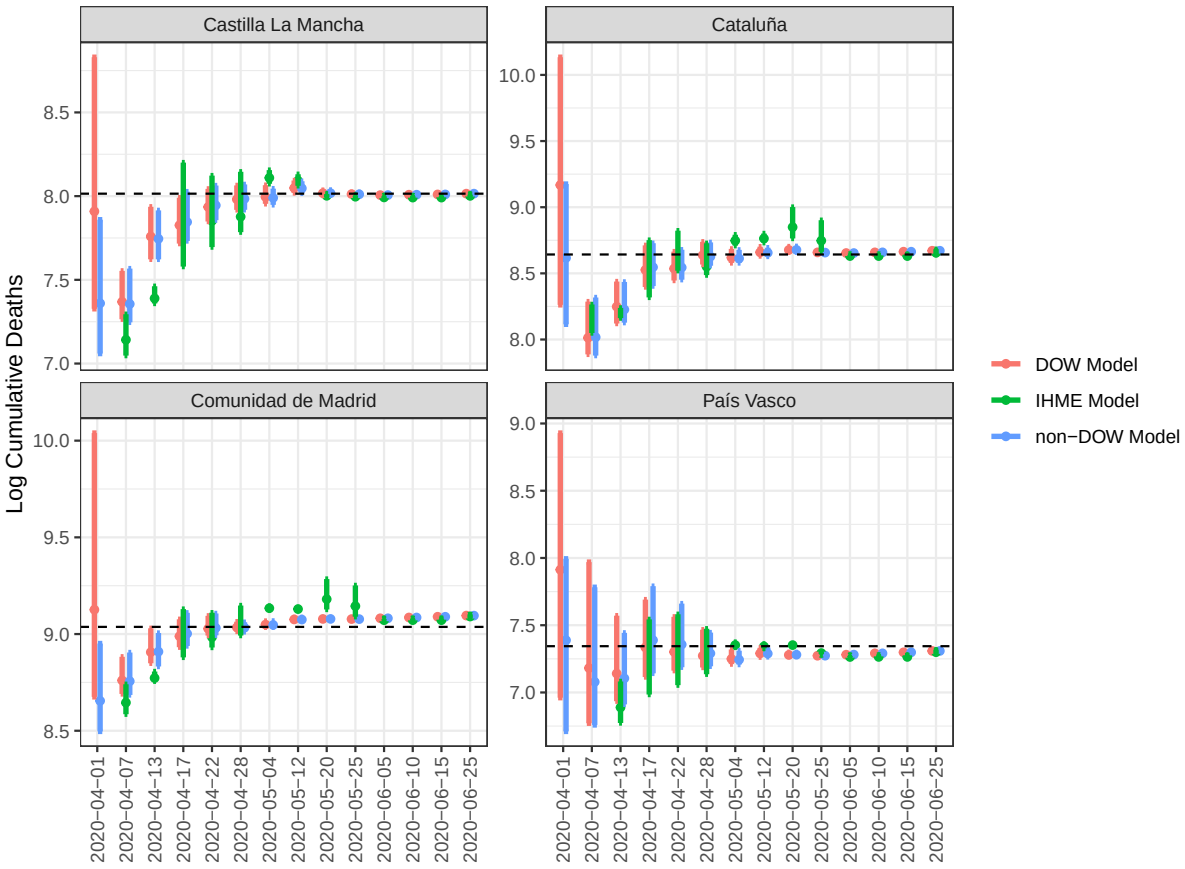


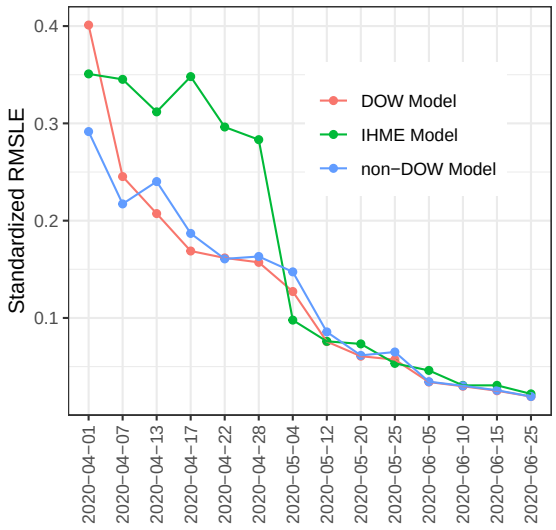
Figure 2.9: Cumulative mortality projections for June 30th made at 14 different dates starting April 1st 2020. The log of the cumulative mortality counts for June 30th are represented by the dashed line. Results are shown for the four Spanish Autonomous Communities with the most COVID-19 deaths.

is likely because many regions had started their second uprising in daily deaths just prior to June 30th, causing models to slightly underestimate deaths in the short term. The DOW and non-DOW models seem to do very well early on in the epidemic, capturing over 90% of true values on April 17th. The IHME model tends to do better in the 1-month projection range. All models tend to perform poorly for very short-term projections (i.e < 2 weeks).

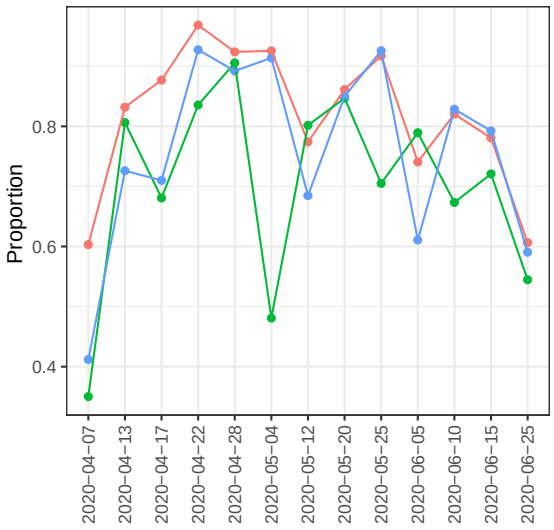
Although the IHME model appears to be better suited in 1-month projections based on our accuracy metric, Figure 2.10d shows that this is likely due to the increased interval width. The IHME's intervals are approximately two times as wide when making 1-month projections. Despite the IHME's wider intervals at almost every time point, the DOW model outperforms the IHME projections in terms of consistency (i.e interval overlap), and is more accurate for longer-term projections. Note that we have not validated these projections for the second uprising in deaths, so although our model outperforms the IHME up until June 30th, an extension of our methodology is likely required to accurately forecast COVID-19 mortality in the second "wave" and beyond.

2.4 Discussion

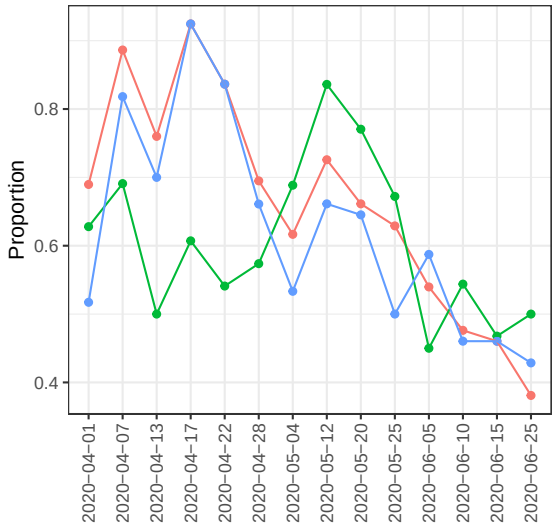
One interpretation of the day of the week effects estimated by our model is that people are more likely to die from COVID-19 on certain days of the week. It may be the case that people are more likely to contract the disease on given days of the week (e.g weekdays), which may cause them to pass away at higher rates in the following days. It is also possible that on certain days of the week, hospitals are more crowded and have fewer available resources, which could increase mortality on those days. However, it is more likely the case that deaths are equally likely to occur on any day-of-the-week, but are simply more likely to be reported on certain days due to hospital administration procedures. Hospital administrative workers likely work less on weekends, so deaths on the weekends may be reported several days after they occur, resulting in lower deaths on Sundays and Mondays. In either case, our results show that regardless of whether or not these are true day-of-the-week effects or are simply an artifact created by inconsistent reporting, accounting for day-of-the-week effects is important when projecting mortality during epidemics. Our non-



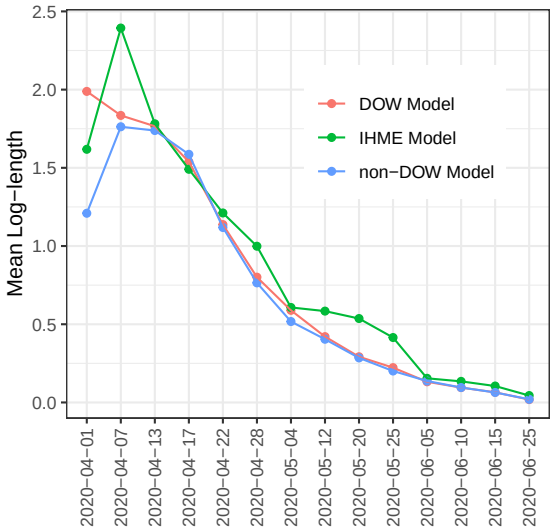
(a) Standardized Root Mean Squared Log Error



(b) Mean proportion of overlap between successive intervals



(c) Proportion of intervals that contained the true value for June 30th.



(d) Mean log-length of intervals

Figure 2.10: Comparing the DOW, non-DOW, and IHME models based on standardized Root Mean Squared Log Error, how much overlap there was between successive intervals, how often their intervals contained the true value, and their interval lengths

DOW model also outperformed the IHME projections, which shows the potential benefits of using a skew-Normal and/or hierarchicalizing the model parameters.

One limitation of our model is that although we estimate day-of-the-week effects, the forecasts output from our model are projections of the number of COVID-19 deaths reported on certain days, as opposed to the true death count. Further extensions of our model could echo the methodology of Seaman et al. (2020), allowing for the forecasting of the true number of deaths on any given day.

Since June 30th 2020, many regions such as Florida have seen a second increase in COVID-19 deaths. Further extensions to our model need to be explored to account for a second increase in daily deaths. Prediction of these seconds increases could be performed using region-level covariates such as lock-down severity or mask usage in the region. The second peak itself could be potentially be modelled by adding a second skew-Normal, where the location and height of the peak daily deaths are related to the first skew-Normal's parameters. Additionally, some regions that appear to be undergoing a second increase but may just be having multiple epidemics in different subregions (e.g counties in the U.S). Having smaller-area level data could improve model performance because of this.

Another extension to this model could be to have a different day-of-the-week parameter per region, but this is likely only possible for regions that have at least several weeks of data. Hierarchicalizing the day of the week parameters is also an option. For example, the Monday effect for regions with less mature data could be estimated by "borrowing" information from regions where the epidemic has matured. Another potential extension is to have a day-of-the-week effect changing throughout time, as COVID-19 death reporting may have improved since the beginning of the pandemic. However, in the midst of an epidemic, computational efficiency is of the utmost importance, as these models can easily take over a week to run. Obtaining robust estimates quickly can help aid policy decisions which can ultimately save lives, so having projections within a day or two is largely beneficial.

Further analysis is needed to fully assess the predictive capabilities of our model by expanding the number of countries and regions included. Additionally, including more models for comparison would be ideal, such as the projections made at `https://covid19-projections.com/` (Scriby, Inc., 2020). This work focused on presenting our core methodology, and pro-

vide accurate single "wave" projections for four countries that were near or past the peak of their epidemic.

Despite these limitations in the DOW model, it seems to forecast mortality related to COVID-19 in the first "wave" quite well when compared to the alternatives. Our projections made in Section 2.3.1 can be seen as accurate projections assuming that COVID-19 restrictions were never loosened in each region. The day-of-the-week effect was shown to be important when forecasting COVID-19 mortality, as the IHME and non-DOW model seemed to be inconsistent with their projections over time, which could be because the projections were made on different days of the week. Our skew-Normal Bayesian hierarchical model with day-of-the-week effects could be used as the basis for future COVID-19 mortality predictions where the epidemic is less mature, or perhaps for future outbreaks where accurate and consistent mortality projections are needed, where only daily mortality data is available as the input.

2.5 Bibliography

Anastassopoulou, C., Russo, L., Tsakris, A., and Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS one*, 15(3):e0230405.

Berument, H. and Kiymaz, H. (2001). The day of the week effect on stock market volatility. *Journal of economics and finance*, 25(2):181–193.

Brown, P., Jha, P., et al. (2020). Mortality from COVID-19 in 12 countries and 6 states of the United States. *medRxiv*.

Bukhari, Q., Jameel, Y., Massaro, J. M., D'Agostino, R. B., and Khan, S. (2020). Periodic oscillations in daily reported infections and deaths for coronavirus disease 2019. *JAMA Network Open*, 3(8):e2017521–e2017521.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Chakraborty, T. and Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, page 109850.

Crevecoeur, J., Antonio, K., and Verbelen, R. (2019). Modeling the number of hidden events subject to observation delay. *European Journal of Operational Research*, 277(3):930–944.

Friedman, J., Liu, P., and Gakidou, E. (2020). Predictive performance of international COVID-19 mortality forecasting models. *medRxiv*.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Institute for Health Metrics and Evaluation (2020). Covid-19 projections.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.

Perc, M., Gorišek Miksić, N., Slavinec, M., and Stožer, A. (2020). Forecasting COVID-19. *Frontiers in Physics*, 8:127.

Petropoulos, F. and Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PloS one*, 15(3):e0231236.

Sarkar, K., Khajanchi, S., and Nieto, J. J. (2020). Modeling and forecasting the COVID-19 pandemic in India. *Chaos, Solitons & Fractals*, page 110049.

Scriby, Inc. (2020). The coronavirus app.

Seaman, S., Samartsidis, P., Kall, M., and De Angelis, D. (2020). Nowcasting COVID-19 deaths in England by age and region. *medRxiv*.

Chapter 3

Capturing Spatial Dependence of COVID-19 Case Counts with Cellphone Mobility Data

Abstract

Spatial dependence is usually introduced into spatial models using measure of physical proximity. When analyzing COVID-19 case counts, this makes sense as regions that are close together are more likely to have more people moving between them, spreading the disease. However, using the actual number of trips between each region may explain COVID-19 case counts better than physical proximity. In this chapter, we investigate the efficacy of using telecommunications-derived mobility data to induce spatial dependence in spatial models applied to two Spanish communities' COVID-19 case counts. We do this by extending Besag York Mollié (BYM) models to include both a physical adjacency effect, alongside a mobility effect. The mobility effect is given a Gaussian Markov random field prior, with the number of trips between regions as edge weights. We leverage modern parametrizations of BYM models to conclude that the number of people moving between regions better explains variation in COVID-19 case counts than physical proximity data. We suggest that this data should be used in conjunction with physical proximity data when developing spatial models

for COVID-19 case counts.

3.1 Introduction

Spatial analyses of COVID-19 case data were first published as early as March of 2020 (Huang et al., 2020; Arab-Mazar et al., 2020; Giuliani et al., 2020), in an attempt to characterize, predict, and attenuate the severity of the pandemic. Subsequent studies have noted substantial spatial dependence in COVID-19 case counts (Kang et al., 2020; Bilal et al., 2020). This makes sense as regions that are close to each other likely have more people moving between them, spreading the disease to nearby regions.

Many groups have attempted to model COVID-19 case counts as a function of climate (Liu et al., 2020; Shi et al., 2020; Briz-Redón and Serrano-Aroca, 2020), healthcare quality (Sugg et al., 2021), socioeconomic factors (Baum and Henry, 2020) and more. More recently, mobility data has become more abundant and popular for modeling COVID-19 transmission. This makes sense because the disease spreads through human contact, meaning that case counts are likely to be a function of the number of people moving around. Such mobility data has been used to model the evolution of the epidemic in Spain (Aràndiga et al., 2020; Iacus et al., 2020), assess the effectiveness of the Spanish lockdown (Orea and Álvarez, 2020), monitoring the epidemic in Switzerland (Persson et al., 2021), identify at-risk populations in France during a lockdown (Pullano et al., 2020), individual-level infection tracing in China (Kraemer et al., 2020), assess the timing of stay-home orders (Audirac et al., 2020), and evaluating the effectiveness of social distancing in the United States (Badr et al., 2020). This data can be found in many forms, but is commonly found in the form of aggregated areal *mobility matrices*. If we denote a mobility matrix $\boldsymbol{M}$, $[\boldsymbol{M}]_{ij}$ corresponds to the number of trips from region i to region j, and $\boldsymbol{M}_{ii}$ represents the number of trips within region i.

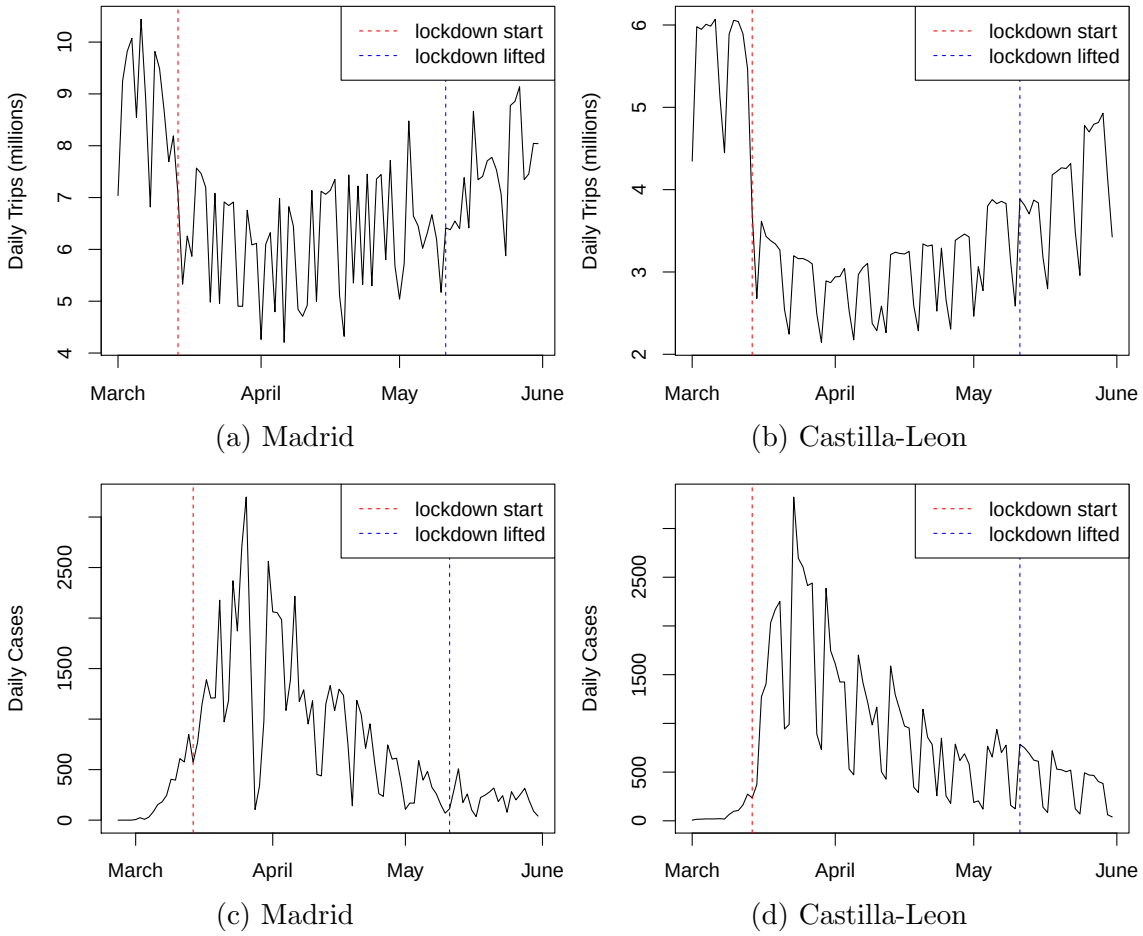
These data have been applied in a variety of different models to answer numerous questions, but lack of available methods makes it difficult for researchers to use this data to its full potential. In this chapter, we demonstrate a novel method for analyzing this data, whereby the mobility data is used as edge weights in a Gaussian Markov random field (network) model. Previous work using network models have been applied to mobility data in

the form of a network compartment model (Chang et al., 2021) which was used to conduct inference regarding societal inequities, and inform reopening. This work does not aim to make such claims, but rather demonstrate the efficacy of mobility data in modern parametrizations of Besag, York, and Mollié (BYM) models (Besag et al., 1991) and their extensions.

BYM models have been used frequently in the spatial analysis literature due to their effectiveness and computational efficiency. In these models, the spatial component is comprised of Conditional Autoregressive (CAR) (Besag, 1974) models and conventional random effects. This means that the spatial effect of region i depends only on its "neighbours". Neighbours could be defined by any quantity the analyst has access to, but is most often defined by physical adjacency, i.e. if two regions share a common border, they are considered neighbours. Several ICAR/BYM models have been applied to COVID-19 data with neighbours defined in this way (DiMaggio et al., 2020; Huang and Brown, 2021; Brainard et al., 2020). Although these spatial model components based on physical adjacency are powerful and computationally efficient, it makes more sense to use mobility between regions to induce spatial dependence in COVID-19 models because the disease spreads via person-to-person contact.

In this chapter, we build a BYM model where mobility data is used to induce spatial dependence between regions. Using mobility data within two Communities in Spain, Madrid and Castilla-Leon, we demonstrate the value of mobility data for COVID-19 spatial modeling applications. Furthermore, we extend modern parametrizations of BYM models to account for both physical adjacency and mobility simultaneously, and show that mobility data captures spatial variation in COVID-19 case counts much more accurately than physical adjacency alone.

This chapter is organized as follows. Section 3.2 presents the data and the modeling strategy based on particular parametrizations of BYM models. The results come in Section 3.3, and the chapter ends with a final discussion in Section 3.4.



(a) Madrid (b) Castilla-Leon

(c) Madrid (d) Castilla-Leon

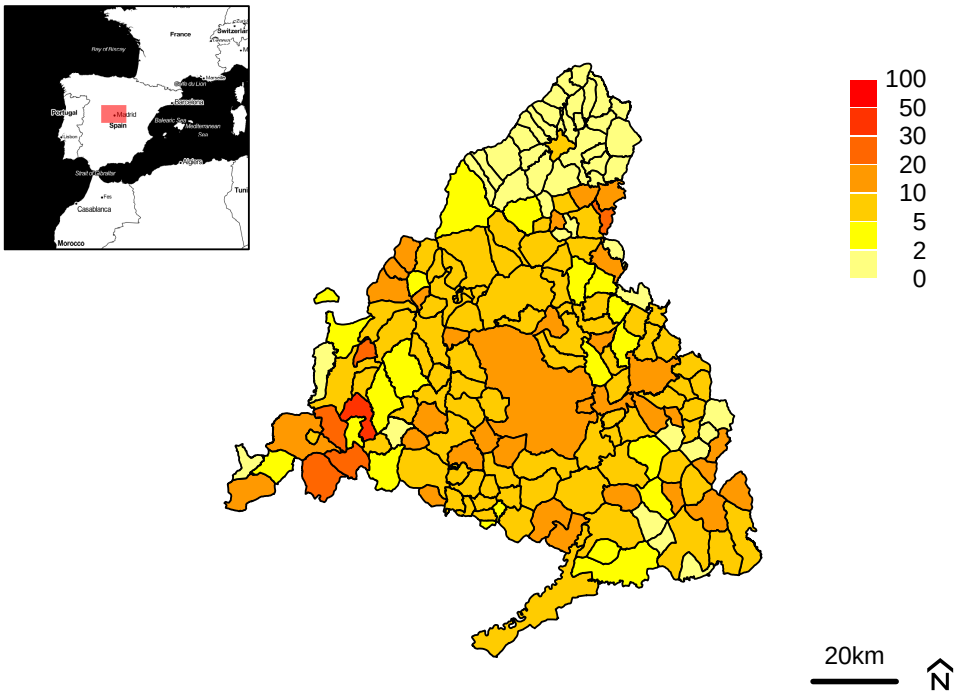
Figure 3.1: Number of trips greater than 500 metres (a and b) and daily case counts (c and d) in the two Communities of Spain from March to June 2020.

3.2 Methods

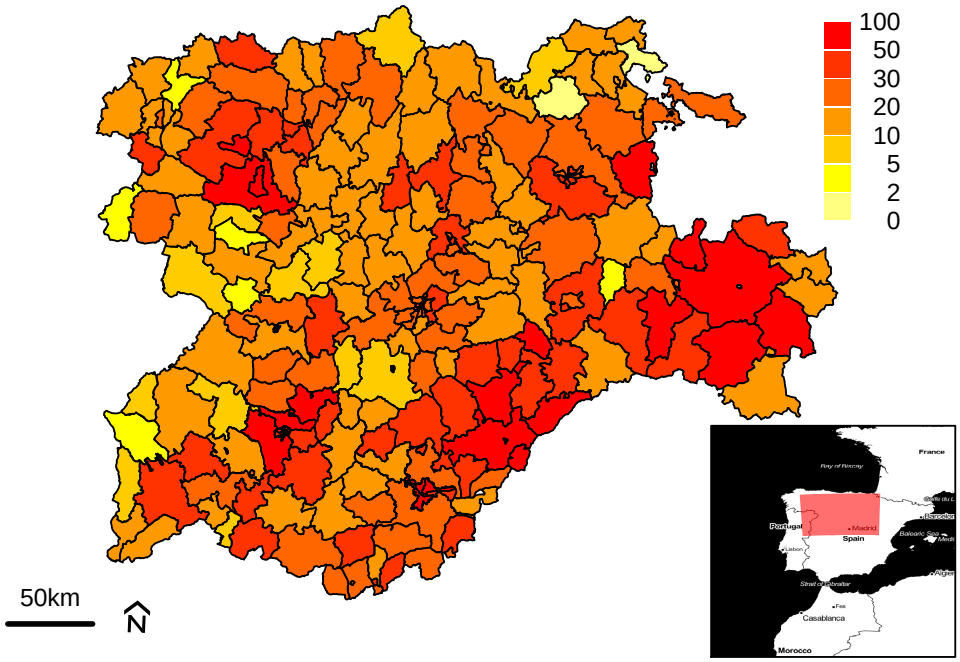
3.2.1 Data

This chapter is focused on two regions in Spain. Castilla-Leon is the largest Community in Spain by area and is located in the northwest part of Spain, with a population of 2.5 million. The Community of Madrid is located in the central part of Spain and has a population of around 6.8 million, and it is home of the capital of the country, Madrid City, with 3.3 million inhabitants.

The human mobility data was obtained from Barcelona Supercomputing Center Flowmap dashboard (Valencia, 2021). Trips within Madrid and Castilla-Leon were extracted from over 13 million phone records provided by a Spanish cellphone company. Both passive



(a) Madrid



(b) Castilla-Leon

Figure 3.2: COVID-19 cases per thousand, up to May 31 2020 for two communities in Spain. Background map ©Stamen Design.

(GPS) and active (text messages, calls etc.) data were aggregated to construct daily movement matrices in each of the Communities, prior to the authors acquisition of the data. Given that trips were only recorded from one cellphone company, adjustment was made to estimate the number of total trips between each region. As a result, the entries of the mobility matrices are non-integer values.

Figures 3.1a and 3.1b show the total daily movement between regions in Madrid, and Castilla-Leon, respectively. There is a sharp drop in the number of trips around March 14th 2020, which corresponds to a nation-wide lockdown. Lockdown restrictions began to ease around May 11th, where the number of trips slowly began to rise. Figures 3.1c and 3.1d show the number of cases of COVID-19 cases in both Communities. COVID-19 daily cases data were retrieved from the open data portal of Castilla-Leon (General Directorate of Information Systems, Quality and Pharmaceutical Provision, 2021) and from the Epidemiological Surveillance Network of Madrid (Ministry of Economic Affairs and Digital Transformation, 2021). Notice that the movement drops as cases rise, because a lockdown was implemented in response to the increasing severity of the epidemic. In order to avoid this potential "reverse causality" problem, we will only use movement data in the first week of March. Our justification for this is that there is a time lag between when the virus spreads and the resulting COVID cases are confirmed. That is, the "first wave" of the epidemic was likely influenced mostly by the movement that occurred prior to the peak in cases, and less by the movement that occurred during it.

