Coupled Markov Switching Models for Spatio-temporal Infectious Disease Counts

Dirk Douwes-Schultz

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec
August 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

For Mom, Dad, Katja, Robbie, Bailey, Maggie, Lovers, Oreo and Blackjack.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor, Dr. Alexandra M. Schmidt, for her mentorship. This work would not have been possible without her. I also wish to express gratitude to my professors at McGill: Dr. James Hanley, Dr. Erica Moodie, Dr. Yi Yang, Dr. Masoud Asgharian, Dr. Abbas Khalili, Dr. Johanna Nešlehová, Dr. Mamadou Yauck, Dr. Andrea Benedetti, Dr. Mathieu Maheu-Giroux and Dr. Joanna-Trees Merckx. Not only did their teaching influence many of the decisions made in this work, but also the general philosophy towards statistical and epidemiological modeling that we adopt. Finally, I would like to acknowledge some of my peers at McGill: Paritosh Kumar Roy, Armando Turchetta, Shuo Mila Sun, Victoire Michal and Mingchi Xu, for their companionship and many stimulating and interesting discussions about statistics.

Much of this work, particularly the simulation studies, was made possible by the computing cluster provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). Dr. Laís Picinini Freitas and Dr. Marilia Sá Carvalho provided useful feedback on the third manuscript. This work was funded in part by support from IVADO and the Canada First Research Excellence Fund/Apogée (PhD Excellence Scholarship 2021-907037534), as well as by my PhD supervisor.

Lastly, I want to thank my family for their continuous support. This thesis is dedicated to them.

Preface

This manuscript-based thesis contains six chapters: an introduction in Chapter 1; a literature review in Chapter 2, meant to be an introduction to the general domain of the thesis; three stand-alone manuscripts in Chapters 3, 4 and 5; and a conclusion in Chapter 6. Each stand-alone manuscript is preceded by a preamble that describes the contributions of the work and how it fits within the bigger picture. A complete bibliography is presented after the appendices.

All three stand-alone manuscripts were conceptualized by Dirk Douwes-Schultz (DDS) and edited by Alexandra M. Schmidt (AMS). DDS wrote each manuscript and conducted all analyses. AMS supervised the works and provided useful suggestions and feedback on the analyses. David Buckeridge and Yannan Shen provided the data for the second manuscript, edited the second manuscript and offered useful suggestions concerning some of the analyses in the second manuscript.

Chapters 1, 2 and 6, as well as the preambles, were written by DDS and edited by AMS.

Abstract

Spatio-temporal infectious disease counts are often subject to abrupt and dramatic changes in behavior associated with different epidemiological events. For example, a disease might go temporarily extinct in an area causing cases to drop to zero for several weeks, or an outbreak might emerge causing cases to rise rapidly. In this thesis, we propose several novel Bayesian coupled Markov switching models to account for these types of shifts in behavior. Our approach in general is to assume the disease moves between different epidemiological periods or states in each area, such as disease absence or outbreak. When the disease is in a certain state in an area, the corresponding time series follows an appropriate statistical submodel, e.g., a degenerate 0 distribution if the disease is absent or an autoregressive model with high autocorrelation if there's an outbreak. We switch between the states through a first-order Markov chain in each area where the transition probabilities can depend on covariates and the states in neighboring areas, to account for disease spread between the areas. Inference is performed under the Bayesian paradigm, and we develop efficient Markov chain Monte Carlo methods based on jointly sampling the hidden state indicators.

In the first manuscript, motivated by the abundance of spatio-temporal disease counts that contain many zeroes, we focus on models where the disease switches between periods of presence and absence in each area. Since we describe the absence state with a degenerate 0 distribution, our approach has similarities to more traditional zero-inflated models like ZIP regression. However our framework has several advantages: it naturally accounts for long periods of disease presence and absence (many consecutive zeroes); it can examine if the effects of disease spread between neighboring areas depend on certain covariates, e.g., if there is a barrier between the areas like a river; and we allow each covariate to have a separate effect on the reemergence (absence to presence) and persistence (presence to presence) of the disease, which is often epidemiologically motivated. We illustrate these advantages by comparing our model with several zero-inflated and non-zero-inflated alternatives on spatio-

temporal dengue counts in Rio de Janeiro.

In the second manuscript, we propose models where the disease switches between absence, endemic and outbreak periods in each area. The endemic and outbreak periods are described by autoregressive count models distinguished by a higher level of transmission during outbreaks. Markov switching models that switch between endemic and outbreak states have a long history. However, our approach has several advantages: it can account for long strings of zeroes, it allows for outbreaks in neighboring areas to affect the probabilities of outbreak emergence and persistence (along with covariates) and it prevents rapid switching between the states by enforcing minimum endemic and outbreak state durations. We apply our model, along with alternatives, to COVID-19 hospital admissions across Quebec and to simulated data where the outbreaks are known.

In the third manuscript, we deal with the case of multiple diseases switching between periods of presence and absence. We are only interested in comparing the transmission dynamics of the diseases and so we assume the cases of the present diseases in an area jointly follow a multinomial distribution. Our proposal represents an interesting leap as all existing zero-inflated multinomial models have assumed independent multivariate observations. We apply the model to spatio-temporal counts of dengue, Zika and chikungunya in Rio de Janeiro.

Many existing statistical models cannot account for, study, detect or forecast the abrupt and dramatic shifts in behavior often observed in spatio-temporal infectious disease counts. Therefore, this thesis represents an important contribution to the area of spatio-temporal infectious disease modeling.

Abrégé

Les cas de maladies infectieuses spatio-temporels ont souvent des changements de comportement brusques et spectaculaires associés à des événements épidémiologiques. Par exemple, une maladie peut disparaître temporairement dans une région, avec une chute des cas à zéro pendant plusieurs semaines, ou une épidémie peut apparaître, avec une hausse rapide des cas. Dans cette thèse, nous proposons des nouveaux modèles de Markov couplés bayésiens pour s'adapter à ces changements. Nous supposons que dans chaque région, la maladie évolue en différentes périodes, ou états épidémiologiques, tels qu'absence ou épidémie. Lorsque la maladie est dans un certain état, la série chronologique correspondante suit un sous-modèle approprié, comme une distribution dégénérée à 0 si la maladie est absente, ou un modèle autorégressif à autocorrélation élevée en cas d'épidémie. Dans chaque région, on change d'état via une chaîne de Markov de premier ordre, dont les probabilités de transition dépendent de covariables et des états voisins, pour tenir compte de la propagation de la maladie entre les régions. L'inférence suit le cadre bayésien et nous proposons des méthodes de Monte Carlo par chaîne de Markov efficaces avec un échantillonnage conjoint des indicateurs d'état cachés.

Dans le premier manuscrit, vu l'abondance de cas spatio-temporels avec de nombreux zéros, nous étudions des modèles où la maladie alterne entre des périodes de présence et d'absence dans chaque région. Puisque l'état d'absence est décrit par une distribution dégénérée à 0, notre approche ressemble aux modèles gonflés à zéro plus communs, comme la régression ZIP. Cependant, notre cadre a plusieurs avantages : il s'ajuste naturellement aux longues périodes de présence et d'absence (nombreux zéros consécutifs) ; il examine si les effets de propagation d'une maladie entre voisins dépendent de covariables, par exemple s'il y a une barrière entre les régions comme une rivière ; et chaque covariable peut avoir un effet distinct sur la réémergence (d'absence à présence) et la persistance (de présence à présence), ce qui est souvent épidémiologiquement sensé. Nous illustrons ces avantages en comparant notre

modèle à des alternatives, gonflées à zéro et non gonflées à zéro, sur les cas de dengue à Rio de Janeiro.

Dans le deuxième manuscrit, nous proposons des modèles où la maladie alterne entre des périodes d'absence, d'endémie et d'épidémie. Les périodes d'endémie et d'épidémie sont décrites par des modèles autorégressifs, avec un niveau d'autorégression plus élevé en cas d'épidémie. Les modèles de Markov alternant entre les états endémiques et épidémiques ont une longue histoire. Cependant, notre approche a plusieurs avantages : elle s'adapte à de longues périodes de zéros, elle permet aux épidémies dans les régions voisines d'impacter les probabilités d'émergence et de persistance de l'épidémie (avec les covariables), et elle empêche un changement rapide d'état en imposant une durée minimale d'endémie et d'épidémie. Nous appliquons notre modèle, ainsi que des alternatives, aux hospitalisations liées à la COVID-19 au Québec, et à des données simulées où les éclosions sont connues.

Dans le troisième manuscrit, nous analysons plusieurs maladies alternant entre présence et absence. Nous souhaitons comparer les dynamiques de transmission entre les maladies, et supposons donc que les cas des maladies présentes dans une région suivent conjointement une loi multinomiale. Notre proposition est intéressante car les modèles existants de loi multinomiale gonflée à zéro supposent l'indépendance des observations multivariées. Nous appliquons le modèle aux cas spatio-temporels de dengue, Zika et chikungunya à Rio de Janeiro.

De nombreux modèles existants ne peuvent s'ajuster, étudier, détecter ou prédire les changements brusques communs pour les cas spatio-temporels de maladies infectieuses. Cette thèse est donc une contribution importante à la modélisation spatio-temporelle des maladies infectieuses.

Table of contents

1 Introduction **1**

2 Literature review **8**

2.1 Markov Switching Models 8

2.1.1 General formulation 9

2.1.2 Bayesian inference 12

2.2 Coupled Markov Switching Models 15

2.2.1 General formulation 17

2.2.2 Bayesian inference 19

2.3 Zero-state Markov Switching Models and Zero-inflated Models 20

2.4 Epidemiology of Dengue, Zika and Chikungunya 23

2.5 Summary 25

3 Zero-state coupled Markov switching count models for spatio-temporal infectious disease spread **27**

3.1 Introduction 31

3.1.1 Motivating example: dengue fever in Rio de Janeiro 34

3.2 A Zero-state Coupled Markov Switching Negative Binomial Model 36

3.3 Inferential Procedure 40

3.3.1 Temporal prediction 43

3.3.2 Fitted values . . . 44
3.4 Analysis of the Dengue Fever Data in Rio de Janeiro . . . 46
3.4.1 Model specification and fitting . . . 46
3.4.2 Results . . . 50
3.4.3 Fitted values and predictions . . . 56
3.5 Concluding Remarks . . . 60

4 A three-state coupled Markov switching model for COVID-19 outbreaks across Quebec based on hospital admissions 64

4.1 Introduction . . . 69
4.2 A Three-state Coupled Markov Switching Model . . . 75
4.2.1 Identifiability constraints . . . 78
4.2.2 Prior specification . . . 79
4.2.3 Clone states . . . 80
4.3 Inferential Procedure . . . 82
4.3.1 Outbreak detection, forecasting and historical retrospection . . . 84
4.3.2 Model comparison . . . 85
4.4 Simulation Study . . . 86
4.5 Application to COVID-19 Outbreaks Across Quebec . . . 88
4.5.1 Model specification and fitting . . . 88
4.5.2 Results . . . 93
4.5.3 Retrospective evaluation and comparison . . . 97
4.5.4 Real-time evaluation and comparison . . . 99
4.6 Discussion . . . 101

5 Markov switching zero-inflated space-time multinomial models for comparing multiple infectious diseases 106

5.1 Introduction . . . 111

5.2 An Autoregressive Multinomial Model for Comparing Transmission Dynamics 115
5.2.1 Incorporating zero-inflation . . . 119
5.3 Inferential Procedure . . . 122
5.3.1 Fitted values . . . 125
5.4 Application to Counts of Dengue, Zika and Chikungunya in Rio . . . 127
5.4.1 Model specification and fitting . . . 127
5.4.2 Results . . . 131
5.4.3 Model fit . . . 136
5.4.4 Sensitivity analysis to control for differences in the susceptible populations . . . 139
5.5 Discussion . . . 140

6 Conclusion **143**
6.1 Summary . . . 143
6.2 Potential Applications . . . 147
6.3 Limitations and Avenues for Future Work . . . 150

Appendices **155**

A Appendix to Manuscript 1 **156**
A.1 The Blocked Forward Filtering Backward Sampling (bFFBS) Algorithm . . . 156
A.1.1 Validating the algorithm . . . 159
A.2 Simulation Study . . . 159
A.3 Monte Carlo Approximations for the Posterior Predictive Distributions . . . 163
A.4 Pearson Residuals . . . 167

B Appendix to Manuscript 2 **169**
B.1 The Individual Forward Filtering Backward Sampling (iFFBS) Algorithm . . . 169
B.1.1 Validating the algorithm . . . 173

B.2 Simulation Study to Assess Parameter Recovery . . . 174
B.3 Widely Applicable Information Criterion (WAIC) . . . 180
B.3.1 Formulation . . . 180
B.3.2 Simulation study . . . 181
B.4 Endemic-epidemic Model . . . 184
B.5 Temporal Predictions . . . 188
B.6 Further Results from the Simulation Study to Quantify and Compare State Estimation (Section 4.4 of the Main Text) . . . 191
B.6.1 Simulated counts from 2 areas (SM Figure B.4) . . . 191
B.6.2 Primary results from the simulation study (SM Table B.4) . . . 192
B.7 Further Results for the Application to COVID-19 Outbreaks Across Quebec 192
B.7.1 Difference in the posterior probability that an outbreak is currently happening between the Full Coupled and Non-coupled models averaged across all hospitals (SM Figure B.5) . . . 192
B.7.2 Map of the likely state in each catchment area during the last week of the study period (SM Figure B.6) . . . 193
B.7.3 Analysis of real-time false alarms . . . 193

C Appendix to Manuscript 3 **197**

C.1 Reed-Frost Derivation of the ARMN Model . . . 197
C.1.1 Extensions to account for disease interactions and geographical spread 201
C.1.2 The difference between multinomial and multivariate Poisson approaches 202
C.2 Simulation Study to Investigate Correlations Induced by the Random Effects 203
C.3 Forward Filtering Backward Sampling (FFBS) Algorithm . . . 205
C.4 Widely Applicable Information Criterion (WAIC) . . . 208
C.5 Estimates From the Comparison Models . . . 208
C.6 Map of Temperature in Rio . . . 212

References 213

List of Tables

3.1 Model comparison using WAIC for models fitted to the dengue data. Note, $\emptyset$ refers to no covariates, or a homogeneous model of disease spread, that its, $\phi_{01,j\to i}^{(t-1)\to t} = \zeta_0^{(c)}$ or $\phi_{11,j\to i}^{(t-1)\to t} = \eta_0^{(c)}$. The bolded model is our final ZS-CMSNB model as we do not accept models that are more complex and do not improve the WAIC substantially (by 10 units). NP = Neighboring prevalence $= \log(y_{j(t-1)}/pop_j + 1)$, G = Gravity term $= \log(pop_i \times pop_j)$, LC = Local cases $= \log(y_{i(t-1)} + 1)$, NN = Number of neighbors $= |NE(j)|$. 48

3.2 Posterior means and 95% posterior credible intervals (in squared brackets) for the estimated parameters from the Markov chain part of the fitted ZS-CMSNB model. Intercepts shifted so they represent the probabilities of dengue reemergence and persistence with 0 neighbors infected, pop=10,000 and all other covariates fixed at mean or median (if very skewed) values. Other rows represent odds ratios, i.e. $\exp(\zeta_k)$ and $\exp(\eta_k)$ for k = pop=1,..., = $\log(y_{i(t-1)} + 1) = K$. The reemergence column of the Neighborhood Presence row shows $\exp(\phi_{01,j\to i}^{(t-1)\to t})$ calculated at median values of neighboring pop/prevalence and pop=10,000 for the home area, the persistence column shows $\exp(\eta_0^{(c)})$. The rows beneath the Neighborhood Presence row show, respectively, unmodified estimates of $\zeta_0^{(c)}/\eta_0^{(c)}$, $\zeta_1^{(c)}$ and $\zeta_2^{(c)}$. 53

3.3 Posterior means and 95% posterior credible intervals (in squared brackets) from the negative binomial part of the fitted ZS-CMSNB model. $\bar{r} = 1/N \sum_{i=1}^{N} r_i$ and $\sigma_r = \sqrt{1/(N-1) \sum_{i=1}^{N} (r_i - \bar{r})^2}$ 56

4.1 Shows the WAIC of the 4 considered models from Section 5.1 fitted to the Quebec hospitalizations. The best fitting model, the one with the lowest WAIC, is bolded. 92

4.2 Posterior means and 95% posterior credible intervals (in parentheses) for the estimated parameters from the Markov chain part of the fitted CMSNB(1,2,4) model. The intercept row shows the transition probabilities at an average level of mobility and beds, no new variant and assuming no neighboring outbreaks. Other rows show the (relative) odds ratios of the corresponding covariate. Units are given in parentheses after the covariates, for reference, .56 is the average weight ω_{ji} between two neighboring areas. The units for beds and mobility are equal to one standard deviation. Odds ratios whose 95% posterior credible intervals do not contain 0 are bolded. 94

4.3 Posterior means and 95% posterior credible intervals (in parentheses) from the count part of the fitted CMSNB(1,2,4) model. The intercepts and covariate effects are exponentiated so that they represent rates and rate ratios. Rate ratios whose 95% posterior credible intervals do not contain 0 are bolded. The units for beds and mobility are equal to one standard deviation. 96

5.1 Shows the WAIC of the 4 considered models from Section 5.4.1 fitted to the Rio data. The best fitting model, the one with the lowest WAIC, is bolded. . 131

5.2 Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted MS-ZIARMN model. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Section 5.4.1 for an explanation of the covariates. 132

5.3 Posterior means and 95% posterior credible intervals (in parentheses) for the estimated parameters from the Markov chain part of the fitted MS-ZIARMN model. The intercept row shows the probability of Zika presence and the logit of the probability of chikungunya presence, assuming no previous disease presence in the area, no cases in neighboring areas and at average values of the other covariates. The Zika column shows the ratio of $p_{2it}/(1 - p_{2it})$ (odds of Zika presence) corresponding to a unit increase in the covariate. The chik. column shows the untransformed estimates (not exponentiated, see text). All covariates are standardized except the neighboring prevalences which are given in units of 1 case per 25,000 (average combined size of neighboring areas) at small prevalences. 135

B.1 Shows the results of model comparison using the WAIC for 24 replications of the Non-spatial Model and 24 replications of the Spatial Model. Each row represents the true model from which the 24 replications were produced. . . 183

B.2 Shows the WAIC of Endemic-epidemic models with different maximum lags p. The best fitting model, the one with the lowest WAIC, is bolded. 186

B.3 Posterior means and 95% posterior credible intervals (in parenthesis) from the baseline and epidemic components of the fitted Endemic-epidemic model. The intercepts and covariate effects are exponentiated so that they represent rates and rate ratios. Rate ratios whose 95% posterior credible intervals do not contain 0 are bolded. The units for beds and mobility are equal to one standard deviation. 187

B.4 Shows the area under the ROC curve (AUC), sensitivity, specificity and timeliness (last 3 calculated with a 50 percent threshold) of three outbreak state estimates: (a) retrospective probabilities, $P(S_{it} = 3|\boldsymbol{y})$ from fitting the models to the full simulated data set (b) real-time detection probabilities, $P(S_{iT} = 3|\boldsymbol{y})$ from fitting the models up to time T for $T = 100, \ldots, 120$ (20 separate fits) and (c) real-time one week ahead forecasts, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$ from fitting the models up to time $T-1$ for $T = 101, \ldots, 120$. All outbreak state estimates were evaluated on the simulated data set described in Section 4.4 of the main text. The best criteria for each outbreak state estimate are bolded. 192

C.1 Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted Zeng (2022) and ARMN models. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Section 5.4.1 of the main text for an explanation of the covariates. 209

C.2 Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted ZIAMRN model. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Section 5.4.1 of the main text for an explanation of the covariates. 210

C.3 Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted sensitivity analysis model from Section 5.4.4 of the main text. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Sections 5.4.1 and 5.4.4 of the main text for an explanation of the covariates. 211

List of Figures

1.1 (a) Shows bi-weekly Zika cases in the northern Rio de Janeiro neighborhood of Higienópolis in Brazil, between 2015 and 2016. (b) Shows weekly COVID-19 hospitalizations in Hôpital Anna-Laberge, Quebec, Canada, between May 2021 and May 2022. The red arrows illustrate examples of the kinds of shifts in the behavior of the counts that we are interested in modeling in this thesis. 2

2.1 Shows simulated time series from models (2.1) (a), (2.2) (b) and (2.3) (c). . 12

2.2 Shows simulated time series from models (2.7) (a) and (2.8) (b) without covariates. 22

3.1 Monthly cases of dengue fever for 2 districts. Summer/winter seasons highlighted in red/light blue. In (b), the 3 red arrows point to the only 3 months with 0 cases in this district during the study period. 35

3.2 (top) Average monthly probability of dengue reemergence based on reported cases, i.e., $\frac{\#(\text{cases}=0\rightarrow\text{cases}>0)}{\#(\text{cases}=0\rightarrow\text{cases}>0)+\#(\text{cases}=0\rightarrow\text{cases}=0)}$. (bottom) Average monthly probability of dengue persistence based on reported cases, i.e., $\frac{\#(\text{cases}>0\rightarrow\text{cases}>0)}{\#(\text{cases}>0\rightarrow\text{cases}>0)+\#(\text{cases}>0\rightarrow\text{cases}=0)}$. 36

3.3 Posterior means (solid lines) and 95% posterior credible intervals (dashed lines) of (top) $\exp\left(\zeta_0^{(c)}+\zeta_1^{(c)}\log(.08+1)+\zeta_2^{(c)}\log(10\times pop_j)\right)$ versus pop_j and (bottom) $\exp\left(\zeta_0^{(c)}+\zeta_1^{(c)}\log(cases_j/26+1)+\zeta_2^{(c)}\log(10\times 26)\right)$ versus $cases_j$ (see (3.17)), where, 10 is the population of a small area, 26 is the median population and .08 (per 1000) is the median prevalence of dengue. 51

3.4 Posterior means (solid lines) and 95% posterior credible intervals (dashed lines) of the estimated monthly probabilities of dengue reemergence and persistence in a small district (pop=10,000). We assumed average monthly temperature and rainfall values, and 2 cases (median) reported in the previous month for the persistence probabilities. 54

3.5 Posterior probability that dengue is present for December 2017 (time T). Districts that have reported cases are in grey to distinguish them from districts where dengue may be undetected. A blue arrow is drawn from $j \in NE(i)$ to i if $\frac{\Omega(S_{i(T+1)}=1|S_{jT=1})}{\Omega(S_{i(T+1)}=1|S_{jT=0})}$ is greater than 1.2 with probability .75, *a posteriori*, where $\Omega(A) = P(A)/(1-P(A))$. 57

3.6 Posterior summaries of the coupled one month ahead fitted values for 4 districts. (top graphs) Coupled one month ahead fitted values of cases versus observed cases. (bottom graphs) Coupled one month ahead fitted values of dengue presence risk versus observed presence (0=0 reported cases which may not correspond to the actual absence of the disease). Posterior means (solid lines), 95% posterior credible intervals (dashed lines) and observed (points). Summer/winter seasons highlighted in red/light blue in the bottom graphs. 58

3.7 Summaries of the 1-12 month ahead posterior predictive distributions of the cases (top graphs) and dengue presence risk (bottom graphs) for 2 districts. Posterior predictive means (solid lines) and 95% posterior predictive credible intervals (dashed lines). Solid circles are from the last 7 years (2011-2017) used to fit the model and open circles are future observed values. 59

4.1 The left graph shows a map of the part of Quebec where the 30 hospitals (solid circles) included in the study are located. Borders separate counties. Each line in the right graph gives the number of hospitalizations in one of the 30 hospitals included in the study. 70

4.2 (a) An illustration of endemic and outbreak periods in the weekly COVID-19 hospitalizations for Pavillon Sainte-Marie in Trois–Rivières. (b) An illustration of absence, endemic and outbreak periods in the weekly COVID-19 hospitalizations for Hôpital Anna–Laberge in Châteauguay. The vertical red lines are drawn at the introduction of the Alpha and Omicron variants for all of Quebec. The "?" reflects the fact that the periods are not fully observable in our framework. 71