Figure 3.2 shows the spatial distribution of the COVID-19 case rates up until May 31, 2020. The cases per thousand people range from (approximately) $0 - 30$ in Madrid, and $0 - 100$ for Castilla-Leon. We can see that there is substantial variation in the case rates within each of these Communities. Note that the extreme values in these plots are mostly small regions, which makes sense since the variance of case rates is higher when population is small. In the north of Madrid, there is a cluster of municipalities that have very low case rates. In Castilla-Leon, case rates are highest near the southeast border, which is the border to Madrid.

Figure 3.3 shows the number of trips to, from, and within each Municipality of Madrid (there are 179 of these small regions), and Castilla-Leon (there are 245 health zones).

Madrid and Castilla-Leon are considered separately throughout this chapter. Although they are adjacent, data on movements between the two communities are not available. In Madrid, there is a lot of movement in and around Madrid City, and less movement in the more rural areas. Castilla-Leon shows a less predictable movement pattern, as there is not a single capital city that accounts for most of the movement. This movement data will be used to induce spatial correlation between regions, as described in Section 3.2.3.

3.2.2 Spatial autoregressive models

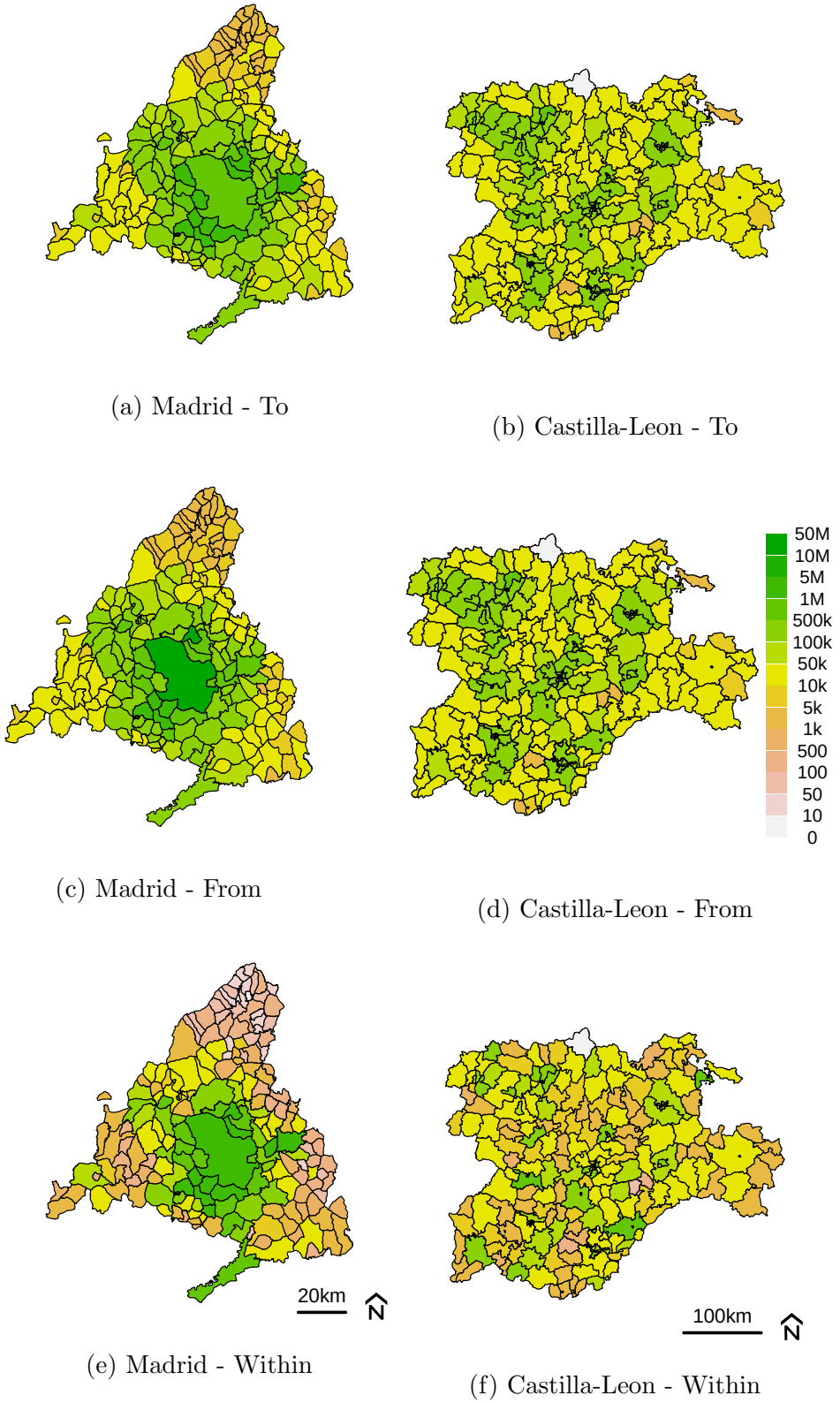
Besag, York, and Mollié (BYM) models (Besag et al., 1991) are widely used in spatial epidemiology and disease mapping due to their simplicity and computational efficiency. They assume the incidence of disease in region i follows a Poisson distribution

$$Y_i \sim \text{Pois}(E_i \lambda_i)$$

where Y_i is the number of infected cases in region i, and E_i is some form of expected count or offset, which could be the at-risk population, exposure time, etc. The log-relative risk, λ_i, is often modeled as

$$\begin{aligned} \log(\lambda_i) &= \mu + \beta X + \phi_i + \theta_i \\ \phi_i | \boldsymbol{\phi}_{-i} &\sim N\left(\frac{1}{\sum_j w_{ij}} \sum_j w_{ij}\phi_j,\ \frac{\sigma_\phi^2}{\sum_j w_{ij}} \right) \\ \theta_i &\overset{i.i.d}{\sim} N(0, \sigma_\theta^2) \end{aligned} \tag{3.1}$$

where μ is the overall intercept, β is the effect of spatial covariates, ϕ_i is the structured spatial random effect, and θ_i is the unstructured spatial random effect which allows for overdispersion in the response. In the spatial formulation of the BYM model, $w_{ij} = 1$ when regions i and j share a common border, and 0 otherwise. That is, region i's structured spatial effect is only conditionally dependent on its neighbours, given all other regions. The distributions $\{\phi_i | \boldsymbol{\phi}_{-i}\}_{i=1}^n$ are known as the *full conditionals*, where $\boldsymbol{\phi}_{-i}$ is short hand for the set $\{\phi_1, \phi_2, ...\phi_{i-1}, \phi_{i+1}, ...\phi_n\}$. We can see from (3.1) that $E(\phi_i | \boldsymbol{\phi}_{-i})$ is a weighted average



(a) Madrid - To

(b) Castilla-Leon - To

(c) Madrid - From

(d) Castilla-Leon - From

(e) Madrid - Within

(f) Castilla-Leon - Within

Figure 3.3: Number of trips (incoming, outgoing, and within) the 179 regions of Madrid, and 245 health zones of Castilla-Leon, for the period March 1 to March 7 2020.

of its neighbours, resulting in spatial smoothing. These full conditionals correspond to the joint distribution of the ϕ's being a Gaussian Markov random field (GMRF) (Rue and Held, 2005), with

$$\boldsymbol{\phi} \sim \boldsymbol{MVN}(\mathbf{0}, \boldsymbol{Q}^{-1})$$
$$\boldsymbol{Q} = \sigma_{\phi}^{-2}\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{W})$$

where $\boldsymbol{W}$ is a matrix of weights such that $w_{ij} > 0$ for $i \neq j$ and $w_{ii} = 0$, and σ^2 is a variance parameter to be estimated. $\boldsymbol{D}$ is a diagonal matrix such that $\boldsymbol{D}_{ii} = \sum_j w_{ij}$. This definition ensures that the precision matrix, $\boldsymbol{Q}$, is both symmetric and positive definite. In addition to the 0-1 weights based on regions being adjacent, other weighting schemes, such as inverse of Euclidean distance between regions, have been used. For a comparison of common weighting schemes, see (Duncan et al., 2017). When we specify $\boldsymbol{Q}$ in this way, we refer to this as an Intrinsic Autoregressive (ICAR) model for $\boldsymbol{\phi}$. The joint density function has a computationally convenient form with

$$p(\boldsymbol{\phi}) \propto \exp\Big[-\frac{1}{2\sigma_{\phi}}\sum_{i<j} w_{ij}(\phi_i - \phi_j)^2\Big]$$

which is sometimes referred to as *the pairwise difference formula*. Notice that this density is invariant to the addition of a constant to each ϕ_i, leaving the spatial random effects unidentifiable up to a constant. This is typically remedied by imposing the constraint $\sum_i \phi_i = 0$ (Duncan et al., 2017). We will now modify this BYM model to account for movement between regions, in addition to physical adjacency.

3.2.3 Movement augmented BYM model

In order to extend the BYM model to allow for spatial correlation based on movement data, a second ICAR term, γ_i, with dependence structure governed by the movement data is added to the model. We also retain an adjacency-determined spatial effect ϕ_i in order to infer the relative importance of mobility-based and adjacency-based spatial dependence in

determining COVID-19 case counts. The resulting model is

$$\begin{aligned}
\log(\lambda_i) &= \mu + \beta X_i + \phi_i + \gamma_i + \theta_i \\
\phi_i|\boldsymbol{\phi}_{-i} &\sim N\left(\frac{1}{\sum_j w_{ij}} \sum_j w_{ij}\phi_j, \frac{\sigma_\phi^2}{\sum_j w_{ij}}\right) \\
\gamma_i|\boldsymbol{\gamma}_{-i} &\sim N\left(\frac{1}{\sum_j v_{ij}} \sum_j v_{ij}\gamma_j, \frac{\sigma_\gamma^2}{\sum_j v_{ij}}\right) \\
\theta_i &\sim N(0, \sigma_\theta^2)
\end{aligned}$$

where ϕ_i and γ_i are the spatial random effects with priors based on the physical data and movement data respectively. The geographically-defined process ϕ_i has weights $w_{ij} = 1$ if regions i and j share a common border and are 0 otherwise, while the movement-defined process γ_i has weights v_{ij} representing the number of trips between regions i and j. Using mobility as edge weights in network models has shown to be effective in the context of infectious diseases (Schrödle et al., 2012; Volkova et al., 2010; Geilhufe et al., 2014). (Schrödle et al., 2012) used mobility weights in an autoregressive term, which allowed the weights matrices to be asymmetric. However, given that our mobility data is being used in a Gaussian prior for a random effect, the precision matrices of $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, Q_ϕ and Q_γ, must be symmetric. Therefore we require $w_{ij} = w_{ji}$ and $v_{ij} = v_{ji}$. While the first equality will always be true, the mobility matrices are not perfectly symmetric, thus symmetry was induced by defining v_{ij} as the sum of the numbers of trips from i to j and from j to i. The GRMF does not account for the movement within a region, so the movement within a region was included in the model as a spatial covariate X_i (fixed effect). That is, X_i was computed as

$$X_i = \frac{\frac{v_{ii}}{E_i} - \underset{j}{\text{mean}}(\frac{v_{jj}}{E_j})}{\underset{j}{\text{sd}}(\frac{v_{jj}}{E_j})}$$

where v_{ii}/E_i is the number of trips per person within a region, and $\text{mean}(v_{jj}/E_j)$ and $\text{sd}(v_{jj}/E_j)$ are the mean and standard deviations of the trips per person in all other regions. This model was run on both the Madrid and Castilla-Leon data.

There are two main drawbacks with the formulations of BYM models presented thus

far. Firstly, the interpretation of the parameters σ_γ and σ_ϕ depend on the average number of neighbours and the total number of trips for each region, and hence their magnitudes are not comparable (Sørbye and Rue, 2014). Secondly, $\sigma_\phi, \sigma_\gamma$, and σ_θ are hard to estimate without very careful choices of hyperpriors (Leroux et al., 2000). We will now address these shortcoming via reparametrizations.

3.2.4 Reparametrizations and Priors

In order to solve issues with comparability, interpretability, and estimation, we apply a reparameterization of our model that is inspired by Riebler et al. (2016) with

$$\begin{aligned}
\sigma^2 &\approx \text{Var}(\phi_i + \gamma_i + \theta_i) \\
\phi_i^* | \boldsymbol{\phi}_{-i}^* &\sim N\left(\frac{1}{\sum_j w_{ij}} \sum_j w_{ij}\phi_{ij}^*,\ \frac{\rho_\phi \sigma^2}{s_\phi \sum_j w_{ij}}\right) \\
\gamma_i^* | \boldsymbol{\gamma}_{-i}^* &\sim N\left(\frac{1}{\sum_j v_{ij}} \sum_j v_{ij}\gamma_{ij}^*,\ \frac{\rho_\gamma \sigma^2}{s_\gamma \sum_j v_{ij}}\right) \\
\theta_i &\sim N(0, \rho_\theta \sigma^2)
\end{aligned}$$

where $\rho_\phi + \rho_\gamma + \rho_\theta = 1$ and $0 < \rho_\gamma, \rho_\phi, \rho_\theta < 1$. The priors for σ and $\boldsymbol{\rho}$ are

$$\begin{aligned}
\sigma &\sim N_+(0, 1) \\
\boldsymbol{\rho} &\sim \text{Dirichlet}(1, 1, 1)
\end{aligned}$$

Note that

$$\begin{aligned}
\phi_i^* &= \sigma\left(\sqrt{\rho_\phi / s_\phi}\right)\phi_i \\
\gamma_i^* &= \sigma\left(\sqrt{\rho_\gamma / s_\gamma}\right)\gamma_i.
\end{aligned}$$

Here, σ^2 is the combined variance of the spatial effects, and the ρ's are mixing parameters, interpreted as the proportion of the combined spatial variance explained by each model component. Note that $\rho_\theta = 1$ reduces the spatial component to purely overdispersion, $\rho_\phi = 1$ reduces the spatial component of the model to an adjacency ICAR model for the

spatial effects, and $\rho_\gamma = 1$ reduces the spatial component to a mobility ICAR model. Most importantly, if $\rho_\gamma > \rho_\phi$ then this means that the mobility data better explains variation in COVID-19 case counts than the adjacency data. As long as the spatial weights matrix and the mobility weights matrix are linearly independent, then having both spatial and mobility terms in our model present no issues with identifiability (Rodrigues and Assunçao, 2012). Finally, s_γ and s_ϕ are scaling factors, such that the geometric means of $s_\gamma^{-1}\text{Var}(\gamma_i)$ and $s_\phi^{-1}\text{Var}(\phi_i)$ are both ≈ 1 for each i, meaning that γ_i^* and ϕ_i^* are the log relative risk contributions from the movement data and physical data respectively (Sørbye and Rue, 2014). Scaling is absolutely necessary in order to conduct inference on the ρ's. We compute the scaling factors as follows

$$s = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log[\boldsymbol{Q}^-]_{ii}\right)$$

where $\boldsymbol{Q}^-$ is the generalized inverse of the $n \times n$ precision matrix (Freni-Sterrantino et al., 2018). In order to scale the precision matrices of the spatial effects, the generalized inverse for sparse matrices from Rue et al. (2017) was used. The diagonal elements, $[\boldsymbol{Q}^-]_{ii}$, of $\boldsymbol{Q}^-$ are referred to as the *marginal variances* of the structured spatial effects, i.e $\text{var}(\phi_i) = [\boldsymbol{Q}_\phi^-]_{ii}$ and $\text{var}(\gamma_i) = [\boldsymbol{Q}_\gamma^-]_{ii}$.

As was the case with the ICAR model in (3.1), we can derive the full conditionals of the combined spatial effect, $\tau_i = \phi_i^* + \gamma_i^* + \theta_i^*$, for the model described in Section 3.2.3

$$\tau_i|\boldsymbol{\tau}_{-i} \sim N\left[\frac{\sum_j(\frac{\rho_\phi}{s_\phi}w_{ij} + \frac{\rho_\gamma}{s_\gamma}v_{ij})\tau_j}{\frac{\rho_\phi}{s_\phi}\sum_j w_{ij} + \frac{\rho_\gamma}{s_\gamma}\sum_j v_{ij} + \rho_\theta}, \frac{\sigma^2}{\frac{\rho_\phi}{s_\phi}\sum_j w_{ij} + \frac{\rho_\gamma}{s_\gamma}\sum_j v_{ij} + \rho_\theta}\right] \tag{3.2}$$

These full conditionals can help provide some intuition as to the mechanism by which this model provides spatial smoothing. As $\rho_\gamma \to 1$, τ_i is simply the weighted sum of the other regions, where the weights are the proportion of region i's total movement between each other region. If $\rho_\phi \to 1$, the conditional mean of τ_i reduces to the arithmetic average of the spatial effects of its neighbours. If $\rho_\theta \to 1$, then the conditional mean shrinks to 0 (remember that $\rho_\phi + \rho_\gamma + \rho_\theta = 1$). Given that ρ_θ is positive, the conditional mean is always shrunk towards 0, resulting in spatial smoothing. In practice, the conditional mean will

be a weighted average of the estimates smoothed by the movement GMRF, the physical GMRF and 0. It is important to note here that the w_{ij}/s_ϕ and v_{ij}/s_γ are relative measures due to the scaling factors. That is, doubling the total amount of movement has no effect on the conditional mean or variance of τ_i. This is in contrast to the combined spatial effects in the commonly used Leroux model Leroux et al. (2000). Additionally, the variance of $\tau_i|\boldsymbol{\tau}_{-i}$ is lower when region i has a lot of movement or many neighbours, relative to the other regions.

3.2.5 Inference, computation, and validation

Four chains each with 3000 iterations of No U-Turn Sampling were used for parameter estimation within Stan Stan Development Team (2021). The first 1500 iterations were used as a warm-up, the 1500 remaining iterations from each chain were thinned by a factor of 10, leaving 600 total posterior samples to perform inference. As mentioned in Section 3.2.2, we require $\sum_i \phi_i = 0$. In practice, we use the soft constraint

$$\sum_i \phi_i \sim N(0, 0.001)$$

for computational reasons (as recommended by the Stan team (Morris et al., 2019)). To complete the model, priors for β and μ were $N(0, 1)$. To ensure the robustness of our results, we also ran BYM models using the adjacency data and the movement data separately. That is, for both Madrid and Castilla-Leon, we ran a model where we assumed $\rho_\gamma = 0$, and a separate model where $\rho_\phi = 0$. The results of these four models are presented in Section 3.3.2.

Our code and posterior samples are posted at `https://github.com/cghr-toronto/public/tree/master/covid/spain_public_code`.

	Parameter	Madrid Est (95% CrI)	Castilla-Leon Est (95% CrI)
$\boldsymbol{\rho}$	Movement	0.76 (0.54, 0.89)	0.88 (0.66, 0.98)
	Neighbour	0.13 (0.01, 0.39)	0.09 (0.01, 0.30)
	Independent	0.10 (0.02, 0.25)	0.02 (0.00, 0.09)
	μ	-5.36 (-5.51, -5.24)	-3.75 (-3.78, -3.73)
	β	0.12 (0.05, 0.20)	-0.01 (-0.04, 0.02)
	σ	0.65 (0.55, 0.78)	0.72 (0.63, 0.83)

Table 3.1: Posterior medians, and 95% credible intervals for $\boldsymbol{\rho}$ in BYM models using movement and physical (adjacency) data in the same model.

3.3 Results

3.3.1 Joint model

Table 3.1 shows posterior medians and credible intervals for the mixing parameters for the model with both movement and adjacency spatial effects. For both Madrid and Castilla-Leon, the proportion of spatial variation explained by $\boldsymbol{\gamma}$ is much higher than that of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The posterior probability that $\rho_\gamma > \rho_\phi$ was 0.997 for Madrid, and 0.998 for Castilla-Leon. However, $\boldsymbol{\phi}$ does seem to account for a non-trivial amount of spatial variation in both Madrid and Castilla-Leon. This means that although movement data is likely more explanatory, adjacency data can help with explaining variation in COVID-19 cases. Additionally, there is a substantial amount of spatial variation explained by the unstructured spatial effect for Madrid. This is not the case for Castilla-Leon, as most of the mass of the posterior of ρ_θ is near 0. This makes sense given that Madrid has a large metropolitan centre surrounded by a mix of suburbs and rural areas, so there are probably spatial confounders that our model is missing. For a plot of the posterior densities of $\boldsymbol{\rho}$, see Appendix 3.A.

Figures 3.4a through 3.4d show the spatial distribution $\boldsymbol{\gamma}^*$ and $\boldsymbol{\phi}^*$, plotted using the same colour scale for comparability. We can see that $\boldsymbol{\gamma}$'s log-relative risks have a lot more spatial variation in both Commmunities. The log-relative risks for $\boldsymbol{\phi}$ tend to have smooth spatial gradients, while $\boldsymbol{\gamma}$ tends to identify clusters of regions as high-risk areas. As seen in equation 3.2, the expectation of the combined spatial effects are a weighted average of these spatial effects, and 0 (notice that the numerator can be rewritten as $\sum_j (\frac{\rho_\phi}{s_\phi} w_{ij} + \frac{\rho_\gamma}{s_\gamma} v_{ij} + \rho_\theta \cdot 0)\tau_j$ where $\rho_\theta > 0$). Figures 3.4e and 3.4f show the predicted cases per 1000 people per region,

	Parameter	Madrid Est (95% CrI)	Castilla-Leon Est (95% CrI)
ρ	movement	0.82 (0.66, 0.91)	0.95 (0.89, 0.98)
	neighbour	0.56 (0.22, 0.83)	0.77 (0.58, 0.91)
μ	movement	-5.34 (-5.48, -5.23)	-3.75 (-3.78, -3.73)
	neighbour	-5.18 (-5.30, -5.09)	-3.74 (-3.78, -3.70)
β	movement	0.12 (0.05, 0.18)	-0.02 (-0.05, 0.02)
	neighbour	0.13 (0.01, 0.24)	-0.01 (-0.05, 0.04)
σ	movement	0.63 (0.55, 0.76)	0.74 (0.65, 0.83)
	neighbour	0.66 (0.56, 0.83)	0.58 (0.51, 0.66)

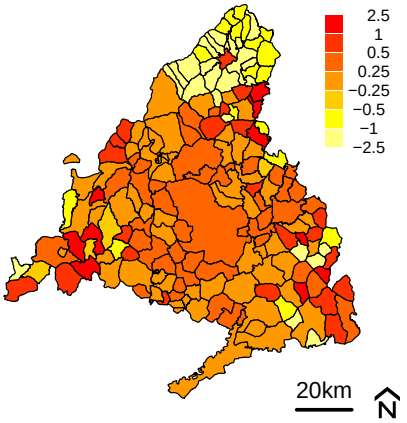
Table 3.2: Posterior medians, and 95% credible intervals for ρ in BYM models using movement and physical (adjacency) data in separate models.

showing highly similar patterns to the observed values in Figure 3.2.

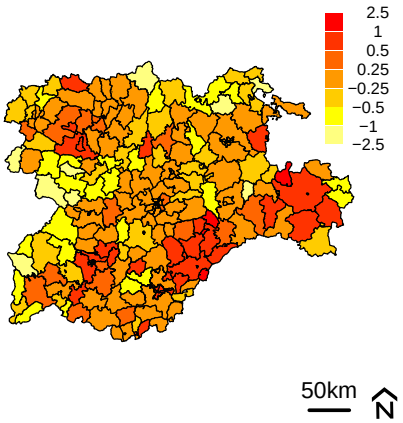
The standard deviation was slightly larger for Castilla-Leon than it was for Madrid. Figure 3.B.2 shows the the spatial distribution of the standard deviation of the cases per thousand people in both communities. Here, we can see that the standard deviation is pretty small in and around Madrid City, because the movement to and from Madrid City is causing a high-degree of spatial smoothing in the surrounding area. The effect of movement within regions, β, is associated with larger case counts in Madrid, but this is not the case for Castilla-Leon. This small covariate effect could result in more variance being attributable to the random effects, potentially contributing to the larger σ in Castilla-Leon.

3.3.2 Model Validation - Individual models

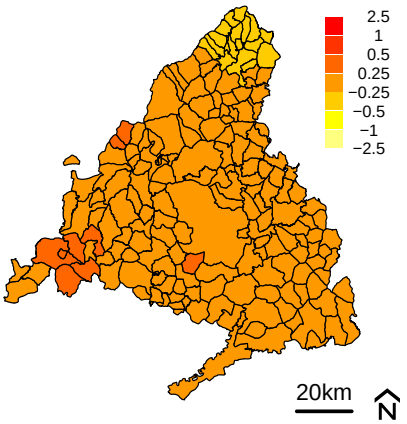
Table 3.2 shows posterior medians and credible intervals for the ρ parameter from the movement and physical BYM models described in Section 3.2.5, fit separately to Madrid and Castilla-Leon (four models total). In both regions, the model where spatial smoothing is induced by population movement explains a higher proportion of the variation in the outcome, indicated by the posterior density of ρ having more mass near 1. Additionally, the BYM model that used physical adjacency as a spatial smoother had a much wider credible interval for ρ, indicating more model uncertainty. Both models show more uncertainty in the region of Madrid than for Castilla-Leon, likely due to the fact that Madrid is more heterogeneous in terms of population density and other factors. For full posterior densities of the ρ parameter, see Appendix 3.A.2.



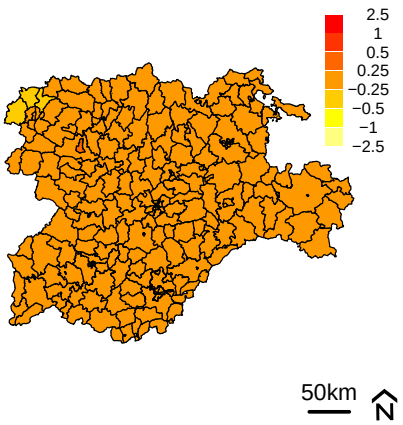
(a) $\boldsymbol{\gamma}^*$ - Madrid



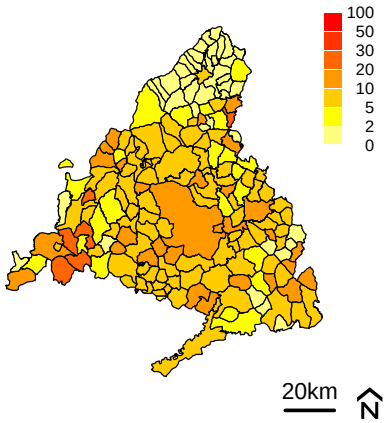
(b) $\boldsymbol{\gamma}^*$ - Castilla-Leon



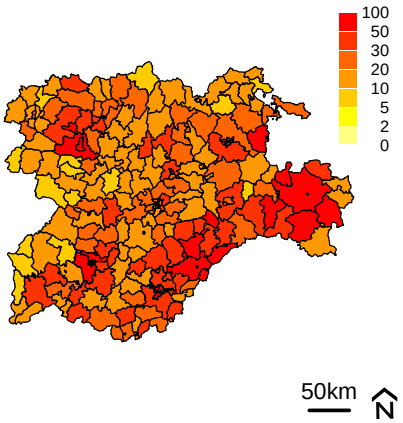
(c) $\boldsymbol{\phi}^*$ - Madrid



(d) $\boldsymbol{\phi}^*$ - Castilla-Leon



(e) $\boldsymbol{\lambda}^*1000$ - Madrid



(f) $\boldsymbol{\lambda}^*1000$ - Castilla-Leon

Figure 3.4: Log-relative risk contributions (a-d) from the movement effects ($\boldsymbol{\gamma}^*$) and spatial effects effects ($\boldsymbol{\phi}^*$). The predicted cases per thousand people are also presented (e-f).

3.4 Discussion

In this chapter, we have demonstrated that there is much value in using mobility data in combination with geographical proximity for defining correlation structures in COVID-19 incidence data. We showed that even while using only one week of movement data, we were able to explain the spatial variation in COVID-19 counts better than using the classic BYM model. Additionally, we showed that the model can be re-parametrized so that the means by which smoothing occurs in these mobility models is intuitive.

A key limitation of this work is that the models presented in this chapter do not serve as individual-level infectious disease models, as correlation is induced by a latent effect rather than direct dependence between the counts. However, this will be a natural extension of this work and would require the addition of many more parameters, including multiple mobility network components at various time points. This will ultimately pose a computational challenge as well.

An additional limitation of this work is that the availability and structure of mobility data will vary across data sources, and may only be available in higher income countries. Furthermore, there is selection bias in the movement data, as it only tracks those who actually have a cellphone, which may tend to be younger and more economically advantaged individuals. Given potential differences in quality of these data, its efficacy in spatial models may need to be assessed on a case by case basis.

Furthermore, the models presented in this chapter may suffer from overfitting. A potential remedy for this would be to put a penalized complexity prior (Simpson et al., 2017) on the mixing parameters, which may improve inference by shrinking ρ_γ (and perhaps ρ_ϕ) towards 0. An interesting area for future work would be to combine Dirichlet and penalized complexity priors to specify a joint prior for the mixing parameters as described in Fuglstad et al. (2020), which can be implemented using the *makemyprior* R package (Hem et al., 2021). This was deemed unnecessary for this work, as we were mainly interested in comparing ρ_γ to ρ_ϕ, and felt that our prior should not favour either one of these terms.

Despite these limitations, this work demonstrates the value of mobility data and provides the foundation for various extensions and future work. This data is only becoming more

abundant as time passes, and methods that allow for efficient use of this data are essential to model the current epidemic, and any spatial epidemiological application where population movement is likely a predictor of disease.

3.5 Bibliography

Arab-Mazar, Z., Sah, R., Rabaan, A. A., Dhama, K., and Rodriguez-Morales, A. J. (2020). Mapping the incidence of the covid-19 hotspot in iran – implications for travellers. *Travel Medicine and Infectious Disease*, 34:101630.

Aràndiga, F., Baeza, A., Cordero-Carrión, I., Donat, R., Martí, M. C., Mulet, P., and Yáñez, D. F. (2020). A spatial-temporal model for the evolution of the COVID-19 pandemic in Spain including mobility. *Mathematics*, 8(10):1677.

Audirac, M., Tec, M., Meyers, L. A., Fox, S., and Zigler, C. (2020). How timing of stay-home orders and mobility reductions impacted first-wave COVID-19 deaths in US counties. *medRxiv*. https://doi.org/10.1101/2020.11.24.20238055.

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., and Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254.

Baum, C. F. and Henry, M. (2020). Socioeconomic factors influencing the spatial spread of COVID-19 in the United States. *Preprints with The Lancet*. http://dx.doi.org/10.2139/ssrn.3559569.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Bilal, U., Barber, S., Tabb, L., and Diez-Roux, A. V. (2020). Spatial inequities in COVID-19 testing, positivity, incidence and mortality in 3 US cities: a longitudinal ecological study. *medRxiv*. https://doi.org/10.1101/2020.05.01.20087833.

Brainard, J. S., Rushton, S., Winters, T., and Hunter, P. R. (2020). Spatial risk factors for pillar 1 COVID-19 case counts and mortality in rural eastern England, U.K. *medRxiv*. https://doi.org/10.1101/2020.12.03.20239681.

Briz-Redón, Á. and Serrano-Aroca, Á. (2020). A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Science of the Total Environment*, 728:138811.

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87.

DiMaggio, C., Klein, M., Berry, C., and Frangos, S. (2020). Blacks/African Americans are 5 times more likely to develop COVID-19: spatial modeling of New York city zip code-level testing results. *MedRxiv*, 14:2020.

Duncan, E. W., White, N. M., and Mengersen, K. (2017). Spatial smoothing in bayesian models: a comparison of weights matrix specifications and their impact on inference. *International Journal of Health Geographics*, 16(1):1–16.

Freni-Sterrantino, A., Ventrucci, M., and Rue, H. (2018). A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and spatio-temporal epidemiology*, 26:25–34.

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15(4):1109–1137.

Geilhufe, M., Held, L., Skrøvseth, S. O., Simonsen, G. S., and Godtliebsen, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal*, 56(3):363–382.

General Directorate of Information Systems, Quality and Pharmaceutical Provision (2021). Open Data of Castile and Leon. https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html. Accessed: Jan 10, 2021.

Giuliani, D., Dickson, M. M., Espa, G., and Santi, F. (2020). Modelling and predicting the spatio-temporal spread of coronavirus disease 2019 (COVID-19) in Italy. *Preprints with The Lancet*. http://dx.doi.org/10.2139/ssrn.3559569.

Hem, I. G., Fuglstad, G.-A., and Riebler, A. (2021). makemyprior: Intuitive construction of joint priors for variance parameters in r. *arXiv preprint arXiv:2105.09712*.

Huang, G. and Brown, P. E. (2021). Population-weighted exposure to air pollution and COVID-19 incidence in Germany. *Spatial Statistics*, 41:100480.

Huang, H., Wang, Y., Wang, Z., Liang, Z., Qu, S., Ma, S., Mao, G., and Liu, X. (2020). Epidemic features and control of 2019 novel coronavirus pneumonia in Wenzhou, China. *Preprints with The Lancet*. http://dx.doi.org/10.2139/ssrn.3550007.

Iacus, S. M., Santamaria, C., Sermi, F., Spyratos, S., Tarchi, D., and Vespe, M. (2020). Human mobility and COVID-19 initial dynamics. *Nonlinear Dynamics*, 101(3):1901–1919.

Kang, D., Choi, H., Kim, J.-H., and Choi, J. (2020). Spatial epidemic dynamics of the COVID-19 outbreak in China. *International Journal of Infectious Diseases*, 94:96–102.

Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490):493–497.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

Liu, J., Zhou, J., Yao, J., Zhang, X., Li, L., Xu, X., He, X., Wang, B., Fu, S., Niu, T., et al. (2020). Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China. *Science of the Total Environment*, 726:138513.

Ministry of Economic Affairs and Digital Transformation (2021). Epidemiological surveillance network of Madrid. https://datos.gob.es. Accessed: Jan 10, 2021.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in Stan. *Spatial and Spatio-temporal Epidemiology*, 31:100301.

Orea, L. and Álvarez, I. C. (2020). How effective has the Spanish lockdown been to battle COVID-19? a spatial analysis of the coronavirus propagation across provinces. *Documento de trabajo FEDEA*, 3:1–33.

Persson, J., Parie, J. F., and Feuerriegel, S. (2021). Monitoring the COVID-19 epidemic with nationwide telecommunication data. *arXiv preprint arXiv:2101.02521*.

Pullano, G., Valdano, E., Scarpa, N., Rubrichi, S., and Colizza, V. (2020). Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study. *The Lancet Digital Health*, 2(12):e638–e649.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Rodrigues, E. C. and Assunçao, R. (2012). Bayesian spatial models with a mixture neighborhood structure. *Journal of Multivariate Analysis*, 109:88–102.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Schrödle, B., Held, L., and Rue, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics*, 68(3):736–744.

Shi, P., Dong, Y., Yan, H., Zhao, C., Li, X., Liu, W., He, M., Tang, S., and Xi, S. (2020). Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Science of the Total Environment*, 728:138890.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

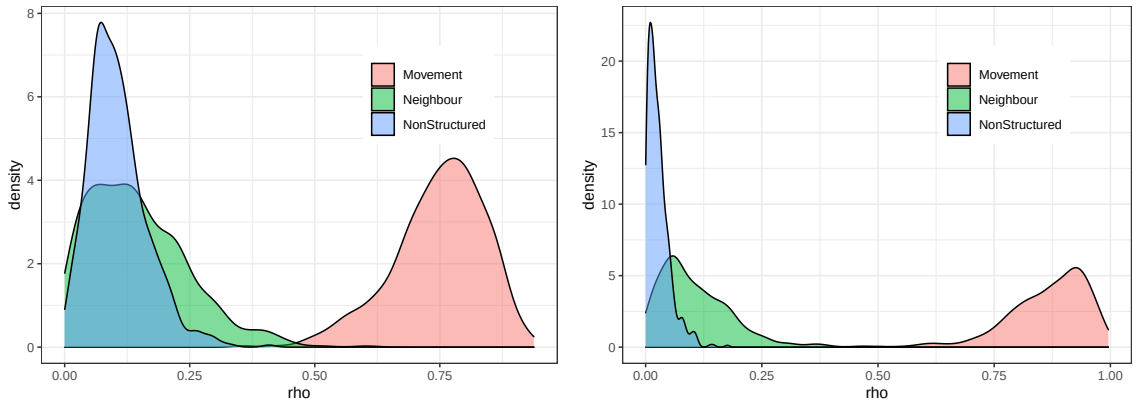
Stan Development Team (2021). Stan Modeling Language Users Guide and Reference Manual, version 2.26. https://mc-stan.org.

Sugg, M. M., Spaulding, T. J., Lane, S. J., Runkle, J. D., Harden, S. R., Hege, A., and Iyer, L. S. (2021). Mapping community-level determinants of COVID-19 transmission in nursing homes: A multi-scale approach. *Science of the Total Environment*, 752:141946.

Valencia, A. (2021). COVID-19 Flow Maps. https://flowmaps.life.bsc.es/flowboard/. Accessed: Jan 10, 2021.

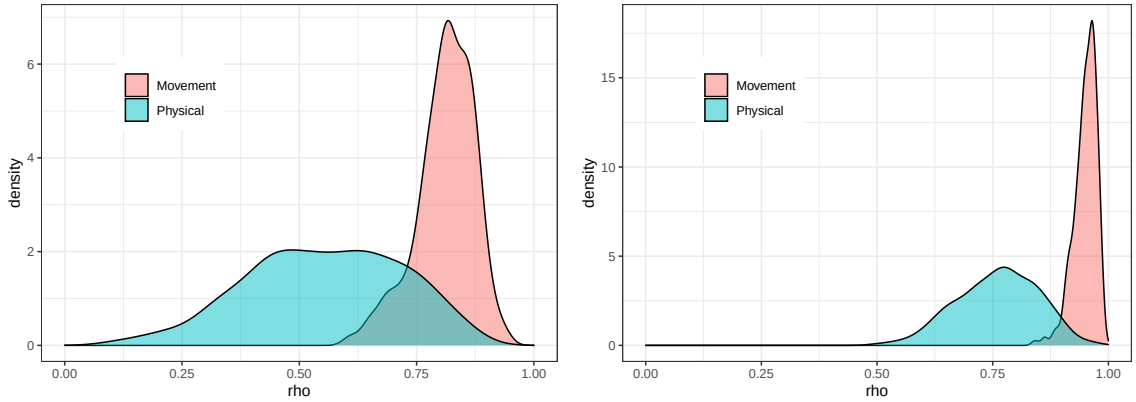
Volkova, V. V., Howey, R., Savill, N. J., and Woolhouse, M. E. (2010). Sheep movement networks and the transmission of infectious diseases. *PloS one*, 5(6):e11185.

3.A Appendix Posterior Densities of ρ for various models



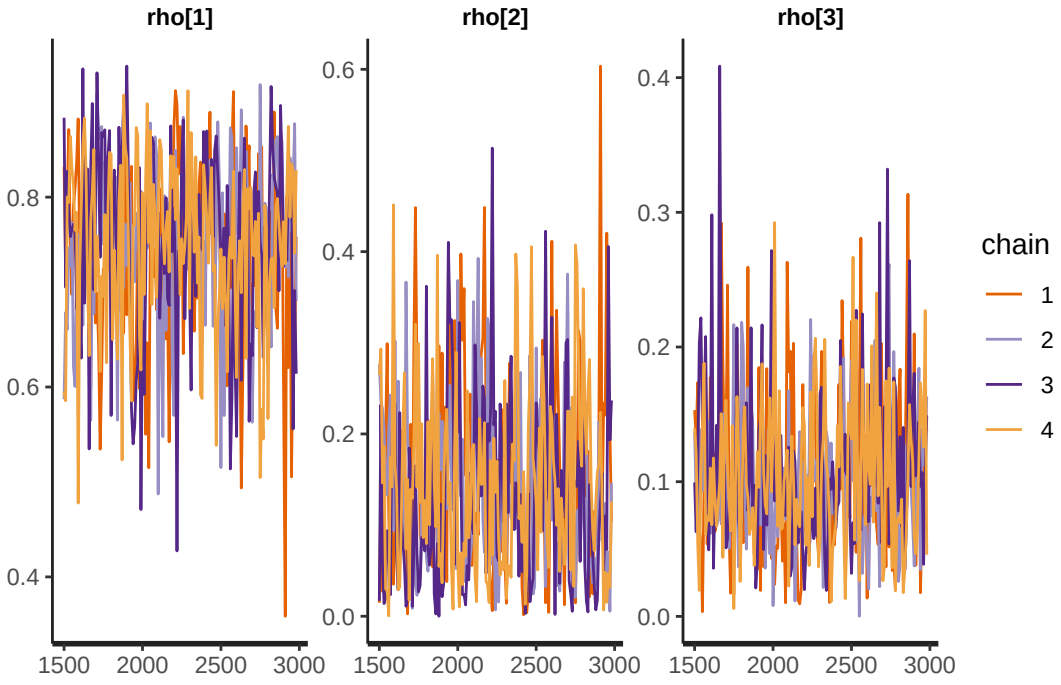
(a) Madrid (b) Castilla-Leon

Figure 3.A.1: Posterior Density of the proportion of variance explained by each of the 3 spatial parameters when adjacency and movement data are included in the same model

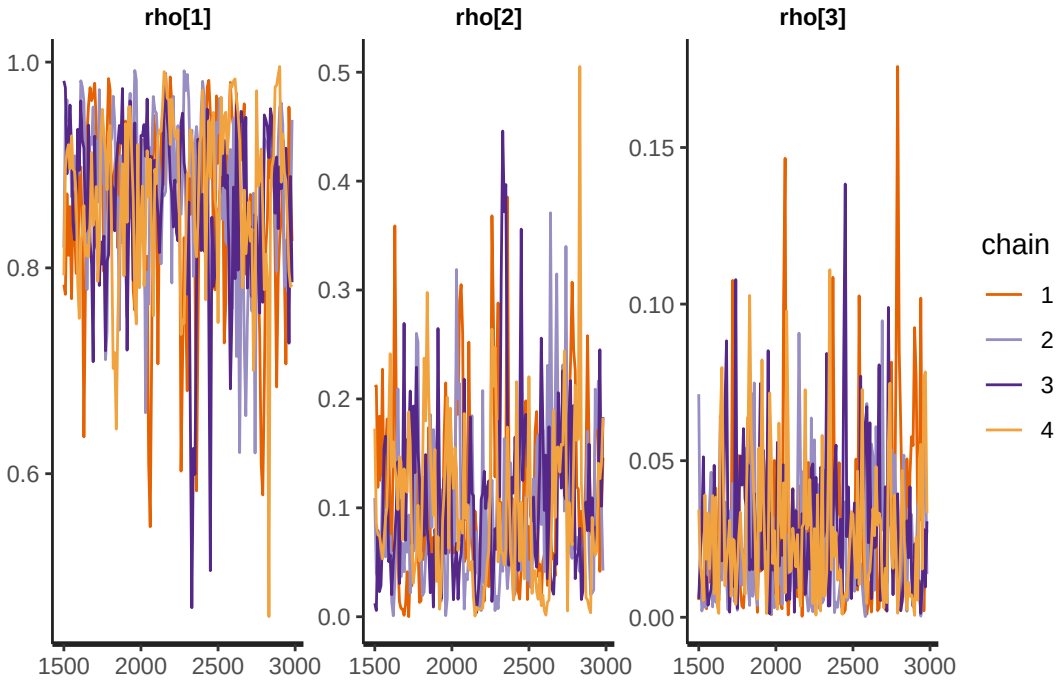


(a) Madrid (b) Castilla-Leon

Figure 3.A.2: Posterior Density of the proportion of variance explained by spatial components when adjacency and movement data are used in separate models (model validation).



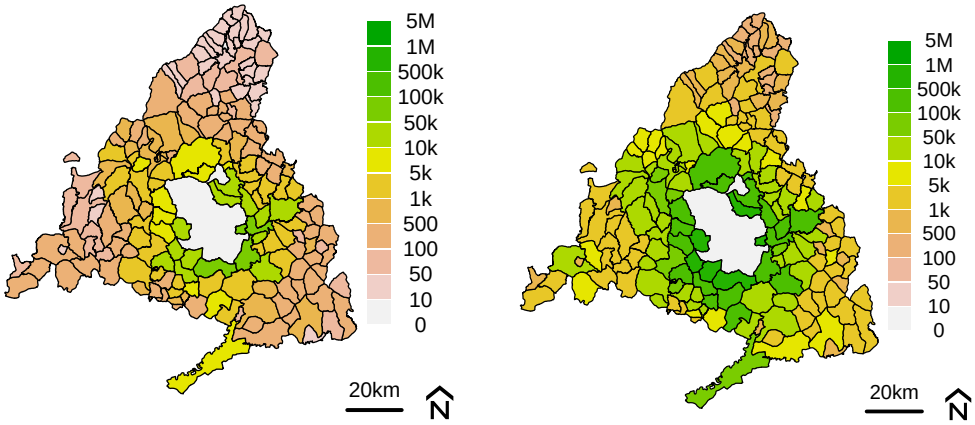
(a) Madrid



(b) Castilla-Leon

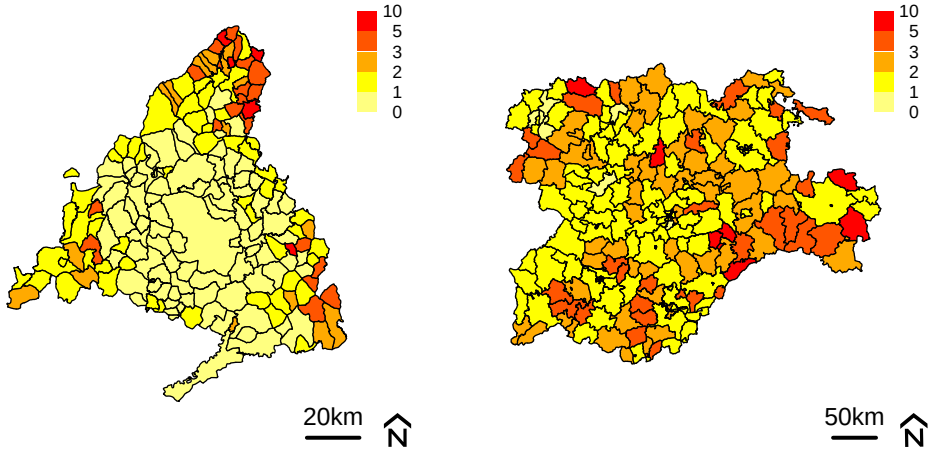
Figure 3.A.3: Traceplots of $\boldsymbol{\rho}$

3.B Additional Spatial plots



(a) To (b) From

Figure 3.B.1: Number of trips to and from Madrid City (white).



(a) Madrid (b) Castilla-Leon

Figure 3.B.2: Standard deviations of predicted cases per thousand people.

Chapter 4

Leveraging mobility networks to assess COVID-19 travel risk

Abstract

Since the beginning of the COVID-19 pandemic, public health authorities across the globe have implemented policies, such as lockdowns, in an attempt to reduce population mobility, and consequently, person-to-person contacts. It is well known that lockdowns reduce mobility, but to what extent does this reduction in mobility lead to lower infection rates? In this chapter, we extend the Endemic-Epidemic modeling framework in a principled manner, incorporating temporally changing mobility network data and quantifying the risk associated with travelling throughout the first year of the pandemic in two Spanish Communities.

4.1 Introduction

Since the beginning of the COVID-19 pandemic, public health authorities across the globe have implemented policies such as lockdowns, with the intention of reducing population mobility and, consequently, person-to-person contacts. Countless studies have attempted to quantify the effectiveness of mobility reductions by using a variety of data sources and statistical methods. Cellphone-derived mobility data is well-suited this purpose, as we can use it to quantify the severity of a lockdown as well as relate it to case counts via a statistical

model such as a generalized linear model or infectious disease model.

Slater et al. (2021b) showed that mobility data better captures spatial heterogeneity in COVID-19 case counts than spatial proximity in Bayesian spatial models. However, the temporal relationship between mobility and case counts poses great modeling challenges, as the correlation between the two changes in each wave (Gottumukkala et al., 2021). We argue that since mobility affects the reproduction rate of infectious diseases (as opposed to the absolute counts), we can indeed infer the impact of mobility on case counts using a spatio-temporal infectious disease model.

In the last two decades, a class of infectious disease models known as *Endemic - Epidemic* (EE) models have gained popularity due to their simplicity and forecasting ability (Held et al., 2005). A simple version of these models can be written as:

$$\begin{aligned} Y_t | Y_{t-1} &\sim \text{Pois}(\mu_t) \\ \mu_t &= v_t + \alpha Y_{t-1} \end{aligned}$$

where Y_t is the number of cases, v_t is the "endemic" component which describes new cases that are not explained by previous cases, and αY_{t-1} is the "epidemic component" which describes new cases that are directly attributable to previous cases. These models have since then been extended to include temporally changing α (Held et al., 2006), multiple diseases (Paul et al., 2008), random effects (Paul and Held, 2011), seasonal effects (Held and Paul, 2012), serial interval distributions of disease (Bracher and Held, 2020) and more. EE models overcome the computational difficulties of fitting classic compartmental (SIR) models, and are an attractive alternative when an abundance of data is available (Wakefield et al., 2019).

An example of a multi-region EE model is

$$\begin{aligned} Y_{it} | \boldsymbol{Y}_{-i,t-1} &\sim \text{Pois}(\mu_{it}) \\ \mu_{it} &= v_{it} + \alpha \sum_j w_{ji} Y_{j,t-1} \end{aligned} \tag{4.1}$$

where i and j are region indicators and w_{ji}'s represent (potentially asymmetric) weights

between regions j and i. Typically these weights are row-normalized (sum to 1) but this is not necessary. The most common form of weights is some function of physical distance or proximity, such as

$$w_{ji} = \frac{1}{|i \sim j|}$$

or

$$w_{ji} = (o_{ji} + 1)^{-\rho}$$

where $|i \sim j|$ is the number of regions sharing a border (neighbors) with region j, o_{ji} is the minimum number of region borders you would have to cross to get from region j to i, and ρ is a parameter to be estimated. These weights tend to work well because they are good proxy for the number of people moving between regions, and resultingly, contact rates between infectious and susceptible people. More interestingly, these weights have been combined with or replaced by other data sources to more accurately estimate the contact rates between individuals of different regions. For instance, Schrödle et al. (2012) used assymetric mobility weights to model the spread of Coxiellosis in Swiss cows. Geilhufe et al. (2014) used mobility data to estimate the relationship between distance and mobility, and define their weights based on this relationship. Meyer and Held (2017) estimate contact rates between age groups using external data and combine these data with spatial proximity weights and used this as an estimate for contact rates between age groups across various regions. Fritz and Kauermann (2022) build weights based on estimated social connectedness via social media data. Grimée et al. (2021) combine border closure data with proximity weights to assess the effectiveness of lockdowns during the COVID-19 pandemic, and estimate case counts under counterfactual scenarios. Celani and Giudici (2022) incorporated mobility weights to assess the effectiveness of containment measures in Italy. Each of these works show that proximity weights can be supplemented or replaced with external data to improve forecasting or inference.

Much of the methodological progress surrounding EE models aims to improve forecasting ability based on the framework presented in Gneiting and Raftery (2007). Consequently, the applications of these models tend to lack interpretability. When the goal is learning about the biological phenomenon, we must make every effort to ensure our model parameters have

clear meanings, and that our results are biologically plausible. Covariates introduced should be done so carefully, and should effect model parameters in a way that are consistent with infectious disease dynamics.

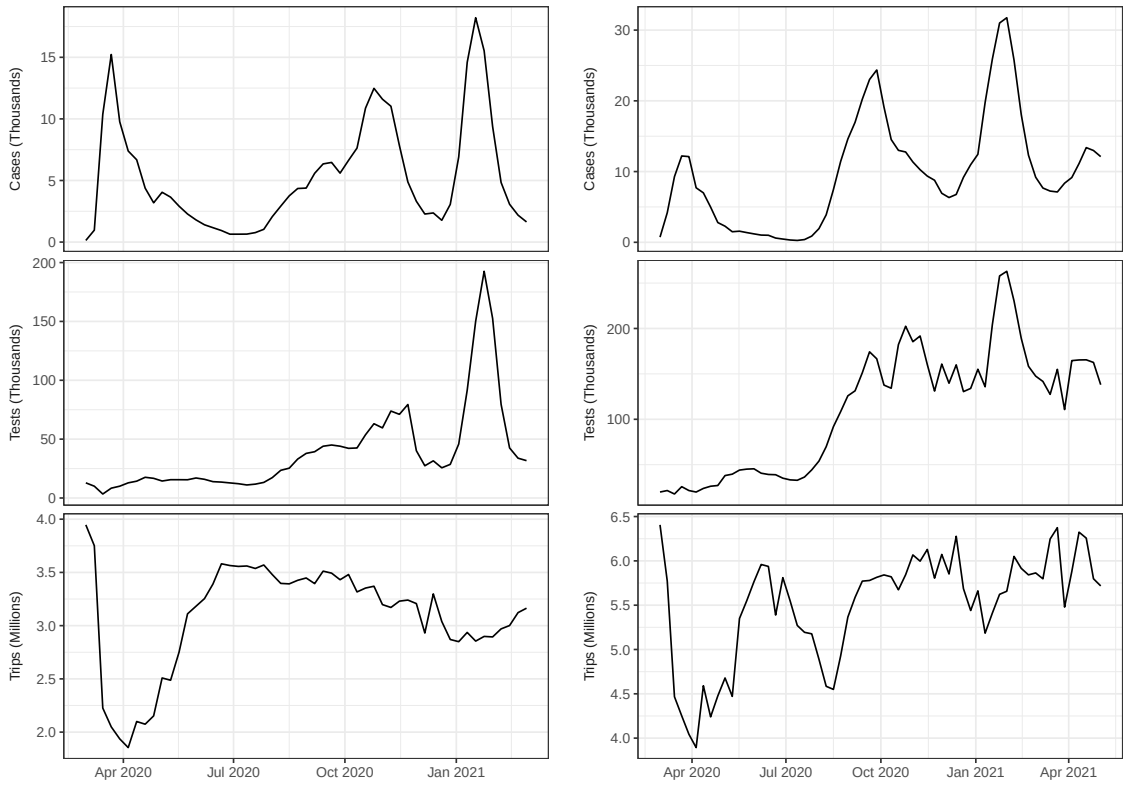
In this chapter, we derive a mobility extended spatio-temporal EE model where contact rates are a temporally changing function of mobility. In doing so, we ensure interpretability of our important parameters, and carefully specify the functional form of the reproduction number via data exploration methods. We use this model to infer the risk associated with travelling during the first 12-15 months of the COVID-19 pandemic in two Spanish Autonomous Communities using high resolution areal mobility networks derived from cellphone GPS signals.

This chapter is structured as follows. We introduce the data that motivated this work in Section 4.2, and present our model and methods in Section 4.3. In Section 4.4, we apply our model to two Spanish Communities, inferring the risk associated with travelling in both. We end with a discussion of our model results, limitations, and future work.

4.2 Data

This chapter focuses on Madrid and Castilla-Leon, two Communities in Spain. Madrid, with a population of approximately 6.8 million, is home to Madrid City, the capital of Spain. Castilla-Leon is geographically the largest Community in Spain, with a population of 2.5 million, and is thus much more rural than the Community of Madrid. Each community is divided into smaller subregions (Madrid has 179 subregions, Castilla-Leon has 245). We obtain weekly mobility network data for the trips between and within each of these subregions, alongside COVID-19 cases. The mobility data used in this chapter is a temporal extension of that used in Slater et al. (2021b). These data can be downloaded from Ministerio de Transportes Movilidad Y Agends Urbana, Gobierno de España (2022), and are described in detail in Ponce-de Leon et al. (2021). Although daily mobility data is available, we aggregated it by week to match the resolution of the case data, avoiding the well-known day-of-the-week effect of COVID-19 case reporting (Slater et al., 2021a).