4.3 Top graph shows simulated counts from the last 20 weeks in one of the areas from the simulation study. Bottom graph solid lines show the real-time outbreak detection probabilities, $P(S_{iT} = 3|\boldsymbol{y})$, versus T. Bottom graph dashed lines show the one-week ahead outbreak forecasts from the previous week, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$, versus T. The red dotted line indicates the exact predetermined start time of outbreak 4 in the area. The Spatial Model is in orange and the Non-spatial Model is in purple. 87

4.4 Posterior means (solid lines) and 95% posterior credible intervals (dashed lines) of some of the estimated transition probabilities versus some of the covariates. Other covariates were fixed at either their average values, for beds and mobility, or at 0, for new variant and the weighted neighborhood outbreak sum. 95

4.5 Top graphs show the posterior means (solid lines) and 95% posterior credible intervals (shaded areas) of the endemic state distribution (blue) and the outbreak state distribution (red) versus the observed hospitalizations (points). The bottom graphs show the posterior probability that the disease was in the absence state (green line), the endemic state (blue line) and the outbreak state (red line) during each week of the study period, that is, $P(S_{it} = s|\boldsymbol{y})$ for $t = 1, \ldots, T$ and $s = 1$ (absence) in green, $s = 2$ (endemic) in blue and $s = 3$ (outbreak) in red, see Section 4.3.1. 98

4.6 Top graphs show the observed hospitalizations for the last 30 weeks of the study period where we conducted the real-time evaluation. Bottom graphs solid lines show the posterior probabilities that an outbreak is currently happening, that is, $P(S_{iT} = 3|\boldsymbol{y})$, versus T. Bottom graphs dashed line shows the one-week ahead outbreak forecasts from the previous week, that is, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$, versus T. The Full Coupled Model is in orange and the Non-coupled Model is in purple. The dotted red lines are drawn at the introduction of the Omicron variant for all of Quebec. 100

5.1 Bi-weekly reported case counts of dengue (a), Zika (b), and chikungunya (c) in Praça Seca, a west neighborhood of Rio de Janeiro, Brazil, between 2015-2016. Summer/winter seasons highlighted in red/light blue. 113

5.2 Shows the posterior means (black solid line) and 95% posterior credible intervals (black dashed lines) of λ_{kit} versus temperature for $k = 2$ (Zika) (left) and $k = 3$ (chikungunya) (right). We assumed an average area, other covariates fixed at their average values and a null space-time random effect. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. The red horizontal dashed lines are drawn at temperatures below and above which λ_{kit} is significantly different from 1. . . 133

5.3 Shows the posterior mean of $\bar{\lambda}_{ki} = \left(1/\sum_{t=2}^{T} S_{kit}\right) \sum_{t=2}^{T} \lambda_{kit} S_{kit}$ for $k = 2$ (Zika) (top map) and $k = 3$ (chik.) (bottom map). If $\bar{\lambda}_{ki} > 1$ ($\bar{\lambda}_{ki} < 1$) then the transmission of disease k was favored (disfavored) on average over dengue when it was present in the neighborhood, and the share of disease k tended to grow (shrink) relative to dengue in the neighborhood. 137

5.4 Top graphs show the posterior means (solid lines) and 95% posterior credible intervals (shaded areas) of the fitted values y'_{kit}, see Section 5.3.1. The bottom graphs show the posterior probability the disease was present in each bi-week, $P(S_{kit} = 1|\boldsymbol{y})$ versus t. "MS-ZIARMN no random effect" refers to the MS-ZIARMN model fit without including ϕ_{kit} from Equations (5.3)-(5.4). 138

A.1 Comparison of the posterior means of $\boldsymbol{S}$ produced by the binary and bFFBS2 samplers. 160

A.2 Averages and 95% quantiles for the sampling distributions (100 replications) of the posterior means from a simulation study with 50 percent and 80 percent 0s. Horizontal lines drawn at true parameter values. 162

A.3 ACF of the Pearson residuals for the 4 districts from Figure 3.6 of the main text. 168

B.1 Comparison of the posterior means of $\boldsymbol{S}$ produced by the one-at-a-time and iFFBS samplers when applied to the No Absence/Clone State Model from Section 4.5 of the main text. $S_{it} = 0$ indicates the endemic state and $S_{it} = 1$ indicates the outbreak state. 173

B.2 Shows simulated hospitalizations in 2 areas from a single replication of (B.5). 175

B.3 Shows the sample mean (circles) and 95% quantile (caps) of the posterior medians from fitting 237 replications of (B.5) using the strong constraints (SC) (see text around figure) and an additional 284 replications of (B.5) using the weak constraints (WC), with our hybrid Gibbs sampling algorithm. The horizontal lines are drawn at the true parameter values. 179

B.4 Shows simulated counts in 2 areas from the simulation study described in Section 4.4 of the main text. The red dotted lines indicate the exact pre-determined start time of each outbreak in the area. 191

B.5 (a)(top) Each line gives the number of hospitalizations in one of the 30 hospitals included in the study. (a)(bottom) shows the difference in the posterior probability that an outbreak is currently happening between the Full Coupled and Non-coupled models averaged across all hospitals, that is, $1/30 \sum_{i=1}^{30} P(S_{iT} = 3|\boldsymbol{y}, \text{Full Coupled Model}) - P(S_{iT} = 3|\boldsymbol{y}, \text{Non-coupled Model})$, versus T. (b) and (c) show posterior means (solid lines) and 95% posterior credible intervals (shaded areas) of the probabilities of outbreak emergence, (b), and outbreak persistence, (c), versus the number of outbreaks in neighboring areas assuming average connectivity and other covariates fixed at their average values. For (b) and (c), the Full Coupled Model is in orange, the Non-coupled Model is in purple, and the models were fit up to $T = 84$. . . . 193

B.6 A map of the part of Quebec where the 30 hospitals (points) included in the study are located. Borders separate counties. The color of the points represents the likely state in the catchment area of the hospital during the last week of the study period, that is, red if $P(S_{iT} = 3|\boldsymbol{y}) > .5$, blue if $P(S_{iT} = 2|\boldsymbol{y}) > .5$ and green if $P(S_{iT} = 1|\boldsymbol{y}) > .5$, where $T = 113 =$ 2022-05-09. From the Full Coupled Model with new variant as a covariate. 194

B.7 Top graphs show the hospitalizations for the last 30 weeks of the study period where we conducted the real-time evaluation. Bottom graphs: solid lines show the posterior probabilities that an outbreak is currently happening, that is, $P(S_{iT} = 3|\boldsymbol{y})$, versus T. The dotted red lines are drawn at the introduction of the Omicron variant for all of Quebec. Shows results from the Full Coupled Model for the two hospitals where we found evidence of false alarms during the real-time evaluation. 195

B.8 Shows the retrospective state estimates from the Full Coupled Model, without new variant as a covariate, for the two hospitals where we found evidence of false alarms during the real-time evaluation. Follows the same graphical structure as Figure 4.5 from the main text. 196

C.1 The black solid line shows corr$(y_2, y_3|\text{total})$ versus ρ, marginalizing the random effects, based on simulations of (C.9). The red horizontal dashed line is drawn at corr$(y_2, y_3|\text{total})$ assuming the random effects are fixed at 0, i.e., $\phi_2 = \phi_3 = 0$ in (C.9). The red vertical dashed line is drawn at the value of ρ estimated in Section 5.4 of the main text. 204

C.2 Shows the average weekly maximum temperature across Rio neighborhoods between 2015-2016. 212

Abbreviations

ACF autocorrelation function

AR autoregressive

ARMN autoregressive multinomial

AUC area under the receiver operating characteristic curve

bFFBS blocked forward filtering backward sampling

CMSNB coupled Markov switching negative binomial

COVID-19 coronavirus disease 2019

EE endemic-epidemic

FFBS forward filtering backward sampling

GNP gross national product

HMM hidden Markov model

iFFBS individual forward filtering backward sampling

MCMC Markov chain Monte Carlo

MS-ZIARMN Markov switching zero-inflated autoregressive multinomial

OR odds ratio

pCAR proper conditional autoregressive

SIR susceptible-infectious-recovered

SM supplementary material

VB variational Bayes

WAIC widely applicable information criterion

ZIARMN zero-inflated autoregressive multinomial

ZIB zero-inflated binomial

ZIC zero-inflated count

ZINB zero-inflated negative binomial

ZINBRE zero-inflated negative binomial random effect

ZIP zero-inflated Poisson

ZS-CMSNB zero-state coupled Markov switching negative binomial

ZS-CMSP zero-state coupled Markov switching Poisson

Chapter 1

Introduction

In this thesis, we are interested in modeling infectious disease counts collected across related areal units over time, for example, monthly cases of an infectious disease reported across the neighborhoods of a city. A challenge in the statistical modeling of these counts is that they often undergo abrupt and dramatic shifts in behavior due to different epidemiological events. For example, an outbreak could emerge in one of the areas causing cases to rise extremely rapidly, which may be difficult to capture for many standard statistical models (Cliff et al., 1987). Another example we are interested in is that an infectious disease, especially a vector-borne disease, can often go temporarily extinct in an area causing cases to drop to 0 for several weeks or months before potentially reemerging (Bartlett, 1957; Coutinho et al., 2006). The resulting large number of zeroes in the counts can be difficult to account for with many popular spatio-temporal Poisson or negative binomial regression models (Arab, 2015). Figure 1.1 shows examples of the types of shifts in behavior we are interested in.

In addition to accounting for the shifts, we also want to study the shifts and how they are related to certain factors, as this is often of considerable epidemiological interest. For instance, Walter et al. (2016) was interested in whether certain ecological covariates, such as distance to a major water body, were associated with the temporary extinction and

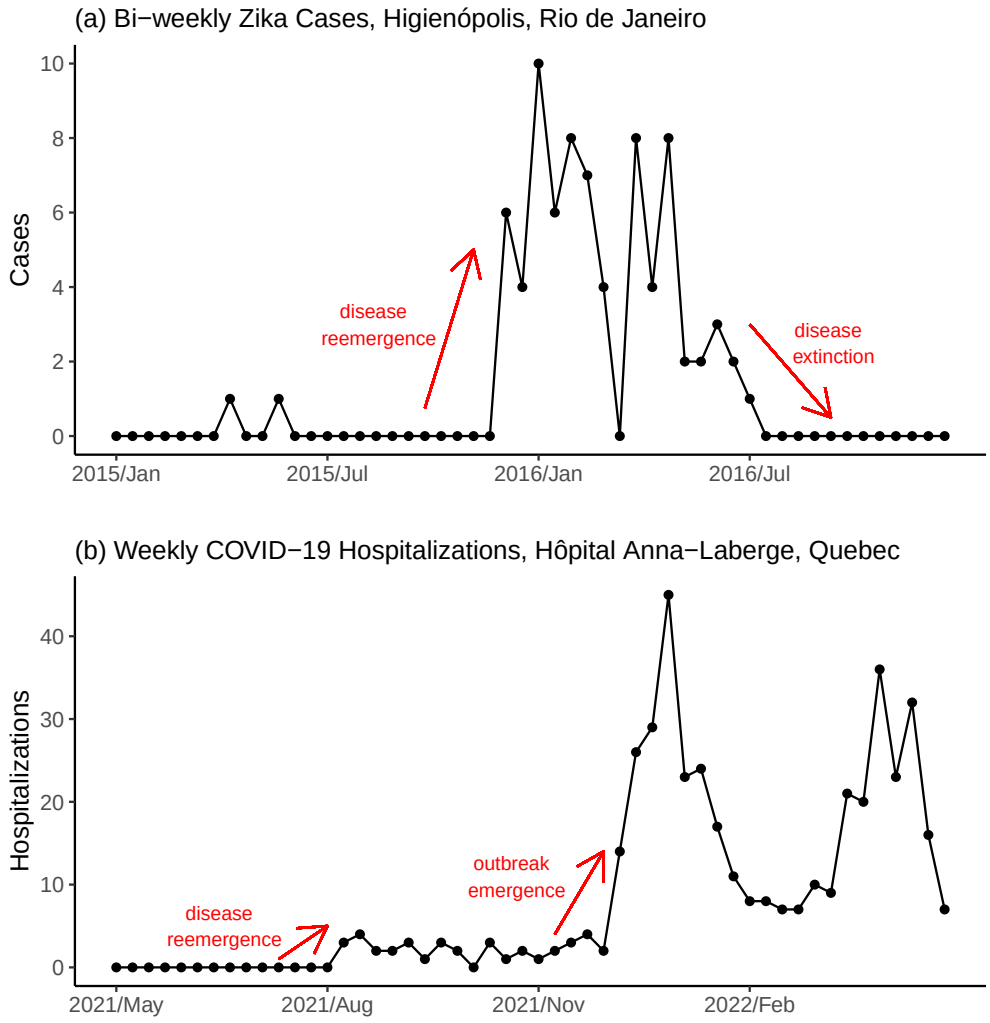


Figure 1.1: (a) Shows bi-weekly Zika cases in the northern Rio de Janeiro neighborhood of Higienópolis in Brazil, between 2015 and 2016. (b) Shows weekly COVID-19 hospitalizations in Hôpital Anna-Laberge, Quebec, Canada, between May 2021 and May 2022. The red arrows illustrate examples of the kinds of shifts in the behavior of the counts that we are interested in modeling in this thesis.

reemergence of Lyme disease in Connecticut towns. Descloux et al. (2012) was interested in how temperature and other climate factors were related to the emergence of dengue outbreaks in New Caledonia. In general, understanding factors contributing to outbreak emergence can help provide early warnings and design preventive measures (Lu et al., 2014). Additionally, we want to detect and forecast the shifts, mainly outbreak emergence. At the very beginning of an outbreak, it is not always clear an outbreak is occurring, and the timely detection of

infectious disease outbreaks is important for mounting an effective response (Buckeridge, 2007). By forecasting the shifts, we mean forecasting the epidemiological event associated with the shift, e.g., outbreak emergence, and not necessarily just the change in the counts. While forecasting an increase in the cases can be useful, if the increase is due to random endemic variation it will be much less concerning to health authorities than if it is due to an outbreak emerging. Providing additional information on the source of the increase, endemic versus epidemic, can help decide if serious measures need to be taken (Unkel et al., 2012).

There are many popular approaches to spatio-temporal infectious disease modeling, such as machine learning approaches (Rahimi et al., 2023) and spatio-temporal autoregressive models (i.e., where the conditional means are modeled as functions of past observations in the area and neighboring areas) (Bracher and Held, 2022). However, most of these approaches cannot be used to study, detect or forecast the epidemiological transitions that we are interested in, like outbreak emergence. For instance, while a machine learning model, like a long short-term memory (LSTM) network (Achterberg et al., 2022), could potentially forecast an abrupt increase in the counts, it cannot determine whether the increase is due to an outbreak emerging, as opposed to endemic variation, and what factors, like mobility, are contributing to it. A spatio-temporal autoregressive model could investigate associations between covariates and overall disease transmission (Ssentongo et al., 2021), however, this does not necessarily address questions about how the covariates are related to the shifts. For instance, smaller areas may experience less intense outbreaks compared to larger areas but in the same number. In this case, there is an association between population size and overall disease transmission but not with outbreak emergence. Finally, the above approaches cannot determine if an observed or forecasted increase in cases is due to endemic variation or an actual outbreak developing, which is an important problem in outbreak detection (Unkel et al., 2012). A more explicit approach is needed that directly models the switches in behavior.

In this thesis, we will take a Markov switching approach to account for, study, detect and forecast abrupt shifts in the behavior of spatio-temporal infectious disease counts. These shifts may be due to, for example, an outbreak emerging or a disease going temporarily extinct. Originally proposed by Souza (1982) (in an epidemiological context, see also Cliff et al. (1987)), this approach assumes the disease can be in one of several epidemiological periods or states in an area. For instance, Souza (1982) assumed the disease could be either in an endemic state, low stable incidence, or in an outbreak state (in this thesis we also consider absence and general presence states, see below). It is assumed the time series in an area follows a different epidemiologically appropriate submodel when the disease is in each state. For example, Conesa et al. (2015) assumed influenza case counts followed an autoregressive process with a high mean during outbreak periods and another autoregressive process with a much lower mean during endemic periods. Switching between the states is usually governed by a first-order Markov chain, whose transition probabilities can depend on covariates (Diebold et al., 1994; Nunes et al., 2013). This means we can investigate associations between covariates and abrupt shifts in the behavior of the counts due to important epidemiological transitions. For instance, in the first manuscript, we investigate how temperature and socioeconomic factors were associated with the reemergence (absence to presence) and persistence (presence to presence) of dengue fever in Rio de Janeiro neighborhoods. Understanding factors contributing to disease reemergence and persistence can help eliminate a disease from an area and prevent it from returning (Walter et al., 2016). Therefore, being able to study the association between important epidemiological transitions, like disease reemergence, and covariates is a useful feature of these models.

Another useful feature of Markov switching models is that they can approximate the posterior probability (i.e., the probability given all observed disease counts) that an observation came from any of the disease states (Frühwirth-Schnatter, 2006). Therefore, we can segment the time series into outbreak and non-outbreak periods (Martínez-Beneito et al., 2008). This can be useful for determining the timing of past outbreaks (Le Strat and Carrat, 1999) and,

more importantly, for real-time outbreak detection and forecasting (Nunes et al., 2013). For instance, we can issue a warning if a new observation likely came from the outbreak state (Martínez-Beneito et al., 2008). It is important to issue a warning for an outbreak early as the rapid rise in cases can be difficult to adjust to, often leading to medical shortages (Shingler and Hendry, 2022).

Finally, a further important use of Markov switching models is that state switching can account for changing behaviors in the time series that other statistical models might struggle to capture. For instance, the rapid rise in cases during the emergence of an outbreak, as illustrated in Figure 1.1. Autoregressive models whose parameters do not change across time, like log-linear autoregressive models (Liboschik et al., 2017) or Endemic-epidemic models (Bracher and Held, 2022), are very popular for modeling infectious disease counts (Dunbar and Held, 2020). These models can only capture such an abrupt increase in disease transmission if it can be explained by observable factors. However, important contributors to outbreak emergence are often difficult to observe directly, for instance, the migration of high-risk individuals (Lu et al., 2014). In an endemic/outbreak Markov switching model, an outbreak can occur at each time point with a certain probability, which allows outbreaks to emerge that are not fully explained by observable factors (Le Strat and Carrat, 1999), see also Figure 2.1. Past studies have found that endemic/outbreak Markov switching models fit infectious disease time series with many outbreaks better than autoregressive models, whose parameters do not change across time (Ansari et al., 2015; Rahmanian et al., 2021). Autoregressive random effects can also be used to allow the parameters of a time series model to change over time (Cargnoni et al., 1997). As shown by Otting et al. (2020), a time series model with an autoregressive random effect can be approximated by a Markov switching model with a large number of states. We prefer the states to represent well-recognized epidemiological periods, e.g., absence, endemic, or outbreak, as it leads to the transition probabilities and state estimates having meaningful important interpretations, as explained in the previous two paragraphs. Although, the two approaches could be combined, as we

briefly discuss in Section 6.3.

Spatio-temporal epidemiological Markov switching models have only been considered by a few authors (most approaches are purely temporal) (Heaton et al., 2012; Amorós et al., 2020). This thesis aims to address the following important limitations of this literature. Firstly, existing approaches have only considered switching between endemic and outbreak states (Amorós et al., 2020), and some transitional states (Lytras et al., 2019). We additionally consider an absence state in our models to account for the long strings of zeroes common in infectious disease counts, see Chen et al. (2019) and Figure 1.1 for an example. In a non-epidemiological context, Malyshkina and Mannering (2010) considered a spatio-temporal Markov switching model that switched with a state described by a degenerate zero distribution, like with our absence state. However, they assumed the states were independent between neighboring areas. Another important contribution of this thesis is that, following the coupled Markov switching literature (Touloupou et al., 2020), we allow the transition probabilities to depend on the states in neighboring areas. For instance, in the second manuscript, we allow the probability of an outbreak emerging in an area to depend on whether outbreaks are occurring in neighboring areas. This is important in our epidemiological context as the disease will spread between areas (Grenfell et al., 2001). Heaton et al. (2012) did let the probability of outbreak emergence depend on neighboring outbreak indicators. However, they used an absorbing state model and did not consider covariates in the transition probabilities. Therefore, their approach cannot analyze data with multiple outbreaks in each area and cannot investigate associations between covariates and outbreak emergence. The models explored in this thesis consider covariates in the transition probabilities and can be applied to data with any number of shifts in each area. Finally, we are not aware of any epidemiological Markov switching models that consider multiple interacting diseases. Comparing the transmission dynamics of multiple diseases is an important problem currently in epidemiology (Freitas et al., 2019). Such models are explored in the third manuscript (see below).

There are other important contributions of this thesis which are detailed in the preambles. Here we just focused on the most substantial ones. The rest of this thesis is organized as follows. Chapter 2 provides an introduction to the general domain that this thesis is interested in. Chapters 3, 4 and 5 each contain a stand-alone manuscript. Chapter 3 considers models where the disease switches between an absence state and a general presence state in each area (we introduce endemic and outbreak states in the second manuscript). Our approach is similar to finite mixture zero-inflated models (Young et al., 2020), and much of the manuscript is devoted to comparing the two. This manuscript was published in the Journal of the Royal Statistical Society Series C (Douwes-Schultz and Schmidt, 2022). Chapter 4 extends Chapter 3 to consider switching between absence, endemic and outbreak states in each area. We show, using simulated data and COVID-19 hospitalizations across Quebec, that including neighboring outbreak indicators in the transition probabilities greatly improves the accuracy and timeliness of outbreak detection and forecasting. This manuscript is undergoing a second round of reviews at the Annals of Applied Statistics. Chapter 5 also extends Chapter 3 to consider multiple diseases switching between periods of presence and absence in each area. We use the model to investigate how certain covariates, such as temperature, were related to differences in the transmission intensity of dengue, Zika and chikungunya in Rio de Janeiro. We plan on submitting this manuscript shortly to an appropriate applied statistics journal. We close the thesis with a summary and conclusion discussing limitations and future work in Chapter 6.

Chapter 2

Literature review

In the introductions of the individual manuscripts, Sections 3.1, 4.1 and 5.1, we give reviews of the literature that are more focused on providing rationale and context for why and how those papers were written. Here, we focus more on providing an introduction to the general domain that this thesis is concerned with. Starting with Section 2.1, we provide a general review of Markov switching models and Bayesian inference. In Section 2.2, we give a review of coupled Markov switching models, which are a type of Markov switching model for multiple interacting time series. Finally, we close with a review of zero-state Markov switching models and zero-inflated models in Section 2.3.

2.1 Markov Switching Models

Markov switching models assume a time series can be described by several submodels, usually called states, periods, or regimes, where switching between submodels is governed by a first-order Markov chain (Hamilton, 1989). For categorical time series, these models are usually called hidden Markov models (HMM) (Rabiner, 1989) and go back to the 1960s (Baum and Petrie, 1966) (we use Markov switching model and HMM interchangeably). For continuous

time series, Goldfeld and Quandt (1973) were the first to consider switching between linear regression models. In a similar vein, Wang and Puterman (1999) proposed switching between Poisson generalized linear models for count time series. Switching between different autoregressive models, i.e., where the conditional mean depends on past observations, was first considered by Poritz (1982) and greatly popularized by Hamilton (1989). Hamilton (1989) derived a filtering and smoothing algorithm for very general Markov switching models, see Section 2.1.1 below, that can be used to calculate the likelihood function and estimate the underlying state sequence. While Hamilton (1989) worked within a frequentist framework, Chib (1996) built on their filter to develop an efficient Bayesian inferential procedure based around jointly sampling the state sequence, see Section 2.1.2.

2.1.1 General formulation

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)^T$ be a multivariate time series, where $\boldsymbol{y}_t$ is p-dimensional (which includes the special case of a univariate time series). Suppose $\boldsymbol{S} = (S_0, \ldots, S_T)^T$ is an underlying state sequence with K possible states, i.e., $S_t \in \{1, \ldots, K\}$. Let $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ be vectors of temporal covariates. The main idea is that S_t follows a first-order Markov chain and determines the submodel that generates $\boldsymbol{y}_t$. To be more concrete, following Frühwirth-Schnatter (2006), a Markov switching model usually makes the following assumptions,

1.

$$p(\boldsymbol{y}_t | S_0, \ldots, S_t, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{\beta}) = p(\boldsymbol{y}_t | S_t, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. That is, $\boldsymbol{y}_t$ only depends on the current state S_t and, potentially, on past observations (autoregression) and covariates.

2.

$$p(S_t|S_0,\ldots,S_{t-1},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t-1},\boldsymbol{z}_t,\boldsymbol{\theta}) = p(S_t|S_{t-1},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t-1},\boldsymbol{z}_t,\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters. That is, S_t only depends on S_{t-1} and, potentially, on covariates and past observations. The probabilities $P(S_t = k|S_{t-1} = j, y_1,\ldots,y_{t-1},\boldsymbol{z}_t,\boldsymbol{\theta}) = p_{jk,t}$ for $j,k = 1,\ldots,K$ are known as the transition probabilities and are stored in the transition matrix $\Gamma(S_t)$, where $\Gamma(S_t)_{jk} = p_{jk,t}$. The transition probabilities must be specified so that the rows of the transition matrix sum to 1, see Diebold et al. (1994) and Spezia (2006) for some possible specifications with covariates. The transition probabilities can depend on past counts, as transitioning between some states may be more likely at certain levels of the counts. For instance, we found in the first manuscript that dengue is more likely to go extinct (presence to absence) when there is a small number of infectious individuals, see Table 3.2. If the transition probabilities do not vary with time, i.e., $p_{jk,t} = p_{jk}$ for all j,k, then the model is known as homogeneous otherwise nonhomogeneous.

3. The distribution of the initial state $p(S_0|\boldsymbol{\theta})$ can depend on $\boldsymbol{\theta}$.

The above assumptions are important for frequentist inference, as they are needed to run the forward filter of Hamilton (1989) to calculate the likelihood function. These assumptions are also required to run the forward filtering (i.e., Hamilton's filter) backward sampling (FFBS) algorithm for efficient Bayesian inference, see Section 2.1.2 below. If the conditions are not met the model can often be reparametrized (Hamilton, 1989).

Now we will give some examples. Firstly, for a count time series, we could consider the

two-state homogeneous Poisson HMM from Le Strat and Carrat (1999),

$$\begin{aligned} &y_t|S_t \sim \text{Poisson}(\lambda_{S_t}), \\ &P(S_t = 2|S_{t-1} = 1) = p_{12}, \qquad P(S_t = 1|S_{t-1} = 2) = p_{21}, \end{aligned} \tag{2.1}$$

where $S_t \in \{1, 2\}$ and $S_0 \sim \text{Bern}(.5) + 1$. Note, $p_{11} = 1 - p_{12}$ and $p_{22} = 1 - p_{21}$. Le Strat and Carrat (1999) used the model for monthly polio cases in the United States between 1970 and 1983. They assumed $S_t = 1$ indicated an endemic period (low stable incidence) and $S_t = 2$ indicated an outbreak of the disease. The unknown parameters are λ_2 (average number of cases during an outbreak month), λ_1 (average number of cases during an endemic month), p_{12} (probability of an outbreak emerging) and p_{21} (probability of an outbreak ending).

Following Zeger and Qaqish (1988), we could extend (2.1) to incorporate an autoregressive component,

$$y_t|S_t, y_{t-1} \sim \text{Poisson}\left(\beta_{S_t}\left(y_{t-1} + 1\right)^{\alpha_{S_t}}\right). \tag{2.2}$$

In (2.2), β mainly controls the mean level of the process while α controls the autocorrelation. Therefore, (2.2) could be appropriate for modeling an autoregressive process that sometimes experiences an abrupt large increase in its mean level. This may happen, for example, due to an outbreak emerging if modeling infectious disease counts. Conesa et al. (2015) applied a model like (2.2) to influenza cases to study and detect (see Section 2.1.2) switching between endemic and outbreak periods.

Finally, following Diebold et al. (1994), we could consider a nonhomogeneous extension of (2.1) where the frequency of switching states can change over time,

$$\text{logit}(p_{12,t}) = \theta_{12,0} + \theta_{12,1}t, \qquad \text{logit}(p_{21,t}) = \theta_{21,0} + \theta_{21,1}t. \tag{2.3}$$

Figure 2.1 shows simulated time series from models (2.1), (2.2) and (2.3), where $\lambda_2 = 3$,

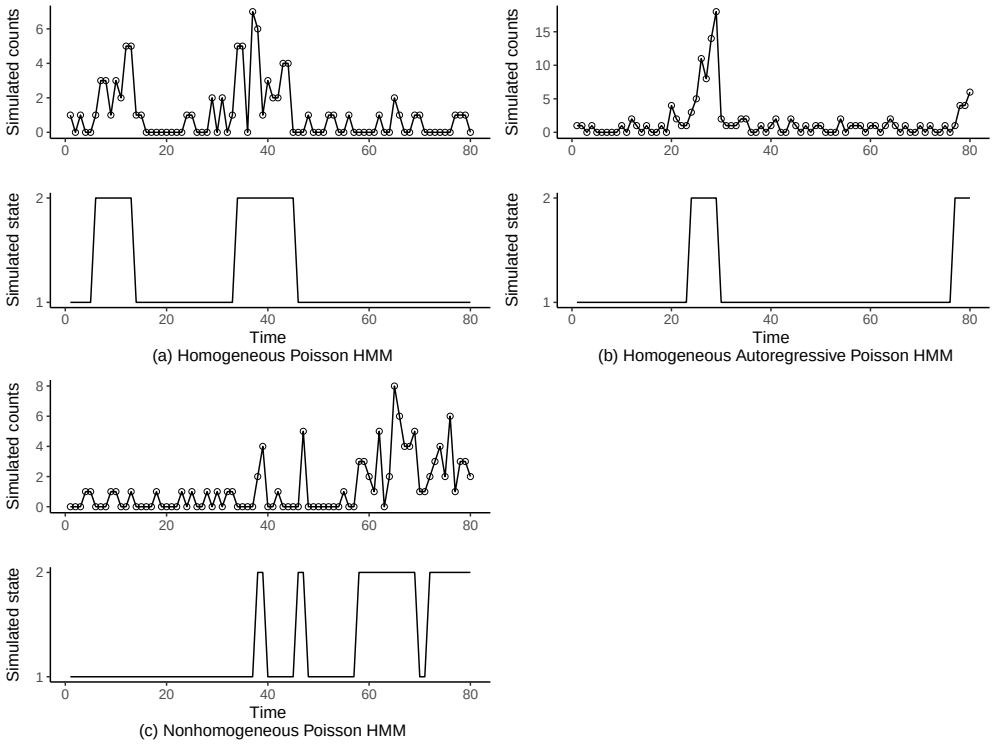


Figure 2.1: Shows simulated time series from models (2.1) (a), (2.2) (b) and (2.3) (c).

$\lambda_1 = .25$, $p_{12} = .05$, $p_{21} = .2$, $\beta_1 = .5$, $\beta_2 = 2$, $\alpha_1 = .5$, $\alpha_2 = .75$, $\theta_{12,0} = -3.5$, $\theta_{12,1} = .05$, $\theta_{21,0} = -1$ and $\theta_{21,1} = -.02$.

2.1.2 Bayesian inference

In this section, we review methods of efficient Bayesian inference for the class of Markov switching models discussed above. Let $\boldsymbol{v} = (\boldsymbol{\beta}, \boldsymbol{\theta})^T$ denote the vector of all model parameters. It is normally assumed that the time series $\boldsymbol{y}$ is observed and the state sequence $\boldsymbol{S}$ is unobserved. Therefore, there is often interest in not only estimating $\boldsymbol{v}$ but also $\boldsymbol{S}$. For example, for (2.1), Le Strat and Carrat (1999) were interested in whether an outbreak was currently happening, i.e., if $S_T = 2$, and when outbreaks had occurred in the past, i.e., when $S_t = 2$ for $t = 1, \ldots, T-1$.

Bayesian inference for a Markov switching model usually follows the Gibbs sampler proposed by Chib (1996), which samples from the joint posterior of $\boldsymbol{S}$ and $\boldsymbol{v}$, $p(\boldsymbol{S}, \boldsymbol{v}|\boldsymbol{y})$. The parameter

vector is first separated into blocks $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_D)^T$ of arbitrary size. Secondly, we specify initial values for the Gibbs sampler, $\boldsymbol{v}^{[0]} = (\boldsymbol{v}_1^{[0]}, \ldots, \boldsymbol{v}_D^{[0]})^T$ and $\boldsymbol{S}^{[0]}$. Then the following steps are repeated for $m = 1, \ldots, Q$, where Q is large enough to ensure convergence,

Step 1: Sample $\boldsymbol{v}_d^{[m]}$ from $p(\boldsymbol{v}_d|\boldsymbol{v}_1^{[m]}, \ldots, \boldsymbol{v}_{d-1}^{[m]}, \boldsymbol{v}_{d+1}^{[m-1]}, \ldots, \boldsymbol{v}_D^{[m-1]}, \boldsymbol{S}^{[m-1]}, \boldsymbol{y})$ for $d = 1, \ldots, D$.

Step 2: Sample $\boldsymbol{S}^{[m]}$ from $p(\boldsymbol{S}|\boldsymbol{v}^{[m]}, \boldsymbol{y})$ using the forward filtering backward sampling (FFBS) algorithm, see below.

After an initial burn-in period of size $M < Q$, the Gibbs sampler draws correlated samples from the joint posterior distribution, $(\boldsymbol{S}^{[m]}, \boldsymbol{v}^{[m]})^T \sim p(\boldsymbol{S}, \boldsymbol{v}|\boldsymbol{y})$. Typically, several iterations of the Gibbs sampler, called chains, are run starting from different random values in the parameter space, to avoid convergence to local modes. The chains are then combined and convergence is checked using the effective sample size and the Gelman-Rubin Statistic, which tries to determine if all chains converged to the same mode (Plummer et al., 2006).

For the first step, note that $p(\boldsymbol{v}_d|\boldsymbol{v}_{-d}, \boldsymbol{S}, \boldsymbol{y}) \propto p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{v})p(\boldsymbol{v})$. The distribution $p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{v})$ represents the likelihood of $\boldsymbol{v}$ given $\boldsymbol{y}$ and $\boldsymbol{S}$ and follows immediately from the assumptions in Section 2.1.1,

$$p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{v}) = p(S_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(\boldsymbol{y}_t|S_t, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{\beta}) p(S_t|S_{t-1}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{z}_t, \boldsymbol{\theta}). \tag{2.4}$$

The prior distribution $p(\boldsymbol{v})$ is set by the modeler and represents *a-priori* beliefs about the true parameter values. For general Bayesian inference, often independent flat priors are used (Reich and Ghosh, 2019). However, the likelihood of a Markov switching model can be highly irregular (containing multiple local modes) in regions of the parameter space that only use a subset of all states (Frühwirth-Schnatter, 2006). For example, for (2.1), where $\lambda_2 \approx \lambda_1$, $p_{12} \approx 0$ or $p_{21} \approx 0$. Therefore, Frühwirth-Schnatter (2006) chapter 4.2.2 recommends using priors that bound the parameters away from these regions, e.g., $p_{12}, p_{21} \overset{ind}{\sim} \text{Beta}(4, 4)$. The full conditional distribution $p(\boldsymbol{v}_d|\boldsymbol{v}_{-d}, \boldsymbol{S}, \boldsymbol{y})$ is proportional to whatever part of $p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{v})p(\boldsymbol{v})$

depends on $\boldsymbol{v}_d$. Once this kernel has been found numerous sampling algorithms can be used to draw from the full conditional, including conjugate, Metropolis-Hastings and Hamiltonian Monte Carlo samplers, see Chapter A.4 of Reich and Ghosh (2019) for a recent review.

For Step 2 we can sample from $p(\boldsymbol{S}|\boldsymbol{v}, \boldsymbol{y})$ using the FFBS algorithm proposed by Chib (1996). See Section C.3 for a step-by-step application of the algorithm applied to the third manuscript. The FFBS algorithm runs Hamilton's forward filter which requires matrix multiplication with the transition matrix $\Gamma(S_t)$. Therefore, it can not be run when the Markov chain has a very large number of states (this is relevant in Section 2.2 below). The FFBS algorithm requires the assumptions in Section 2.1.1 to hold (Frühwirth-Schnatter, 2006). If the assumptions do not hold then we can sometimes replace Step 2 with sampling from $p(S_t|S_0, \ldots, S_{t-1}, S_{t+1}, \ \ldots, S_T, \boldsymbol{v}, \boldsymbol{y})$ for $t = 0, \ldots, T$ (Frühwirth-Schnatter, 2006). This is usually less efficient but can be applied to a more general class of Markov switching models, for example, when y_t depends on $S_0, \ldots, S_t$ (Chen et al., 2019).

Once a sample from the joint posterior $p(\boldsymbol{S}, \boldsymbol{v}|\boldsymbol{y})$ has been obtained, we can approximate posterior means, medians and 95% credible intervals to summarize our estimation of $\boldsymbol{v}$. We can also estimate the values of the state sequence, $P(S_t = k|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} I[S_t^{[m]} = k]$ for $t = 0, \ldots, T$ and $k = 1, \ldots, K$, where $I[\bullet]$ is an indicator function. For example, for (2.1), $P(S_t = 2|\boldsymbol{y})$ represents the posterior probability an outbreak of polio occurred in the United States during month t.

In practice, there are several challenges to applying the above Gibbs sampler. Firstly, if at least two of the states have the same parametric form then label switching can occur (Stephens, 2000). As an example, consider fitting (2.1) to a time series. Assume there is a mode in the likelihood function around $\lambda_1 = 1$, $\lambda_2 = 2$, $p_{12} = .05$ and $p_{21} = .2$. Note the permuted parameter values $\lambda_1 = 2$, $\lambda_2 = 1$, $p_{12} = .2$ and $p_{21} = .05$ define the same model. In both cases, we switch between a Poisson distribution with mean 1 and mean 2 with the same probabilities of transitioning between them. Therefore, there will be a second mode

around the permuted parameter values. The Gibbs sampler can sometimes switch between the modes, known as label switching, which makes the posterior draws difficult to interpret (Frühwirth-Schnatter, 2006).

A popular method for dealing with the label-switching problem is to apply constraints to the parameter space that isolate one of the modes (Jasra et al., 2005). Continuing our earlier example of fitting (2.1), we could apply the constraint $\lambda_2 > \lambda_1$ to hopefully isolate the mode around $\lambda_2 = 2$ and $\lambda_1 = 1$. Arbitrary constraints can fail to remove multimodality from the posterior distribution, especially when the true values of the constrained parameters are very close together (Stephens, 2000; Jasra et al., 2005). Therefore, the constraints need to be chosen carefully based on knowledge of the application (Martínez-Beneito et al., 2008) or by visually examining draws from the unconstrained Gibbs sampler (Frühwirth-Schnatter, 2006).

Relabeling algorithms can also be used to address the label-switching problem (Stephens, 2000). However, they can be challenging to implement and do not remove genuine multimodality (extra modes unexplained by permuting the parameters) from the posterior distribution (Jasra et al., 2005). Genuine multimodality occurs quite often in complex mixture models (Stephens, 2000; Jasra et al., 2005). This is an important issue as Gibbs sampling usually cannot successfully sample from a multimodal distribution (Yao et al., 2022). If the extra modes occur in regions of the parameter space that are *a-priori* infeasible, then they can be removed with appropriate constraints, see Section B.2 from the supplementary materials of the second manuscript.

2.2 Coupled Markov Switching Models

A coupled Markov switching model is a type of Markov switching model for multiple time series, where the underlying state sequences for each time series can interact with one another (Pohle et al., 2021). We will use the early example of Phillips (1991) to illustrate some of

the ideas. Phillips (1991) was interested in modeling the time series of the gross national product (GNP) in two countries, $\boldsymbol{y}_{1(1:T)} = (y_{11}, \ldots, y_{1T})^T$ and $\boldsymbol{y}_{2(1:T)} = (y_{21}, \ldots, y_{2T})^T$. Following Hamilton (1989), they assumed that country i was either in a recession state during time t, indicated by $S_{it} = 1$, or an expansion state, indicated by $S_{it} = 2$, for $i = 1, 2$ and $t = 0, \ldots, T$. They assumed a higher mean level of GNP during the expansion state (modified slightly),

$$y_{it} = \mu_{S_{it},i} + \rho y_{i(t-1)} + \epsilon_{it}, \qquad \epsilon_{it} \sim N(0, \sigma^2),$$

where $\mu_{2,i} > \mu_{1,i}$ for $i = 1, 2$. It is reasonable to assume the state sequences, $\boldsymbol{S}_{1(0:T)} = (S_{10}, \ldots, S_{1T})^T$ and $\boldsymbol{S}_{2(0:T)} = (S_{20}, \ldots, S_{2T})^T$, interact in some way. For instance, one country going into a recession might push the other into a recession, or they might go into a recession simultaneously. To capture such possible interactions, Phillips (1991) modeled the transition matrix of the vector $\boldsymbol{S}_t = (S_{1t}, S_{2t})^T$,

$$\begin{array}{c} \\ (1,1) \\ (2,1) \\ (1,2) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (1,2) & (2,2) \end{array} \\ \begin{bmatrix} p_{11|11} & p_{21|11} & p_{12|11} & p_{22|11} \\ p_{11|21} & p_{21|21} & p_{12|21} & p_{22|21} \\ p_{11|12} & p_{21|12} & p_{12|12} & p_{22|12} \\ p_{11|22} & p_{21|22} & p_{12|22} & p_{22|22} \end{bmatrix} \end{array}, \tag{2.5}$$

Note that Equation (2.5) contains 12 free parameters as the rows of the transition matrix must sum to one. Phillips (1991) showed that (2.5) could capture many possible interactions between the state sequences including complete independence, perfect correlation, one country leading the other or both tending to move simultaneously.

Let $S_t^* \in \{1, \ldots, 4\}$ be an indicator for the possible values of $(S_{1t}, S_{2t})^T$. Assume S_t^* follows a Markov chain with a transition matrix given by (2.5). Then $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)^T$, where $\boldsymbol{y}_t = (y_{1t}, y_{2t})^T$, and $\boldsymbol{S}^* = (S_0^*, S_1^*, \ldots, S_T^*)^T$ follow the assumptions given in Section 2.1.1.

Therefore, the model of Phillips (1991) can be fit using the methods described in Section 2.1.2. Note if we were modeling N countries, the assumptions in 2.1.1 would still be satisfied, but the transition matrix would have dimensions $2^N X 2^N$. As such, if N were large, we would not be able to use the FFBS algorithm and other inference methods would have to be considered. We discuss this more below.

2.2.1 General formulation

Assume we have N time series, $\boldsymbol{y}_{i(1:T)} = (y_{i1}, \ldots, y_{iT})^T$ for $i = 1, \ldots, N$. Assume an underlying state sequence for each time series, $\boldsymbol{S}_{i(0:T)} = (S_{i0}, \ldots, S_{iT})^T$ for $i = 1, \ldots, N$, where $S_{it} \in \{1, \ldots, K\}$, that we would expect to interact. Let $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ be vectors of covariates. Finally, let $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})^T$ and $\boldsymbol{S}_t = (S_{1t}, \ldots, S_{Nt})^T$. To formulate a general class of coupled Markov switching models, we will mostly follow Touloupou et al. (2020) (though we incorporate autoregression) as their proposal allows for Bayesian inference when N is very large. When N is small, more general specifications can be used, see Pohle et al. (2021) for a recent review. Similar to Touloupou et al. (2020), we make the following assumptions,

1.

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{S}_0, \ldots, \boldsymbol{S}_t, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{\beta}) &= p(\boldsymbol{y}_t|\boldsymbol{S}_t, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} p(y_{it}|S_{it}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_{it}, \boldsymbol{\beta}), \end{aligned}$$

   where $\boldsymbol{\beta}$ is a vector of unknown parameters. That is, y_{it} only depends on S_{it} and, potentially, on covariates and past observations, and y_{it} is conditionally independent of y_{jt} $j \neq i$. Note marginal dependence between y_{it} and y_{jt} would be induced by the next assumption.

2.

$$p(\boldsymbol{S}_t|\boldsymbol{S}_0,\ldots,\boldsymbol{S}_{t-1},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t-1},\boldsymbol{z}_t,\boldsymbol{\theta}) = p(\boldsymbol{S}_t|\boldsymbol{S}_{t-1},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t-1},\boldsymbol{z}_t,\boldsymbol{\theta})$$
$$= \prod_{i=1}^{N} p(S_{it}|\boldsymbol{S}_{t-1},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t-1},\boldsymbol{z}_{it},\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters. That is, S_{it} can depend on the state of any time series from the previous time $\boldsymbol{S}_{t-1}$, covariates and past observations, and S_{it} is conditionally independent of S_{jt} $j \neq i$. The conditional independence assumption implies the relationship between the state sequences is lagged and that they do not move simultaneously (Pohle et al., 2021). This is often appropriate for modeling infectious disease spread (Touloupou et al., 2020), but could be relaxed if N is small (Pohle et al., 2021). The conditional transition probabilities are defined as $P(S_{it} = k|S_{i(t-1)} = j, \boldsymbol{S}_{(-i)(t-1)}, \boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t-1},\boldsymbol{z}_{it},\boldsymbol{\theta}) = p_{jk,it}$ for $j,k = 1,\ldots,K$, where $\boldsymbol{S}_{(-i)(t-1)} = (S_{1(t-1)},\ldots,S_{(i-1)(t-1)},S_{(i+1)(t-1)},\ldots,S_{N(t-1)})^T$. These are stored in the conditional transition matrix $\Gamma(S_{it}|\boldsymbol{S}_{(-i)(t-1)})$, which is K by K and the rows must sum to 1. Otranto (2005) proposed a convenient form for the conditional transition probabilities when $K = 2$,

$$\begin{aligned}\text{logit}(p_{12,it}) &= \theta_{12,0} + \sum_{j\neq i}\theta_{12,ji}I[S_{j(t-1)} = 2],\\ \text{logit}(p_{21,it}) &= \theta_{21,0} + \sum_{j\neq i}\theta_{21,ji}I[S_{j(t-1)} = 1],\end{aligned} \tag{2.6}$$

where $I[\bullet]$ is an indicator function. In (2.6), interactions between the Markov chains can be quantified through odds ratios. For example, $e^{\theta_{12,ji}}$ represents the multiplicative change in the odds of the Markov chain in area i transitioning from state 1 to state 2 associated with the Markov chain in area j transitioning from state 1 to state 2 during the previous time (holding the states of all other chains constant) (note the lagged dependence). Due to this convenience, we use the idea of (2.6) throughout the thesis.

3. The initial state distributions are conditionally independent, $p(\boldsymbol{S}_0|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(S_{i0}|\boldsymbol{\theta})$.

2.2.2 Bayesian inference

Let $S_t^* \in \{1, \ldots, K^N\}$ be an indicator for the possible values of $\boldsymbol{S}_t$. Then the assumptions in Section 2.2.1 imply the assumptions in Section 2.1.1 hold for $\boldsymbol{y}$ and $\boldsymbol{S}^*$. Therefore, the inference could follow Section 2.1.2. However, note the transition matrix $\Gamma(S_t^*)$ has dimensions $K^N X K^N$. As such, if N is very large, we would not be able to run the FFBS algorithm as it requires multiplication with the transition matrix.

Touloupou et al. (2020) proposed a modified version of the Gibbs sampler in Section 2.1.1 for coupled Markov switching models with large N,

Step 1: Sample $\boldsymbol{v}_d^{[m]}$ from $p(\boldsymbol{v}_d|\boldsymbol{v}_1^{[m]}, \ldots, \boldsymbol{v}_{d-1}^{[m]}, \boldsymbol{v}_{d+1}^{[m-1]}, \ldots, \boldsymbol{v}_D^{[m-1]}, \boldsymbol{S}^{[m-1]}, \boldsymbol{y})$ for $d = 1, \ldots, D$.