For Castilla-Leon, the weekly case data from March 1, 2020, to March 7, 2021 was



(a) Castilla-Leon (b) Madrid

Figure 4.1: Time series of cases, trips, and tests between March 2020 and March 2021 (Castilla-Leon), and March 2020 and May 2021 (Madrid).

obtained from the open data portal of Castilla-Leon (General Directorate of Information Systems, Quality and Pharmaceutical Provision, 2022). For Madrid, case data from March 1, 2020, to May 9, 2021 was obtained from Epidemiological Surveillance Network of Madrid (Epidemiological Surveillance Network of Madrid, 2022). Note that nearing the end of our Madrid case data, vaccines were being administered to the public, and thus should be accounted for. Country level vaccine data was obtained from Ministerio de Sanidad, Gobierno De España (2022b), where about 30% of the population had been vaccinated prior to the end of our Madrid case data. Although we don't expect this to have a substantial impact on our results, the effect of vaccines should at least be explored.

Daily testing data was acquired from the Government of Spain Ministry of Health website (Ministerio de Sanidad, Gobierno De España, 2022a). Testing data was only available at the community level, which we aggregated by week. The first two weeks of the pandemic

were missing, and were imputed using the fourth and fifth weeks, as this seemed to suit the temporal testing pattern well.

4.3 Methodology

In this section, we start by introducing a single-region version of a mobility extended EE model, a derivation inspired by Bauer and Wakefield (2018). We then extend this model to a multi-region model, and describe reasonable assumptions that make implementing this model computationally feasible. We then describe the careful processes of accounting for delayed reporting, serial intervals, and under reporting, while retaining interpretability of our model. We conclude the section with an explanation of the summary statistics used in this chapter, followed by our inference methodology.

4.3.1 Single region model

In epidemiology, the force of infection at time t, λ_t, is defined as the rate at which susceptible individuals become infected. Mathematically, we write it as (Halloran et al., 2010)

$$\lambda_t = \mathcal{C}_{t-1} \times \mathcal{P}_{t-1} \times \frac{I_{t-1}}{N}$$

where $\mathcal{C}_{t-1}$ is the number of contacts between infectious and susceptibles individuals, $\mathcal{P}_{t-1}$ is the probability of infection given a contact between an infectious and susceptible individual, and I_{t-1} is the number of infectious individuals at time $t-1$ (meaning that $\frac{I_{t-1}}{N}$ is the prevalence at time $t-1$). For simplicity, we will assume $I_{t-1} = y_{t-1}$, that the number of infectious individuals equals the number of cases, but will later relax this assumption. Typically, the number of contacts is assumed to be constant (*frequency dependent*) or proportional to the population size N (*density dependent*). Distinguishing between these is inconsequential in our models as we will see later on. However, in this chapter, we assume C_t is a function of mobility, w, which is the number of trips as described in Section 4.2.

That is, we assume that the contacts function takes the form

$$\mathcal{C}_{t-1}(w) = c^{AR} + \sum_{d=1}^{D} c_d^{\text{mob}} w_{t-d}$$

where the c's are parameters to be estimated, and D is some small integer (i.e 1,2 or 3) chosen by the analyst. The reason for including higher lags of mobility is because cases appearing at y_{t-1} may have been infectious at time $t-2$ or earlier but didn't immediately produce a positive test. If we assume that the per-contact probability of infection is time constant $\mathcal{P}_{t-1} = p$, then our force of infection is

$$\begin{aligned}
\lambda_t &= \big(c^{\text{AR}} + \sum_{d=1}^{D} c_d^{\text{mob}} w_{t-d}\big) \times p \times \frac{y_{t-1}}{N} \\
&= c^{\text{AR}} p \frac{y_{t-1}}{N} + \Big(\sum_{d=1}^{D} c_d^{\text{mob}} p w_{t-d}\Big) \frac{y_{t-1}}{N} \\
&= \alpha^{AR} \frac{y_{t-1}}{N} + \Big(\sum_{d=1}^{D} \alpha_d^{\text{mob}} w_{t-d}\Big) \frac{y_{t-1}}{N}
\end{aligned}$$

If we make the assumption that infected people are equally likely to move as the rest of the population, then α_d^{mob} can be interpretted as the number of infected trips from d time units ago required to cause an infection at time t, and α^{AR} is the number of new infections caused by previous infections, but not related to mobility. This assumption may not be as problematic as it may sound, as people can be infectious several days prior to showing any symptoms (He et al., 2020), and thus likely not to change their behavior in this time.

Bauer and Wakefield (2018) show that when the disease is rare and the susceptible population is close to the total population, the number of infections at time t is approximately Poisson distributed

$$Y_t | Y_{t-1} \sim \text{Pois}(\lambda_t).$$

Furthermore, it is common, and mathematically convenient, to assume that there is some number infections, α^{EX}, that come from outside the region, not related to the previous cases y_{t-1}. In doing so, we arrive at an extension of the univariate Endemic-Epidemic

model (Held et al., 2005)

$$
\begin{aligned}
Y_t | Y_{t-1} &\sim \text{Pois}(\lambda_t^\dagger) \\
\lambda_t^\dagger &= \alpha^{\text{EX}} + \alpha^{\text{AR}} Y_{t-1} + \Big(\sum_{d=1}^{D} \alpha_d^{\text{mob}} w_{t-d} \Big) Y_{t-1}
\end{aligned} \tag{4.2}
$$

This model can be thought of as a branching process with immigration, with reproduction number, $\alpha^{\text{AR}} + \sum_{d=1}^{D} \alpha_d^{\text{mob}} w_{t-d}$, that linearly depends on mobility, and an immigration of α^{EX}. This implies that mobility only effects the reproduction rate of the disease, and does not relate directly to the case counts. This is an attractive property of this model, as mobility can only cause infections in the next generation if infectious people from the previous generation move around. If the effect of mobility is small, then the reproduction number will be almost entirely described by the constant α^{AR}. In other words, α^{AR} can be thought of as an autoregressive term that relates previous cases to current cases, or it can be thought of as the intercept in the line relating the reproduction number to mobility.

α^{EX} represents an influx of cases caused by infectious people outside our dataset infecting susceptibles in our region. When regions are large, this number should be relatively small. Including α^{EX} serves to prevent our branching process from dying out, which will be especially helpful in the multi-region case when there are small subregions with low amounts of mobility. In some applications, this component is appropriately referred to as an "endemic" component, as it may describe predictable yearly fluctations/periodicities in disease incidence. However, even cases that arise in an "endemic" are often still attributable previous cases, but with a more predictable/periodic pattern, and can be thought of as the "background rate of disease" (Gordis, 2013). COVID-19 had not yet reached endemic status, thus estimating the background rate of infection is challenging. Thus we believe the term "exogenous" is more appropriate for our application, and should be viewed as factors influencing the absolute number of cases in a region as opposed to the infectiousness of the disease.

An alternative way to view model (4.2) is using the competing risks framework as in Bauer and Wakefield (2018). That is, we can view the exogenous, autoregressive, and

movement terms as their own Poisson process, and the total force of infection, indicated by †, is the sum of three Possion random variables with mean

$$\lambda_t^{\dagger} = \lambda_t^{\mathrm{EX}} + \lambda_t^{\mathrm{AR}} + \lambda_t^{\mathrm{mob}}.$$

In other words, a susceptible can be infected in one of three ways, all with some positive probability. This tells us that each of the α's should be positive. Furthermore, we can compute the proportion of cases attributable to movement (PCAtM) at time t as $\frac{\lambda_t^{\mathrm{mob}}}{\lambda_t^{\dagger}}$. We will use this measure and its associated uncertainty to assess the association between mobility and infection. We will now extend our model to the multi-region case.

4.3.2 Multi-region model

Now that we are dealing with more than one geographic region (245 for Castilla-Leon, 179 for Madrid), we must define a region-wise force of infection. The force of infection, λ_{jit} is defined as the rate at which infectious individuals in region j, infect susceptible individuals in region i, at time t. Similar to the univariate case, we can write this mathematically as

$$\begin{aligned}
\lambda_{jit} &= \mathcal{C}_{ji}(w_{ji,t-1}) p_{ji,t-1} \frac{y_{j,t-1}}{N_j} \\
&= (c^*_{ji,t-1} p_{ji,t-1} + \sum_{d=1}^{D} c^{\mathrm{mob}}_{ji,t-d} p_{ji,t-1} w_{ji,t-d}) \frac{y_{j,t-1}}{N_j} \\
&= (\alpha^*_{jit} + \sum_{d=1}^{D} \alpha^{\mathrm{mob}}_{ji,d} w_{ji,t-d}) \frac{y_{j,t-1}}{N_j},
\end{aligned}$$

where $\alpha^*_{jit} = c^*_{jit} p_{jit}$ is number of cases in region i attributed to a single case in region j, that is not accounted for by mobility. $\alpha^{\mathrm{mob}}_{jit} = c^{\mathrm{mob}}_{jit} p_{jit}$ is the of cases in region i caused by infected trips from region j to i. As is, the number of model parameters grow at a rate of $O(I^2 \times T)$ where I is the number of regions and T is the number of time points. Given that we will be dealing with hundreds of subregions, we simplify the problem by making the following assumptions:

- We assume that α^*_{jit} is temporally constant, and is equal to the sum of an autoregressive term and a spatial term: $\alpha^*_{jit} = \alpha_i^{\mathrm{AR}} + \alpha_i^{\mathrm{spat}} \sum_j v_{ji}$, where $v_{ji} = \frac{1}{|\mathrm{Ne}(j)|}$, with

$|\text{Ne}(j)|$ being the number regions sharing a border (neighbors) with region j.

- We assume that $\alpha^{\text{mob}}_{jit}$ is temporally constant, and does not depend on the origin j, but only on the destination i: $\alpha^{\text{mob}}_{ji,t-d} = \alpha^{\text{mob}}_{i,d}$.

- For every i, t there are j independent Poisson processes (with mean λ_{jit}) competing to infect susceptibles in region i. Since the sum of Poisson processes is Poisson, we arrive at $\lambda_{it} = \sum_j \lambda_{jit}$.

The number of parameters to be estimated is now $O(I)$, which is much more computationally feasible. Adding an exogenous component, α^{EX}_i, for each region, leads us to an extension of the multi-region Endemic-Epidemic model

$$Y_{it}|\boldsymbol{Y}_{-i,t-1} \sim \text{Pois}(\lambda^{\dagger}_{it})$$

$$\lambda^{\dagger}_{it} = \underbrace{\alpha^{\text{EX}}_i}_{\lambda^{EX}_{it}} + \underbrace{\alpha^{\text{AR}}_i \frac{Y_{i,t-1}}{N_i}}_{\lambda^{\text{AR}}_{it}} + \underbrace{\alpha^{\text{spat}}_i \sum_{j \neq i} v_{ji} \frac{Y_{j,t-1}}{N_j}}_{\lambda^{\text{spat}}_{it}} + \underbrace{\sum_{d=1}^{D} \alpha^{\text{mob}}_{id} \sum_j w_{ji,t-d} \frac{Y_{j,t-1}}{N_j}}_{\lambda^{\text{mob}}_{it}}$$

If the α^{mob}_d's are 0, then this model reduces to a typical EE model as seen frequently in the literature.

4.3.3 Delayed reporting, serial intervals, and incubation periods

The modeling challenges caused by delayed reporting of cases is closely tied with the serial interval of infection and to the incubation period. The serial interval for COVID-19 has been estimated to be between 4 and 7 days, while the incubation period is between 4 and 9 days (Alene et al., 2021). These quantities can vary between individuals, and can be hard to measure due to delayed reporting/testing. García-García et al. (2021) showed that in Spain, cases may have peaked several days before the observed peak in cases, but the delay varied across Spanish provinces. Although we don't attempt to estimate any of these factors individually, we may be able to account for their combination by including additional time lags in our model. Bracher and Held (2020) showed that including cases from several time units in the past improved forecasting ability of EE models in the presence of random serial

intervals. Following their guidance, we assume that the number of cases at time t is a weighted average of cases at s time points in the past. Our force of infection is now:

$$\begin{aligned}\lambda_{jit} &= \left(\alpha_i^{\text{AR}} + \alpha_i^{\text{spat}} v_{ji} + \sum_{d=1}^{D} \alpha_{i,d}^{\text{mob}} w_{ji,t-d}\right) \sum_{s=1}^{S} \rho_s \frac{Y_{j,t-s}}{N_j} \\ &= \left(\alpha_i^{\text{AR}} + \alpha_i^{\text{spat}} v_{ji}\right) \sum_{s=1}^{S} \rho_s \frac{Y_{j,t-s}}{N_j} + \sum_{d=1}^{D} \alpha_{i,d} w_{ji,t-d} \sum_{1 \le s < d} \rho_s \frac{Y_{j,t-s}}{N_j} \end{aligned} \tag{4.3}$$

where $\sum_{s=1}^{S} \rho_s = 1$. Note that in the second term, we exclude terms where the mobility lag is higher than the cases lag (e.g Y_{t-2}, w_{t-1}) as we don't suspect any reporting delay with our mobility data, so if someone tests positive at $t-2$, when they move at time $t-1$, they should no longer be infectious, thus their mobility won't contribute to new cases.

It remains to specify D and S. In determining D, we first consider a univariate model for case counts: $Y_t|Y_{t-1} \sim \text{Poisson}(\lambda_t)$ with $\lambda_t = \phi_t Y_{t-1}$ where ϕ_t is the effective reproduction number at time t. If we solve for ϕ_t, and replace λ_t with Y_t, then we arrive at a crude estimate of R_{eff} (Crude R_{eff}) $\phi_t \approx \frac{Y_t}{Y_{t-1}}$. To determine how many mobility lags to include in our model, we examine the relationship of w_{t-h} with $\frac{Y_t}{Y_{t-1}}$ for various lags $h > 0$. If w_{t-h} has a strong relationship with $\frac{Y_t}{Y_{t-1}}$, then we include this in our model.

Determining S is more challenging, but we can be fairly confident that $S \le 2$, as it is fairy unlikely that a case would take 3 or more weeks from primary case to cause a secondary case. Thus we will investigate values of $S = 1$ and $S = 2$.

4.3.4 Under-reporting

Epidemic curves are well-known to suffer from under-reporting. The number of cases is usually an underestimate of the number of infections, because of testing capacity limitations and asymptomatic or minimally symptomatic cases going undetected. This is troublesome when conducting inference, as our estimate of the α's depends on reporting. There are several methods that correct for underreporting, such as those in Wakefield et al. (2019), but this involves estimating $T \times I$ latent discrete variables, which quickly becomes infeasible. One solution to the under-reporting issue is to assume that the true cases, $Y_{i,t}^*$ in region i

at time t, are reported with some probability $\gamma_{i,t}$

$$\begin{aligned} Y_{i,t}|Y_{i,t-1} &\sim \text{Bin}(Y^*_{i,t}, \gamma_{i,t}) \\ Y^*_{i,t}|Y^*_{i,t-1} &\sim \text{Poisson}(\lambda^\dagger_{i,t-1}) \\ \implies Y_{i,t}|Y_{i,t-1} &\sim \text{Poisson}(\gamma_{i,t}\lambda^\dagger_{i,t-1}) \end{aligned} \tag{4.4}$$

where $Y_{i,t}$ and $Y^*_{i,t}$ are the observed and the actual COVID-19 cases, respectively. That is, the observed number of reported cases is binomially distributed with reporting probability γ_{it}. We must place some restriction on the γ_{it}, otherwise the model is unidentified. In this chapter, we assume the reporting probability $\log(\gamma_{i,t}) = \beta_{\text{test}} \log\left(\frac{x_t}{\max_t(x_t)}\right)$, where x_t is the number of tests at time t.

4.3.5 Summary Statistics

The basic reproduction number, R_0, is a succinct way to describe the infectiousness of a disease, and is defined as the average number of secondary cases caused by an index primary case (Diekmann et al., 2013). When dealing with more than one region (or some other strata), measuring R_0 is nontrivial. Rewriting (4.1) in matrix form, we obtain

$$\boldsymbol{\mu}_t = \boldsymbol{\nu}_t + \boldsymbol{\Lambda}\boldsymbol{Y}_{t-1}$$

with $\boldsymbol{\Lambda}_{ij} = \lambda w_{ij}$. Diekmann et al. (1990) use a limit argument to show that after a large number of generations, the typical number of primary cases given secondary cases is well described by the dominant eigenvalue of $\boldsymbol{\Lambda}$ (assuming $\boldsymbol{\Lambda}$ is irreducible and aperiodic). They thus define R_0 to be this dominant eigenvalue. We don't believe that this argument extends well to the case when $\boldsymbol{\Lambda}$ is temporally changing, since $\boldsymbol{\Lambda}$ only 'acts' on $\boldsymbol{Y}$ for ≈ 1 generation (this is assuming that the generation time is one time unit). However, in keeping with the EE literature, we will present dominant eigenvalues over time where possible, as it is likely a good representation of infectiousness, but we suspect it is biassed and is more noisy than R_0 should be. Where possible, we compute the dominant eigenvalue of the matrix with

entries

$$\alpha_i^{\text{AR}} I_{\{i=j\}} + \alpha_i^{\text{spat}} v_{ji} I_{\{i\neq j\}} + \sum_{d=1}^{D} \left(\alpha_{i,d}^{\text{mob}} w_{ji,t-d}\right) \tag{4.5}$$

where I is the indicator function. We will plot this over time t.

Reproduction numbers measure the number of new cases stemming from old cases, but we also want to quantify the number of new cases stemming from the mobility of infectious people. We summarize the number of new infections per infected trip over time as

$$\frac{\sum_i (\sum_{d=1}^{D} \alpha_{i,d}^{\text{mob}} \sum_j w_{ji,t-d} \frac{Y_{j,t-1}}{N_i})}{\sum_i \sum_{d=1}^{D} \sum_j w_{ji,t-d} \frac{Y_{j,t-1}}{N_i})}$$

Although this may look cumbersone, it is simply a weighted average of the $\alpha_{i,d}^{\text{mob}}$'s over time. Furthermore, we can look at the number of infections per infected trip at the region level by summing over t instead of i. This formula can be easily modified in the presence of a serial interval.

4.3.6 Inference

All model parameters were estimated using Bayesian Markov chain Monte Carlo. In particular, we used the No-U-Turn sampler readily available in Stan (Carpenter et al., 2017) and its associated R package (Stan Development Team, 2021). Four chains with 1000 interations, with the first half being warmup were used for each model. Trace plots were used to visually assess convergence of Markov chains. The scale reduction factor was were also used to confirm an appropriate amount of mixing using a cutoff of $\hat{R} < 1.01$ (Vehtari et al., 2021). Since each of our summary statistics is a function of the model parameters, we can easily obtain credible intervals for each statistic by using draws from the joint posterior.

4.4 Application

In this section, we apply our model to two Spanish Communities separately. In Section 4.4.1, we treat all of Castilla-Leon as a single region, which is mainly used as an exploratory analysis to inform our multi-region (spatial) model. In 4.4.1, we apply our multi-region model to the 245 subregions of Castilla-Leon and quantify the risk associated with travelling

during the pandemic. We then apply our model to the 179 subregions of Madrid in Section 4.4.2.

4.4.1 Assessing the risk associated with travelling in Castilla-Leon

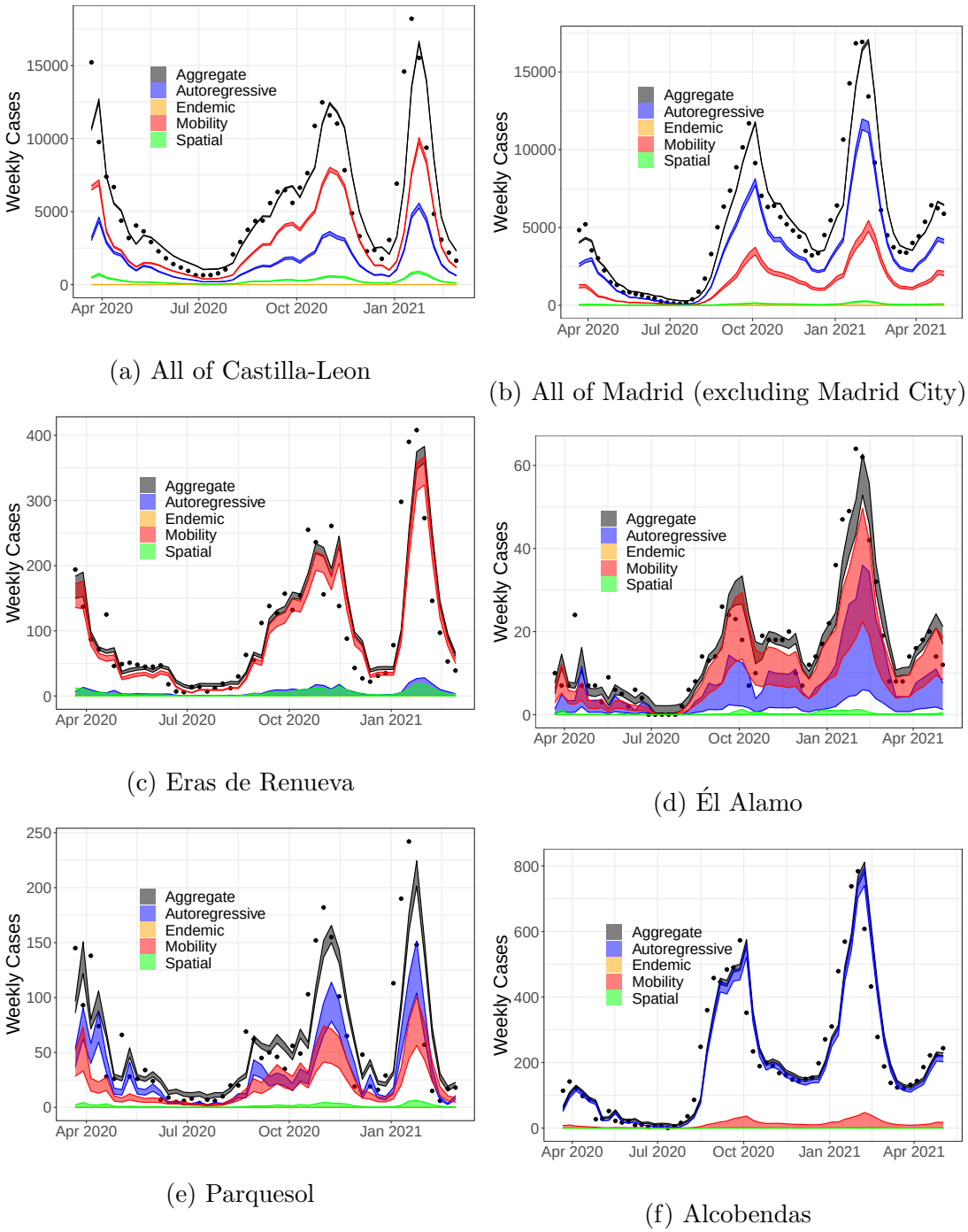
Castilla-Leon - aggregated model

A plot of the case, test, and mobility data for all of Castilla-Leon is shown in Figure 4.1a. As noted by other authors, there is often a large time lag between a peak in mobility and the subsequent peak in cases, and this effect appears to change over time (Gottumukkala et al., 2021). However, mobility should only affect the relative change in the number of infections, as mobility can only affect cases through current infectious individuals coming into contact with susceptibles. To examine the relationship between mobility and infectiousness, we compute the Crude R_{eff} over time and look at the cross correlation between it and mobility. We found that mobility at time $t-2$ and $t-1$ show strong correlation with the crude R_{eff} at time t, followed by a sharp drop in correlation when mobility is lagged by 3 or more time units. For this reason, we will consider the following mean as a starting point:

$$\lambda_t = \alpha^{\text{EX}} + \alpha^{\text{AR}} y_{t-1} + \alpha_1^{\text{mob}} w_{t-1} y_{t-1} + \alpha_2^{\text{mob}} w_{t-2} y_{t-1}$$

The fitted values of this model are shown in Figure A1. This plot suggests that mobility is explaining a very large portion of the case counts. However, this seems too large and warrants investigation. If we plot the Crude R_{eff} versus the mobility (Figure A3), we can see that there are two extremely high leverage points with high mobility and Crude R_{eff}. These two points correspond to the first weeks of March 2020, when people were moving a lot, and there were no mask mandates or policies enacted to slow the spread. As a result of these high leverage points, the effect of mobility (slope of the green line in Figure A3) is too high. Although this plot is an oversimplification of exactly what our model is doing when estimating the effect of mobility, they warrant our attention. If we remove these points, the least squares line becomes much shallower and fits the Crude R_{eff} estimates much better, as seen by the red line in Figure A3.

We now fit the model without the first three weeks of March, the results are shown in



(a) All of Castilla-Leon

(b) All of Madrid (excluding Madrid City)

(c) Eras de Renueva

(d) Él Alamo

(e) Parquesol

(f) Alcobendas

Figure 4.2: Multi-region mobility extended EE model. For both Castilla-Leon (left) and Madrid (right), we present the results for the entire region, alongside a region that showed a strong mobility effect, and a region showing a weaker mobility effect. The 95% credible interval for each model component is presented, alongside their aggregation ($\lambda_t^\dagger$). Observed case counts are shown as black points.

Figure A2, where the autoregressive component is much more substantial relative to the mobility component. For this reason, we will exclude these first three weeks of data when extending our model to multiple regions.

Castilla-Leon – Spatial Model

The model fit from the spatial model is shown in Figure 4.2. The movement component appears to be the strongest, followed by the autoregressive components and the spatial component. In some regions the movement component was very small, while it dominated the infections in others. Given that our results can be sensitive to one or two time points, we suspect that the region-level mobility effects are noisy. However, the aggregation of them is more likely to produce a clear signal.

When adjusting for testing using the method described in Section 4.3.4, we found that the estimated testing probability was very close to 1. We explored several minor alterations of this method with little or no change in the estimate. This is likely a limitation of the data, indicating that we may need region level testing to tease out the potentially spatially heterogenous effect. For this reasons, we did not control for testing in Castilla-Leon, and simply acknowledge this limitation. Although this doesn't affect our PCAtM and dominant eigenvalue summary statistics, it will cause us to underestimate infections per infected trip.

The proportion of cases attributable to movement (PCAtM) is presented in Table 4.1 for four different models

1. No serial interval, two mobility lags, and no testing adjustment
2. No serial interval, two mobility lags, and testing adjustment (method 2)
3. Serial interval of 2 weeks, 3 mobility lags, and no testing adjustment
4. No serial interval, 3 mobility lags, and no testing adjustment

Adjusting for testing had little effect on the PCAtM. Similarly, using a serial interval of 2 weeks (as opposed to 1 week) had little effect on the PCAtM, but the additional mobility lag seems to be accounting for additional cases. For this reason, we present statistical summaries for Model 4 above.

		PCAtM (95% CrI)	ρ_1
Castilla	No SI, No testing	44.96 (43.75, 46.31)	-
	No SI, testing	43.79 (42.55, 44.88)	-
	SI, no testing, additional lag	56.99 (55.95, 58.03)	> 0.999
	No SI, no testing, additional lag	57.01 (55.94, 57.98)	-
Madrid	No SI, with testing	17.00 (16.11, 18.03)	-
	SI, with testing	14.00 (13.12, 15.02)	< 0.001
	SI, no Madrid City, with testing	28.54 (26.76, 30.68)	< 0.001

Table 4.1: Percentage of cases attributable to movement (PCAtM) for various models fit to Castilla-Leon and Madrid data. In models with a serial interval (SI), ρ_1 is presented. Posterior median and 95% CrI's are presented.

The proportion of cases attributable to movement (PCAtM) and the trips per infection for each region is shown in Figures 4.3a and 4.3b. Both the PCAtM and the trips per infection show a high amount of heterogeneity between regions. The temporal variation in trips per infection, averaged across Castilla-Leon are shown in Figure 4.4a. Based on this model, it takes roughly 70 infected trips to see a new infection.

The temporally changing dominant eigenvalues computed from (4.5) are shown in Figure 4.5. The dominant eigenvalue exceeding one seems to correspond with increases in case counts in Castilla, with the exception of the third viral wave. This may be due to properties of the virus at this time (such as a new variant), the drastic increase in testing that we had trouble accounting for, or some other confounding factors.

4.4.2 Assessing the risk associated with travelling in the Community of Madrid

For completeness, we present the results for Madrid with two major caveats: 1) a single large region (Madrid City) contains 49.6% of the Community of Madrid's population and 51% of the COVID-19 cases and 2) the intra-regional mobility in Madrid City (10.2% of the Community of Madrid's mobility) shows a highly different pattern (see Figure B1) than the rest of the mobility in the region, with a peak during the first lockdown. Since the trend in case counts is roughly the same as the rest of the region, but the mobility is highly different, we do not believe that our model accurately captures the relationship between mobility and infectiousness in Madrid City.