Step 2: Sample $\boldsymbol{S}_{i(0:T)}^{[m]}$ from $p(\boldsymbol{S}_{i(0:T)}|\boldsymbol{S}_{1(0:T)}^{[m]}, \ldots, \boldsymbol{S}_{(i-1)(0:T)}^{[m]}, \boldsymbol{S}_{(i+1)(0:T)}^{[m-1]}, \ldots, \boldsymbol{S}_{N(0:T)}^{[m-1]}, \boldsymbol{v}^{[m]}, \boldsymbol{y})$ for $i = 1, \ldots, N$, using the individual forward filtering backward sampling (iFFBS) algorithm (Touloupou et al., 2020).

That is, instead of sampling from $p(\boldsymbol{S}|\boldsymbol{v}, \boldsymbol{y})$ we sample from $p(\boldsymbol{S}_{i(0:T)}|\boldsymbol{S}_{(-i)(0:T)}, \boldsymbol{v}, \boldsymbol{y})$ for $i = 1, \ldots, N$. The iFFBS algorithm can be run under the assumptions in Section 2.1.1 (Touloupou et al., 2020). Due to the recency of this literature, it remains somewhat unclear the degree to which these assumptions could be relaxed. The iFFBS algorithm uses the conditional transition matrix $\Gamma(S_{it}|\boldsymbol{S}_{(-i)(t-1)})$ instead of $\Gamma(S_t^*)$. See Section B.1 for a step-by-step application of the algorithm applied to the second manuscript. Another possibility is to sample from $p(S_{it}|\boldsymbol{S}_{-(it)}, \boldsymbol{v}, \boldsymbol{y})$ for all i and t (Touloupou et al., 2020). However, this can be much less efficient, see Section 4.3 from the second manuscript for example.

For Step 1 we now have that, following the assumptions in Section 2.2.1,

$$p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{v}) = \prod_{i=1}^{N} p(S_{i0}|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_{it}|S_{it}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_{it}, \boldsymbol{\beta}) p(S_{it}|\boldsymbol{S}_{t-1}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{z}_{it}, \boldsymbol{\theta}).$$

For small N, Brand et al. (1997) introduced coupled HMMs for categorical time series, see Pohle et al. (2021) for a review of this literature. Phillips (1991), Otranto (2005) and Pohle et al. (2021) proposed models for continuous time series. For small N, inference usually follows the frequentist paradigm as the transition matrix is small enough to run Hamilton's filter to calculate the likelihood function.

For large N, Sherlock et al. (2013) and Touloupou et al. (2020) considered coupled HMMs for categorical time series. Billio et al. (2016) and Agudze et al. (2022) considered models for continuous time series in econometrics. Billio et al. (2016) derived a special case of the iFFBS algorithm to fit their model. Heaton et al. (2012) considered a coupled Markov switching model for count time series to model outbreak spread. We discuss this proposal extensively in the second manuscript in Chapter 4. Heaton et al. (2012) considered the above Gibbs sampler but used the FFBS algorithm in Step 2 with $\Gamma(S_{it}|\boldsymbol{S}_{(-i)(t-1)})$ instead of the iFFBS. As shown by Touloupou et al. (2020), this will not sample from the correct full conditional distribution $p(\boldsymbol{S}_{i(0:T)}|\boldsymbol{S}_{(-i)(0:T)}, \boldsymbol{v}, \boldsymbol{y})$. For large N, only Bayesian methods have been considered as the transition matrix is too large to run Hamilton's filter.

2.3 Zero-state Markov Switching Models and Zero-inflated Models

Zero-state Markov switching count models switch between a degenerate 0 state and a count state, such as a Poisson (Wang, 2001) or negative binomial (Malyshkina and Mannering, 2010) state. Wang (2001), motivated by count time series with many zeroes, introduced the

following Markov switching model (see Section 2.1),

$$y_t \mid S_t \sim \begin{cases} 0, & \text{if } S_t = 0 \\ \text{Poisson}(\lambda_t), & \text{if } S_t = 1, \end{cases} \tag{2.7}$$

$$P(S_t = 1|S_{t-1} = 0) = p_{01,t}. \qquad P(S_t = 1|S_{t-1} = 1) = p_{11,t},$$

where $S_t \in \{0, 1\}$. They linked λ_t, $p_{01,t}$ and $p_{11,t}$ to covariates through log and logistic link functions. Note $p_{00,t} = 1 - p_{01,t}$ and $p_{10,t} = 1 - p_{11,t}$. If we let $p_{01,t} = p_{11,t} = p_{1,t}$ (i.e., assume the previous state does not influence the current state) then the state sequence follows a series of independent Bernoulli distributions, and we arrive at the following model,

$$y_t \mid S_t \sim \begin{cases} 0, & \text{if } S_t = 0 \\ \text{Poisson}(\lambda_t), & \text{if } S_t = 1, \end{cases} \tag{2.8}$$

$$P(S_t = 1) = p_{1,t},$$

which is the well-known zero-inflated Poisson (ZIP) model of Lambert (1992).

Figure 2.2 shows simulated time series from (2.7) and (2.8) without covariates and with parameter values $\lambda = 3$, $p_{01} = .10$, $p_{11} = .90$ and $p_1 = .5$. For both time series, the zero-state (S_t=0) occurs approximately 50% of the time. However, the Markov switching model can produce far more consecutive zeroes, which are a common feature of count time series (Chen et al., 2019). The main advantage of (2.7) over (2.8) is that the Markov chain accounts for autocorrelation in the state sequence.

However, there are modifications of (2.8) that could also account for this. For example, following Otting et al. (2020) we could consider an autoregressive random effect to capture

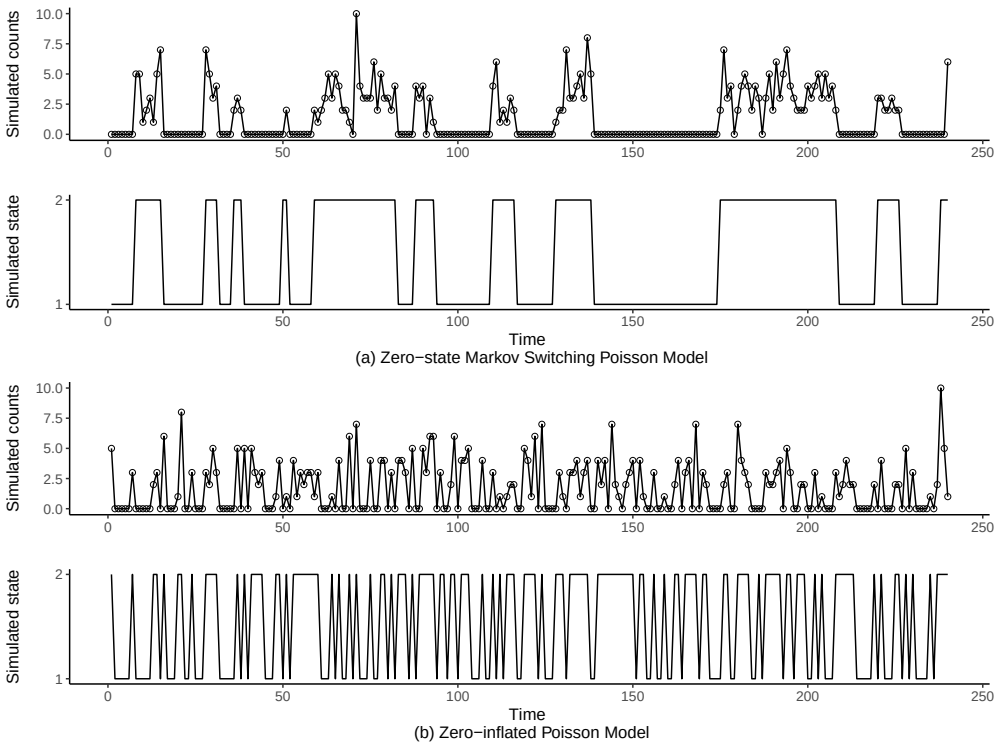


Figure 2.2: Shows simulated time series from models (2.7) (a) and (2.8) (b) without covariates.

serial correlation in the state sequence,

$$\begin{aligned} \text{logit}(p_{1,t}) &= \eta_0 + \boldsymbol{z}_t^T \boldsymbol{\eta} + \phi_t \\ \phi_t &= \rho \phi_{t-1} + \epsilon_t, \end{aligned} \tag{2.9}$$

where $\boldsymbol{z}_t$ is a vector of covariates, $\epsilon_t \sim N(0, \sigma^2)$ and $|\rho| < 1$ (to ensure ϕ_t is stationary). If $\rho = 0$ there is no autocorrelation in the state sequence while if $\rho > 0$ there is positive autocorrelation. In a space-time setting, Hoef and Jansen (2007) (see Section 3.4.1) considered a spatio-temporal ZIP model similar to (2.9), where they used random effects correlated across space and time to capture dependencies in the presence of the count state. See Young et al. (2020) for a recent review of zero-inflated models for temporal and spatio-temporal counts (we also review this literature extensively in the first manuscript in Chapter 3).

To better understand the differences between the approaches in (2.7) and (2.9), note we can

parametrize the transition probabilities in (2.7) as,

$$\text{logit}\left(P(S_t = 1|S_{t-1})\right) = \eta_0 + \eta_1 S_{t-1} + (1 - S_{t-1})\boldsymbol{z}_t^T \boldsymbol{\eta}_{01} + S_{t-1}\boldsymbol{z}_t^T \boldsymbol{\eta}_{11}. \tag{2.10}$$

Therefore, an advantage of (2.7) is that the covariates can have separate effects on the 0 to 1 transition, represented by $\boldsymbol{\eta}_{01}$ in (2.10), and the 1 to 1 transition, represented by $\boldsymbol{\eta}_{11}$. We found this was important in the first manuscript (Chapter 3) as different dynamics can govern the reemergence (absence to presence) and persistence (presence to presence) of an infectious disease. Much of the first manuscript is devoted to comparing spatio-temporal extensions of (2.8), like Hoef and Jansen (2007), with coupled Markov switching, (see Section 2.2), extensions of (2.7). Finally, we note that label switching (see Section 2.1.2) is not an issue for (2.7) and there is no need to constrain the parameters.

2.4 Epidemiology of Dengue, Zika and Chikungunya

Dengue, Zika and chikungunya are vector-borne diseases transmitted between humans and, mainly in Brazil, the *Aedes aegypti* mosquito (Schmidt et al., 2022). The epidemiology of these diseases is complex, and here we will just focus on information relevant to building models and interpreting results in the first and third manuscripts. Firstly, it is important to understand how these diseases are transmitted. Both humans and mosquitoes can be infected (Coutinho et al., 2006). Humans can be infected by being bitten by infectious mosquitoes and mosquitoes can be infected by biting infectious humans. Human-to-human transmission does not occur for either dengue or chikungunya. Zika can be transmitted from human to human through sexual contact, however, human-to-mosquito-to-human transmission is much more prevalent (Counotte et al., 2018).

Since dengue, Zika and chikungunya are primarily transmitted between humans and mosquitoes it is important to understand the biology of their shared vector, the *Aedes aegypti*. *Aedes*

aegypti require standing water to lay their eggs (Marquardt, 2004). Therefore, rainfall and, potentially, human-made water containers are important for the growth of the mosquito population. Schmidt et al. (2011) found an association between the use of water storage containers, which are often used in poorer areas, and dengue incidence in villages in Vietnam. Additionally, the mosquito population is highly dependent on temperature. *Aedes aegypti* generally live longer, lay more eggs, and mature faster at higher temperatures, though these relationships are non-linear and can reverse at very high temperatures (Chen and Hsieh, 2012). Also, the rate at which exposed mosquitoes become infectious with dengue increases with temperature. This is important since the mosquitoes only live around three weeks, depending on temperature, and, therefore, can easily die before being able to pass on the disease. Descloux et al. (2012) found a strong association between temperature and the occurrence of dengue outbreaks in New Caledonia. Since the life cycle of *Aedes aegypti* is highly dependent on temperature and rainfall, the mosquito population can drop to near 0 during the winter or dry seasons (Adams and Boots, 2010). As such, long strings of zeroes are a common feature in the time series of these diseases' case counts, see Figure 1.1 and Chen et al. (2019) for some examples. Finally, we note that there is often a delay between changes in the mosquito population, driven by temperature and rainfall, and changes in dengue incidence (Coutinho et al., 2006). This could be due to the mosquito population needing to reach a critical threshold before transmission is high enough for incidence to increase.

As much of the first manuscript is concerned with modeling the spread of dengue between neighborhoods in Rio de Janeiro, it is important to understand how the disease spreads geographically. Firstly, note that *Aedes aegypti* only travel short distances ($<$100m) (Harrington et al., 2005). Therefore, the long-distance spread of the diseases, e.g., between neighborhoods, is primarily driven by human movement. Either an infectious human can bring the disease to susceptible mosquitoes in a non-infected area, or a susceptible human can travel to an infected area, get bitten by infectious mosquitoes and bring the disease back to susceptible mosquitoes in a non-infected area (Stoddard et al., 2009). Since *Aedes aegypti* primarily

bite during the day, when most humans travel, human movement can spread dengue very quickly across geographical areas like cities (Stoddard et al., 2013). Stoddard et al. (2013) found that travel outside the home to high-risk areas was an important predictor of dengue infection.

Finally, we will provide a brief history of dengue, Zika and chikungunya in Rio de Janeiro. Dengue has been endemic in Rio de Janeiro since 1987. There are outbreaks of dengue every few years in the city on average during the summer months, see Figure 3.1 for example. The first cases of Zika were detected in the city in 2015 and in 2016 for chikungunya (Freitas et al., 2019). Between 2015-2016, there was a triple epidemic of the three diseases in Rio de Janeiro, which is the subject of the third manuscript in Chapter 5. Although dengue, Zika, and chikungunya are transmitted by the same vector, there can be important differences in their transmission. For instance, laboratory studies have shown that *Aedes aegypti* can transmit Zika more effectively at higher temperatures compared to dengue (Tesla et al., 2018). We will explore these differences in more detail in the third manuscript.

2.5 Summary

In this chapter, we introduced the general domain that this thesis is interested in. Firstly, we reviewed Markov switching models in Section 2.1. We showed how these models can account for sudden shifts in the behavior of a single epidemiological time series, due to, for example, an outbreak emerging. In this thesis, we are interested in accounting for these kinds of shifts in behavior across multiple interacting epidemiological time series, such as monthly reported cases of an infectious disease across the neighborhoods of a city. Related to this, in Section 2.2 we reviewed coupled Markov switching models for multiple time series which allow the underlying state sequences to interact. These models assume that the probabilities of transitioning between states for one time series can depend on the states of the other time series. For our models, the states represent different epidemiological periods, such as

disease absence or outbreak. Therefore, following the coupled Markov switching literature reviewed here, we allow the transition probabilities of our models to depend on the states in neighboring areas, to account for geographical disease spread (Smith et al., 2002). For instance, in the second manuscript in Chapter 4, we allow the probability of an outbreak emerging in an area to depend on whether outbreaks are occurring in neighboring areas. We also reviewed zero-state Markov switching and zero-inflated models in Section 2.3. We will use a zero-state in our models to represent the absence of a disease and account for the long strings of zeroes common in infectious disease counts (Chen et al., 2019). Finally, we gave a brief overview of the epidemiology of dengue, Zika, and chikungunya in Section 2.4. In the first manuscript in Chapter 3, we apply our zero-state coupled Markov switching model to counts of dengue cases in Rio de Janeiro neighborhoods. In the third manuscript in Chapter 5, we apply our Markov switching zero-inflated autoregressive multinomial model to cases of dengue, Zika, and chikungunya in Rio neighborhoods. Therefore, it is important to understand the epidemiology of these diseases. Again, we emphasize that this literature review was meant to be a general introduction to these topics. The introductions of the individual manuscripts go into much more specific details about how the existing literature relates to our proposed approaches (also see the preambles).

Chapter 3

Zero-state coupled Markov switching count models for spatio-temporal infectious disease spread

Preamble to Manuscript 1. In this first manuscript, we consider coupled Markov switching models where the disease switches between periods of presence and absence in each area. The absence state, also called the zero state, is described by a degenerate 0 distribution, while the presence state is described by an autoregressive negative binomial model. Non-coupled, meaning the transition probabilities do not depend on the states in neighboring areas, Markov switching models that switch between a zero state and a count state, e.g., Poisson state, (zero-state Markov switching count models) were considered by Wang (2001) and Malyshkina and Mannering (2010). However, these approaches are not appropriate for infectious disease counts as we would expect the presence of the disease in an area to affect its presence in neighboring areas, due to between-area mixing (Smith et al., 2002). For our application to dengue counts in Section 3.4, we found that the fit of our model generally improved as the coupling became more dynamic, see Table 3.1, indicating the importance of

accounting for complex dependencies between the Markov chains.

Zero-state Markov switching count models can be seen as Markov mixture versions of finite mixture zero-inflated count models, such as the popular ZIP regression model of Lambert (1992). Therefore, zero-inflated count models for space-time data, such as those explored in Hoef and Jansen (2007), are important competitors to our approach. As we stress in the manuscript, the main advantages of our approach are (1) we allow for each covariate to have a separate effect on the reemergence (absence to presence) and persistence (presence to presence) of the disease, which is often epidemiologically justified, and (2) we allow the effects of disease spread between neighboring areas to depend on space-time factors, such as the population of the areas (gravity effect (Tuite et al., 2011)). These advantages allow us to gain several interesting insights into the epidemiology of the disease beyond what can be gained from existing zero-inflated approaches, see the second paragraph of Section 3.5. Additionally, on the dengue counts we found our model fit better compared to a reasonably specified alternative finite mixture spatio-temporal zero-inflated model, see Section 3.4.1 and Table 3.1.

In conclusion, by extending the non-coupled zero-state Markov switching count models of Wang (2001) and Malyshkina and Mannering (2010), and by offering several important advantages over finite mixture zero-inflated approaches (Young et al., 2020), this manuscript provides an important contribution to the literature on modeling spatio-temporal counts that contain many zeroes (Arab, 2015).

Zero-state coupled Markov switching count models for spatio-temporal infectious disease spread

Dirk Douwes-Schultz[1], Alexandra M. Schmidt[1].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

This thesis contains the accepted version of the corresponding paper published in the *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (Douwes-Schultz and Schmidt (2022)).

Abstract

Spatio-temporal counts of infectious disease cases often contain an excess of zeros. With existing zero inflated count models applied to such data it is difficult to quantify space-time heterogeneity in the effects of disease spread between areas. Also, existing methods do not allow for separate dynamics to affect the reemergence and persistence of the disease. As an alternative, we develop a new zero-state coupled Markov switching negative binomial model, under which the disease switches between periods of presence and absence in each area through a series of partially hidden nonhomogeneous Markov chains coupled between neighboring locations. When the disease is present, an autoregressive negative binomial model generates the cases with a possible 0 representing the disease being undetected. Bayesian inference and prediction is illustrated using spatio-temporal counts of dengue fever cases in Rio de Janeiro, Brazil.

3.1 Introduction

In epidemiology, counts of infectious disease cases are being increasingly reported in several related areal units across time. A common issue encountered when modeling these counts is the presence of excess zeros (Arab, 2015). That is, there are often many more zeros in the counts than can be predicted by the usual Poisson and negative binomial count models. Lambert (1992) proposed the zero-inflated Poisson (ZIP) model to deal with excess zeros in count data, and this approach, with various extensions (see Young et al. (2020)), has been applied in many fields including epidemiology. We will refer to any model that incorporates zero-inflation into a count distribution as a zero-inflated count (ZIC) model (e.g. ZIP, ZINB, ZIGP, ZICMP models, etc. (Young et al., 2020)). In a disease mapping application of the ZIC model, the presence/absence of the disease is generated through a Bernoulli process and then, when the disease is present, the number of reported cases is generated through a count process, typically negative binomial or Poisson (Fernandes et al., 2009). A zero coming from the count process represents the disease being undetected, while a zero from the Bernoulli process represents the true absence of the disease. A ZIC model can help account for the excess zeros, through the additional Bernoulli zero generation, and also allows us to investigate factors related to the presence/absence of the disease while considering imperfect detection of disease presence (Vergne et al., 2016). ZIC models have become increasingly popular in epidemiology for modeling excess zeros in spatio-temporal infectious disease counts (Fernandes et al., 2009; Aktekin and Musal, 2015; Wangdi et al., 2018).

However, in a spatio-temporal setting, an important issue with existing ZIC approaches is that they only model the probability of overall disease presence. In a spatio-temporal setting we can separate the presence of a disease into two events of epidemiological interest, persistence (presence to presence) and reemergence (absence to presence). Some important epidemiological covariates can affect the reemergence and persistence of an infectious disease

quite differently. An example from our application of dengue fever is temperature, as temperature affects both the hatching rate of vertically infected mosquito eggs, which is more important for the reemergence of the disease, and the life cycle of the infected mosquitoes, which is more important for the persistence of the disease (Coutinho et al., 2006).

A second issue concerns the use of random effects in the Bernoulli process of a ZIC model, which can be used to account for spatio-temporal correlations in the presence of the disease (Hoef and Jansen, 2007; Torabi, 2017; Giorgi et al., 2018). Splines can also be used for this purpose (Ghosal et al., 2020), but are functionally similar so we will refer in general to random effect models. Random effect models are built by specifying relationships between probabilities of disease presence between areas and across time (Torabi, 2017). In contrast, it is more traditional in epidemiology to build models of infectious disease spread by specifying the probability that the disease will be present in an area given its actual presence in the area and neighboring areas previously (Keeling et al., 2001; Smith et al., 2002; Hooten and Wikle, 2010). That is, random effects models condition on the probability of disease presence in neighboring areas while epidemiological models usually condition on the actual presence of the disease in neighboring areas. If a goal of the analysis is to quantify associations between various space-time factors and the effects of disease spread between areas, like in Smith et al. (2002) who examined whether rivers block the spread of rabies, then the traditional epidemiological approach is more appropriate. This is because the disease can only spread from an area if it is present there, while the probability of disease presence does not tell you if the disease is spreading or not. For example, the disease could be present in a neighboring area and spread from there even if its probability of presence is low. Therefore, it is difficult to quantify space-time heterogeneity in the effects of disease spread between areas with existing ZIC models.

To help address these issues we propose a zero-state Markov switching approach to modeling the zero inflation. Markov switching models, first introduced by Goldfeld and Quandt (1973)

and extensively developed by Hamilton (1989), allow a time series to be described by several submodels (states) where switching between submodels is governed by a first-order Markov chain. The motivation behind the Markov switching model, over a finite mixture model, is that the states are often dependent across time and occur consecutively (Goldfeld and Quandt, 1973). Markov switching models that switch between a zero state and a count state were considered in a temporal setting by Wang (2001) and in a spatio-temporal setting by Malyshkina and Mannering (2010). However, Malyshkina and Mannering (2010) assumed the states were independent between spatial units and neither of these applications dealt with infectious disease data.

In our framework, following traditional disease mapping ZIC models (Fernandes et al., 2009), we assume the existence of a zero state, representing the absence of the disease, and a negative binomial state, representing the presence of the disease. Then we allow the disease to switch between the presence and absence state in each area through a nonhomogeneous (Diebold et al., 1994) Markov chain. The Markov chains allow for switching between long periods of disease presence and absence, and allow for each covariate to have a separate effect on the reemergence of the disease compared to the persistence. As the disease can spread between areas, we extend the zero-state Markov switching count model to a coupled Markov switching model (Pohle et al., 2021), by coupling the Markov chains between neighboring areas. In a coupled Markov switching model, the transition probabilities of a Markov chain can depend on the states of other Markov chains. We use a collection of coupling parameters to account for heterogeneity in the effects of neighboring disease spread across space and time. The coupling parameters allow us to quantify associations between the effects of neighboring disease spread and various space-time factors related to either area.

This paper is structured as follows. In Section 3.1.1 we introduce our motivating example of dengue fever cases in Rio de Janeiro and lay out the goals for our analysis. In Section 3.2 we introduce our proposed model, the zero-state coupled Markov switching negative binomial

model. In Section 3.3 we describe Bayesian inference using data augmentation through Markov chain Monte Carlo (MCMC) methods. Here we discuss efficient sampling strategies for the parameters and unknown state indicators. In Section 3.4 we apply the model to Rio's dengue fever data. Finally, we close with a general discussion in Section 3.5.

3.1.1 Motivating example: dengue fever in Rio de Janeiro

Dengue fever is endemic in Rio de Janeiro, Brazil, and there have been several major epidemics there since 1987. We focus on modeling monthly cases of dengue fever in the 160 districts of Rio de Janeiro between 2011-2017, as reported by the Health Secretary of the city (http://www.rio.rj.gov.br/web/sms/exibeconteudo?id=2815389). Figure 3.1 shows monthly dengue fever cases for two districts, a relatively small district (a) and a relatively large district (b). In the smaller district, there are both long periods of dengue presence (9 months longest period) and long periods of dengue absence (14 months longest period). Additionally, in the smaller district there is clearly a recurring seasonal pattern to the persistence and reemergence of the disease. The disease has a lower chance of persisting during the winter, where it often goes extinct, and then dengue often reemerges in the summer. This could be due to changes in rainfall and temperature, which often lead to large fluctuations in the mosquito population (Marquardt, 2004). Additionally, temperature and rainfall play an important role in the life cycle of vertically infected eggs, which are important for the reemergence of the disease (Coutinho et al., 2006). The same patterns in disease presence are not seen in the larger district in Figure 3.1 (b), where dengue persists for much longer on average compared to the smaller district, rarely goes extinct and reemerges quickly. Figure 3.2 similarly shows that there is significant spatial variation in the district level average, across time, probabilities of dengue reemergence and persistence (although these are based on reported cases and do not account for imperfect detection of disease presence). This could be explained by differences in population between districts (Bartlett, 1957), but also possibly differences in socioeconomic factors as mosquitoes often

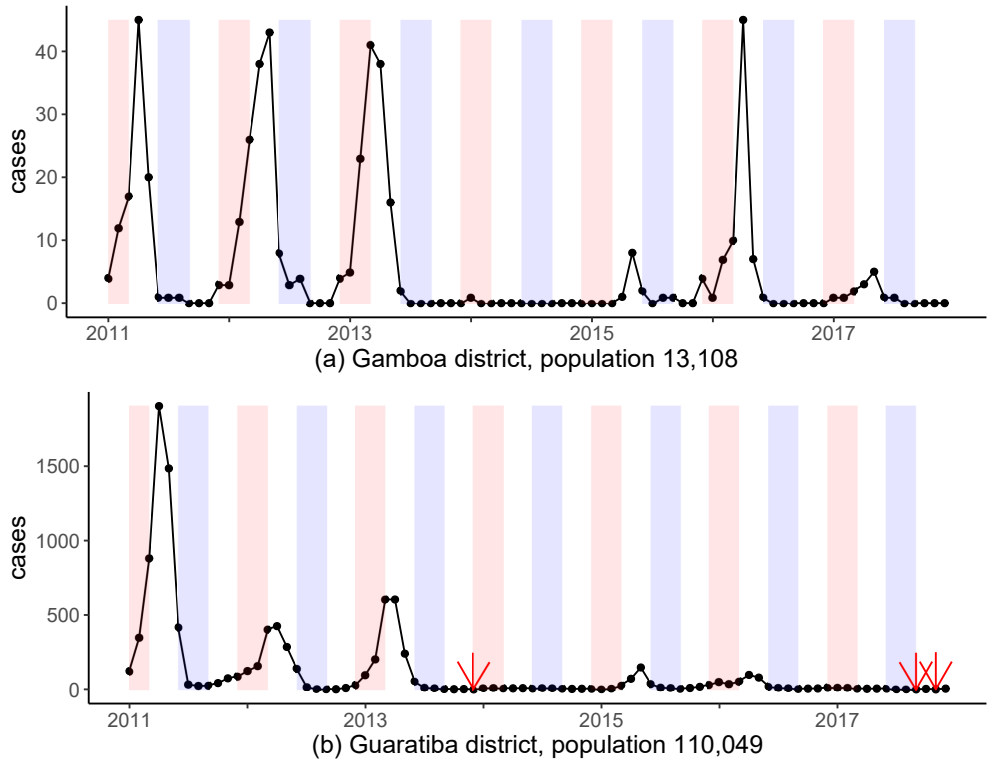


Figure 3.1: Monthly cases of dengue fever for 2 districts. Summer/winter seasons highlighted in red/light blue. In (b), the 3 red arrows point to the only 3 months with 0 cases in this district during the study period.

use human made water containers to lay eggs (Schmidt et al., 2011).

A statistical model can help quantify some of the above patterns. Specifically, the goal of our analysis is to investigate how certain risk factors (such as rainfall, temperature, socioeconomic factors and population size) are related to the reemergence and persistence of the disease, and if any effects are different for the reemergence compared to the persistence. Additionally, as Stoddard et al. (2013) showed that many individuals are infected by dengue outside their homes, we want to quantify the risk of dengue spreading between neighboring districts. We want to investigate whether certain space-time factors are associated with an increased risk of dengue spreading from a neighboring district. The number of reported cases in the neighboring district should be important in determining the risk of spread, but also, individuals are known to move more through high population areas (Grenfell et al., 2001). Finally, we wish to design some useful warning/forecasting systems for policy makers.

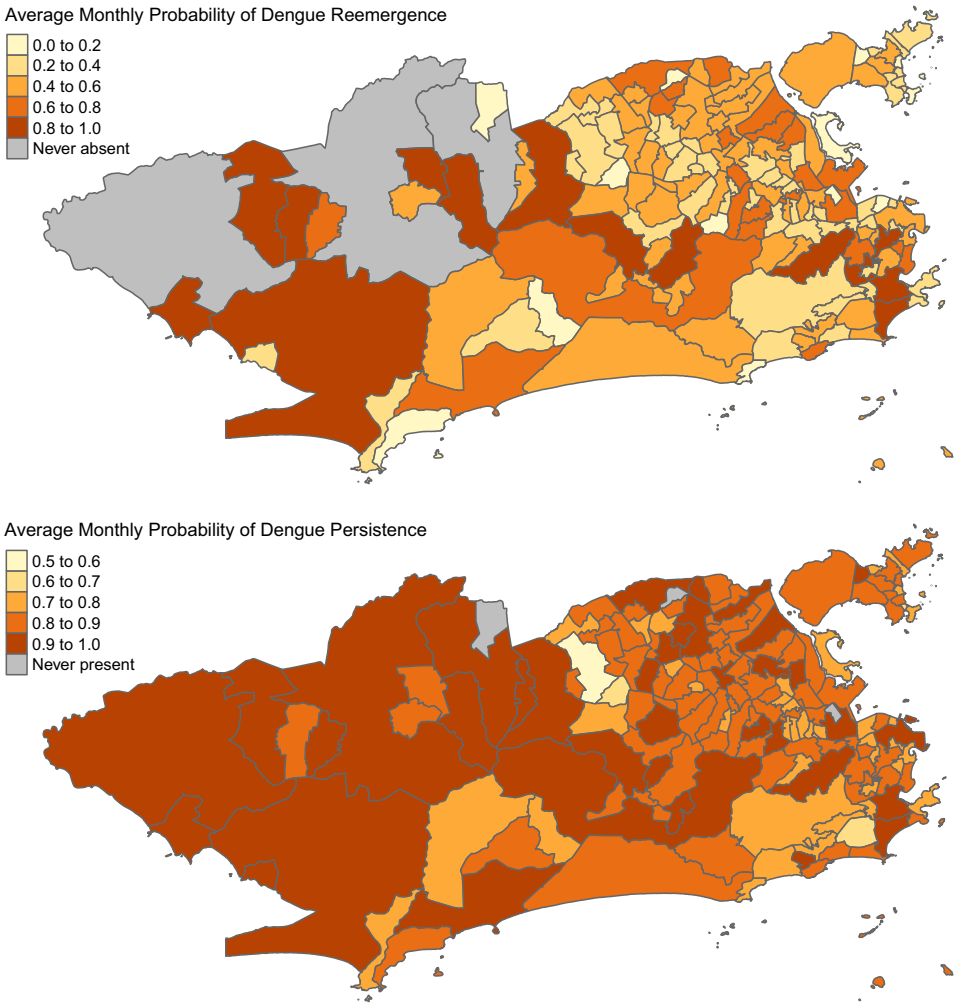


Figure 3.2: (top) Average monthly probability of dengue reemergence based on reported cases, i.e., $\frac{\#(\text{cases}=0\rightarrow\text{cases}>0)}{\#(\text{cases}=0\rightarrow\text{cases}>0)+\#(\text{cases}=0\rightarrow\text{cases}=0)}$. (bottom) Average monthly probability of dengue persistence based on reported cases, i.e., $\frac{\#(\text{cases}>0\rightarrow\text{cases}>0)}{\#(\text{cases}>0\rightarrow\text{cases}>0)+\#(\text{cases}>0\rightarrow\text{cases}=0)}$.

3.2 A Zero-state Coupled Markov Switching Negative Binomial Model

Assume we have areal data with $i = 1, ..., N$ areas across $t = 1, ..., T$ time periods. Let y_{it} be the reported case count from area i at time t. Let S_{it} be a binary indicator for the true presence of the disease, meaning $S_{it} = 1$ if the disease is present in area i during time t and $S_{it} = 0$ if the disease is absent. Finally, let $\boldsymbol{S}_{(-i)(t-1)} = \left(S_{1(t-1)}, ..., S_{(i-1)(t-1)}, S_{(i+1)(t-1)}, ..., S_{N(t-1)}\right)^T$

be the vector containing the state indicators of all areas excluding area i at time $t-1$; and let $\boldsymbol{y}^{(t-1)} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_{t-1})^T$ be the vector of all case counts up to time $t-1$, where $\boldsymbol{y}_t = (y_{1t}, ..., y_{Nt})^T$.

We assume that when the disease is present the reported cases are generated by a negative binomial distribution and that when the disease is absent no cases will be reported, that is,

$$y_{it} \mid S_{it}, \boldsymbol{y}^{(t-1)} \sim \begin{cases} 0, & \text{if } S_{it} = 0 \\ NB(\lambda_{it}, r_{it}), & \text{if } S_{it} = 1, \end{cases} \tag{3.1}$$

where λ_{it} is the expected number of reported cases given the disease is present and r_{it} is an overdispersion parameter such that $Var\left(y_{it} \mid S_{it} = 1, \boldsymbol{y}^{(t-1)}\right) = \lambda_{it}\left(1 + \lambda_{it}/r_{it}\right)$. Zero reported cases may arise when the disease is present due to the disease being undetected by the surveillance system. Therefore, when $y_{it} = 0$ we do not observe S_{it} as the disease could be absent or present but undetected. In general we assume the following forms for λ_{it} and r_{it},

$$\lambda_{it} = g\left(\boldsymbol{\delta}, \boldsymbol{x}_{it}, \boldsymbol{y}^{(t-1)}\right) \text{ and } r_{it} = h\left(\boldsymbol{\gamma}, \boldsymbol{w}_{it}, \boldsymbol{y}^{(t-1)}\right), \tag{3.2}$$

where $g(\cdot)$ and $h(\cdot)$ are positive valued functions, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are vectors of unknown parameters (including possibly random effects), and $\boldsymbol{x}_{it}$ and $\boldsymbol{w}_{it}$ are vectors of covariates. Therefore, λ_{it} and r_{it} can depend on past values of the case counts, to possibly account for temporal autocorrelation due to transmission of the disease, but not on past values of the state indicators, which is a common assumption in Markov switching models that greatly simplifies model fitting (Frühwirth-Schnatter, 2006).

To model the switching between periods of disease presence and absence in area i, we assume that S_{it} follows a two state nonhomogeneous Markov chain conditional on $\boldsymbol{S}_{(-i)(t-1)}$ and

$\boldsymbol{y}^{(t-1)}$. We propose the following conditional transition matrix for the Markov chain, for $t = 1, ..., T$,

$$\Gamma\left(S_{it}|\boldsymbol{S}_{(-i)(t-1)}, \boldsymbol{y}^{(t-1)}\right) = \begin{array}{c} \textbf{State} \\ S_{i(t-1)}{=}0\ \textbf{(absence)} \\ S_{i(t-1)}{=}1\ \textbf{(presence)} \end{array} \begin{array}{c} \begin{array}{cc} S_{it}{=}0\ \textbf{(absence)} & S_{it}{=}1\ \textbf{(presence)} \end{array} \\ \begin{bmatrix} 1 - p01_{it} & p01_{it} \\ 1 - p11_{it} & p11_{it} \end{bmatrix} \end{array}, \tag{3.3}$$

where,

$$p01_{it} = P\left(S_{it} = 1|S_{i(t-1)} = 0, \boldsymbol{S}_{(-i)(t-1)}, \boldsymbol{y}^{(t-1)}\right) \quad \text{(probability of disease reemergence)},$$
$$p11_{it} = P\left(S_{it} = 1|S_{i(t-1)} = 1, \boldsymbol{S}_{(-i)(t-1)}, \boldsymbol{y}^{(t-1)}\right) \quad \text{(probability of disease persistence)}.$$

The probability of disease reemergence in area i during time t, $p01_{it}$, is allowed to depend on a K-dimensional vector of space-time covariates $\boldsymbol{z}_{it} = (z_{it1}, ..., z_{itK})^T$ as well as the presence of the disease in neighboring areas during the previous time period,

$$\text{logit}(p01_{it}) = \zeta_0 + \boldsymbol{z}_{it}^T\boldsymbol{\zeta} + \sum_{j \in NE(i)} \phi_{01,j\to i}^{(t-1)\to t} S_{j(t-1)}, \tag{3.4}$$

where $NE(i)$ is the set of all neighboring locations of location i and $\phi_{01,j\to i}^{(t-1)\to t}$ is a coupling parameter that represents the effect that the disease spreading from neighboring area j has on the reemergence of the disease in area i at time t, the specification of which we will give shortly. The probability of disease persistence, $p11_{it}$, is defined similarly but all parameters can differ,

$$\text{logit}(p11_{it}) = \eta_0 + \boldsymbol{z}_{it}^T\boldsymbol{\eta} + \sum_{j \in NE(i)} \phi_{11,j\to i}^{(t-1)\to t} S_{j(t-1)}. \tag{3.5}$$

The transition probabilities can depend on $\boldsymbol{y}^{(t-1)}$ through $\boldsymbol{z}_{it}$, which may contain transformed lagged values of the reported cases (e.g. $\log(y_{i(t-1)} + 1)$). This can be justified by the fact

that the disease will only go extinct when there is a small number of infectious individuals. We also need to specify an initial state distribution for the Markov chain in each area, i.e. $p(S_{i0})$ for $i = 1, ..., N$.

In order to account for space-time heterogeneity in the effects of neighboring disease spread and to quantify associations between the effects of neighboring disease spread and various space-time factors we use linear models for the coupling parameters (Smith et al., 2002),

$$\begin{aligned} \phi_{01,j\to i}^{(t-1)\to t} &= \zeta_0^{(c)} + \boldsymbol{z}_{01,ijt}^{(c)T}\boldsymbol{\zeta}^{(c)} \\ \phi_{11,j\to i}^{(t-1)\to t} &= \eta_0^{(c)} + \boldsymbol{z}_{11,ijt}^{(c)T}\boldsymbol{\eta}^{(c)}, \end{aligned} \tag{3.6}$$

where $\boldsymbol{z}_{01,ijt}^{(c)}$ and $\boldsymbol{z}_{11,ijt}^{(c)}$ are vectors of space-time covariates, possibly different, related to either area, for example, $z_{01,ijtk}^{(c)} = \log(y_{j(t-1)}/pop_j+1)$ (reported prevalence in the neighboring area j) or $z_{01,ijtk}^{(c)} = \log(pop_i \times pop_j)$ (a gravity term (Tuite et al., 2011)). This is justified as effects of disease spread between areas often vary across space and time with observed factors, some examples being: rabies is less likely to spread between areas separated by rivers (Smith et al., 2002); hand, foot, and mouth disease is more likely to spread between farms with many cattle (Keeling et al., 2001); and measles is more likely to spread from large cities (Grenfell et al., 2001). We allow the effect of neighboring disease spread on the reemergence of the disease to be different from its effect on the persistence of the disease as these represent two distinct epidemiological processes. The disease may spread to non-infected areas from neighboring areas (modeled through $\phi_{01,j\to i}^{(t-1)\to t}$ in (3.4)); but also, neighboring areas can supply infectious individuals needed to maintain the persistence of the disease in infected areas (modeled through $\phi_{11,j\to i}^{(t-1)\to t}$ in (3.5)) (Okano et al., 2020).

As the coupling terms in (3.4)-(3.5) account for disease spread from neighboring areas, the $\zeta_0+\boldsymbol{z}_{it}^T\boldsymbol{\zeta}$ and $\eta_0+\boldsymbol{z}_{it}^T\boldsymbol{\eta}$ terms account for long distance dispersal of the disease as well as within area disease dynamics, i.e. spread from human and/or vector populations within the area. We assume the effect of covariate z_{itk} on the reemergence of the disease, represented by ζ_k in

(3.4), can be different from its effect on the persistence of the disease, represented by η_k in (3.5). The advantage of considering differing effects for each covariate on the reemergence and persistence of the disease is that many important epidemiological covariates can interact with the reemergence and persistence of an infectious disease quite differently, like temperature in our motivating example.

A final important motivation of the Markov chain model (3.3) for S_{it} is that it can account for both many consecutive periods of disease presence and absence. The probability the disease will be absent in location i during time t given it was absent at time $t-1$ is given by $1-p01_{it}$, that is, $1-p01_{it}$ is the probability of having a consecutive period of disease absence. Therefore, as the probability of disease reemergence, $p01_{it}$, approaches 0, consecutive periods of disease absence become more likely. Similarly, as the probability of disease persistence, $p11_{it}$, approaches 1, consecutive periods of disease presence become more likely. Therefore, when $p01_{it} << p11_{it}$ the Markov chain can model switching between long periods of disease absence and long periods of disease presence which is often observed for infectious diseases in smaller areas, see Figure 3.1(a) and Adams and Boots (2010).

We will refer to the model defined by (3.1)-(3.6) as the zero-state coupled Markov switching negative binomial (ZS-CMSNB) model. Finally, we note that a Poisson version of the ZS-CMSNB model is a special case as r_{it} goes to infinity, which we will refer to as a zero-state coupled Markov switching Poisson (ZS-CMSP) model.

3.3 Inferential Procedure

Let $\boldsymbol{S}_t = (S_{1t}, ..., S_{Nt})^T$ be the vector of all state indicators at time t and let $\boldsymbol{S} = (\boldsymbol{S}_0, ..., \boldsymbol{S}_T)^T$ be the vector of all state indicators. The vector $\boldsymbol{S}_t$ forms a first order Markov chain, conditional on $\boldsymbol{y}^{(t-1)}$, with state space $\{0,1\}^N$ and $2^N X 2^N$ transition matrix $\Gamma\left(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)}\right)$. An

element of $\Gamma\left(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)}\right)$ is given by,

$$\begin{aligned} &P\left(S_{1t}=s_{1t},...,S_{Nt}=s_{Nt}|S_{1(t-1)}=s_{1(t-1)},...,S_{N(t-1)}=s_{N(t-1)},\boldsymbol{y}^{(t-1)}\right) \\ &\quad=\prod_{i=1}^{N}P\left(S_{it}=s_{it}|\boldsymbol{S}_{t-1}=\boldsymbol{s}_{t-1},\boldsymbol{y}^{(t-1)}\right). \end{aligned} \tag{3.7}$$

Therefore, rewriting the ZS-CMSNB model in terms of $\Gamma(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)})$ and $p(\boldsymbol{y}_t|\boldsymbol{S}_t,\boldsymbol{y}^{(t-1)}) = \prod_{i=1}^{N}p(y_{it}|S_{it},\boldsymbol{y}^{(t-1)})$ shows that it is a Markov switching model as defined by Frühwirth-Schnatter (2006). However, the transition matrix has an exceptionally large dimension $(2^N X 2^N)$ which necessitates some changes to the inferential procedure compared to more traditional Markov switching models (Frühwirth-Schnatter, 2006), as we will discuss next.

The likelihood of $\boldsymbol{v}=(\boldsymbol{\theta},\boldsymbol{\beta})^T$, where $\boldsymbol{\theta}=(\zeta_0,\eta_0,\boldsymbol{\zeta},\boldsymbol{\eta},\zeta_0^{(c)},\eta_0^{(c)},\boldsymbol{\zeta}^{(c)},\boldsymbol{\eta}^{(c)})^T$ and $\boldsymbol{\beta}=(\boldsymbol{\delta},\boldsymbol{\gamma})^T$, given $\boldsymbol{y}=(\boldsymbol{y}_1,...,\boldsymbol{y}_T)^T$ and $\boldsymbol{S}$ is given by,

$$\mathrm{L}(\boldsymbol{y},\boldsymbol{S}|\boldsymbol{v})=\prod_{i=1}^{N}\prod_{t=1}^{T}p(y_{it}|S_{it},\boldsymbol{y}^{(t-1)},\boldsymbol{\beta})\prod_{i=1}^{N}p(S_{i0})\prod_{t=1}^{T}p(S_{it}|\boldsymbol{S}_{t-1},\boldsymbol{y}^{(t-1)},\boldsymbol{\theta}). \tag{3.8}$$

When N is large it is not possible to marginalize out $\boldsymbol{S}$ from (3.8), as doing so requires matrix multiplication with $\Gamma(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)})$ (Frühwirth-Schnatter, 2006). Therefore, we estimate the unknown elements of $\boldsymbol{S}$ along with $\boldsymbol{v}$ by sampling both from their joint posterior distribution which, from Bayes' theorem, is proportional to,

$$p(\boldsymbol{S},\boldsymbol{v}|\boldsymbol{y})\propto\mathrm{L}(\boldsymbol{y},\boldsymbol{S}|\boldsymbol{v})p(\boldsymbol{v}), \tag{3.9}$$

where $p(\boldsymbol{v})$ is the prior distribution of $\boldsymbol{v}$. We specify independent uninformative normal and gamma priors for all low-level parameters. As the joint posterior is not available in closed form, we resort to Markov chain Monte Carlo methods, in particular, we used a hybrid Gibbs sampling algorithm with some steps of the Metropolis-Hastings algorithm to sample from it. We sampled most elements of $\boldsymbol{v}$ individually, using an adaptive random walk Metropolis step

(Shaby and Wells, 2010). The parameter vectors $(\eta_0, \eta_0^{(c)}, \boldsymbol{\eta}^{(c)})^T$ and $(\zeta_0, \zeta_0^{(c)}, \boldsymbol{\zeta}^{(c)})^T$ showed high posterior correlations and, therefore, we jointly sampled them using automated factor slice sampling (Tibbits et al., 2014). This doubled the efficiency (minimum effective sample size per hour) of the Gibbs sampler in our application.