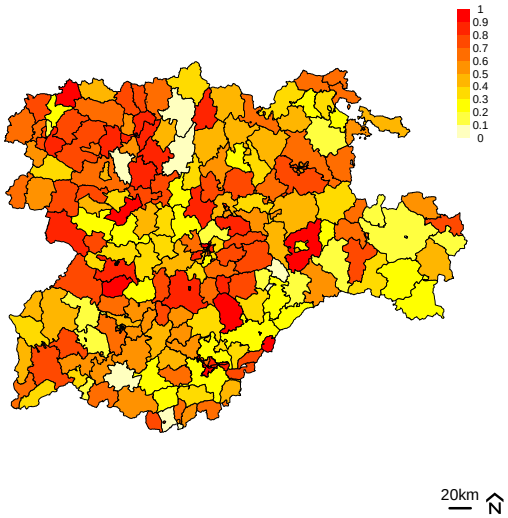
Figure 4.1b displays time series of weekly trips, tests, and cases aggregated across the community of Madrid. After removing the first three weeks of data (as with Castilla-Leon) and correcting for changes in testing, we find that our assumption regarding the reproduction number being a linear function of mobility is reasonable (see Figure B2). Furthermore, we adjusted the per-contact-probability of infection for vaccinations (as described in Appendix 4.C), but found no substantial difference in our results.

The model fit for the spatial Madrid model is shown in Figure 4.2. Mobility accounts for a substantial proportion of the cases, but the autoregressive term explains the most. The proportion of cases attributable to movement is shown in Table 4.1, for a model with no serial interval, and one with a serial interval of two weeks. In the model with the serial interval, ρ_1 was very close to zero, indicating that the model with the serial interval is more appropriate. However, this may have occurred due to the lag one mobility effect being very small, and our model is avoiding including that term.

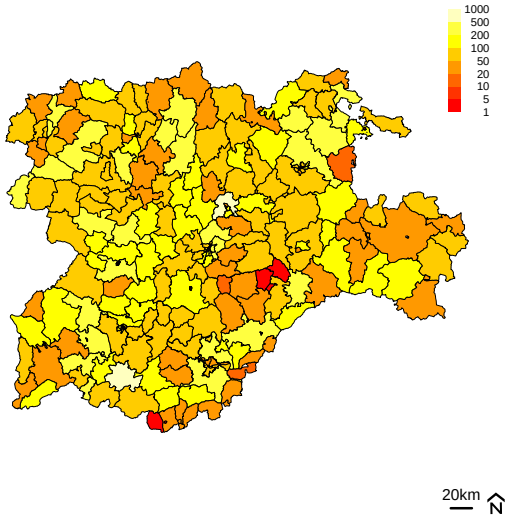
The spatial distribution of the PCAtM is shown in Figure 4.3c. Note that the regions with a low PCAtM tend to be very close to Madrid City, while the regions with high PCAtM don't show a spatial pattern. The number of infections per infected trip is shown both spatially and temporally in Figures 4.3d and 4.4b. Figure 4.3d suggests that the trips required for a new infection are spatially correlated, indicated by the clusters of regions of the same colour. Figure 4.4b suggests that, excluding Madrid City, roughly 140 infected trips are required for a new infection to arise.

4.5 Discussion

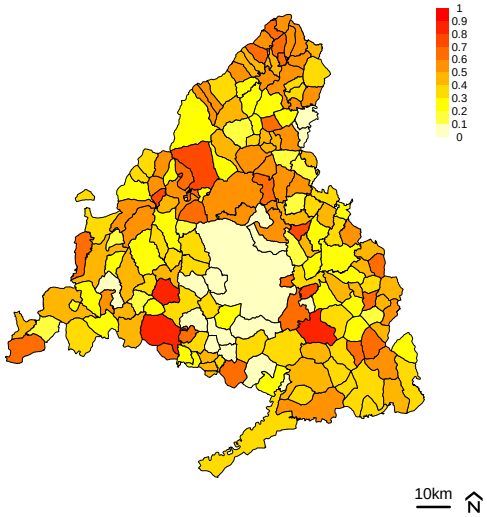
In this chapter we developed an infectious disease model where the number of contacts between people is a linear function of trips between regions. We showed that this model is an extension of Endemic-Epidemic models frequently found in the literature. We applied this model to two Spanish Communities with the intention of quantifying the risk associated with travelling in each Community. In Castilla-Leon, we found that we could relate just over half of the trips to our cellphone mobility data, while this was much lower in Madrid. Our model appears to work better when regions are small, as our cellphone mobility data



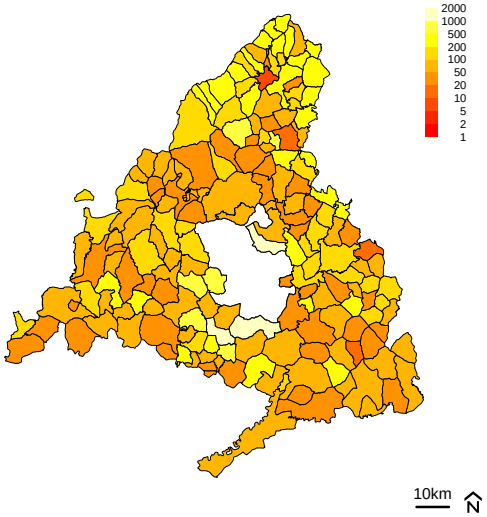
(a) PCAtM



(b) Trips per infection

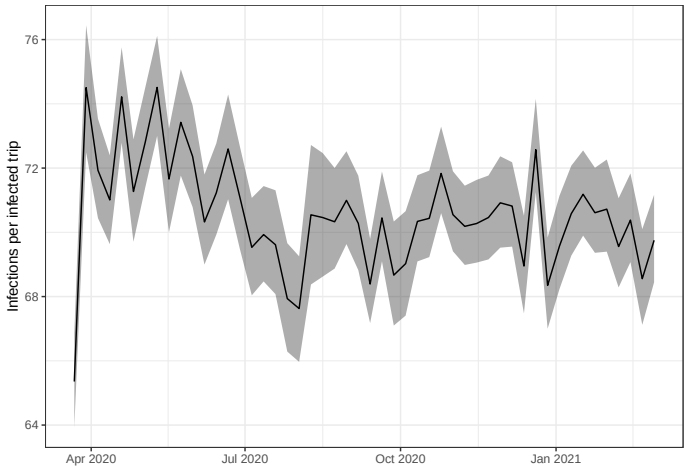


(c) PCAtM

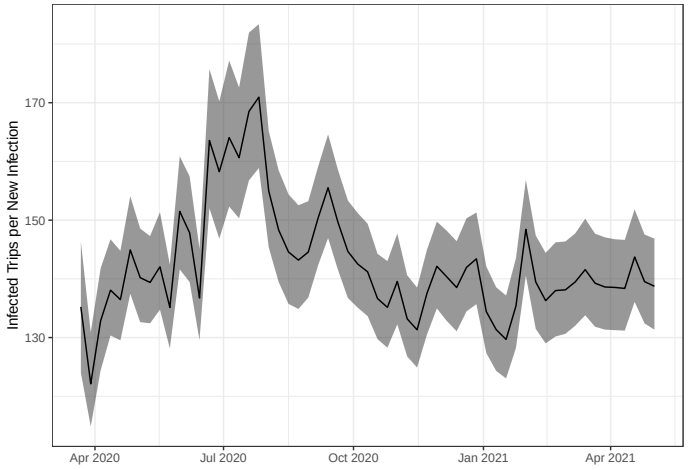


(d) Trips per infection

Figure 4.3: Spatial distribution of proportion of cases attributable to movement (PCAtM) and the number of trips associated with one new infection. The trips per infection in Madrid City (white region in 3d) was calculated to be 3753.



(a) Castilla-Leon



(b) Madrid

Figure 4.4: Temporal variation of number of trips associated with one new infection. Madrid City was excluded from this analysis, as the data quality issues caused this number to be implausibly high. The posterior median, alongside 95% credible intervals are presented.

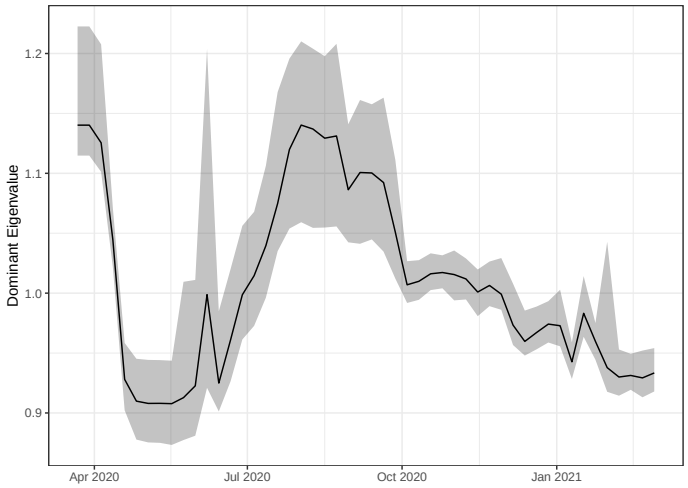


Figure 4.5: Posterior median and 95% credible interval of dominant eigenvalue in Castilla-Leon. A dominant eigenvalue > 1 will generally lead to an increase in cases.

is more informative.

We found that this class of models is sensitive to large changes in case counts that occurred early in the pandemic, as well as rapid changes in testing capacity. Although we took great care in specifying each model component, developing robust methods for modeling the infectiousness of the disease when using this class of models should be researched further. We stress the importance that exploratory and diagnostic plots can greatly improve inference and interpretation when using Endemic-Epidemic, or any infectious disease model.

One strength of this work is that we utilize rich mobility data and spatial data to model disease spread through a carefully parametrized infectious disease model. In doing so we were able to assign a number to the risk associated with travelling during a pandemic.

A further strength of this work is that it was done during a time period prior to mass vaccinations and the introduction of the major COVID-19 variants, which could have confounded our analysis. This could also be viewed as a limitation, as we could have introduced a changepoint when the Delta and Omnicron variants arose, and could easily account for higher vaccination rates using the methodology from this chapter. In our analysis of the Community of Madrid, too few people had been vaccinated for it to make any major difference in our results. Ideally, we would have mobility data over the course of the entire pandemic, so that we could see how the risk associated with travelling changes with new

variants and increasing levels of immunity in the population.

A limitation of our work that we must emphasize is that we cannot associate individual trips to individual infections, and thus cannot infer causality. Although we are confident that mobility is required for COVID-19 to spread, we cannot be sure that the trips recorded in our data are causing cases according to our model specification, as there may be confounding factors associated with between-region mobility and case counts.

A further limitation of this work is that we rely on the rare disease assumption. If the study period extended later into 2021, we would have to relax this assumption, leading to an alternate model formulation that would not lead to an Endemic-Epidemic model. Of course an Endemic-Epidemic model could still be used, but it would not have as nice of an interpretation.

This work opens the door for many avenues of future research. Firstly, robust methods for modeling infectiousness as a function of mobility (or any covariate) would be extremely useful. For instance, a method utilizing quantiles would be insensitive to rapid changes in the observed cases. Furthermore, we need to rethink how to compute temporally changing reproductive numbers from this class of models, especially as the model becomes more complex.

Although this study has focussed on COVID-19, we want to emphasize that the model and associated principles can be extended to a wide variety of infectious diseases, and various forms of network data. Extensions and simplifications should be made on a case-by-case basis, and should be guided by careful data exploration.

4.6 Acknowledgements

JJS is supported by the Natural Sciences and Engineering Research Council of Canada (PGSD3-559264-2021). JSR is supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2019-04142). PEB is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2022-05164).

4.A Treating Castilla-Leon as a single region

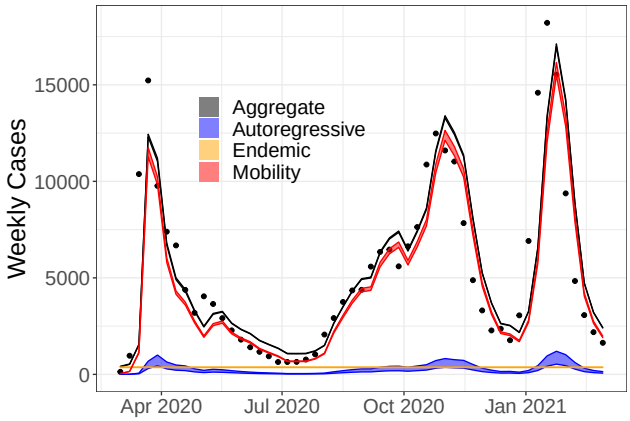


Figure A1: Single region, mobility-extended EE model fit to aggregate Castilla-Leon data, separated into components.

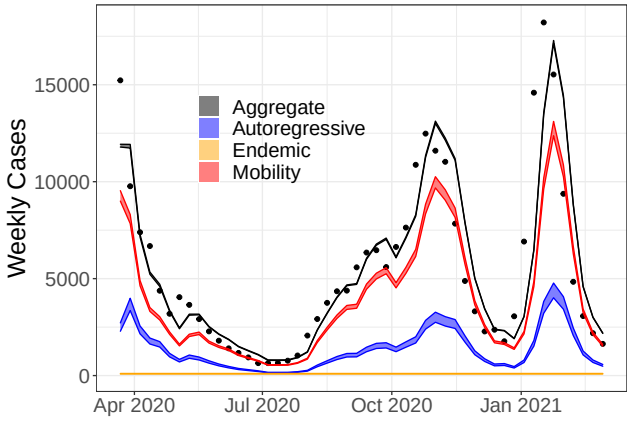


Figure A2: Single region, mobility-extended EE model fit to aggregate Castilla-Leon data with the first three weeks of data removed.

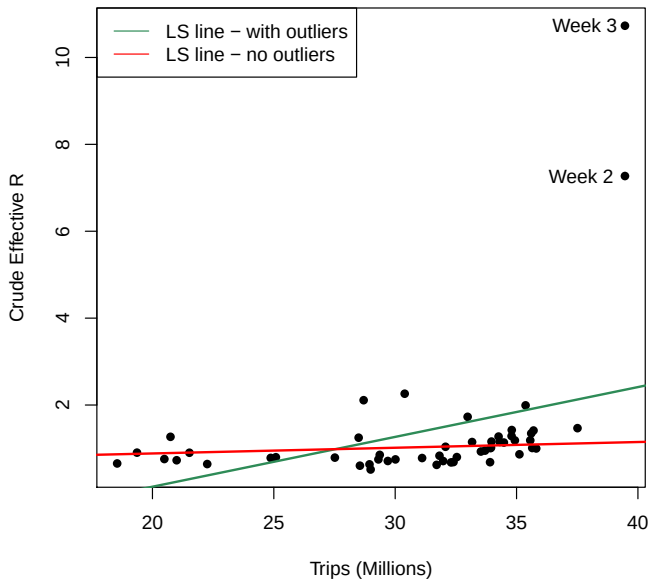
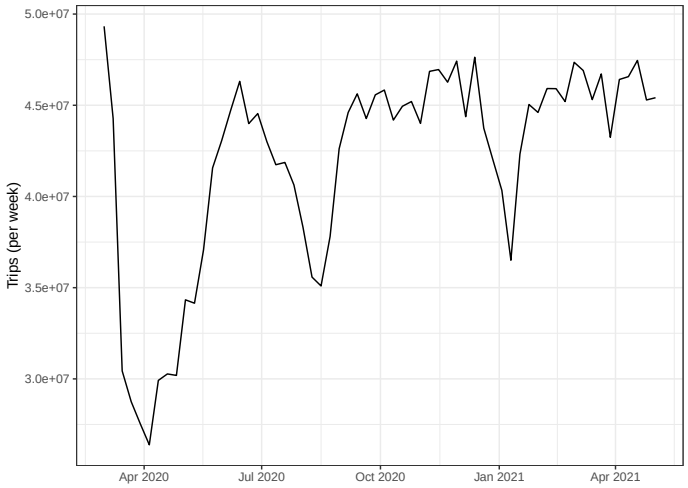


Figure A3: Crude R_{eff} vs number of trips. There are two high leverage points which correspond to the first three weeks of the pandemic. These have a strong influence on the effect of mobility and cause the green line to be much steeper than it should be. The red line is the least squares line with the two influential points removed, and visually fits the data much better.

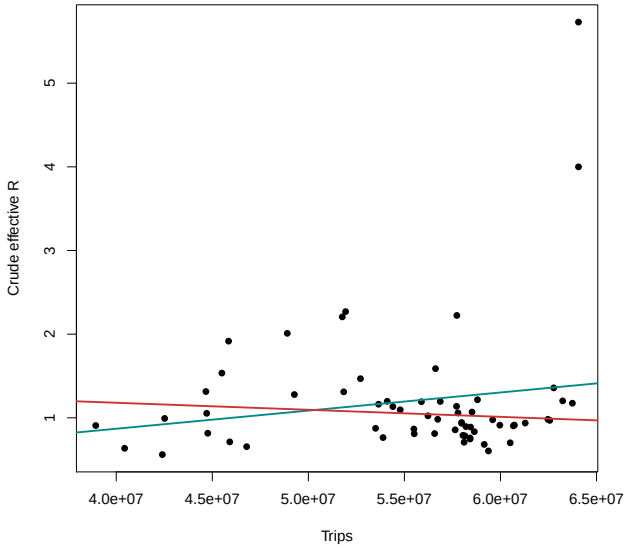
4.B Madrid: Supplementary plots regarding modelling decisions

(a) Comparing the number of trips within Madrid City with the incoming/outgoing mobility of Madrid City

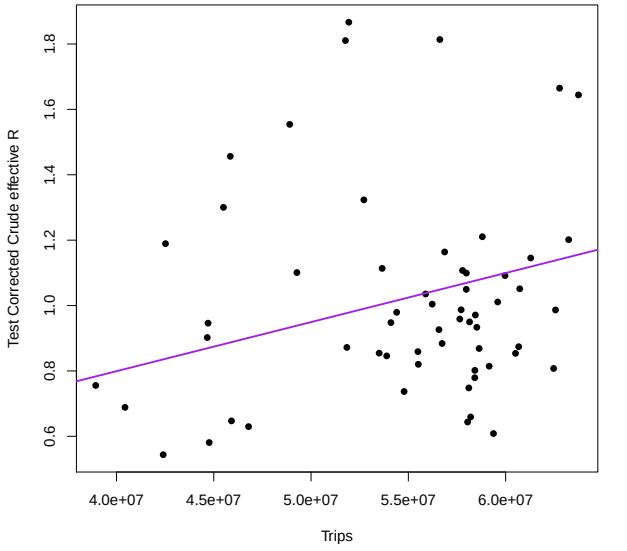


(b) Madrid - all mobility not involving Madrid City

Figure B1



(a) Community of Madrid - Trend lines for R_{eff} vs. lag-one mobility including (green line) and excluding (red line) the first 3 weeks of data



(b) Community of Madrid - Trend lines for test corrected R_{eff} vs. lag-one mobility, excluding the first three weeks of data.

Figure B2

4.C Accounting for Vaccinations in the Community of Madrid

In Spain, vaccines started to reach the general public in January of 2021, with about 30% of the public having at least one dose of the vaccine by May 2021. Thus, this may impact our Madrid results and should be explored. The immunity induced by vaccines should reduce the per-contact-probability of infection. Using the univariate model with $D = 1$ as an example, our force of infection is now

$$\lambda_t^{\text{vacc}} = \mathcal{C}(w_{t-1}) \times p(u_{t-1}) \times \frac{y_{t-1}}{N}$$

where

$$g[p(u_{t-1})] = p_0 - \tau u_{t-1}$$

with τ being a reduction in infection probability due to vaccination, g being a link function, and u_t is the proportion of the population that is vaccinated at time t. The force of infection becomes

$$\lambda_t^{\text{vacc}} = (c^{\text{AR}} + c^{\text{mob}} w_{t-1}) \cdot g^{-1}(p_0 - \tau u_{t-1}) \frac{y_{t-1}}{N}.$$

The identity link would lead to

$$\lambda_t^{\text{vacc}} = \left(\alpha^{\text{AR}} + \alpha^{\text{mob}} w_{t-1} - (c^{\text{AR}} + c^{\text{mob}} w_{t-1}) \tau u_{t-1}\right) \frac{y_{t-1}}{N}$$

which allows for potentially negative values of λ_t without some numerically unstable constraints. Furthermore, this would assume a linear relationship between proportion vaccinated and infection probability, which seems unrealistic. Instead, we used a log link leading to

$$\lambda_t^{\text{vacc}} = \left(\alpha^{\text{AR}} + \alpha^{\text{mob}} w_{t-1}) e^{-\tau u_{t-1}}\right) \frac{y_{t-1}}{N}.$$

4.4 Bibliography

Alene, M., Yismaw, L., Assemie, M. A., Ketema, D. B., Gietaneh, W., and Birhan, T. Y. (2021). Serial interval and incubation period of COVID-19: a systematic review and

meta-analysis. *BMC Infectious Diseases*, 21(1):1–9.

Bauer, C. and Wakefield, J. (2018). Stratified space–time infectious disease modelling, with an application to hand, foot and mouth disease in China. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1379–1398.

Bracher, J. and Held, L. (2020). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting*, 38(3):1221–1233.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Celani, A. and Giudici, P. (2022). Endemic–epidemic models to understand COVID-19 spatio-temporal evolution. *Spatial Statistics*, 49:100528.

Diekmann, O., Heesterbeek, H., and Britton, T. (2013). *Mathematical tools for understanding infectious disease dynamics*, volume 7. Princeton University Press.

Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, 28(4):365–382.

Epidemiological Surveillance Network of Madrid (2022). COVID 19-TIA by Municipalities and Districts of Madrid.

Fritz, C. and Kauermann, G. (2022). On the interplay of regional mobility, social connectedness and the spread of COVID-19 in Germany. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(1):400.

García-García, D., Vigo, M. I., Fonfría, E. S., Herrador, Z., Navarro, M., and Bordehore, C. (2021). Retrospective methodology to estimate daily infections from deaths (REMEDID) in COVID-19: the Spain case study. *Scientific reports*, 11(1):1–15.

Geilhufe, M., Held, L., Skrøvseth, S. O., Simonsen, G. S., and Godtliebsen, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal*, 56(3):363–382.

General Directorate of Information Systems, Quality and Pharmaceutical Provision (2022). Open Data of Castile and Leon. https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html. Accessed: March 10, 2022.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gordis, L. (2013). *Epidemiology*. Elsevier Health Sciences, 5th edition.

Gottumukkala, R., Katragadda, S., Bhupatiraju, R. T., Kamal, A. M., Raghavan, V., Chu, H., Kolluru, R., and Ashkar, Z. (2021). Exploring the relationship between mobility and COVID-19 infection rates for the second peak in the United States using phase-wise association. *BMC Public Health*, 21(1):1–14.

Grimée, M., Dunbar, M. B.-N., Hofmann, F., Held, L., et al. (2021). Modelling the effect of a border closure between Switzerland and Italy on the spatiotemporal spread of COVID-19 in Switzerland. *Spatial statistics*, page 100552.

Halloran, M. E., Longini, I. M., and Struchiner, C. J. (2010). R_0 and deterministic models. In *Design and Analysis of Vaccine Studies*, pages 85–102. Springer.

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5):672–675.

Held, L., Hofmann, M., Höhle, M., and Schmid, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics*, 7(3):422–437.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199.

Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal*, 54(6):824–843.

Meyer, S. and Held, L. (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*, 18(2):338–351.

Ministerio de Sanidad, Gobierno De España (2022a). Datos abiertos de pruebas realizadas.

Ministerio de Sanidad, Gobierno De España (2022b). Estrategia de vacunacion COVID-19 en España.

Ministerio de Transportes Movilidad Y Agends Urbana, Gobierno de España (2022). Evolución de la movilidad diara. www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data. Accessed: Oct 26, 2022.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, 30(10):1118–1136.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27(29):6250–6267.

Ponce-de Leon, M., Del Valle, J., Fernandez, J. M., Bernardo, M., Cirillo, D., Sanchez-Valle, J., Smith, M., Capella-Gutierrez, S., Gullón, T., and Valencia, A. (2021). COVID-19 flow-maps an open geographic information system on COVID-19 and human mobility for Spain. *Scientific Data*, 8(1):1–16.

Schrödle, B., Held, L., and Rue, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics*, 68(3):736–744.

Slater, J. J., Brown, P. E., and Rosenthal, J. S. (2021a). Forecasting subnational COVID-19 mortality using a day-of-the-week adjusted Bayesian hierarchical model. *Stat*, 10(1):e328.

Slater, J. J., Brown, P. E., Rosenthal, J. S., and Mateu, J. (2021b). Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. *Spatial Statistics*, page 100540.

Stan Development Team (2021). RStan: the R interface to Stan. R package version 2.21.3.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718.

Wakefield, J., Dong, T. Q., and Minin, V. N. (2019). Spatio-temporal analysis of surveillance data. In *Handbook of Infectious Disease Data Analysis*, pages 455–475. Chapman and Hall/CRC.

Chapter 5

A Bayesian approach to estimating COVID-19 incidence and infection fatality rates

Abstract

Naive estimates of incidence and infection fatality rates (IFR) of COVID-19 suffer from a variety of biases, many of which relate to preferential testing. This has motivated epidemiologists from around the globe to conduct serosurveys that measure the immunity of individuals by testing for the presence of SARS-CoV-2 antibodies in the blood. These quantitative measures (titre values) are then used as a proxy for previous or current infection. However, statistical methods that use this data to its full potential have yet to be developed. Previous researchers have discretized these continuous values, discarding potentially useful information. In this chapter, we demonstrate how multivariate mixture models can be used in combination with poststratification to estimate cumulative incidence and IFR in an approximate Bayesian framework without discretization. In doing so, we account for uncertainty from both the estimated number of infections and incomplete deaths data to provide estimates of IFR. This method is demonstrated using data from the Action to Beat Coronavirus (Ab-C) serosurvey in Canada.

5.1 Introduction

As of April 1, 2022, there have been close to 500 million confirmed cases of COVID-19 worldwide (World Health Organization, 2022). However, the general consensus is that this number is an underestimate of the true cumulative incidence of the disease, as this estimate is largely dependent on the number of tests being administered, the accuracy of testing (Burstyn et al., 2020a,b), and to whom these tests are being issued. If testing is extensive enough, and a correction is made for underreporting of asymptomatic cases, then a test-based case fatality rate may be a reasonable proxy for the infection fatality rate (IFR) (Luo et al., 2021). However, given that the testing early in the pandemic was sparse, and estimating IFR accurately is of the utmost importance, epidemiologists across the globe are conducting serosurveys that measure immunity of individuals by testing for the presence of SARS-CoV-2 antibodies in the blood (Chen et al., 2021). This quantitative measure (which we will call a *titre value*) is then used as a proxy for previous or current infection. However, how exactly this data should be used to accurately estimate important epidemiological quantities (like incidence and IFR) is an active area of research.

The standard approach is to label everyone who has a titre value above some threshold as "infected", and consider everyone else not infected. This leads to the problem of selecting the cutoff, which can be made based on known cases/controls and analysis of the Receiver Operating Characteristic (ROC) Curve. The ROC plots the true positive rate (sensitivity) vs the false positive rate (1-specificity) and it is typical to select the cutoff that results in the highest Youden Index (sensitivity + specificity - 1) (Krzanowski and Hand, 2009). Gelman and Carpenter (2020) suggest that the uncertainty in sensitivity and specificity can be considered parameters to be estimated in a Bayesian hierarchical model assuming that informative priors are used for the sensitivity and specificity. Although this method accounts for uncertainty in the sensitivity and specificity, it still suffers from the loss of information in the discretization process. Particularly in COVID-19 applications, a subject with an extremely high level of antibodies should have a lower probability of being a false-positive than someone who is just barely above the threshold. This could be partially remedied by allowing sensitivity and specificity to be a function of covariates, but ideally

methods that avoid these issues all together are preferable.

Mixture models are a natural choice to overcome the limitations of using a fixed cutoff, as they allow infection status and associated uncertainty to depend on the magnitude of individuals' titre values. Mixture models have been widely applied when studying the prevalence of infectious diseases in animals (Ødegård et al., 2003, 2005; Nielsen et al., 2007) and in humans (Vink et al., 2015, 2016; Kyomuhangi and Giorgi, 2022). There are several other papers that have modeled the COVID-19 antibody levels directly to infer cumulative incidence through the use of mixture models. Bouman et al. (2021) showed that mixture models can outperform the methods of Gelman and Carpenter (2020) for estimation of cumulative incidence of COVID-19. Furthermore, Bottomley et al. (2021) apply mixture models to Kenyan serosurvey data and show that mixture of skew normal distributions more accurately estimates cumulative incidence than methods based on thresholds. However, the applications of these models thus far has been rather limited. For instance, some unexplored questions include: how do we use these mixture models to account for survey bias and get cumulative incidence rates for the general population? How do we incorporate multiple titre values per person? How do we estimate cumulative incidence in the presence of vaccinated individuals? How do we use these mixture models to estimate IFR while accounting for uncertainty in both the number of infections and deaths?

In this chapter, we demonstrate how mixture models can be used to estimate cumulative incidence in an approximate Bayesian framework without discretization. Specifically, we apply a mixture of multivariate t-distributions to the log of the titre values, using a logistic regression model for the mixing parameter to account for covariates. We then use poststratification to obtain estimates of cumulative incidence and its associated uncertainty. Furthermore, we estimate the number of COVID-19 related deaths using partially complete data, and use this in combination with incidence estimates to estimate the IFR across Canada.

5.1.1 Data

Dry blood spot (DBS) samples were collected from participants of the Action to Beat Coronavirus (Ab-C) study (https://www.abcstudy.ca/). This chapter is concerned with

the first two *phases* of the study. In Phase 1, DBS samples from 9123 participants were collected from June to November 2020 and roughly corresponding to the first viral wave (April 1 to July 31, 2020). In Phase 2, DBS samples from 7299 were collected from December 2020 to May 2021 and roughly correspond to the second viral wave (October 1, 2020 to March 1, 2021). These blood spots were tested for prevalence of Immunoglobin G (IgG) antibodies, measured using three antigens: Spike (SmT1), RBD, and nucleocapsid (NP). Two different versions of the SmT1 antigen test were used on the Phase 1 blood spots, while all three were applied to Phase 2 blood spots. All three titres will show larger values for participants who have been exposed to COVID-19, but only SmT1 and RBD will show larger values for mRNA vaccinated individuals. This is because the mRNA vaccines do not contain the nucleocapsid (NP) protein. Therefore, people who received an mRNA vaccine and did not have a history of prior infection, will not develop anti-NP antibodies. Those that were previously infected, regardless of vaccination status, will have anti-NP antibodies (Houlihan and Beale, 2020). This will be helpful for distinguishing between vaccinated and infected individuals in Section 5.3.3. In Phase 1, 8919 people had one SmT1 measurement, and 8704 had two SmT1 titre measurements, along with complete covariate information. In Phase 2, 7065 had all three measurements, along with complete covariate information. Of those 7065, 624 joined the study in Phase 2 (6441 participants had complete Phase 1 and Phase 2 data). These data have been previously analyzed by Tang et al. (2022) using a simpler model. Additional medical details regarding these antigen tests can be found in their paper. Tang et al. (2022) also investigated the representativeness of study participants when compared to the Canadian population. They found that the study population tended to be older, more university educated, more likely to be indigenous, etc. See eTable 3 in their paper for further reading.

Although serosurveys are a proven way to accurately measure seroprevalence, the notion of seroprevalence itself has several drawbacks. Firstly, there is a chance that participants got infected and returned their blood spots soon after. Antibodies generally take between 7 and 14 days to be measurable from the onset of infection (Centre for Disease Control and Prevention, 2022). This may cause a slight under-estimation of incidence. Secondly, antibodies wane slowly over time. However, they have been shown to remain elevated for

many months after infection. In a study (Alfego et al., 2021) evaluating 39,086 individuals with confirmed positive COVID-19 infection by RT-PCR between March 2020 to January 2021, the anti-NP antibody maintained a rate of 68.2% [95% Cl: 63.1-70.8%] after 293 days, while anti-SmT1 antibody maintained a rate of 87.8% [95% Cl: 86.3-89.1%] through 300 days. Note that the majority of people in our study were likely infected far less than 300 days prior to submitting their blood spots, so the maintenance rate in our study was likely higher than those in Alfego et al. (2021). At this point, we simply note these limitations of seroprevalence, and examine the potential impact of waning immunity on our results in Appendix 5.E.

Population demographics (age, sex, province, ethnicity, education, and long-term care residency) were obtained from 2016 census data from Statistics Canada (Statistics Canada, 2016). We are using the 2016 Census data because as the 2021 Census data is not yet complete. Although the total population of Canada has increased by about 5%, the geographic and age distributions seem to be similar between 2016 and 2021 based on the data that we do have available. This information will be used for poststratification as described in Section 5.2.3. The long-term care (LTC) COVID-19 deaths were obtained from https://ltc-covid19-tracker.ca (Samir et al., 2022) between Sept 2020 and March 2021 for each province. The total deaths for each province by age and sex were obtained from the different provincial governments (Ontario, Alberta, and Quebec). For additional provinces, where deaths by age and sex could not be obtained, we used the distribution of nearby provinces to approximate those deaths. The age/sex distribution of deaths in Alberta was used to infer the distribution of deaths in British Columbia and Saskatchewan. The age/sex distribution of deaths in Quebec was used to infer the distribution for the Atlantic region (New Brunswick, Nova Scotia, Newfoundland, and Prince Edward Island). Manitoba reported different age groups than Ontario, but seemed to have a similar distribution. Thus we used Ontario data to infer Manitoba's age/sex deaths for the different age groups. This means that although the aggregate IFR estimates for the Atlantic region, Manitoba, British Columbia, and Saskatchewan are likely valid, the estimates by age/sex should be treated with caution due to the imputations noted above.

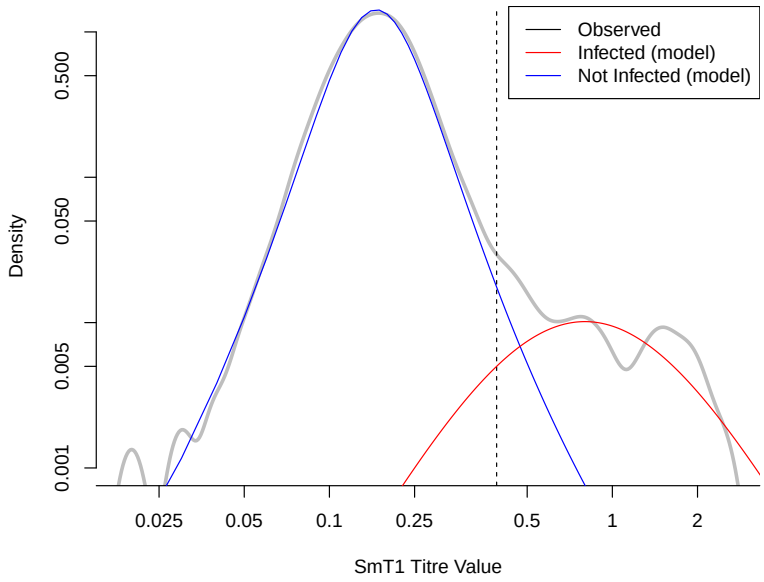


Figure 5.1: Mixture of t-distributions for the Phase 1 univariate model fit to the SmT1 titre values. The posterior median for each parameter is used. The vertical dashed line represents the cutoff used in Tang et al. (2022). Keep in mind that this plot does not display uncertainty in the model parameters of the t-distributions.

5.2 Methods

Our first goal is to estimate the cumulative incidence of SARS-CoV-2 in Canada. We define *cumulative incidence* in Phase 1 to be the number of SARS-CoV-2 infections up until September 30th 2020, divided by the population size. The cumulative incidence in Phase 1 and 2 has the cumulative number of infections up until March 31st 2021 as the numerator. We define the *incidence proportion* in Phase 2 to be the number of infections from Oct 1st 2020 to March 31st 2021, divided by the population size. We recognize that the terms cumulative incidence and incidence proportion are used interchangeably in the epidemiology literature, and we are avoiding the term "cumulative" when presenting estimates of incidence in Phase 2 alone. We estimate incidence in two steps. First, we will fit a Bayesian mixture model to the titre values, relating an individual's infection status, a latent variable, to their measured covariates via a logistic regression model. Second, we will use poststratification to account for the disparity between the population of survey

responders versus the general Canadian population. This will yield an estimate of the number of infections in Canada for each covariate combination, and hence, an estimate of the cumulative incidence.

Our second goal is to estimate the *Infection Fatality Rate*, which is defined as the number of COVID-19 related deaths divided by the number of infections. This will be estimated in Phase 1, Phase 1 and 2, and Phase 2 alone with the same time periods as mentioned previously. We do this by building a Bayesian model for the number of deaths in Canada by age/sex/province group, and dividing this by the estimated number of infections. This will allow for estimates of IFR in any age/sex/province category that we want, accounting for uncertainty in both the deaths and the infections.

5.2.1 Notation

Lower case Latin letters are used to represent (potentially vector-valued) observed data; x are observed covariates, w is observed titre values, and d is observed deaths. The exception is p, which is an unknown probability of infection. Upper-case Latin letters represent latent variables ("missing data"), such as the unknown number of infections Y, an unknown number of deaths D, and the latent infection status Z of an individual. Greek letters will be used for model parameters.

5.2.2 Mixture models

In this subsection we will introduce three mixture models that will be used to infer cumulative incidence. First, we will introduce a univariate (one titre value), two-component ("not infected" and "infected") mixture model, relating each study participant's covariates to their probability of infection. We will then extend this model to the bivariate case with two titre values in 5.2.2. These two models will be fit to the Phase 1 data. We will then present a trivariate, three-component ("unvaccinated, not infected", "unvaccinated, infected", and "vaccinated, not infected") mixture model that will be fit to the Phase 2 data. Note that the "infected" group here contains both vaccinated and unvaccinated people as our titres values are not precise enough to determine vaccination status if a person is infected. This is likely inconsequential as we will explain shortly.

Univariate mixture of t-distributions - Phase 1.

The infectivity status, Z_i, of an individual i is latent and is measured through an antibody lab test (titre), which is a quantitative measure. The density of the logged Phase 1 SmT1 titre values is shown in Figure 5.1. Notice that there is an approximately symmetric mound around 0.15 which is likely to be comprised of individuals who never had COVID-19. Previously, Gaussian distributions were used to model the logged titre values in non-infected individuals (Bottomley et al., 2021). However, we expected a heavier-tailed distribution would be needed, and employ a t-distribution for both the negative and positive individuals.

The univariate, two-component version of our mixture model can be written as follows:

$$\begin{aligned} \log(w_i)|Z_i = k &\sim f_1(\mu_k, \sigma_k, \nu_k), k = 0, 1 \\ Z_i|x_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= \beta^T x_i \end{aligned} \tag{5.1}$$

where w_i is the titre value of invidual i, Z_i is the latent variable indicating SARS-CoV-2 infection ($Z_i = 1$) or non-incidence ($Z_i = 0$), x_i is a $m \times 1$ vector of covariates, β is a $1 \times (m + 1)$ vector of regression coefficients which will be used for poststratification as described in Section 5.2.3, f_1 is the univariate (shifted and scaled) t-density, and $p_i = \text{logit}^{-1}(\beta^T x_i)$ is the probability that individual i has been infected with COVID-19. That is, the probability that someone had COVID-19 is a function of their covariates, but the parameters of the t-distributions are not. The covariates used in our mixture models were age ($<$ 20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+), sex (male, female), province (Alberta, Atlantic Region, British Columbia, Manitoba, Ontario, Quebec, Saskatchewan), ethnicity (white, indigenous, not white or indigenous), and education (university degree, college degree, less than college degree), meaning that $m = 18$.

Since Z_i is a latent discrete variable, certain MCMC software programs cannot sample it directly. However, we can marginalize Z_i out to obtain the following likelihood:

$$\pi(\log(w_i); \beta, \boldsymbol{\xi}, x_i) = [1-\text{logit}^{-1}(\beta^T x_i)] f_1[\log(w_i)|\mu_0, \sigma_0, \nu_0] + \text{logit}^{-1}(\beta^T x_i) f_1[\log(w_i)|\mu_1, \sigma_1, \nu_1]$$

where $\boldsymbol{\xi} = \{\mu_0, \mu_1, \sigma_0, \sigma_1, \nu_0, \nu_1\}$ is a vector of parameters which need to be estimated, but are not used to infer incidence directly.

For both Phase 1 and Phase 2, we have continuous values for multiple titres, and thus will now extend this univariate mixture model to a mixture of multivariate t-distributions.

A bivariate mixture model for Phase 1.

For Phase 1, we have two measurements of SmT1 for each sample. Using both titres should improve our ability to identify individuals who were infected. Our model naturally extends to the bivariate case by replacing the univariate t-distribution by a bivariate t-distribution (f_2):

$$\begin{aligned} \log(\boldsymbol{w}_i)|Z_i = k, x_i &\sim f_2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k), k = 0, 1 \\ Z_i|x_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= \beta^T x_i \end{aligned} \tag{5.2}$$

where $\boldsymbol{\mu}_k$ is a vector of length 2, Σ_k is a 2x2 covariance matrix, and the rest of the parameters are the same as Section 5.2.2. Note that the logistic regression model for Z_i in the second level is still univariate. This allows the model to accomodate multiple titre values per person without the number of parameters getting out of control. We fit this bivariate model on the two Phase 1 titre values using MCMC to obtain posterior samples of β which will be used later for poststratification.

A trivariate, three-component mixture model for Phase 2.

In Phase 1, vaccinations had not yet been made available and Z_i could only take on two values: "infected" or "not infected". However, during Phase 2, a non-negligible proportion ($\approx$ 2.5%) had claimed to have been vaccinated. Given that vaccinated people are distinguishable from infected people based on the three titre values that we have available, we now have three mutually exclusive values for Z_i: "unvaccinated, not infected","unvaccinated, infected", and "vaccinated, not infected". We did not include a fourth group "vaccinated, infected", as there were likely to be very few participants in this category. Note that we can

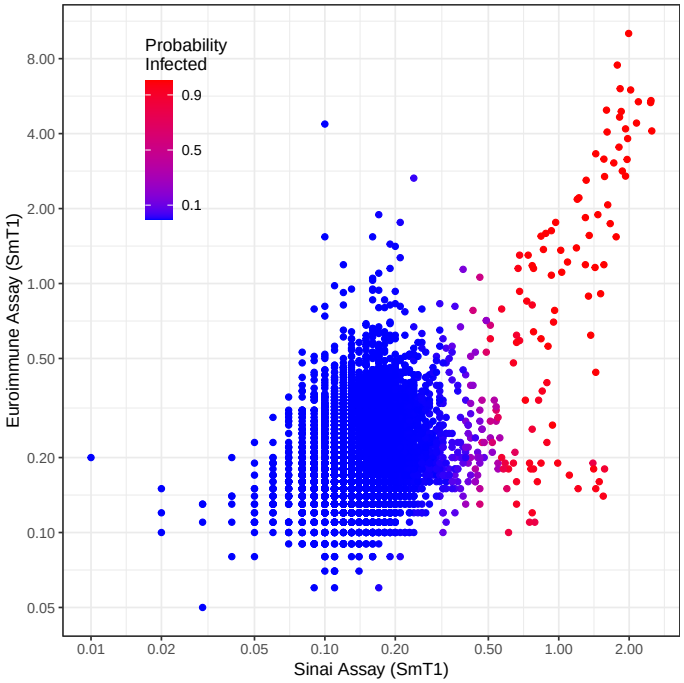
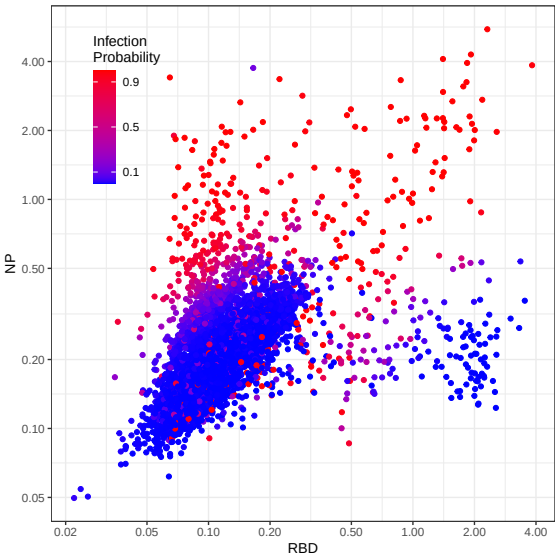


Figure 5.2: Probability of infection given each individual's titre values using the bivariate mixture of t-distributions in Phase 1. Each dot represents a participant in the Ab-C study. On the x-axis is the titre value that was used in the univariate model. On the y-axis is an second SmT1 protein assay. A red dot indicates that this model predicts a high probability of infection, with blue being a low probability of infection, and purple being indeterminate.

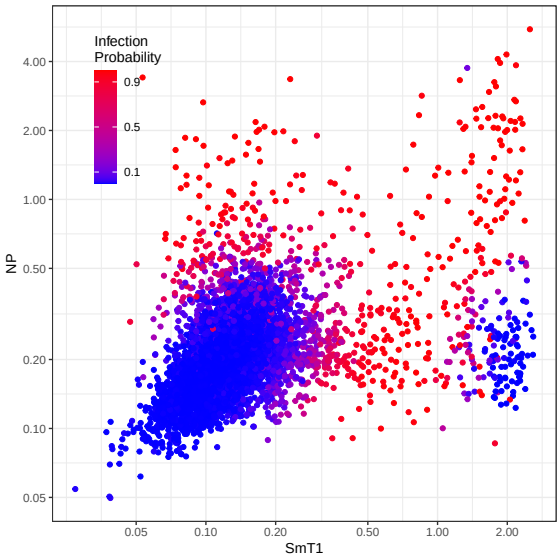
differentiate between "vaccinated, not infected" and "unvaccinated, infected" individuals because infected individuals will tend to have high titre values for all three titres, while vaccinated individuals should not have an elevated titre value for NP. That is, if a participant shows a high value of SmT1 and RBD, and a low value for NP, it should predict a small probability of infection. If a participant has a large value for all three, then the model should predict a large probability of infection.

Furthermore, we decided not to use self reported vaccination status as data, as only about half of the participants who claimed to be vaccinated were showing large values of SmT1 and RBD. This may be because they had only received one dose, or perhaps they had provided their blood spot less than two weeks since their second dose. Either way, we want the data (titre values) to determine SARS-CoV-2 incidence, rather than rely on self-reported claims of vaccination.

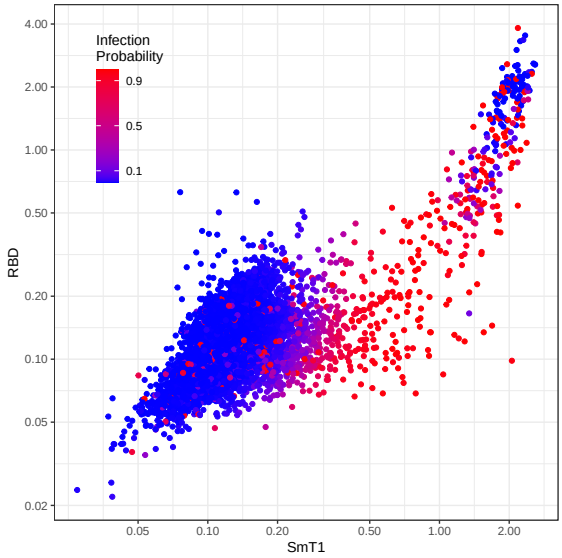
In addition to having three infected statuses, we also now have three titre values which



(a) NP vs. RBD



(b) NP vs SmT1



(c) RBD vs SmT1

Figure 5.3: Probability of infection given each individual's titre values using the trivariate mixture of t-distributions in Phase 2. A red dot indicates that this model predicts a high probability of infection, with blue being a low probability of infection, and purple being indeterminate. In theory, participants who have never been infected or vaccinated should have low values for all three titres. Vaccinated, but never infected individuals should have high SmT1 and RBD, but low NP, and infected individuals have high values for all three.

we can use to define a mixture of three trivariate t-distributions (f_3). The likelihood for this trivariate model is:

$$\begin{aligned}\pi(\log(\boldsymbol{w}_i);\beta,\boldsymbol{\xi},x_i) = (1-\rho)[1-\text{logit}^{-1}(\beta^T x_i)]f_3(\log(\boldsymbol{w}_i)|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\nu_0)\\ + \text{logit}^{-1}(\beta^T x_i)f_3(\log(\boldsymbol{w}_i)|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\nu_1)\\ + \rho[1-\text{logit}^{-1}(\beta^T x_i)]f_3(\log(\boldsymbol{w}_i)|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2,\nu_2)\end{aligned}$$

where $\rho = \text{Prob}(y_i = 2|y_i \neq 1)$. Here, $\text{Prob}(y_i = 0) = \text{Prob}(y_i = 0|y_i \neq 1)\text{Prob}(y_i \neq 1) = (1-\rho)(1-\text{logit}^{-1}(\beta^T x_i))$. We fit this trivariate model to Phase 2 data using Bayesian MCMC to obtain posterior samples of β which will be used for poststratification.

5.2.3 Estimating incidence using poststratification

Incidence is defined as the number of people with an infection in a given time frame, divided by the population. We estimate incidence of COVID-19 in a subgroup of Canadians G by taking posterior samples of I_G where

$$I_G = \frac{\sum_{h\ell j \in G} Y_{h\ell j}}{\sum_{h\ell j \in G} n_{h\ell j}} = \frac{Y_G}{n_G},$$

h is ethnicity/education, ℓ is age/sex, j is province, $p_{h\ell j}$ is the probability of COVID-19 infection (as in Equation 5.2) for a person with covariate combination $h\ell j$, $Y_{h\ell j}$ is the number of people in Canada with covariate combination $h\ell j$ who were infected with COVID-19, and $n_{h\ell j}$ is the number of people in Canada with covariate combination $h\ell j$. To obtain samples of I_G we first fit the mixture models presented in Section 5.2.2 to obtain T posterior samples of $p_{h\ell j}$. We then use poststratification (Little, 1993) to generalize these results to the Canadian population. That is, we draw one sample from

$$Y_{h\ell j}^{(t)} \sim \text{Bin}(n_{h\ell j}, p_{h\ell j}^{(t)})$$

for each $t = 1...T$. We then compute

$$I_G^{(t)} = \frac{\sum_{h\ell j \in G} Y_{h\ell j}^{(t)}}{\sum_{h\ell j \in G} n_{h\ell j}}$$

for $t = 1...T$, which are then used to obtain point estimates and credible intervals for cumulative incidence in Phase 1 and Phase 1 and 2 combined. The incidence proportion in Phase 2 is estimated by computing these two cumulative incidence estimates for each t, then taking the difference.

5.2.4 Estimating infection fatality rates outside of long-term care homes

The infection fatality rate (IFR) is a measure of the deadliness of a disease. It is defined as

$$\text{IFR} = \frac{\text{Number of deaths from disease}}{\text{Number of infected individuals}}.$$

The methods described in Sections 5.2.2 and 5.2.3 provide estimates of the denominator with associated uncertainty, but we still need to estimate the number of deaths in the numerator. The number of COVID-19 related deaths in Canada are publicly available, but include long-term care (LTC) residents. Our target of inference is the IFR for the "community-dwelling" Canadian population and does not apply to people living in LTC homes. The spread of COVID-19 is substantially different in LTC homes than in the general population and residents of LTC homes are particularly vulnerable to severe illness and death from infection; see Danis et al. (2020). Indeed nearly 80% of the reported deaths from COVID-19 prior to Sept. 2020 in Canada were in LTC homes (Samir et al., 2022). Modeling the spread and mortality of COVID-19 within LTC homes will require unique approaches and should be considered in a separate analysis; see the recommendations of Pillemer et al. (2020). The Ab-C study excludes residents of LTC and thus we need to exclude this population from our numerator as well. To do this, we will extend our poststratified mixture models to estimate the deaths outside of long-term care homes, using publicly available COVID-19 deaths data and long-term care deaths data described in Section 5.1.1.

In the rest of this section, we describe the extended mixture model and algorithm used

to estimate IFR in this chapter. We start by displaying the full model with a description of each component. We then provide a Directed Acyclic Graph (DAG) that displays the relationship between all quantities in the model. We then provide a full factorization of the posterior distribution and explain how our algorithm approximates this posterior.

The complete model.

The full model is shown in Equations 5.3a-5.3h, followed by a description of each component. Equations 5.3a-5.3c represent the mixture model and post-stratification described previously, and will be referred to as "Module 1" of our IFR model. Equations 5.3d-5.3h represent the model extension to estimate the number of deaths outside of long-term care, and will be referred to as "Module 2". Left aligned are the model components, right aligned are the nomenclature used in the posterior factorization in Section 5.2.4.

$$\log(\boldsymbol{W}_i)|Z_i = k, x_i \sim f_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) \qquad \pi(\boldsymbol{W}|\boldsymbol{\xi}, Z) \qquad (5.3a)$$

$$\text{Prob}(Z_i = 1|x_i, \beta) = p_{h\ell j[i]} = \text{logit}^{-1}(\beta^T x_i) \qquad \pi(Z|\beta, x) \qquad (5.3b)$$

$$Y_{h\ell j} \sim \text{Bin}(n_{1h\ell j}, p_{h\ell j}) \qquad \pi(Y|\beta, x) \qquad (5.3c)$$

$$D_{1\ell j} \sim \text{Bin}(Y_{\cdot\ell j}, \eta_{\ell j}) \qquad \pi(D|Y, \eta) \qquad (5.3d)$$

$$d_{\ell j} \sim \text{Pois}(\lambda_{1\ell j} + \lambda_{2lj}) \qquad \pi(d|Y, \eta, \theta) \qquad (5.3e)$$

$$d_{2\cdot j} \sim \text{Pois}\Big(\sum_l \lambda_{2\ell j}\Big) \qquad \pi(d_2|\theta) \qquad (5.3f)$$

$$\lambda_{1\ell j} = Y_{\cdot\ell j}\eta_{\ell j} \qquad (5.3g)$$

$$\lambda_{2\ell j} = n_{2\ell j}\theta_{\ell j} \qquad (5.3h)$$

- Indices: h, ℓ, and j represent education/ethnicity, age/sex, and province groups respectively. Subscripts 1 and 2 are used to distinguish between quantities outside and within long-term care respectively.

- 5.3a: The log of the titre values, $\boldsymbol{w_i}$, of individual i follow a (shifted and scaled) multivariate t-distribution, with parameters that depend on the infectious status $Z_i = k$

of that individual. k=0: "unvaccinated, not infected", k=1: "unvaccinated, infected", k=2: "vaccinated, not infected" (for Phase 2 only).

- 5.3b: an individual's infection status, Z_i, depends on the infection probability corresponding to that individual's covariate combination, $p_{h\ell j[i]}$.

- 5.3c: The number of infections in Canada with covariate combination $h\ell j$ is determined by the number of people in Canada with that covariate combination, $n_{h\ell j}$, and the probability, $p_{h\ell j}$, that a person with that covariate combination was infected.

- 5.3d: The number of deaths outside long-term care in age/sex/province group ℓj, $D_{1\ell j}$, depends on the number of infections in that group, $Y_{\cdot\ell j}$, and the infection fatality rate in that group, $\eta_{\ell j}$. Note that we do not attempt to estimate the deaths by education and ethnicity, which is why we sum over h in $Y_{\cdot\ell j}$.

- 5.3e: The total number of COVID-related deaths in age/sex/province group ℓj, $d_{\ell j}$, has death rate equal to the sum of the death rates outside long-term care, $\lambda_{1\ell j}$, and the death rate inside long-term care, $\lambda_{2\ell j}$.

- 5.3f: Outside long-term care, we only know the death rates aggregated by province (the age/sex distribution is unknown). If we assume that the number of deaths outside long-term care in age/sex group ℓ and province j follows an independent Poisson process with mean $\lambda_{2\ell j}$, then the deaths aggregated by province, $d_{2\cdot j}$, will be Poisson distributed with mean $\sum_\ell \lambda_{2\ell j}$. Note that if we knew $d_{2\ell j}$, there would be no need for Module 2.