It is straightforward to sample each unknown element of $\boldsymbol{S}$ one at a time. When $y_{it} = 0$, the full conditional of S_{it} is given by,

$$\begin{aligned} &P\left(S_{it} = 1 | \boldsymbol{y}, \boldsymbol{v}, \{S_{jt}\}_{j \neq i, t \in \{0,\ldots,T\}}\right) \\ &= \frac{P(y_{it} = 0 | S_{it} = 1) P(S_{it} = 1 | \boldsymbol{S}_{t-1}) \prod_{j=1}^{N} p(S_{j(t+1)} | S_{it} = 1, \boldsymbol{S}_{(-i)(t)})}{\sum_{s_{it}=0}^{1} P(y_{it} = 0 | S_{it} = s_{it}) P(S_{it} = s_{it} | \boldsymbol{S}_{t-1}) \prod_{j=1}^{N} p(S_{j(t+1)} | S_{it} = s_{it}, \boldsymbol{S}_{(-i)(t)})}, \end{aligned} \tag{3.10}$$

where $P(y_{it} = 0 | S_{it} = s_{it}) = \left(\frac{r_{it}}{r_{it}+\lambda_{it}}\right)^{r_{it} s_{it}}$ and the dependence of the densities on $\boldsymbol{y}^{(t-1)}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are suppressed to reduce the size of the equation. However, one at a time sampling is known to lead to poor mixing in the Markov switching literature due to the strong posterior correlations between the unknown state indicators (Scott, 2002). Better mixing can be achieved by sampling all of $\boldsymbol{S}$ jointly from $p(\boldsymbol{S}|\boldsymbol{y}, \boldsymbol{v})$ (Chib, 1996). However, this is not computationally feasible with our model when N is large, as it again involves matrix multiplication with $\Gamma(\boldsymbol{S}_t | \boldsymbol{y}^{(t-1)})$ (Frühwirth-Schnatter, 2006). As an alternative, we propose to block sample $\boldsymbol{S}$ with each block containing all the state indicators in a different collection of locations. More specifically, assume we have put the locations into $k = 1, ..., B$ disjoint blocks and let $\boldsymbol{b}_k = (b_{k1}, ..., b_{kn_k})^T$ denote the vector of locations in block k ordered numerically. We assume $\sum_{k=1}^{B} n_k = N$ so that every location is contained in some block. We will adopt the following notation, let $\boldsymbol{S}_{(\boldsymbol{b}_k)t} = (S_{(b_{k1})t}, ..., S_{(b_{kn_k})t})^T$ be the vector of all state indicators at time t whose location is in block k, let $\boldsymbol{S}_{(\boldsymbol{b}_k)(t_1:t_2)} = (\boldsymbol{S}_{(\boldsymbol{b}_k)t_1}, ..., \boldsymbol{S}_{(\boldsymbol{b}_k)t_2})^T$, let $\boldsymbol{S}_{(\boldsymbol{b}_k)} = \boldsymbol{S}_{(\boldsymbol{b}_k)(0:T)}$ be the vector of all state indicators whose location is in block k, let $\boldsymbol{S}_{(-\boldsymbol{b}_k)t}$ be $\boldsymbol{S}_t$ with $\boldsymbol{S}_{(\boldsymbol{b}_k)t}$ removed, let $\boldsymbol{S}_{(-\boldsymbol{b}_k)(t_1:t_2)}$ be $(\boldsymbol{S}_{t_1}, ..., \boldsymbol{S}_{t_2})^T$ with $\boldsymbol{S}_{(\boldsymbol{b}_k)(t_1:t_2)}$ removed and let $\boldsymbol{S}_{(-\boldsymbol{b}_k)} = \boldsymbol{S}_{(-\boldsymbol{b}_k)(0:T)}$ so that $\boldsymbol{S}_{(-\boldsymbol{b}_k)}$ is the vector of all state indicators whose location is outside of block k. The

idea is to jointly sample $\boldsymbol{S}_{(\boldsymbol{b}_k)}$ from its full conditional distribution, given by,

$$p(\boldsymbol{S}_{(\boldsymbol{b}_k)}|\boldsymbol{S}_{(-\boldsymbol{b}_k)}, \boldsymbol{y}, \boldsymbol{v}) = p(\boldsymbol{S}_{(\boldsymbol{b}_k)T}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:T)}, \boldsymbol{y}, \boldsymbol{v}) \times \prod_{t=0}^{T-1} p(\boldsymbol{S}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(\boldsymbol{b}_k)t+1}, \boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)}, \boldsymbol{y}^{(t)}, \boldsymbol{v}). \tag{3.11}$$

A blocked forward filtering backward sampling (bFFBS) algorithm is needed to sample from (3.11), which we provide in the Supplementary Material (SM) Section A.1. The bFFBS algorithm only requires matrix multiplication with a $2^{n_k} X 2^{n_k}$ matrix. Our algorithm can be seen as an extension of the individual forward filtering backward sampling (iFFBS) algorithm recently proposed by Touloupou et al. (2020), who considered $n_k = 1$ for all k.

Our hybrid Gibbs sampler was implemented using the R package Nimble (de Valpine et al., 2017). Nimble comes with built in Metropolis-Hastings, automated factor slice and binary (equivalent to one at a time sampling for the unknown presence/absence indicators) samplers. The bFFBS samplers were implemented using Nimble's custom sampler feature. All Nimble R code, including for the custom bFFBS samplers, are provided on github (`https://github.com/Dirk-Douwes-Schultz/ZS_CMSP_code`). Nimble was chosen as it is extremely fast (C++ compiled) and only requires the coding of new samplers. In the SM Section A.2, we provide a simulation study which shows that our proposed Gibbs sampler can recover the true parameters of the ZS-CMSNB model.

3.3.1 Temporal prediction

For arbitrary K step ahead temporal predictions, we use a simulation procedure (Frühwirth-Schnatter, 2006) to draw samples from the posterior predictive distributions. Algorithm 1 in the SM will obtain realizations from the posterior predictive distribution of the cases, $y^{[m]}_{i(T+k)} \sim p(y_{i(T+k)}|\boldsymbol{y})$, and the presence of the disease, $S^{[m]}_{i(T+k)} \sim p(S_{i(T+k)}|\boldsymbol{y})$, for $i = 1, ..., N$, $k = 1, ..., K$ and $m = M+1, ..., Q$, where the superscript $[m]$ denotes a draw from the posterior distribution of the variable, M is the size of the burn-in sample and Q is the total MCMC

sample size. However, as $S^{[m]}_{i(T+k)}$ can only take two values, 0 or 1, it is difficult to interpret the uncertainty around this prediction for the presence of the disease. Therefore, instead of using summaries of $S^{[m]}_{i(T+k)}$ we use summaries of $P(S_{i(T+k)} = 1|S^{[m]}_{i(T+k-1)}, \boldsymbol{S}^{[m]}_{(-i)(T+k-1)}, \boldsymbol{y}^{[m]}_{T+k-1}, \boldsymbol{\theta}^{[m]})$. See the SM Section A.3 for the Monte Carlo approximations of the posterior predictive distributions.

3.3.2 Fitted values

Comparing the predictions, for a fixed K, to the observed values is not practical, as it requires fitting the model to multiple time points. This would result in long computational times, with MCMC methods, and unstable estimates for earlier times. Assume in this section $t \in \{1, ..., T\}$ and $i \in \{1, ..., N\}$. There are three types of fitted values for a Markov switching model: one step ahead, filtered and smoothed (Hamilton, 1993). In our Bayesian setting, the smoothed fitted values are given by,

$$S^{*[m]}_{it} = S^{[m]}_{it} \text{ drawn from } p(S_{it}|\boldsymbol{y}), \tag{3.12}$$

$$y^{*[m]}_{it} \text{ drawn from } p(y_{it}|S^{[m]}_{it}, \boldsymbol{y}^{(t-1)}, \boldsymbol{\beta}^{[m]}), \tag{3.13}$$

for $m = M + 1, ..., Q$. Note, $y^{*[m]}_{it} \sim p(y^*_{it}|\boldsymbol{y})$, where $y^*_{it}|S_{it}, \boldsymbol{y}^{(t-1)}, \boldsymbol{\beta} \sim NB(S_{it}\lambda_{it}, r_{it})$. That is, y^*_{it} represents a new case count generated by the same parameters, state and past counts that generated y_{it}. The smoothed fitted value of the state is drawn automatically by the Gibbs sampler used to fit the model, which is an advantage in using data augmentation to fit a Markov switching model. Note that $P(S_{it} = 1|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} S^{[m]}_{it}$. When $y_{it} = 0$, the $P(S_{it} = 1|\boldsymbol{y})$ represents the posterior probability the disease is present and thus undetected. Therefore, a map of the posterior mean of S_{it} for i where $y_{it} = 0$ can be used by policy makers as a warning system to identify areas reporting 0 cases that have a high chance of the disease being undetected.

Clearly, smoothed fitted values represent fits given all information in the data about the

states, including from present and future counts. If the model is not good at predicting the state one step ahead, which is largely based on the estimated transition probabilities of the Markov chain, then the smoothed fitted values may not reveal that as much of the information about S_{it} can come from y_{it}. The one step ahead distribution of the states is given by, $p(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)}, \boldsymbol{v})$, and of the cases is given by, $p(\boldsymbol{y}_t|\boldsymbol{y}^{(t-1)}, \boldsymbol{v}) = \sum_{\boldsymbol{s}_t \in \{0,1\}^N} p(\boldsymbol{y}_t|\boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{y}^{(t-1)}, \boldsymbol{\beta}) P(\boldsymbol{S}_t = \boldsymbol{s}_t|\boldsymbol{y}^{(t-1)}, \boldsymbol{v})$ (Frühwirth-Schnatter, 2006). As $\boldsymbol{v}$ is unknown, we can draw from the posterior of the one step ahead distributions. That is, the one step ahead fitted value of the state is given by, $S_{it}^{1*[m]}$ drawn from $p(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)}, \boldsymbol{v}^{[m]})$, and of the counts is given by, $y_{it}^{1*[m]}$ drawn from $p(y_{it}|S_{it}^{1*[m]}, \boldsymbol{y}^{(t-1)}, \boldsymbol{\beta}^{[m]})$, for $m = M+1, ..., Q$, where we use the superscript $1*[m]$ to denote a draw from the posterior of the one step ahead distributions. The advantage of the one step ahead fitted values is that information from $(\boldsymbol{y}_t, ..., \boldsymbol{y}_T)^T$ is only used to learn about $\boldsymbol{v}$ and not the states. Therefore, they should more accurately reflect the predictions compared to the smoothed values and should be able to diagnose a lack of fit in the Markov chain component of the model. The distribution $p(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)}, \boldsymbol{v}^{[m]})$ can be calculated using a Hamiltonian forward filter algorithm (Frühwirth-Schnatter, 2006). However, the filter requires matrix multiplication with $\Gamma(\boldsymbol{S}_t|\boldsymbol{y}^{(t-1)})$ and so is not computationally feasible to run when N is large.

As an alternative for large N, we propose coupled one step ahead fitted values. Assume we wish to calculate the one step ahead fitted values for location i when N is large. We will put location i in a block with locations $\boldsymbol{b}_{NE(i)} = (i, i_1, ..., i_{d_i})^T$ where $i_1, ..., i_{d_i} \in NE(i)$. Then, borrowing notation from (3.11), the coupled one step ahead fitted values are given by,

$$S_{it}^{c1*[m]} \text{ drawn from } p(S_{it}|\boldsymbol{S}^{[m]}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{v}^{[m]}), \tag{3.14}$$

$$y_{it}^{c1*[m]} \text{ drawn from } p(y_{it}|S_{it}^{c1*[m]}, \boldsymbol{y}^{(t-1)}, \boldsymbol{\beta}^{[m]}), \tag{3.15}$$

for $m = M+1, ..., Q$, where we use the superscript $c1*[m]$ to denote a draw from the posterior of the coupled one step ahead distributions. The distribution $P(S_{it} = 1|\boldsymbol{S}^{[m]}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{v}^{[m]})$

can be calculated using the forward filtering part of the bFFBS algorithm given in the supplementary material. We do not use quantiles of $S_{it}^{c1*[m]}$ as it can only take the two values 0 or 1. Instead we use quantiles of $P(S_{it} = 1|\boldsymbol{S}^{[m]}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{v}^{[m]})$, which represents a draw from the posterior distribution of the coupled one step ahead fitted probability of disease presence. We only use information from $(\boldsymbol{y}_t, ..., \boldsymbol{y}_T)^T$ to learn about $\boldsymbol{S}_{(-\boldsymbol{b}_{NE(i)})(0:t)}$. Therefore, if we block neighboring locations together there should be minimal information from $(\boldsymbol{y}_t, ..., \boldsymbol{y}_T)^T$ used in the state estimation.

3.4 Analysis of the Dengue Fever Data in Rio de Janeiro

3.4.1 Model specification and fitting

The ZS-CMSNB model has two components, a nonhomogeneous Markov chain which switches dengue between periods of presence and absence through modeling the reemergence and persistence of the disease, and a negative binomial component that generates the reported cases when dengue is present. For the negative binomial component, we use an endemic/epidemic specification for λ_{it} (Bauer and Wakefield, 2018), as it can account for temporal autocorrelation due to the onward transmission of the disease,

$$\lambda_{it} = \lambda_{it}^{AR} y_{i(t-1)} + \lambda_{it}^{EN}. \tag{3.16}$$

The auto-regressive rate λ_{it}^{AR} is meant to represent the transmission intensity of dengue in district i during month t. We model λ_{it}^{AR} as, $\log\left(\lambda_{it}^{AR}\right) = \beta_{0i}^{AR} + \beta_1 \text{Rain}_{t-1} + \beta_2 \text{Temp}_{t-1}$, where Temp_{t-1} is the maximum temperature in Rio during the previous month, Rain_{t-1} is the millimeters of rainfall and $\beta_{0i}^{AR} \sim N(\beta_0^{AR}, \sigma_{AR})$ are random intercepts accounting for between district differences in transmission intensity, where σ_{AR} is the standard deviation of the random effects. Rainfall and temperature are entered in a lagged manner since there is usually a delay separating changes in the mosquito population and dengue incidence (Coutinho et al.,

2006). The endemic risk λ_{it}^{EN} accounts for incidence not due to within area transmission. We compared both space-time and space varying endemic risk using the widely applicable information criterion (WAIC) (Gelman et al., 2014) (results not shown) and settled on a purely spatial varying risk, $\log\left(\lambda_{it}^{EN}\right) = \beta_{0i}^{EN}$, where $\beta_{0i}^{EN} \sim N(\beta_0^{EN}, \sigma_{EN})$. Examining the reported dengue counts it appears the overdispersion varies by district, therefore, we specified $r_{it} = r_i$. We used the weakly informative log-normal prior from Bauer and Wakefield (2018) for r_i which assumes the overdispersion is likely between 1.5 and 5 times the mean count in the district.

We considered $\boldsymbol{z}_{it} = (\text{pop}_i, \text{HDI}_i, \text{Rain}_{t-1}, \text{Temp}_{t-1}, \log(y_{i(t-1)}+1))^T$, where pop_i is the population of district i and HDI_i is the human development index of district i, as covariates possibly affecting the reemergence and persistence (see equations (3.4) and (3.5)) of dengue. The monthly rainfall and temperature data were obtained from the National Institute for Space Research (`http://bancodedados.cptec.inpe.br/`) and the district level human development indexes were obtained from ipeadata (`http://www.ipeadata.gov.br/Default.aspx`). Since we condition on the first observation, we specified $p(S_{i1}) \sim Bern(.5)$ if $y_{i1} = 0$ and $p(S_{i1})$ is degenerate 1 if $y_{i1} > 0$, as the initial state distributions.

As for covariates potentially associated with the effects of dengue spreading from a neighboring district, i.e. $\boldsymbol{z}_{01,ijt}^{(c)}$ and $\boldsymbol{z}_{11,ijt}^{(c)}$ in (3.6), we considered; $\log(y_{j(t-1)}/pop_j + 1)$, the prevalence of the disease in the neighboring district; $\log(pop_i \times pop_j)$, a gravity term reflecting the fact that individuals are more likely to move between high population areas; $|NE(j)|$, the number of neighbors of area j which can account for the effects of disease spread potentially being proportioned among neighbors; and, finally, $\log(y_{i(t-1)} + 1)$ for the persistence only, to potentially account for crowding out effects. Due to the high computational cost per covariate added to the coupling parameters we started with a simple homogeneous model of disease spread and added coupling covariates sequentially, removing them if they did not improve the WAIC substantially (by 10 units); this process is illustrated in Table 3.1. Based

Table 3.1: Model comparison using WAIC for models fitted to the dengue data. Note, $\emptyset$ refers to no covariates, or a homogeneous model of disease spread, that its, $\phi_{01,j\to i}^{(t-1)\to t} = \zeta_0^{(c)}$ or $\phi_{11,j\to i}^{(t-1)\to t} = \eta_0^{(c)}$. The bolded model is our final ZS-CMSNB model as we do not accept models that are more complex and do not improve the WAIC substantially (by 10 units). NP = Neighboring prevalence $= \log(y_{j(t-1)}/pop_j + 1)$, G = Gravity term $= \log(pop_i \times pop_j)$, LC = Local cases $= \log(y_{i(t-1)} + 1)$, NN = Number of neighbors $= |NE(j)|$.

Model	$\boldsymbol{z}_{01,ijt}^{(c)}$	$\boldsymbol{z}_{11,ijt}^{(c)}$	**WAIC**
endemic/epidemic	–	–	69,939
ZINB	–	–	67,798
ZINBRE	–	–	67,782
ZS-CMSNB	$\emptyset$	$\emptyset$	67,729
	$(\text{NP})^T$	$\emptyset$	67,652
	$\boldsymbol{(\text{NP},\text{G})^T}$	$\boldsymbol{\emptyset}$	**67,632**
	$(\text{NP},\text{G})^T$	$(\text{NP})^T$	67,631
	$(\text{NP},\text{G})^T$	$(\text{NP},\text{LC})^T$	67,628
	$(\text{NP},\text{G},\text{NN})^T$	$\emptyset$	67,648
ZS-CMSP	$\emptyset$	$\emptyset$	90,458

on Table 3.1 we decided on $\boldsymbol{z}_{01,ijt}^{(c)} = \left(\log(y_{j(t-1)}/pop_j + 1), \log(pop_i \times pop_j)\right)^T$ and a homogeneous effect of neighboring disease spread for the persistence of the disease. That is, we let $\phi_{01,j\to i}^{(t-1)\to t} = \zeta_0^{(c)} + \zeta_1^{(c)} \log\left(y_{j(t-1)}/pop_j + 1\right) + \zeta_2^{(c)} \log\left(pop_i \times pop_j\right)$ and $\phi_{11,j\to i}^{(t-1)\to t} = \eta_0^{(c)}$ for the final model.

Each ZS-CMSNB model in Table 3.1, and the ZS-CMSP model as it is a special case, was fitted using our proposed Gibbs sampler from Section 3.3, to the monthly Rio dengue cases for 2011-2017 ($t = 1, ..., 84$) and all districts in the city ($i = 1, ..., 160$). We ran the Gibbs sampler for 80,000 iterations on 3 chains with an initial burn in of 30,000 iterations. All sampling was started from random values in the parameter space to avoid convergence to local modes. Convergence was checked using the Gelman-Rubin statistic (all estimated parameters <1.05) and the minimum effective sample size (>1000) (Plummer et al., 2006). For the final bolded model in Table 3.1, we compared the efficiency (minimum effective

sample size per hour) of 3 candidate samplers for the state indicators: a one at time sampler (binary) and two bFFBS samplers with block sizes 1 (iFFBS) and 2 (bFFBS2). For the bFFBS2 sampler, we blocked 71 neighboring locations together leaving 15 locations that could not be matched to a neighbor that were put in single location blocks (remaining 3 locations had all positive observations and did not need to be sampled). The iFFBS sampler was 10% more efficient than the binary sampler and the bFFBS2 sampler was 5% more efficient. Therefore, there does not appear to be much if any gain to joint sampling the hidden states in our application.

We compared the fit of our model to a nested model without the Markov chain, i.e. a model for which $\zeta_0 = \eta_0$, $\boldsymbol{\zeta} = \boldsymbol{\eta}$, $\zeta_0^{(c)} = \eta_0^{(c)} = 0$ and $\boldsymbol{\zeta}^{(c)} = \boldsymbol{\eta}^{(c)} = \mathbf{0}$ in equations (3.4)-(3.6). Note that this is a standard zero-inflated negative binomial (ZINB) model (Greene, 1994). However, the ZINB model does not account for spatio-temporal correlations in the presence of the disease. Therefore, we also compared to a model like the one used by Hoef and Jansen (2007). Let $p_{it} = P(S_{it} = 1|\boldsymbol{\theta}_{RE})$, where $\boldsymbol{\theta}_{RE}$ is a vector of random effects, then Hoef and Jansen (2007) fit a model where $\text{logit}(p_{it}) = \alpha_0 + \boldsymbol{q}_{it}^T\boldsymbol{\alpha} + \eta_t + \delta_{it}$, where $\boldsymbol{q}_{it}$ is a vector of space-time covariates, η_t is an overall time trend which followed an autoregressive (AR) process and $\{\delta_{it}\}_{i=1}^N$ followed a proper conditional autoregressive (pCAR) distribution for each time period that was independent across time. As most excess variation in spatio-temporal infectious disease counts occurs temporally and not spatially (Bauer et al., 2016), we flip their proposal so that $\text{logit}(p_{it}) = \alpha_0 + \boldsymbol{q}_{it}^T\boldsymbol{\alpha} + \eta_i + \delta_i(t)$, where η_i is an overall spatial effect and $\delta_i(t)$ represents a time trend for each area. We assume a pCAR distribution for $\{\eta_i\}_{i=1}^N$. Torabi (2017) suggests using an AR process or splines to model $\delta_i(t)$. However, we prefer using autoregression for the spatial-temporal interaction, combining the proposals of Fernandes et al. (2009) and Yang et al. (2013), as it is computationally much simpler and also epidemiologically motivated. That is, we let $\delta_i(t) = \rho_1 \log(y_{i(t-1)} + 1) + \rho_2 S_{i(t-1)}$. The motivation being that for the disease to go extinct all previous cases must fail to pass on the disease but since $y_{i(t-1)}$ represents reported cases there might still be an effect at

$y_{i(t-1)} = 0$ due to the disease being undetected. We take $\boldsymbol{q}_{it} = \{\boldsymbol{z}_{it}\} \setminus \{\log(y_{i(t-1)} + 1)\}$ as $\log(y_{i(t-1)} + 1)$ is already included in the model. We call this model a zero-inflated negative binomial random effect (ZINBRE) model.

Finally, we also compared the fit of our model to a model without zero inflation (endemic/epidemic) and to a Poisson version of our model (ZS-CMSP). Table 3.1 gives the results of the model comparison using WAIC for the five classes of models. The endemic/epidemic, ZINB and ZINBRE models in Table 3.1 were fit using standard MCMC methods in Nimble, the code is available on github (https://github.com/Dirk-Douwes-Schultz/ZS_CMSP_code). From Table 3.1, incorporation of zero inflation and overdispersion significantly improves model fit, and the ZS-CMSNB model is the superior model for the zero inflation.

3.4.2 Results

From Table 3.1, the most relevant factors for determining the strength of dengue spread from a neighboring area are, the reported prevalence of dengue in the neighboring area, and the population of the neighboring area and the home area. To quantify the association between the effects of neighboring dengue spread and these two factors we can use the odds ratio (OR) of dengue reemergence given disease spread from a neighboring area, or mathematically,

$$\frac{\Omega(S_{it} = 1 | S_{i(t-1)} = 0, S_{j(t-1)} = 1)}{\Omega(S_{it} = 1 | S_{i(t-1)} = 0, S_{j(t-1)} = 0)} = \exp\left(\zeta_0^{(c)} + \zeta_1^{(c)} \log(y_{j(t-1)}/pop_j + 1) + \zeta_2^{(c)} \log(pop_i \times pop_j)\right), \tag{3.17}$$

for $j \in NE(i)$, where $\Omega(A) = P(A)/(1-P(A))$. From Table 3.2, which shows estimates from the Markov chain component of the model, the posterior mean and 95% posterior credible interval of $\zeta_1^{(c)}$ is 5.63 [3.33-8.67] and of $\zeta_2^{(c)}$ is .22 [.09-.36]. Therefore, the effect of dengue spreading from a neighboring area is positively associated with the reported prevalence of dengue in the neighboring area and the population of both areas, especially with the preva-

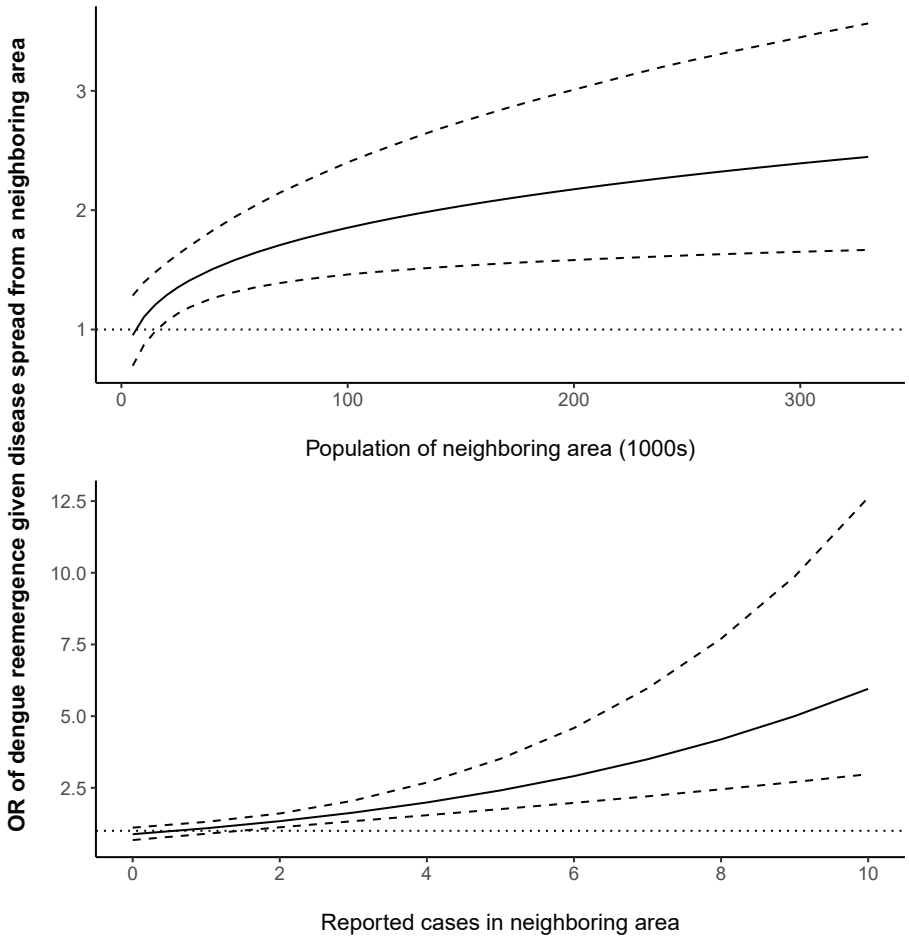


Figure 3.3: Posterior means (solid lines) and 95% posterior credible intervals (dashed lines) of (top) $\exp\left(\zeta_0^{(c)} + \zeta_1^{(c)} \log(.08+1) + \zeta_2^{(c)} \log(10 \times pop_j)\right)$ versus pop_j and (bottom) $\exp\left(\zeta_0^{(c)} + \zeta_1^{(c)} \log(cases_j/26+1) + \zeta_2^{(c)} \log(10 \times 26)\right)$ versus $cases_j$ (see (3.17)), where, 10 is the population of a small area, 26 is the median population and .08 (per 1000) is the median prevalence of dengue.

lence. To help visualize these associations we have plotted, in Figure 3.3, (3.17) versus the population of the neighboring area (top) and the reported cases of dengue in the neighboring area (bottom). In the plots, we fixed the population of the home area at 10,000, to represent a small area, and for the top graph we fixed prevalence in the neighboring area at its median value (.08 per 1000), and for the bottom graph we fixed the population in the neighboring area at its median value (26,000) (exact formulas are given in the caption). From Figure 3.3 (top), the effect of dengue spreading from a neighboring area increases gradually with

the population of the neighboring area and there is likely not a large effect of disease spread from low population areas, at least at median prevalence. These estimates are reasonable as there will be much more travel to high population neighbors compared to low population neighbors. From Figure 3.3 (bottom), the effect of dengue spreading from a neighboring area increases rapidly with the reported prevalence of the disease in the neighboring area. At a prevalence of $10/26 = .38$, where we have cut off the graph to improve visualization, the odds ratio is 5.95 [2.97-12.62], meaning the odds of dengue reemergence is increased 6 (likely lowest 3) times due to the disease spread, a large effect. A prevalence of .38 is about at epidemic levels, 75th percentile, and, therefore, the figure implies that if an epidemic occurs in an area there is a good chance that dengue will spread to neighboring areas where it is absent.

We also wanted to investigate whether dengue spreads between areas undetected. Note that the OR of dengue reemergence given disease spread from a neighboring area where dengue is undetected is given by $\exp\left(\zeta_0^{(c)} + \zeta_2^{(c)} \log(pop_i \times pop_j)\right)$. Through repeated calculations we found that, *a posteriori*, this quantity has a 75% chance of being greater than one if $pop_j > 685.4/pop_i$. Setting $pop_i = 10$ (1000) gives $pop_j > 68.53$, which represents the top 15 percent of districts in terms of population size. Therefore, the disease likely spreads undetected only from the larger districts.

The estimated coefficients for the Markov chain part of the fitted ZS-CMSNB model are given in Table 3.2. From the estimated intercepts, in small areas consecutive periods of disease presence are heavily favored on average, while consecutive periods of disease absence are only moderately favored. This highlights the importance of preventing dengue reemergence in small areas (one effective way would be to prevent disease spread from neighboring epidemic areas as shown in Figure 3.3), as when the disease does reemerge it will likely persist for some time. We shift the intercepts to small districts (pop=10,000) as that is where most of the 0s occur (there are not many 0s in large districts). Population size has a significantly

Table 3.2: Posterior means and 95% posterior credible intervals (in squared brackets) for the estimated parameters from the Markov chain part of the fitted ZS-CMSNB model. Intercepts shifted so they represent the probabilities of dengue reemergence and persistence with 0 neighbors infected, pop=10,000 and all other covariates fixed at mean or median (if very skewed) values. Other rows represent odds ratios, i.e. $\exp(\zeta_k)$ and $\exp(\eta_k)$ for $k =$ pop=1,..., $= \log(y_{i(t-1)} + 1) = K$. The reemergence column of the Neighborhood Presence row shows $\exp(\phi_{01,j\to i}^{(t-1)\to t})$ calculated at median values of neighboring pop/prevalence and pop=10,000 for the home area, the persistence column shows $\exp(\eta_0^{(c)})$. The rows beneath the Neighborhood Presence row show, respectively, unmodified estimates of $\zeta_0^{(c)}/\eta_0^{(c)}$, $\zeta_1^{(c)}$ and $\zeta_2^{(c)}$.

	Probability or Odds Ratio	
Covariate	**Reemergence (absence to presence)**	**Persistence (presence to presence)**
Intercept (shifted)	.33 [.22-.46]	.84 [.78-.90]
pop (1000s)	1.06 [1.02-1.11]	1.02 [1.01-1.03]
HDI (.1)	1.07 [.71-1.52]	.82 [.62-1.09]
Temp_{t-1} (Celc.)	1.24 [1.11-1.39]	1.21 [1.13-1.31]
Rain_{t-1} (10 mm)	1.04 [1-1.09]	1.13 [1.09-1.17]
$\log(y_{i(t-1)} + 1)$	–	5.07 [4.17-6.19]
Neighborhood Presence	1.35 [1.14-1.64] (average effect in small area)	1.12 [1.01-1.23]
Intercept	-1.38 [-2.28- -.51]	.12 [.01-.23]
$\log\left(y_{j(t-1)}/pop_j + 1\right)$	5.63 [3.33-8.67]	–
$\log\left(pop_i \times pop_j\right)$	.22 [.09-.36]	–

larger effect on dengue reemergence compared to persistence (diff=.04 [0-.08]), especially considering population increases the effects of dengue spreading from neighboring areas on the reemergence of the disease, see Figure 3.3. This can explain why, in Figure 3.2, there is more spatial variation observed in the average probabilities of dengue reemergence compared to persistence. Additionally, the strong positive association between population size and dengue reemergence means we would expect much longer periods of disease absence in the smaller districts compared to larger districts, which follows the well-known theories from Bartlett (1957). Interestingly, we found no association between human development index

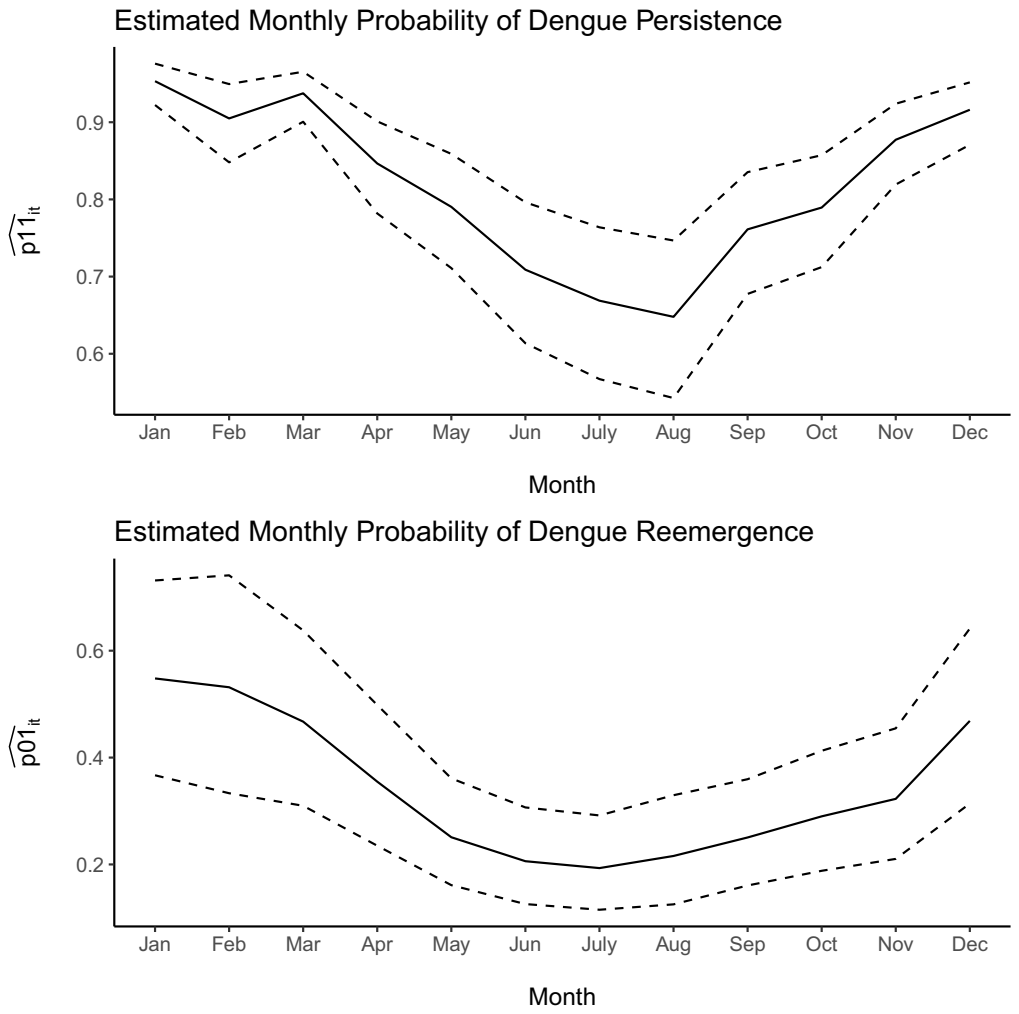


Figure 3.4: Posterior means (solid lines) and 95% posterior credible intervals (dashed lines) of the estimated monthly probabilities of dengue reemergence and persistence in a small district (pop=10,000). We assumed average monthly temperature and rainfall values, and 2 cases (median) reported in the previous month for the persistence probabilities.

and the risk of dengue reemergence or persistence. Both rainfall and temperature have a strong positive association with dengue reemergence and persistence. For example, we estimated that a one degree rise in maximum temperature during the previous month is associated with a 24 [11-39] percent increase in the odds of dengue reemerging and a 21 [13-31] percent increase in the odds of dengue persisting. Therefore, we will have longer periods of disease absence during the winter (at least in the smaller districts) and longer periods of disease presence during the summer, which follows patterns in the mosquito population.

Rainfall has a significantly larger effect on the persistence of the disease (diff=.08 [.02-.15]), which clearly manifests in Figure 3.4, and could be due to the role played by rainfall in the lifecycle of the mosquito egg (Coutinho et al., 2006). When the disease is present water is needed for egg laying and so rainfall helps the mosquito population persist, however, if the disease is absent there are not many mosquitoes and rainfall could wash away vertically infected mosquito eggs. The effect of the previous months cases, $\log(y_{i(t-1)} + 1)$, on the persistence of the disease is quite large and reflects the fact that dengue will only go extinct when there is a small number of infected individuals.

From the Neighborhood Presence row of Table 3.2, we estimated that dengue being present in a neighboring district during the previous month is associated with a 35 [14-64] percent increase in the odds of dengue reemerging in a small area on average and a 12 [1-23] percent increase in the odds of dengue persisting. The average effect of neighboring dengue spread on the reemergence of the disease is much higher than the homogeneous effect on the persistence. Therefore, efforts to restrict cross infection between districts will typically be more effective if one district is not already infected. A possible explanation is a "spark effect", like forest fires, in that only a small amount of disease spread could be needed to start the disease being present in an area but will not contribute much to, say, a large epidemic. Note, however, that the Neighborhood Presence row of Table 3.2 only shows the average effect of neighboring dengue spread on the reemergence of the disease in a small area, i.e. $\exp\left(\zeta_0^{(c)} + \zeta_1^{(c)} \log(.08 + 1) + \zeta_2^{(c)} \log(10 \times 26)\right)$, and the actually effect varies quite a lot with space and time as shown in Figure 3.3, making broad comparisons somewhat challenging.

Figure 3.4 shows posterior summaries of the estimated monthly probabilities of dengue reemergence and persistence for a small district (pop=10,000) without neighboring disease spread. The temporal evolution of the two probabilities are clearly different which mainly reflects the increased effect of rainfall on the persistence of the disease. Additionally, the

Table 3.3: Posterior means and 95% posterior credible intervals (in squared brackets) from the negative binomial part of the fitted ZS-CMSNB model. $\bar{r} = 1/N \sum_{i=1}^{N} r_i$ and $\sigma_r = \sqrt{1/(N-1) \sum_{i=1}^{N} (r_i - \bar{r})^2}$

Covariate	**Parameter**	**Estimate**
Intercept AR	β_0^{AR}	-.37 [-.4- -.33]
Rain_{t-1} (10 mm) AR	β_1	.036 [.03-.04]
Temp_{t-1} (C) AR	β_2	.23 [.22-.24]
Std. dev AR	σ_{AR}	.12 [.09-.14]
Intercept EN	β_0^{AR}	.37 [.29-.45]
Std. dev EN	σ_{EN}	.42 [.35-.5]
Avg. overdispersion	$\bar{r}$	2.27 [2.16-2.39]
Std. dev overdispersion	σ_r	1.08 [.92-1.37]

figure illustrates how the model can recreate the seasonal switching of dengue between long periods of disease absence in the winter and long periods of disease presence in the summer (Adams and Boots, 2010).

Table 3.3 gives the estimated parameters from the negative binomial part of the model. When dengue is present in a district, increases in temperature and rainfall lead to increased transmission of the disease. There is significant between district differences in the transmission rate and endemic risk. The differences in the endemic risk could be driven by either differences in the environment or reporting rates. There is significant overdispersion on average (variance is 11 times the mean on average) and there is significant between district differences in the overdispersion.

3.4.3 Fitted values and predictions

Figure 3.5 shows a map of the posterior probability that dengue is present in the districts of Rio de Janeiro during December 2017 (time T). If $y_{iT} > 0$ then the posterior probability that dengue is present in district i during time T is 1 (grey districts on the

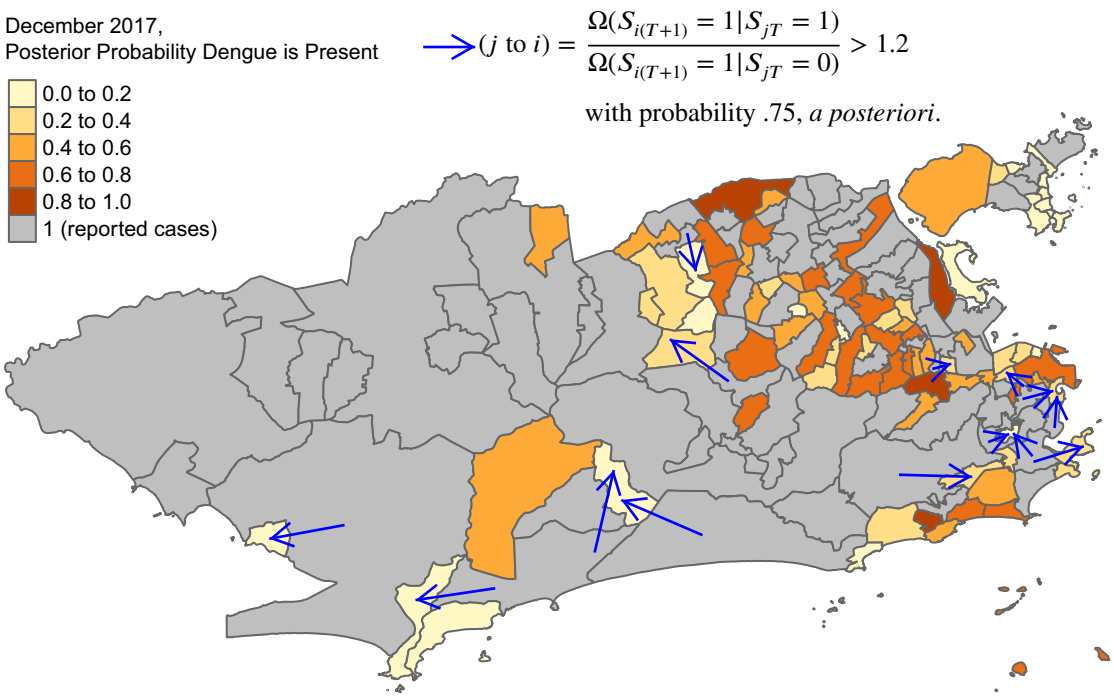


Figure 3.5: Posterior probability that dengue is present for December 2017 (time T). Districts that have reported cases are in grey to distinguish them from districts where dengue may be undetected. A blue arrow is drawn from $j \in NE(i)$ to i if $\frac{\Omega(S_{i(T+1)}=1|S_{jT=1})}{\Omega(S_{i(T+1)}=1|S_{jT=0})}$ is greater than 1.2 with probability .75, *a posteriori*, where $\Omega(A) = P(A)/(1-P(A))$.

map). Otherwise, if $y_{iT} = 0$ the disease may be undetected and we can approximate $P(S_{iT} = 1|\boldsymbol{y}) \approx \frac{1}{Q-M}\sum_{m=M+1}^{Q} S_{iT}^{[m]}$. A blue arrow is drawn from $j \in NE(i)$ to i if $\frac{\Omega(S_{i(T+1)}=1|S_{jT=1})}{\Omega(S_{i(T+1)}=1|S_{jT=0})} = S_{iT}\exp\left(\phi_{11,j\rightarrow i}^{T\rightarrow(T+1)}\right) + (1-S_{iT})\exp\left(\phi_{01,j\rightarrow i}^{T\rightarrow(T+1)}\right)$ is greater than 1.2 with probability .75, *a posteriori*, where $\Omega(A) = P(A)/(1-P(A))$. That is, if the current effect of disease spread from the neighboring area is likely greater than the effect of a one degree rise in temperature (see Table 3.2). The map shows where dengue currently is in the city and where it is spreading to and, therefore, would be of interest to policy makers. For example, the Santa Teresa district has 3 arrows coming out of it indicating it is an important source of dengue spread in the city currently.

Panels of Figure 3.6 show posterior summaries of the coupled one month ahead fitted values, see equations (3.14)-(3.15), of dengue cases (top graphs) and dengue presence (bottom

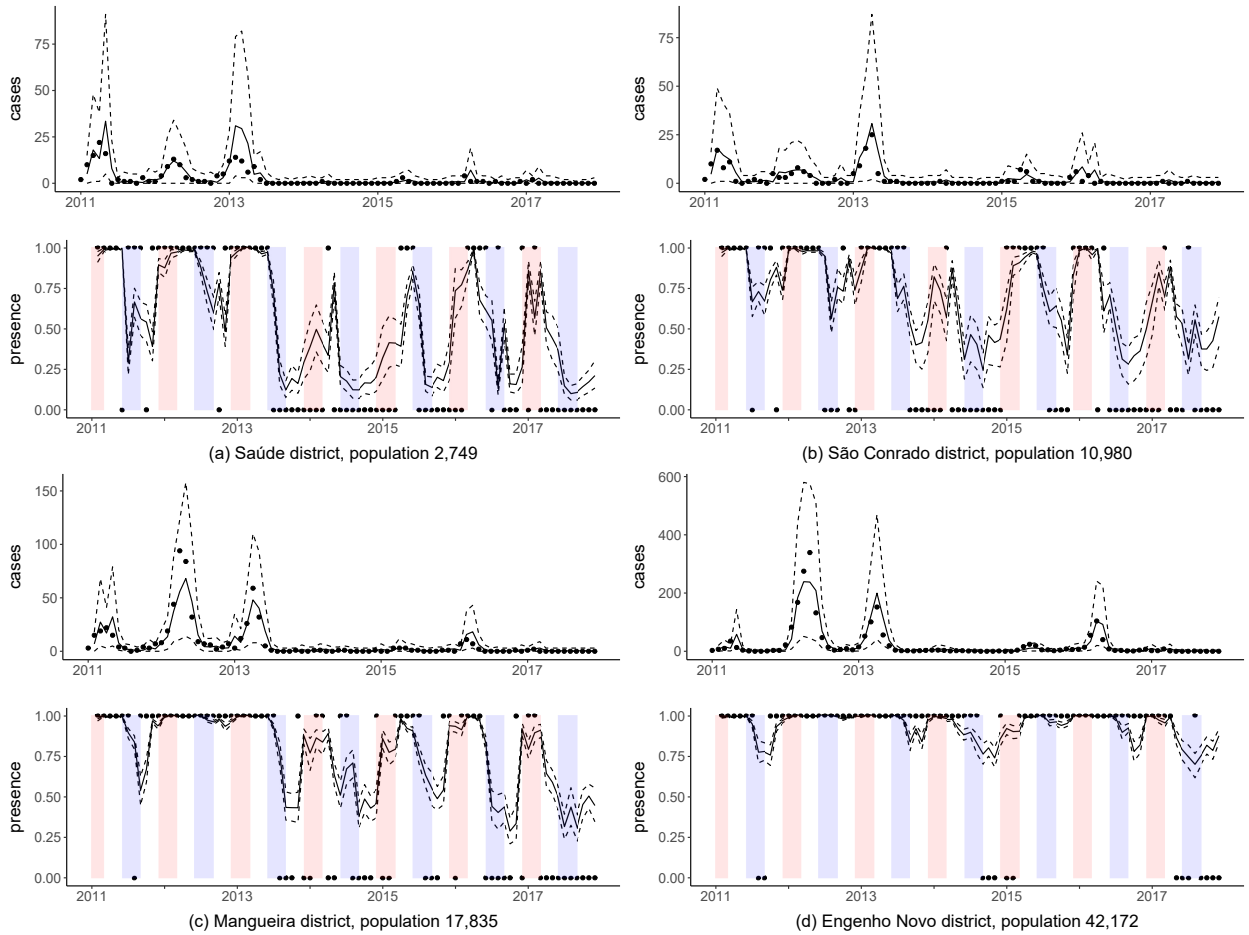


Figure 3.6: Posterior summaries of the coupled one month ahead fitted values for 4 districts. (top graphs) Coupled one month ahead fitted values of cases versus observed cases. (bottom graphs) Coupled one month ahead fitted values of dengue presence risk versus observed presence (0=0 reported cases which may not correspond to the actual absence of the disease). Posterior means (solid lines), 95% posterior credible intervals (dashed lines) and observed (points). Summer/winter seasons highlighted in red/light blue in the bottom graphs.

graphs), for 4 districts. These were constructed by running the blocked forward filter in blocks with 2 neighboring locations, with each MCMC draw, as explained in Section 3.3.2. For Figure 3.6, we chose mostly small districts that had a good mix of zeros and positive counts to illustrate both the zero-inflated and count components of the model. Generally, the model is able to predict, in sample, the presence of the disease one month ahead well when the disease is observed to be present. It is difficult to assess the presence fits when the disease is reportedly absent, as dengue could be undetected and actually present. The fit to the cases is also reasonable, however, there does appear to be some overestimation of