- 5.3g: In each age/sex/province group, the mean number of deaths (death rate) outside long-term care, $\lambda_{1\ell j}$, is the product of the number of infections outside of long-term care $Y_{\ell j}$, and the infection fatality rate outside long-term care, $\eta_{\ell j}$.

- 5.3h: In each age/sex/province group, the mean number of deaths (death rate) within long-term care, $\lambda_{2\ell j}$, is the product of the number of people in Canada in long-term care $n_{2\ell j}$, and the COVID-19 death rate in long-term care, $\theta_{\ell j}$.

Approximating the Bayesian posterior.

Figure 5.4 displays the model represented in Equations 5.3a-5.3h as a Directed Acyclic Graph (DAG). Based on this DAG, the full posterior can be factored as follows:

$$\begin{aligned}
&\pi(Y, D, \eta, \beta, \boldsymbol{\xi}, \theta, Z | x, \boldsymbol{W}, d, d_2) \\
&\propto \pi(D|Y, \eta)\pi(Y|\beta, x, d)\pi(\boldsymbol{W}, d, d_2|\eta, \beta, \boldsymbol{\xi}, \theta, Z, x)\pi(\eta, \beta, \boldsymbol{\xi}, \theta, Z) \\
&= \underbrace{\pi(Y|\beta, x, d)\pi(\boldsymbol{W}|\boldsymbol{\xi}, Z)\pi(Z|\beta, x)\pi(\beta)\pi(\boldsymbol{\xi})}_{\text{Module 1}} \cdot \underbrace{\pi(D|Y, \eta)\pi(d|Y, \eta, \theta)\pi(d_2|\theta)\pi(\eta)\pi(\theta)}_{\text{Module 2}} \quad (5.4)
\end{aligned}$$

However, sampling from this posterior poses a computational challenge, as Y and D are both discrete latent variables, and all three terms in $\pi(D|Y, \eta)$ are unknown. Instead, we sample from the "cut distribution" (Plummer, 2015), which is the same as Equation 5.4 but the dependence on d in $\pi(Y|\beta, x, d)$ is dropped. The removal of this dependence is sometimes referred to as "cutting feedback". Since we are not allowing our deaths data to influence our infection estimates, we are only approximating Bayesian inference when computing IFR. The cut distribution has been shown to give more sensible results than the full posterior in some scenarios where certain portions (modules) of the model are misspecified, or data quality is poor (Lunn et al., 2009). It is important to note that our serosurvey data is very high quality individual level data, but our deaths data is partially imputed and is from an unofficial source. The cut model allows us to base our estimates of incidence solely on the serosurvey data (and census data), while still utilizing all data sources to estimate IFR. We sample from the cut distribution using the following two step algorithm:

1) We first sample from the joint posterior of the parameters in the first module:

$$\begin{aligned}
\pi(Y, \beta, \boldsymbol{\xi}, Z | x, \boldsymbol{W}) &\propto \pi(Y|\beta, x)\pi(\boldsymbol{W}|, \boldsymbol{\xi}, Z)\pi(Z, \boldsymbol{\xi}, \beta) \\
&= \pi(Y|\beta, x)\pi(\boldsymbol{W}|\boldsymbol{\xi}, Z)\pi(Z|\beta, x)\pi(\beta)\pi(\boldsymbol{\xi})
\end{aligned}$$

which is the same as the Module 1 portion of Equation 5.4 but with the dependence of d dropped in the first term. We sample from this distribution by obtaining T (post burn-in) posterior samples of each parameter using $\pi(\beta, \boldsymbol{\xi}, Z|x, \boldsymbol{W}) = \pi(\boldsymbol{W}|\boldsymbol{\xi}, Z)\pi(Z|\beta, x)\pi(\beta)\pi(\boldsymbol{\xi})$

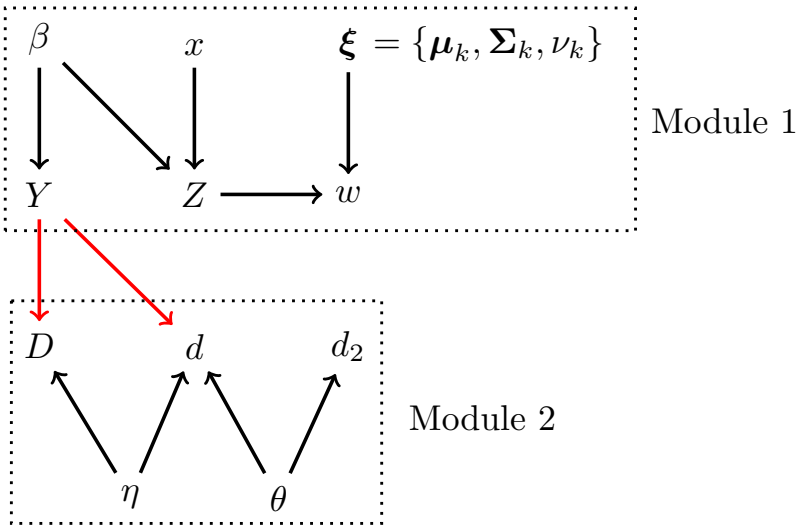


Figure 5.4: Directed acyclic graph corresponding to the model presented in equations 5.3a-5.3h, with subscripts omitted. Lower case Latin letters are known, all other terms are unknown. Module 1 is the portion of the model concerned with estimating infections. Module 2 is the portion of the model concerned with estimating deaths. The red arrows indicate a one-directional flow of information, and are the reason we are sampling from the cut distribution as opposed to the Bayesian posterior. β is the effect of covariates, x, on the log(odds) of infection; Z is infection status, w represents titre values from the serosurvey; $\boldsymbol{\xi}$ are the parameters of the multivariate t-distributions; Y is the number of infections outside of long-term care; D is the number of deaths outside long-term care; d is the total number of deaths by age/sex/province; d_2 is the number of deaths inside long-term care by province; η is the population average probability of death given infection; θ is the COVID-19 death rate in long-term care.

as a target distribution in MCMC. We then draw a sample, $Y^{(t)}$, from $\pi(Y|\beta^{(t)}, x)$ for $t = 1...T$.

2) For each $t = 1...T$, we use MCMC to obtain 1 post burn-in sample from the posterior of Module 2. To do this, we first obtain one post burn-in sample using $\pi(d|Y^{(t)}, \eta, \theta)\pi(d_2|\theta)\pi(\eta)\pi(\theta)$ as the target in MCMC for each $t = 1...T$. We then sample $D^{(t)}$ from $\pi(D|Y^{(t)}, \eta^{(t)})$ for $t = 1...T$.

We used this algorithm for both Phase 1 and Phase 2 data, obtaining T samples of $(Y_{\cdot\ell j}, D_{1\ell j})$ from $\pi_{\text{cut}}(Y, D)$. We then estimate IFR by computing samples from $\pi_{\text{cut}}(\text{IFR}_G)$ for any subgroup of Canadians G outside of long-term care:

$$\text{IFR}_G^{(t)} = \frac{\sum_{\ell j \in G} D_{1\ell j}^{(t)}}{\sum_{\ell j \in G} Y_{\cdot\ell j}^{(t)}} \tag{5.5}$$

for each $t = 1...T$. We can then compute point estimates with uncertainty for all of Canada, and any age/sex/province combination that we so please. We compute the IFR_G for various age/sex/province combinations using univariate and bivariate models to estimate the denominators for the Phase 1 data, and the multivariate model for Phase 1 and 2 combined. We do not attempt to estimate IFR by education/ethnicity, so we sum over h in $Y_{\cdot\ell j}$.

Since individuals who were likely to be positive in Phase 1 were also likely to be positive in Phase 2, estimating incidence and deaths just based on Phase 2 data will also include people who were likely infected in Phase 1. In order to estimate the new infections and deaths (and as a result, IFR) in just Phase 2, we found posterior samples of Y from the multivariate model and subtracted the posterior samples from the bivariate model to get the denominator. The same was done for the deaths D for each posterior sample, allowing us to calculate IFR for any subgroup we desire.

5.2.5 Priors

In all three mixture models, a weakly informative prior of $N(0, 1)$ was used for each β. This will stabilize estimates in groups with a small amount of data, and have little effect on those that have a lot of data. A weakly informative penalized complexity prior was put on the degrees of freedom in all three models (see Appendix 5.A). In the multivariate

cases, informative priors were used to overcome well-known computational challenges of fitting Bayesian mixture models as noted in the Stan documentation (Betancourt, 2017). We describe our informative priors and their justifications in detail in Appendix 5.D.1. In the reproducible example that we provide in the supplemental materials, we show that our results are not too sensitive to "mis-specified" informative priors on the mixture components. We also note that it is primarily the estimation of β's that influence the results of this chapter. A weakly informative prior was used on Σ as recommended by Section 1.13 of the Stan User's Guide (Stan Development Team, 2021). A complete list of priors for all models is presented in Appendix 5.D.

5.2.6 Inference

Each model was run using No-U-Turn sampling, a form of Hamiltonian Monte Carlo that is readily available in the Stan software (Carpenter et al., 2017; Stan Development Team, 2021). Four chains with 1000 iterations, with the first half being warmup, were used for each model component. Traceplots were used to visually assess convergence of Markov chains, alongside values of Rhat < 1.01 confirming an appropriate amount of mixing (Vehtari et al., 2021). Point estimates are taken to be the 50th percentile of the (approximate) posterior distributions, and credible intervals (CrI's) are computed using the 2.5th and 97.5th quantiles.

5.3 Results

5.3.1 Univariate model - Phase 1

Estimated cumulative incidence and IFR by age group is presented in Figure 5.5. Using the univariate model, the overall estimated cumulative incidence in Phase 1 (Feb - Sept 2020) is 1.79% (95% CrI: 1.21%, 2.66%), which is similar to the estimate presented in Tang et al. (2022) of 1.9% (95% CI: 0.7%, 4.7%). Using this model for the denominators in the IFR calculation leads to an estimated infection fatality rate of 0.35% (95% CrI: 0.24%, 0.52%) for all Canadians outside of long-term care homes. This is, again, consistent with the estimates presented in Tang et al. (2022) of 0.373 (95% CI: 0.153%, 1.024%).

When we look at the age distribution of cumulative incidence, we see a general downward trend with increasing age, with estimates for the age group 70+ being the smallest at 0.71% (95% CrI: 0.24%,1.74%). However, the credible intervals all overlap which suggests that incidence is similar between age groups. We see an upward trend in IFR with increasing age, with non-overlapping credible intervals. This is to be expected, as COVID-19 is now known to be much deadlier in older populations (Williamson et al., 2020).

A plot of the two univariate t-distributions is shown in Figure 5.1. Notice that the density plot for the positive group has mass to the left of the cutoff used by Tang et al. (2022), and the negative group has mass to the right of the cutoff. Large values of titres (> 2) will show high probability of SARS-CoV-2 incidence from our model, but this is not true for titre values around 0.5. If these values had been discretized using a fixed cutoff, participants with very large titre values would be indistinguishable from those with values of ≈ 0.5, thus would have the same probability of being false positives. Although this univariate case works well to demonstrate our method, we will use the results from the bivariate model when computing estimates for Phase 1.

5.3.2 Bivariate model - Phase 1

Figure 5.5 presents estimated cumulative incidence and infection fatality rates for the bivariate model in Phase 1 using both SmT1 titres. The overall cumulative incidence for Canada was 1.60% (95% CrI: 1.15%, 2.23%). This point estimate is somewhat consistent (slightly lower) with the univariate results, with a smaller credible interval. This is reassuring, since our uncertainty should decrease as more data is used in the model. Our Phase 1 estimates are comparable with the estimate for seroprevalence in Canada from O'Driscoll et al. (2021) of 1.4% (CI: 1.16%, 1.68%, as of September 1st 2020). The estimated overall infection fatality rates for residents outside of long-term care homes was 0.39% (95% CrI: 0.27%, 0.56%), which is also consistent with our univariate results. We will use the bivariate results for Phase 1 going forward.

When broken down by age, we see very similar trends in both cumulative incidence and IFR as with the univariate model. We also see slightly reduced uncertainty in all age groups, which is to be expected since we are adding more information (an extra titre value)

into the model. The decrease in uncertainty is small, suggesting that the additional assay didn't provide much additional information when predicting infection. We can investigate which titre value had more influence on the probability of infection by computing

$$\text{Prob}(Z_i = 1|\boldsymbol{w}_i) = \frac{\text{Prob}(\boldsymbol{w}_i|Z_i = 1)\text{Prob}(Z_i = 1)}{\text{Prob}(\boldsymbol{w}_i)}$$

That is, we compute the probability of infection given the titre values, which are easily computed based on results from (5.2).

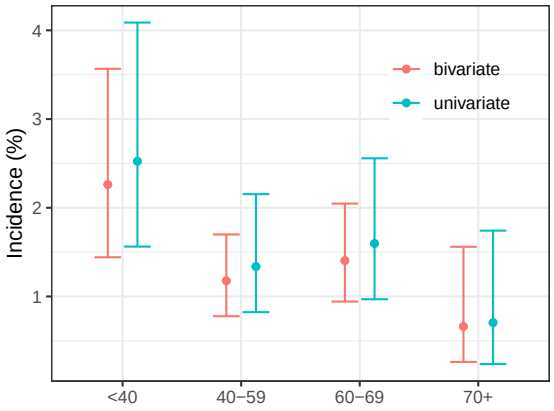
Figure 5.2 shows the probability of infection given each individual's titre values using the Bivariate mixture of t-distributions. Our model seems to "trust" the Sinai titre value more, given that it predicts a high probability when the Sinai value is high, even if the Euroimmune titre value is low. Our model seems to be indeterminate around the cutoff (Sinai titre value ≈ 0.5) that was chosen by Tang et al. (2022), which implies some agreement between the two methods.

5.3.3 Trivariate model - Phase 2

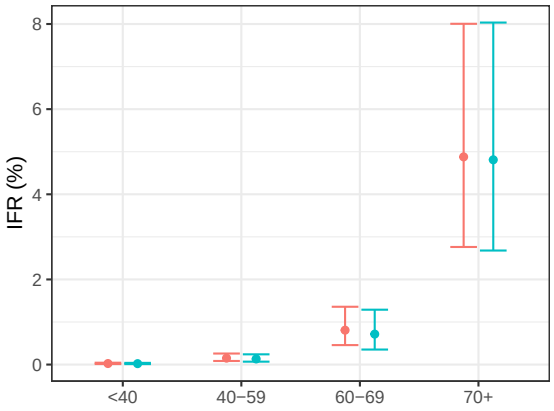
Estimates of cumulative incidences and infection fatality rates in Phase 2 are presented in Figures 5.5c and 5.5d. Using a trivariate mixture of t-distributions with three latent groups and poststratification, the estimated incidence proportion in Phase 2 was 6.81% (95% CrI: 5.35%, 8.42%). This is obviously much higher than our estimates in Phase 1, which is to be expected. The estimated infection fatality rate in Phase 2 was 0.31% (95% CrI 0.25%, 0.39%), which is slightly lower than Phase 1. This is comparable, but slightly lower than other estimates for Canadian IFR ($\sim 0.65\%$ from O'Driscoll et al. (2021)), which is unsurprising since our study excluded those in nursing homes.

The incidence proportion in Phase 2 was comparable across age groups, with the IFR again trending upwards with age. In Phase 2, see that each age category had a lower IFR than Phase 1. Our estimates of IFR by age were highly comparable to international estimates (see Table S3 of O'Driscoll et al. (2021)).

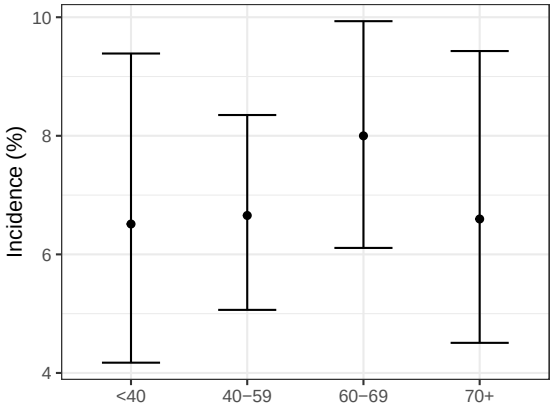
The cumulative incidence and IFR's for Phase 1 and Phase 2 combined are shown in Figures 5.5e and 5.5f. The cumulative incidence estimate is 8.41% (95% CrI: 7.04%, 9.92%),



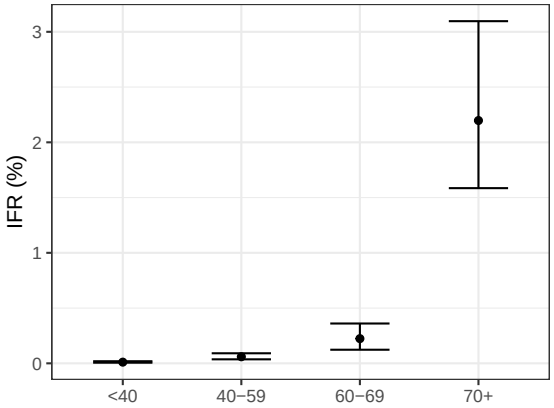
(a) Phase 1 Cumulative Incidence



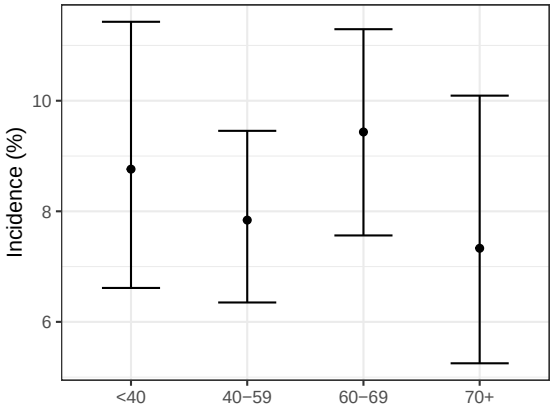
(b) Phase 1 IFR



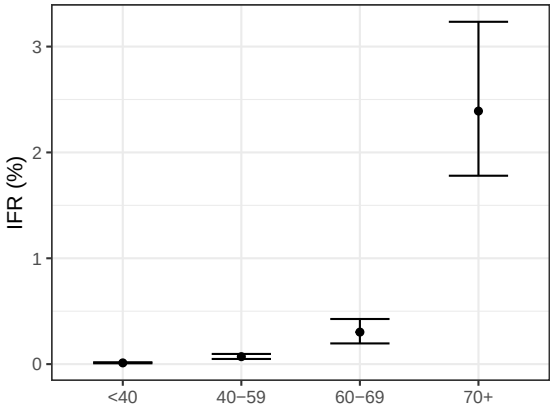
(c) Phase 2 Incidence Proportion



(d) Phase 2 IFR



(e) Phase 1 and 2 Cumulative Incidence



(f) Phase 1 and 2 IFR

Figure 5.5: Incidence/IFR by age (years) for each time period. Posterior medians are used as point estimates, and the 2.5th and 97.5th posterior quantiles define the error bars.

with an IFR of approximately 0.31% (95% CrI: 0.27%, 0.37%). The patterns in incidence and IFR by age are highly similar to those in Phase 2 alon. The probabilities of infection given the titre values of each participant are shown in Figure 5.3. Since our outcome is three-dimensional, three separate plots are required. Blue dots in the bottom right corner of Figures 5.3a and 5.3b, and the top right corner of Figure 5.3c, identify participants that are likely showing immunity due to being vaccinated, as vaccinated individuals should be low on NP and high on the other two. We see that our model tends to "trust" the NP and SmT1 titres more when predicting infection. People who are high on NP or SmT1 tend to have higher probabilities, while people with only high RBD values tend to have a low probability of infection.

5.3.4 Cumulative incidence and IFR by province

One advantage to the methods presented in this chapter, is that once we have posterior samples for infections and deaths outside of long-term care, we can break the results down by any covariate combination that we so please. Figure B2 shows the cumulative incidence and infection fatality rates by province in both phases. In Phase 1, Ontario had the highest point estimate for cumulative incidence, and Quebec had the highest IFR. Our estimated IFR in Ontario was 0.27% (95% CrI: 0.19%, 0.41%) in Phase 1, which is much lower than the estimate given by Public Health Ontario at the time (2.8% as of May 17, 2020 (Public Health Ontario, 2020)). Although these numbers aren't directly comparable, as our estimates do not include people in nursing homes, this likely doesn't account for all of the disparity. Public Health Ontario's number was estimated based on IFR numbers obtained using individual-level data from China (Verity et al., 2020), and was adjusted to match the age distribution of Ontario. We therefore remain somewhat skeptical of the numbers presented in Public Health Ontario (2020). When comparing our overall estimate to the estimate in Verity et al. (2020) (0.657%, CI 0.389% - 1.33%), our number is much more comparable.

In Phase 1, Quebec had a very high reported number of deaths, which was not proportional to the number of long-term-care home deaths, resulting in a high IFR. In Phase 2 Quebec's incidence went up substantially, while the IFR dropped significantly. In Phase 2,

the credible intervals for both cumulative incidence and IFR overlap between provinces.

Estimates by age group in each province are shown in Figure B1. In all provinces, incidence in Phase 1 was highest in 18 to 39 year olds, and lowest in 70+ year olds. With the exception of Alberta, this pattern did not hold in Phase 2, as incidence seems to be less predictable as a function of age. In each province and phase, IFR reliably trends upwards with age.

Estimates of incidence by ethnicity in each province are shown in Appendix 5.C. In both phases, the white and indigenous groups have comparable incidences in each province. The "not white or indigenous" group (NWoI) has relatively high incidence in Ontario and British Columbia in both phases, and low incidence in the Atlantic region and Saskatchewan in Phase 2. Note that estimates of IFR are not reported by ethnicity, as we do not have (even aggregate) COVID-19 deaths data by ethnicity.

5.4 Discussion

In this chapter, we developed an approximate Bayesian approach to estimate cumulative incidence and IFR using a multivariate mixture of t-distributions. We used data from the Ab-C serosurvey to estimate the probability that individuals were infected with COVID-19 based on their titre values and covariate combinations, and used poststratification to generalize our results to the Canadian population that resides outside of long-term care. Our Phase 1 cumulative incidence estimates were slightly lower than previous estimates based on fixed cutoffs. Our Phase 2 estimate was higher than the one in the literature. Furthermore, our method accounts for uncertainty in both the number of infections and the number of deaths, and is essentially a cut model where we do not allow the deaths data to affect the estimation of the number of infections.

Estimates of incidence by age do not show any noteworthy patterns other than a slight upward trend in Phase 1. In both Phase 1 and Phase 2, IFR increased with age. Furthermore, IFR was higher in Phase 2 than Phase 1 in each age group, although the overall IFR was the same.

The main strength of our approach is that it uses the exact titre values as outcomes in

our model, as opposed to a discretized version which discards information. Furthermore, we can leverage multiple titre values in a multivariate model to improve estimated probabilities of infection, while being able to differentiate between previously infected and vaccinated individuals. An additional strength of our study is that error is correctly accounted for in both the calculation of the number of infections and deaths outside long-term care, and consequently, IFR. We have not considered under-reporting of COVID-19 deaths, and we acknowledge this could be a potential issue. One way to accommodate this would be to make an assumption that a known proportion of COVID-19 deaths go unreported and include draws of unreported deaths in each posterior sample of the IFR. In the absence of information of what this proportion should be, we have treated the reported death counts as correct with the caveat that the estimated IFRs only refer to deaths directly attributed to COVID-19.

A methodological limitation of this study is that we are assuming that both the infected and uninfected groups follow a multivariate t-distribution. This may not be the most appropriate distribution for these data, and perhaps a distribution that allows for skewness may be more appropriate. Although our model makes no direct assumption about sensitivity and specificity, these two quantities are directly related to the length of the tails of the t-distributions for any given cutoff. However, the parameters of the multivariate t-distribution are estimated from the data, so our method is analogous to a non-discretized version of the methodology presented in Gelman and Carpenter (2020), where sensitivity and specificity are parameters to be estimated in the model.

A second limitation is that some responses to the survey happened before the end of the survey, such that they could have returned a "negative" dry blood spot sample and subsequently gotten infected. This would lead to slightly underestimating incidence (overestimating IFR). On the other hand, there is a time lag between infection and death, so if we counted infections up until the end of September 2020, then those infected people could experience death several weeks later and not be recorded. However, given that the vast majority of participants returned their blood samples study more than two weeks prior to each Phase's end date (see Figure F1), we figured that accounting for this time lag was not necessary.

A third limitation of our methodology is that we were unable to incorporate information regarding Phase 1 infection probabilities (from SmT1 protein) into our Phase 2 estimates of incidence. Although Phase 1 and Phase 2 SmT1 protein titre values are not directly comparable (due to the assays being calibrated slightly differently), we recognize that there is some potential to treat the SmT1 titre longitudinally from Phase 1 to Phase 2. However, we figured that this would require a drastic reworking of our current model and inference framework, and thus we deemed it out of the scope of this chapter. The potential consequence of this is a slight underestimate of cumulative incidence at the end of Phase 2, as some "infected" individuals in Phase 1 would have their immunity wane. However, Tang et al. (2022) show that roughly 80% of people retain their "seropositivity" status from Phase 1 to Phase 2. Furthermore, the exploratory analysis presented in Appendix 5.E suggests that waning may not be a large issue.

A direction for future work will be to apply these methods to upcoming Phase 3 and Phase 4 data that includes a much larger vaccinated population, as well as breakthrough infections in people who have been vaccinated. Furthermore, we will have to account for reinfections as the populations' immunity wanes and new variants emerge. This could involve a longitudinal mixture model or Hidden Markov Model. Furthermore, an improved serosurvey design and associated statistical methodology that allowed for estimation of incidence (and consequently, IFR) in real-time would be an ambitious and highly interesting area of future research.

This study only looks at humoral immune response, but cellular immunity also plays an important role in the immune response to SARS-CoV-2. Other studies have evaluated the effects of T-cell response in infected people (Guo et al., 2022; Moss, 2022). An interesting line of future work would be to develop similar methods to incorporate T-cell response data into estimates of incidence and IFR.

Although we focused on SARS-CoV-2 infections and deaths in this chapter, the methods presented can be applied to a variety of outcomes for any infectious disease of interest in which serosurvey data is available. There are plenty of potential extensions to this model that can be implemented to suit a variety of problems in epidemiology and biostatistics.

5.5 Bibliography

Alfego, D., Sullivan, A., Poirier, B., Williams, J., Grover, A., Gillim, L., Adcock, D., and Letovsky, S. (2021). A population-based analysis of the longevity of SARS-CoV-2 antibody seropositivity in the United States. *EClinicalMedicine*, 36:100902.

Betancourt, M. (2017). Identifying bayesian mixture models. `https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html`. Accessed Sept 18, 2022.

Bottomley, C., Otiende, M., Uyoga, S., Gallagher, K., Kagucia, E., Etyang, A., Mugo, D., Gitonga, J., Karanja, H., Nyagwange, J., et al. (2021). Quantifying previous SARS-CoV-2 infection through mixture modelling of antibody levels. *Nature Communications*, 12(1):1–7.

Bouman, J. A., Riou, J., Bonhoeffer, S., and Regoes, R. R. (2021). Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: Exploiting cutoff-free approaches. *PLoS Computational Biology*, 17(2):e1008728.

Burstyn, I., Goldstein, N. D., and Gustafson, P. (2020a). It can be dangerous to take epidemic curves of COVID-19 at face value. *Canadian Journal of Public Health*, 111(3):397–400.

Burstyn, I., Goldstein, N. D., and Gustafson, P. (2020b). Towards reduction in bias in epidemic curves due to outcome misclassification through Bayesian analysis of time-series of laboratory test results: Case study of COVID-19 in Alberta, Canada and Philadelphia, USA. *BMC Medical Research Methodology*, 20(1):1–10.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Centre for Disease Control and Prevention (2022). Interim guidelines for COVID-19 antibody testing. `https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antibody-tests-guidelines.html`. Accessed Sept 30, 2022.

Chen, X., Chen, Z., Azman, A. S., Deng, X., Sun, R., Zhao, Z., Zheng, N., Chen, X., Lu, W., Zhuang, T., et al. (2021). Serological evidence of human infection with SARS-CoV-2: a systematic review and meta-analysis. *The Lancet Global Health*, 9(5):e598–e609.

Danis, K., Fonteneau, L., Georges, S., Daniau, C., Bernard-Stoecklin, S., Domegan, L., O'Donnell, J., Hauge, S. H., Dequeker, S., Vandael, E., et al. (2020). High impact of COVID-19 in long-term care facilities, suggestion for monitoring in the EU/EEA, May 2020. *Eurosurveillance*, 25(22):2000956.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1269–1283.

Guo, L., Wang, G., Wang, Y., Zhang, Q., Ren, L., Gu, X., Huang, T., Zhong, J., Wang, Y., Wang, X., et al. (2022). SARS-CoV-2-specific antibody and T-cell responses 1 year after infection in people recovered from COVID-19: a longitudinal cohort study. *The Lancet Microbe*, 3(5):e348–e356.

Houlihan, C. F. and Beale, R. (2020). The complexities of SARS-CoV-2 serology. *The Lancet Infectious Diseases*, 20(12):1350–1351.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Crc Press.

Kyomuhangi, I. and Giorgi, E. (2022). A threshold-free approach with age-dependency for estimating malaria seroprevalence. *Malaria Journal*, 21(1):1–12.

Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88(423):1001–1012.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(1):19–38.

Luo, G., Zhang, X., Zheng, H., and He, D. (2021). Infection fatality ratio and case fatality ratio of COVID-19. *International Journal of Infectious Diseases*, 113:43–46.

Moss, P. (2022). The t cell immune response against sars-cov-2. *Nature immunology*, 23(2):186–193.

Nielsen, S. S., Toft, N., Jørgensen, E., and Bibby, B. M. (2007). Bayesian mixture models for within-herd prevalence estimates of bovine paratuberculosis based on a continuous ELISA response. *Preventive Veterinary Medicine*, 81(4):290–305.

Ødegård, J., Jensen, J., Madsen, P., Gianola, D., Klemetsdal, G., and Heringstad, B. (2003). Detection of mastitis in dairy cattle by use of mixture models for repeated somatic cell scores: A Bayesian approach via Gibbs sampling. *Journal of Dairy Science*, 86(11):3694–3703.

Ødegård, J., Madsen, P., Gianola, D., Klemetsdal, G., Jensen, J., Heringstad, B., and Korsgaard, I. (2005). A Bayesian threshold-normal mixture model for analysis of a continuous mastitis-related trait. *Journal of Dairy Science*, 88(7):2652–2659.

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A., Azman, A. S., Paireau, J., Fontanet, A., Cauchemez, S., and Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, 590(7844):140–145.

Pillemer, K., Subramanian, L., and Hupert, N. (2020). The importance of long-term care populations in models of COVID-19. *JAMA*, 324(1):25–26.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43.

Public Health Ontario (2020). COVID-19 case fatality, case identification, and attack rates in Ontario. `https://www.publichealthontario.ca/-/media/documents/ncov/epi/2020/06/covid19-epi-case-identification-age-only-template.pdf?sc_lang=en`. Accessed Sept 18, 2022.

Samir, S. K., Doherty, R., McCleave, R., and Dunning, J. (2022). NIA Long Term Care COVID-19 Tracker. https://ltc-covid19-tracker.ca/.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.

Stan Development Team (2021). RStan: the R interface to Stan. R package version 2.21.3.

Statistics Canada (2016). Table 13-10-0809-01 Canadians' health and COVID-19, by region, age, gender and other characteristics. https://doi.org/10.25318/1310080901-eng.

Tang, X., Sharma, A., Pasic, M., Brown, P., Colwill, K., Gelband, H., Birnboim, H. C., Nagelkerke, N., Bogoch, I. I., Bansal, A., et al. (2022). Assessment of SARS-CoV-2 seropositivity during the first and second viral waves in 2020 and 2021 among Canadian adults. *JAMA Network Open*, 5(2):e2146798–e2146798.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bü"rkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6):669–677.

Villa, C. and Rubio, F. J. (2018). Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula. *Computational Statistics & Data Analysis*, 124:197–219.

Vink, M. A., Berkhof, J., van de Kassteele, J., van Boven, M., and Bogaards, J. A. (2016). A bivariate mixture model for natural antibody levels to human papillomavirus types 16 and 18: baseline estimates for monitoring the herd effects of immunization. *PLOS One*, 11(8):e0161109.

Vink, M. A., van de Kassteele, J., Wallinga, J., Teunis, P. F., and Bogaards, J. A. (2015). Estimating seroprevalence of human papillomavirus type 16 using a mixture model with smoothed age-dependent mixing proportions. *Epidemiology*, 26(1):8–16.

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., et al. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821):430–436.

World Health Organization (2022). WHO Coronavirus (COVID-19) dashboard. https://covid19.who.int/.

5.A Penalized complexity prior on degrees of freedom

As mentioned in 5.2.2, we noticed that a Normal distribution is likely not heavy-tailed enough to accurately model the log(titre) of the non-infected group. The t-distribution adds a degrees of freedom parameter, ν, which controls how heavy-tailed the t-distribution is relative to the Normal distribution. The t-distribution reduces to a Normal distribution as $\nu \to \infty$. Therefore we can view ν in this case as a parameter that adds complexity to a base model, the Normal model. The closer ν is to 1, the more "complex" the model is. Simpson et al. (2017) outlines a framework for penalizing model component complexity as a function of the distance to a base model. We used a penalized complexity (PC) prior on ν that will encourage ν to be large (closer to the Normal model) unless there is appropriate evidence in the data.

Rather than putting a prior on ν itself, Simpson et al. suggest putting a prior on the root Kullback-Leibler (KL) distance:

$$\delta(\nu) = \sqrt{2 \cdot D_{\text{KL}}[t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})]} \tag{5.6}$$

where t_ν and $\mathcal{N}$ denote the multivariate t and normal densities respectively, and D_{KL} is the KL divergence. Note that the shifting ($\boldsymbol{\mu}$) and scaling ($\boldsymbol{\Sigma}$) parameters cancel out, and hence D_{KL} is only a function of ν (Villa and Rubio, 2018). Unfortunately, D_{KL} in Equation (5.6) has no closed form that the authors are aware of, so we computed it numerically as described in Appendix 5.A.

(Villa and Rubio, 2018) showed that the Kullback-Liebler Divergence between two d-dimensional Multivariate-t distributions, $f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, and $f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu')$, is

$$\log \frac{K(d,\nu)}{K(d,\nu')} - \frac{\nu + d}{2} \text{E}_f \Big[\log \Big(1 + \frac{\boldsymbol{x}^T \boldsymbol{x}}{\nu} \Big) \Big] + \frac{\nu' + d}{2} \text{E}_f \log \Big(1 + \frac{\boldsymbol{x}^T \boldsymbol{x}}{\nu'} \Big)$$

where

$$K(d,\nu) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{(\pi\nu)^d}}$$

The two expectations are shown to be

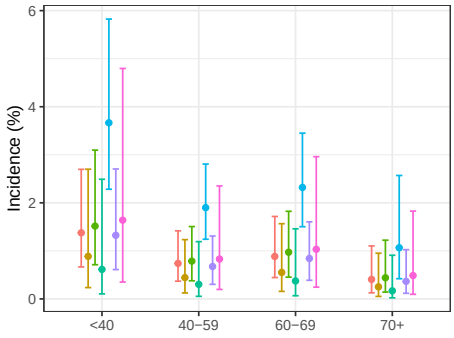
$$\mathrm{E}_f\Big[\log\Big(1+\frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu}\Big)\Big] = \Psi\Big(\frac{\nu+d}{2}\Big) - \Psi\Big(\frac{\nu}{2}\Big)$$
$$\mathrm{E}_f\Big[\log\Big(1+\frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu'}\Big)\Big] = K(d,\nu)\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}\int_0^\infty \Big(1+\frac{t}{\nu}\Big)^{-\frac{\nu+d}{2}} t^{\frac{d}{2}-1}\log\Big(1+\frac{t}{\nu'}\Big)dt$$

Meaning that the d-dimensional integral can be reduced to one dimensional integral. Since we are interested in the KLD between a multivariate T and a multivariate normal, we substitute $\nu' = 200$, and compute this integral numerically as a function of ν. We then approximate the distance, $\delta(\nu) = \sqrt{2\cdot \mathrm{D_{KL}}}$ with a polynomial. For example, $\delta(\nu)$ for the bivariate model was $\delta(\nu) \propto \nu^{-1.3}$. We then say that

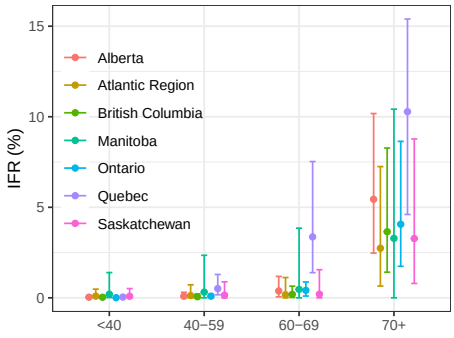
$$\pi(\delta(\nu)) \sim \exp(\lambda)$$

with $\lambda = -\log(\alpha)/\delta(U)$ where α and U are chosen such that our prior belief is that there is a 50% chance that ν is greater than 30.

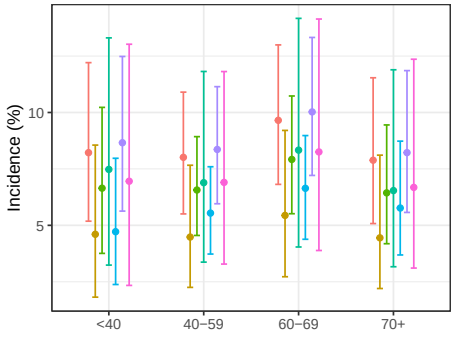
5.B Estimates by age and Province



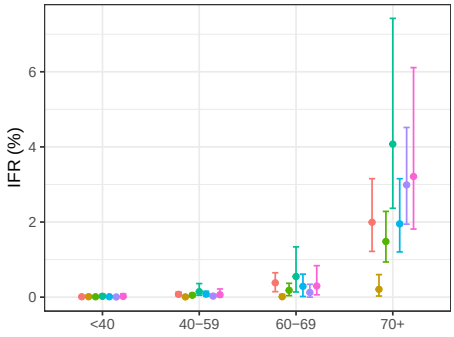
(a) Phase 1 Cumulative Incidence



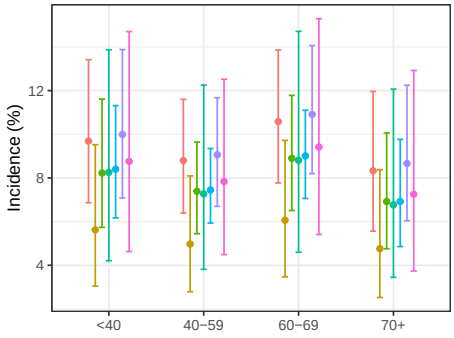
(b) Phase 1 IFR



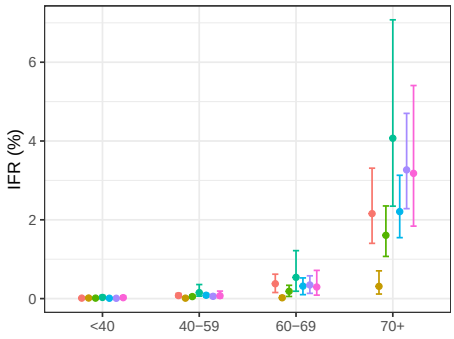
(c) Phase 2 Incidence Proportion



(d) Phase 2 IFR

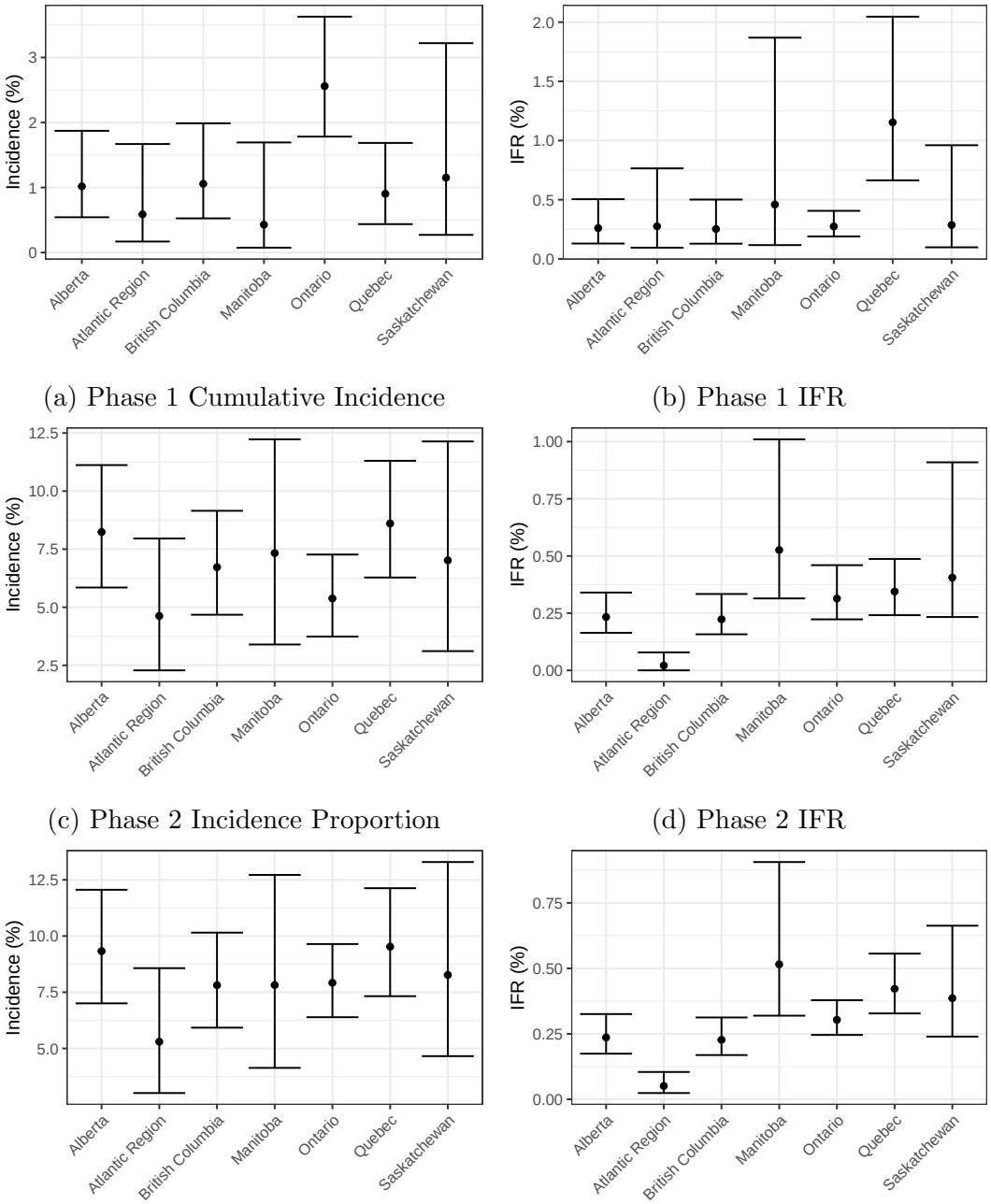


(e) Phase 1 and 2 Cumulative Incidence



(f) Phase 1 and 2 IFR

Figure B1: Incidence/IFR by age (years) in each province. Posterior medians are used as point estimates, and the 2.5th and 97.5th posterior quantiles define the error bars.



(a) Phase 1 Cumulative Incidence

(b) Phase 1 IFR

(c) Phase 2 Incidence Proportion

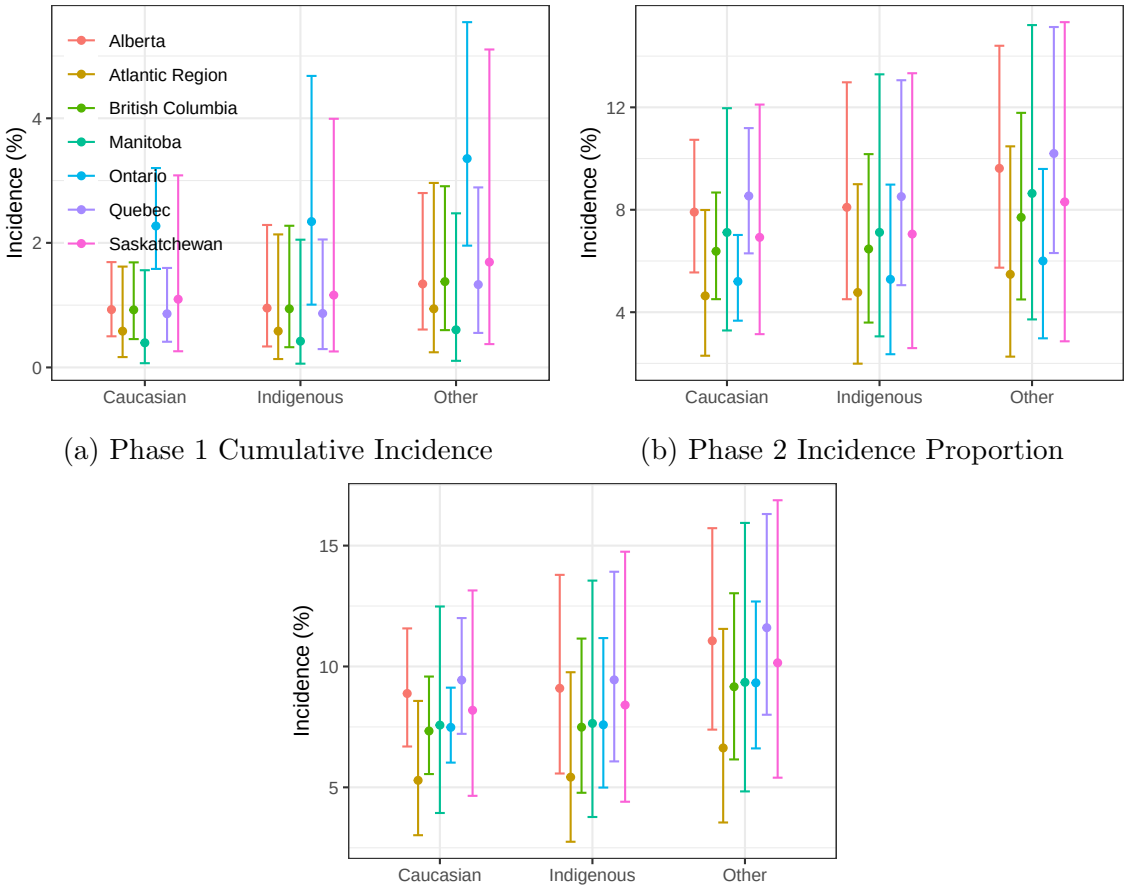
(d) Phase 2 IFR

(e) Phase 1 and 2 Cumulative Incidence

(f) Phase 1 and 2 IFR

Figure B2: Incidence/IFR by province. Posterior medians are used as point estimates, and the 2.5th and 97.5th posterior quantiles define the error bars.

5.C Estimates by province and ethnicity



(a) Phase 1 Cumulative Incidence

(b) Phase 2 Incidence Proportion

(c) Phase 1 and 2 Cumulative Incidence

Figure B3: Incidence by ethnicity in each province. Posterior medians are used as point estimates, and the 2.5th and 97.5th posterior quantiles define the error bars.

5.D Prior distributions

Parameter	Prior
μ_0, μ_1	$N(0, 10)$
σ_0, σ_1	$N_+(0, 10)$
β	$N(0, 1)$
ν_k	$\text{Prob}(\nu > 10) = 0.5$

Table D1: Priors used in Phase 1 univariate model

Parameter	Prior
μ_0	$MVN\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}\right)$
μ_1	$MVN\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}\right)$
β	$N(0, 1)$
ν_k	$\text{Prob}(\nu > 10) = 0.5$
$\mathbf{\Sigma}_k = \text{diag}(\tau) \times \Omega \times \text{diag}(\tau)$	
τ	$\text{Cauchy}_+(0, 1)$
Ω	$\text{LKJCorr}(2)$

Table D2: Priors used in Phase 1 bivariate model

5.D.1 Phase 2 model prior justification

As mentioned in the main text, we require informative priors for computational reasons. In this Section, we justify our choices of informative priors for the Phase 2 trivariate model. We note that these priors are not very sensitive to

- μ_0 corresponds to the means of the "not infected" group. The first element of μ_0 corresponds to the mean NP titre values in "not infected individuals". Alongside the NP titre values collected from the survey, the lab also provided us with "control" samples of known negatives. We found that the vast majority of the control samples fell between -2.5 and -1 on the log scale. Therefore we are very confident that the *mean* of NP titre values from "not infected" people should be in this range. Therefore we applied the conservative but informative prior $N(-1.75, 0.25)$. Similar reasoning was used for the prior on the second element of μ_0, corresponding to the mean of RBD titre values in "not infected" people.

Parameter	Prior
μ_0	$MVN\left(\begin{bmatrix}-1.75\\-2.4\\-1.918\end{bmatrix}, \begin{bmatrix}0.25 & 0 & 0\\0 & 0.2 & 0\\0 & 0 & 0.03\end{bmatrix}\right)$
μ_1	$MVN\left(\begin{bmatrix}-0.5\\0\\-0.065\end{bmatrix}, \begin{bmatrix}0.2 & 0 & 0\\0 & 0.1 & 0\\0 & 0 & 0.07\end{bmatrix}\right)$
μ_2	$MVN\left(\begin{bmatrix}-\\0.6\\0.6\end{bmatrix}, \begin{bmatrix}- & - & -\\- & 0.2 & 0\\- & 0 & 0.2\end{bmatrix}\right)$
β	$N(0,1)$
ν_k	$\text{Prob}(\nu_k > 30) = 0.5$
ρ	$N_+(0.015, 0.0025)$
$\mathbf{\Sigma}_k = \text{diag}(\tau) \times \Omega_k \times \text{diag}(\tau)$	
τ	$\text{Cauchy}_+(0,1)$
Ω_k	$\text{LKJCorr}(0.5) \prod_c N(c\|m_c, s_c)$

Table D3: Priors used in Phase 2 mixture model

Parameter	Prior
η	$N_+(0.004, 0.05)$
θ	$N_+(0.01, 0.1)$

Table D4: Priors used in deaths module (Section 5.2.4)

- When setting priors for the "not vaccinated, not infected" and "infected" groups based on Smt1 titre values, we used the corresponding posterior distributions from Phase 1. Although the tests are calibrated slightly differently, and there will be a small amount of waning between phases, we do expect these values to be somewhat similar.

- To determine the posterior of the mean of the infected group for NP titre values (first element of μ_1), we consider the fact that any titre value above mean+3SDs is likely a previous infection (this is how the cutoff was chosen in Tang et al.). We then ensure that the bulk of the prior distribution for the positive N group was above this value, with some overlap. We used similar reasoning for the RBD positive group.

- To determine the prior for the mean RBD/SmT1 titre values in the vaccinated groups, we used similar reasoning as above, trying to ensure that the prior has most of it's mass above that of the infected group's with some overlap.

- We used a weakly informative prior for Ω_k using the the LKJ distribution with

shape=0.5. This provides a roughly uniform distribution across positive-semidefinite 3x3 matrices. We then add additional information for each off-diagonal by multiplying by normal densities. For instance, if we suspect that the correlation between two parameters should be positive (i.e off-diagonal element c of Ω_k is positive), we multiply the prior for c by $N(c|0.5, 0.2)$ which gently encourages the the correlation to be positive, but still has mass below 0.

5.E Potential waning immunity

It is well known that antibodies decay over time, but how much this effects our results is unclear. Unfortunately, we can't simply compare antibody results from Phase 1 to Phase 2, as these numbers are not directly comparable. Instead, we compared the Phase 1 and Phase 2 probabilities of participants who had a high probability of infection in Phase 1. A comparison of these predicted probabilities is shown in Figure E1. It appears that those with large predicted probabilities in Phase 1 still had large predicted probabilities in Phase 2. This is largely because in Phase 2, we see relatively lower parameter estimates for the means of the infected group. This likely will also make estimates of infection noisier, as the variance will also increase. So although our model does not appear to be underestimating Cumulative Incidence due to waning, waning likely does cause more uncertainty when predicting infection. More work needs to be done to confirm this assertion.

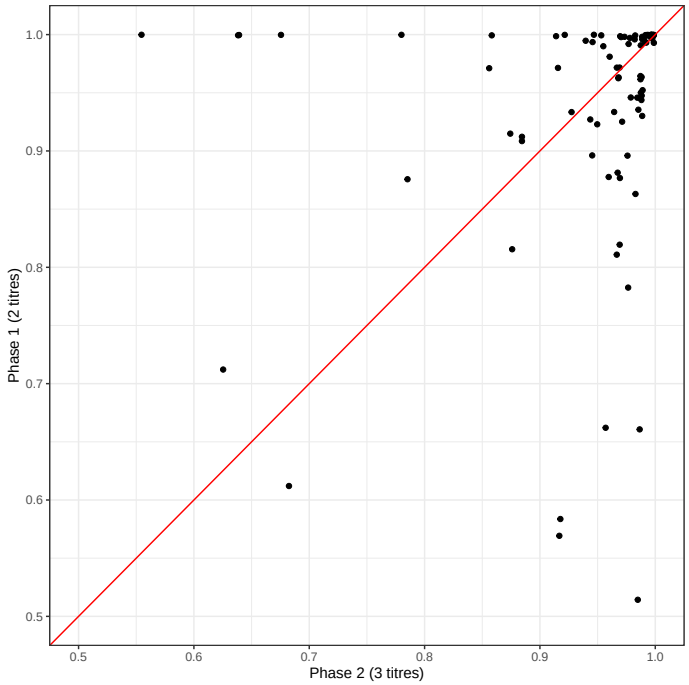


Figure E1: Phase 1 vs Phase 2 predicted probabilities for participants who had large predicted probabilities in Phase 1. Points above the red line indicate that Phase 1 predicted probability was higher.

5.F Date distributions of samples received

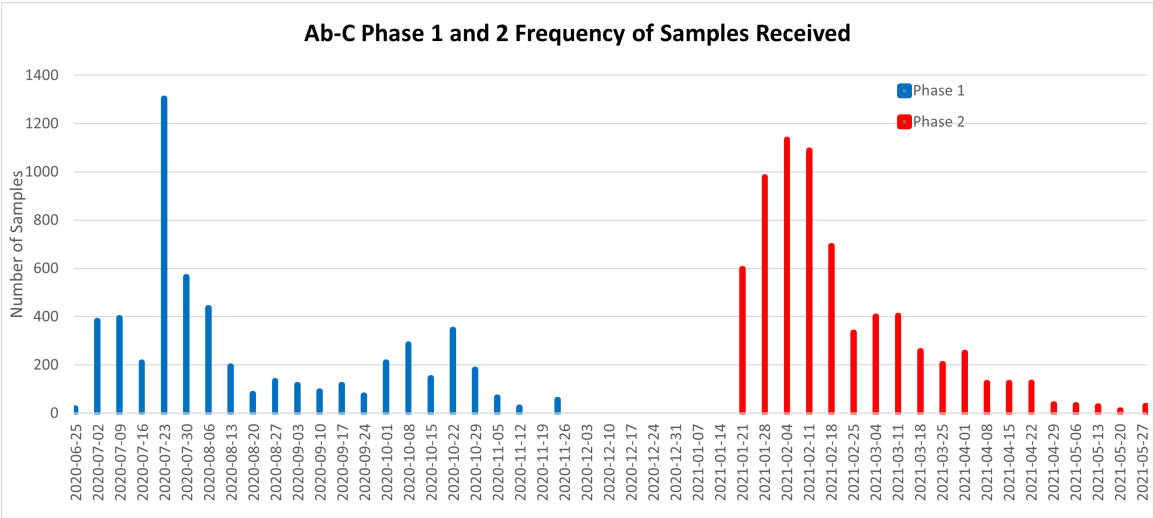


Figure F1: Distribution of dates of samples received for Phase 1 and Phase 2.