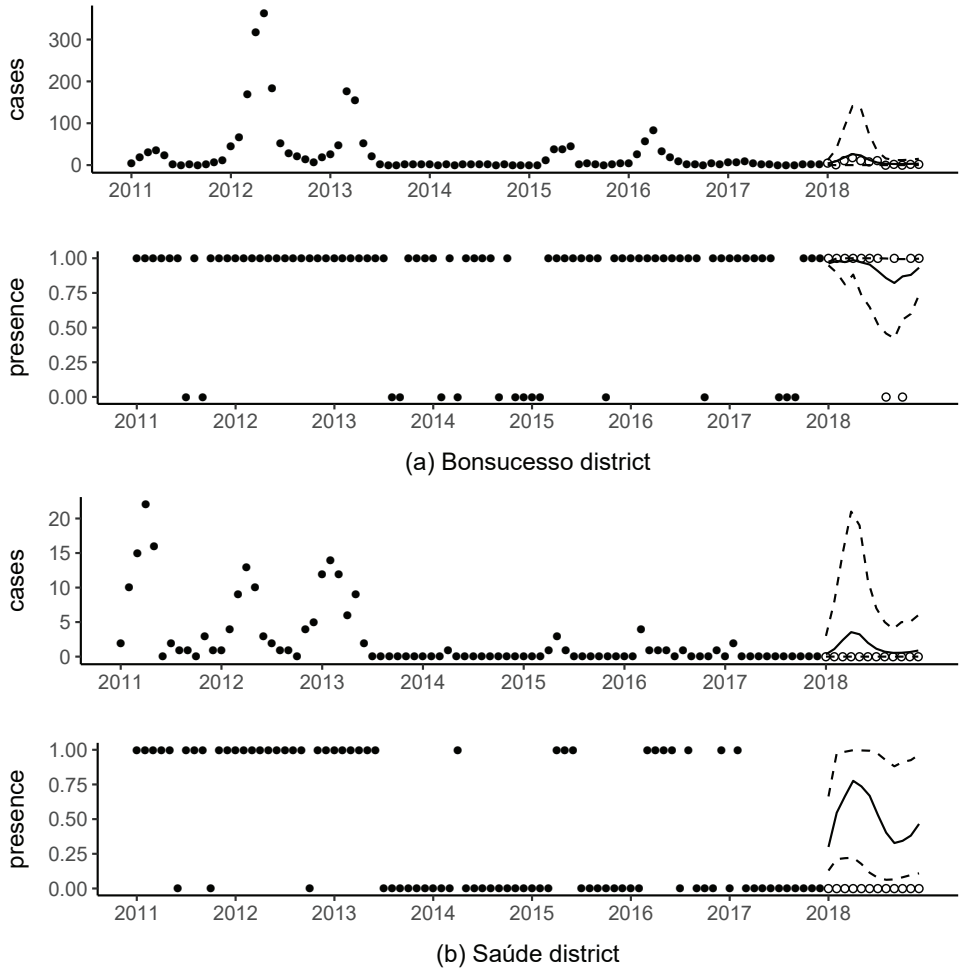


Figure 3.7: Summaries of the 1-12 month ahead posterior predictive distributions of the cases (top graphs) and dengue presence risk (bottom graphs) for 2 districts. Posterior predictive means (solid lines) and 95% posterior predictive credible intervals (dashed lines). Solid circles are from the last 7 years (2011-2017) used to fit the model and open circles are future observed values.

uncertainty at high counts. This is due to the quadratic mean variance relationship of the negative binomial distribution. There is an advantage to this mean variance relationship, however, for estimating the Markov chain part of the model as most of the weight is placed at low counts which is where the reemergence and persistence of the disease occurs. As an additional goodness of fit measure we plot the autocorrelation function of the Pearson residuals in the SM Section A.4 for the 4 districts of Figure 3.6. There are not many significant autocorrelations, suggesting that there is no structure left in the residuals.

Panels of Figure 3.7 show summaries of the 1-12 month ahead posterior predictive distributions of dengue cases (top graphs) and dengue presence risk (bottom graphs) in two districts. These were calculated using Algorithm 1 in the SM with $K = 12$. There was very little dengue activity in 2018 and we chose to show Bonsucesso as it was one of the few districts that experienced a small epidemic while Saúde represents a more typical district. We assumed average monthly temperature and rainfall values in the calculations, although other forecasting scenarios could be considered, such as high temperature scenarios, or forecasts of the meteorological variables could be used. Both predictions of dengue presence and the case counts show a clear seasonal pattern, peaking in the summer and declining during the winter. The uncertainty in the predictions of dengue presence is large in the Saúde district. A reasonable explanation for this large uncertainty is that Saúde reported 0 cases for the last time point, meaning there is uncertainty around whether dengue is present in Saúde at time $T = 84$. As Saúde is a small district (pop=2,000) there is a large difference in the probabilities of disease reemergence and persistence (due to the larger effect of population size on the reemergence of the disease), and so uncertainty around the presence of the disease previously leads to wide confidence intervals for the future predictions of disease presence.

3.5 Concluding Remarks

We have proposed a zero-state coupled Markov switching negative binomial (ZS-CMSNB) model for general spatio-temporal infectious disease counts that contain an excess of zeros. Our approach can model nonhomogeneous switching between long periods of disease presence and absence while accounting for space-time heterogeneity in the effects of disease spread between areas. The main difference with existing ZIC models used in spatio-temporal disease mapping is that the disease switches between periods of presence and absence in each area through a series of nonhomogeneous Markov chains coupled between neighboring locations, as opposed to the presence of the disease forming a series of conditionally independent

Bernoulli random variables, which is presently the most popular approach (Young et al., 2020). The difference is essentially between a finite mixture and finite Markov mixture model (Frühwirth-Schnatter, 2006), although we also introduce some spatial dependence in the states by coupling the partially hidden Markov chains between neighboring areas.

There are several alternatives to our coupled Markov chains, within the framework of finite mixture ZIC models (Young et al., 2020), for accounting for spatio-temporal correlations in the presence of the disease, mainly: random effects (Hoef and Jansen, 2007), splines (Ghosal et al., 2020) and autoregression (Fernandes et al., 2009; Yang et al., 2013). The main advantages of our approach are (1) we allow for each covariate, and between area disease spread, to have a separate effect on the reemergence of the disease compared to the persistence, which is epidemiologically justified in many instances and (2) it is much easier to quantify space-time heterogeneity in the effects of disease spread between areas with our approach as we allow the effects of neighboring disease spread to depend on a vector of space-time covariates related to either area. In our application of dengue fever our model allowed for several interesting insights into the epidemiology of the disease beyond what existing ZIC methods can provide, some examples being: dengue is more likely to spread from high population areas controlling for prevalence, if an epidemic of dengue occurs in an area the disease is very likely to spread to neighboring areas, there is typically a larger effect of neighboring dengue spread on the reemergence of the disease compared to the persistence, and rainfall has a larger effect on the persistence of dengue which leads to clear differences in the temporal evolutions of the two probabilities (see Section 3.4 for these and more examples). Additionally, our model fit the dengue data better than a reasonably specified alternative existing ZIC model that combined random effects and autoregression, although there are many possible specifications for such a model.

Although we have applied the ZS-CMSNB model to spatio-temporal infectious disease counts, it could be applied to data in other fields as well. For example, Malyshkina and Mannering

(2010) considered a zero-state Markov switching count model, without coupling, for modeling traffic accidents across 335 highway segments in Indiana between 1995-1999. In their application, the zero state represented a low risk of accidents and the count state represented a high risk of accidents. It could be that a highway segment in the high risk state signals that nearby highway segments are also unsafe, and by coupling the chains between neighboring highways we could borrow strength between them to help determine the states.

There are also some limitations with our approach. We use a negative binomial distribution to account for overdispersion, but other count distributions could be used, such as the generalized Poisson distribution (Joe and Zhu, 2005). The negative binomial distribution places a lot of weight on low counts which is helpful when modeling the reemergence and persistence of a disease, as these events mainly occur at low counts, but has the disadvantage of leading to an overestimation of uncertainty during epidemic periods. Additionally, our approach relies heavily on having good covariates that reflect knowledge of the disease. We could incorporate random effects or splines into $\text{logit}(p01_{it})$ or $\text{logit}(p11_{it})$ to account for additional heterogeneity when known covariates are not sufficient. Finally, we are only explicitly considering disease spread between neighboring areas and, especially in a city, people often move around far outside their neighboring areas. We could allow for disease spread from all areas and incorporate distance into the coupling parameters, although this would likely come at a great computational cost that would need to be overcome to make it a viable approach.

Acknowledgements

This work is part of the PhD thesis of D. Douwes-Schultz under the supervision of A. M. Schmidt in the Graduate Program of Biostatistics at McGill University, Canada. Schmidt is grateful for financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant RGPIN-2017-04999). Douwes-Schultz is grate-

ful for financial support from IVADO and the Canada First Research Excellence Fund / Apogée (PhD Excellence Scholarship 2021-9070375349). This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). The authors thank the Editor, Associate Editor and two reviewers whose comments improved the presentation of the paper.

Chapter 4

A three-state coupled Markov switching model for COVID-19 outbreaks across Quebec based on hospital admissions

Preamble to Manuscript 2. In this second manuscript, we consider coupled Markov switching models where the disease switches between absence, endemic and outbreak periods in each area. The outbreak and endemic periods are described by autoregressive count models distinguished by a higher level of disease transmission during outbreaks. Like in the first manuscript in Chapter 3, the absence state is described by a degenerate zero distribution. Therefore, the models explored in this paper can be seen as three-state extensions of those considered in the first manuscript.

Non-coupled Markov switching models that switch between endemic and outbreak states have been considered since the early 1980s (Souza, 1982; Amorós et al., 2020). By non-coupled we mean the transition probabilities, such as the probability of an outbreak emerging, do not depend on whether outbreaks are occurring in neighboring areas. Due to human movement, the occurrence of outbreaks in neighboring areas should make outbreak emergence more likely

(Grenfell et al., 2001). Also, Markov switching models that only switch between endemic and outbreak states can struggle with capturing many zeroes in the counts (Zou et al., 2014). For our motivating example to Quebec hospitalizations in Section 4.5, we found that the inclusion of neighboring outbreak indicators in the transition probabilities and the addition of an absence state greatly improves the fit, see Table 4.1, and state estimation, see Figures 4.5 and 4.6, of an outbreak/endemic Markov switching model. Additionally, in a simulation study with spatially synchronized outbreaks (see Section 4.4), we found that the inclusion of neighboring outbreak indicators in the transition probabilities substantially improves sensitivity, specificity and timeliness during real-time outbreak detection and forecasting, see Table B.4 and Figure 4.3.

Heaton et al. (2012) (the only coupled endemic/outbreak Markov switching model to the best of our knowledge) allowed the probability of an outbreak emerging to depend on whether outbreaks were occurring in neighboring areas. However, to prevent rapid switching between the states, they assumed an absorbing outbreak state. This means their model can only be applied to data with at most one outbreak in each area. Also, they did not allow the probability of outbreak emergence to depend on covariates. Therefore, their approach has limited applicability. For instance, in our motivating example, there are 3-4 outbreaks in most areas, and we want to investigate how certain factors, such as mobility, were related to the emergence of the outbreaks. In contrast, the coupled models explored in this manuscript can be applied to space-time data with any number of outbreaks in each area. We also allow the transition probabilities to depend on covariates. To prevent rapid switching, we instead consider clone states, popular in econometrics (Kaufmann, 2018), which enforce minimum endemic and outbreak state durations. For our motivating example, we show that the clone states effectively prevent rapid switching between the outbreak and endemic states, see Figure 4.5.

Finally, outside a Markov switching framework, many methods have been proposed for mod-

eling spatio-temporal infectious disease counts, including epidemiological compartmental models (Crawford et al., 2022) and multivariate autoregressive models (Bracher and Held, 2022). We believe two aspects of our Markov switching approach are unique. Firstly, we can quantify associations, through odds ratio, between covariates and important epidemiological transitions, such as outbreak emergence. Secondly, we can provide the posterior probability that an outbreak is occurring or will occur soon for outbreak detection and forecasting respectively. For our motivating example, we compare to a popular alternative model (outside the Markov switching framework) for spatio-temporal infectious disease counts (Bracher and Held, 2022). We show our approach obtains a better fit and provides more useful interpretations, see Table 4.1 and the associated discussion.

In conclusion, in this manuscript, we make three important contributions to the existing (Amorós, 2017) outbreak/endemic Markov switching literature. We include an absent state, to account for many consecutive zeroes; neighboring outbreak indicators along with covariates in the transition probabilities, to account for geographical outbreak spread; and clone states, to prevent rapid switching between the outbreak and endemic states. Also, our model has useful advantages compared to popular approaches that lay outside a Markov switching framework. Therefore, this manuscript represents an important contribution to the literature on spatio-temporal infectious disease modeling.

This manuscript is currently under a second round of reviews at the Annals of Applied Statistics.

A three-state coupled Markov switching model for COVID-19 outbreaks across Quebec based on hospital admissions

Dirk Douwes-Schultz[1], Alexandra M. Schmidt[1], Yannan Shen[1], David Buckeridge[1].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

Abstract

Recurrent COVID-19 outbreaks have placed immense strain on the hospital system in Quebec. We develop a Bayesian three-state coupled Markov switching model to analyze COVID-19 outbreaks across Quebec based on admissions in the 30 largest hospitals. Within each catchment area, we assume the existence of three states for the disease: absence, a new state meant to account for many zeroes in some of the smaller areas, endemic and outbreak. Then we assume the disease switches between the three states in each area through a series of coupled nonhomogeneous hidden Markov chains. Unlike previous approaches, the transition probabilities may depend on covariates and the occurrence of outbreaks in neighboring areas, to account for geographical outbreak spread. Additionally, to prevent rapid switching between endemic and outbreak periods we introduce clone states into the model which enforce minimum endemic and outbreak durations. We make some interesting findings, such as that mobility in retail and recreation venues had a positive association with the development and persistence of new COVID-19 outbreaks in Quebec. Based on model comparison our contributions show promise in improving state estimation retrospectively and in real-time, especially when there are smaller areas and highly spatially synchronized outbreaks. Furthermore, our approach offers new and interesting epidemiological interpretations, such as being able to estimate the effect of covariates on disease extinction.

4.1 Introduction

Quebec has been the epicenter of the COVID-19 epidemic in Canada with 15,389 deaths and 52,788 hospitalizations through May 2022 (INSPQ, 2022). As a result, immense strain has been placed on Quebec's hospital system. Due to the high demand for hospital beds, many elective surgeries have been delayed or canceled which can adversely affect outcomes in patients with conditions other than COVID-19 (Shingler and Hendry, 2022). Additionally, there have been concerning shortages of essential medical supplies, drugs and staff (Laframboise, 2020; Legault and Blais, 2020).

In this paper we analyze weekly COVID-19 hospital admissions in the 30 largest hospitals in Quebec, a map is given in Figure 4.1 along with the time series for each hospital. Figure 4.2(a) shows the weekly COVID-19 hospitalizations for Pavillon Sainte-Marie in Trois–Rivières, an average-sized hospital among the 30. As illustrated in Figure 4.2(a), and this can also be observed less clearly in Figure 4.1 across all hospitals, the hospital admissions can be characterized by a series of outbreak periods separated by quiescent endemic periods with low levels of hospitalizations. The outbreak periods represent COVID-19 outbreaks in the surrounding communities serviced by the hospital (catchment areas) as we do not include hospital-acquired infections. Clearly, most hospitalizations occur during the outbreak periods. However, this does not reflect the true burden of the outbreaks as the rapid rise in hospitalizations at the beginning of an outbreak can be difficult to adjust to, leading to shortages and a lack of beds (Shingler and Hendry, 2022). To gain a better understanding of how the outbreaks develop, a substantive aim of our analysis is to quantify how certain factors, such as mobility and the introduction of new variants (marked in Figure 4.2(a), see Section 4.5.1 for more details), are associated with the emergence and persistence of the outbreaks. Additionally, we want to detect and forecast the outbreaks so that the hospitals can better plan the allocation of their medical resources.

Many methods have been proposed for modeling infectious disease counts in space and time,

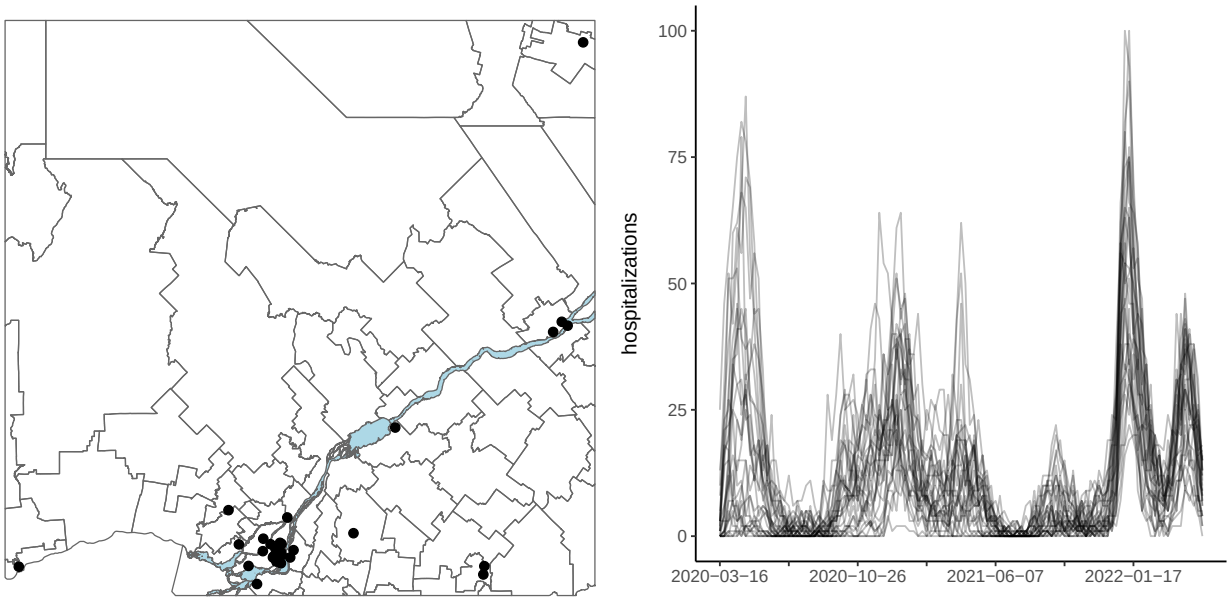


Figure 4.1: The left graph shows a map of the part of Quebec where the 30 hospitals (solid circles) included in the study are located. Borders separate counties. Each line in the right graph gives the number of hospitalizations in one of the 30 hospitals included in the study.

a subject that has gained increased attention in recent years partly due to the COVID-19 pandemic. Among the most popular are epidemiological compartmental models, such as susceptible-infectious-recovered (SIR) models (Keeling and Rohani, 2007; Bjørnstad, 2023). These approaches attempt to model individuals within the population moving through different disease compartments such as susceptible, infectious, recovered and quarantined (Crawford et al., 2022). Multivariate autoregressive count time series models, where the current expected counts are modeled as some function of past counts, are also popular (Ssentongo et al., 2021; Bracher and Held, 2022) and can be motivated from discrete-time SIR models (Bauer and Wakefield, 2018; Mizumoto and Chowell, 2020). When fit with covariates, both of these approaches are often used to investigate associations between certain factors and some measure of disease transmission, e.g., the average number of new cases per each recent previous case (Bauer and Wakefield, 2018; Flaxman et al., 2020; Ssentongo et al., 2021). However, they usually do not clearly distinguish between outbreak and non-outbreak periods. As mentioned above, we are interested in studying the transition between calm endemic

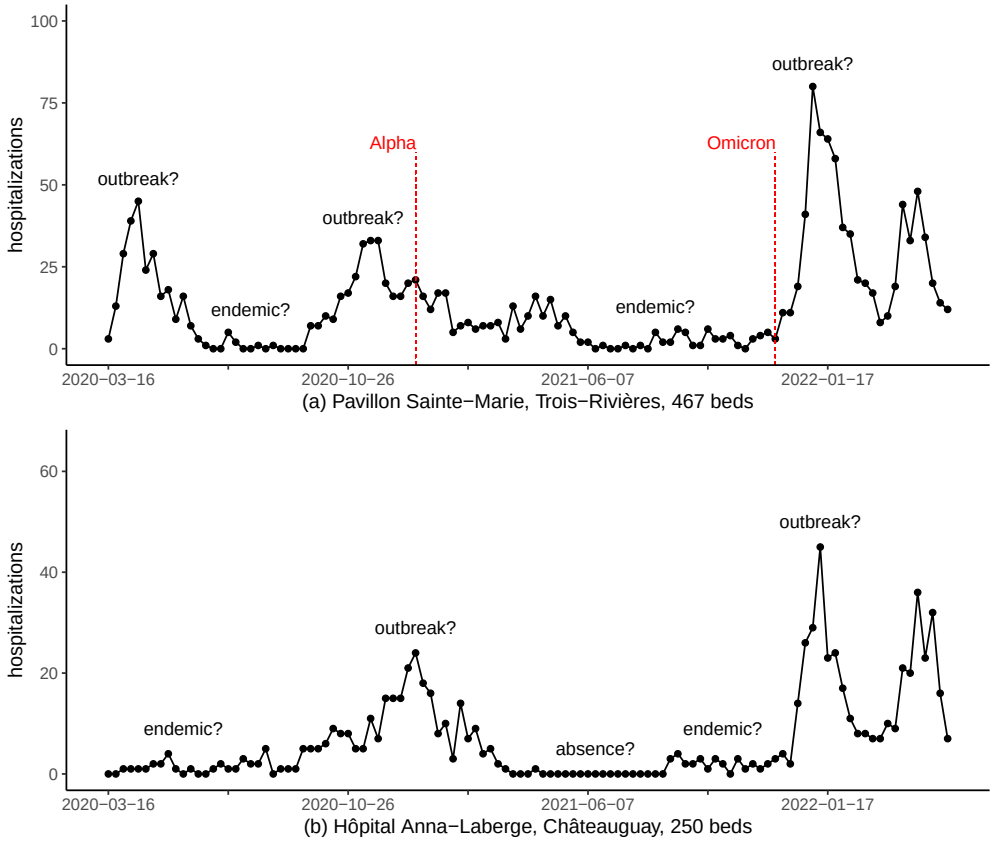


Figure 4.2: (a) An illustration of endemic and outbreak periods in the weekly COVID-19 hospitalizations for Pavillon Sainte-Marie in Trois–Rivières. (b) An illustration of absence, endemic and outbreak periods in the weekly COVID-19 hospitalizations for Hôpital Anna–Laberge in Châteauguay. The vertical red lines are drawn at the introduction of the Alpha and Omicron variants for all of Quebec. The "?" reflects the fact that the periods are not fully observable in our framework.

periods and outbreak periods, and, therefore, we require a precise model definition of those transitions. Also, these methods are not well suited for outbreak detection and forecasting. While a compartmental model could forecast a likely increase in some epidemiological indicator, it cannot distinguish between an increase due to random endemic variation and an actual outbreak developing, which is an important problem in outbreak detection (Unkel et al., 2012). Several other methods that are more focused on producing predictions, such as spline-based (Bauer et al., 2016) or machine learning (Rahimi et al., 2023) methods, are also popular, however, they are often not appropriate for learning about the dynamics of the disease.

In this paper we take a Markov switching approach to modeling the occurrence of the outbreak periods. As we explain below, this approach allows us to study, detect and forecast transitions between endemic and outbreak periods. Markov switching models assume a time series can be described by several submodels, usually called states or regimes, where switching between submodels is governed by a hidden first-order Markov chain (Hamilton, 1989). Markov switching models that switch between endemic and outbreak states have a long history (Souza, 1982; Amorós et al., 2020). In this framework, the transitions between the endemic and outbreak states are usually modeled as a change in the parameters of an autoregressive process, i.e., a change in transmission intensity (Lu et al., 2010). These transitions are assumed unobservable and must be inferred probabilistically from the time series, which is convenient as we usually do not observe exactly when the outbreaks start and end. The probabilities that govern the transitions between the endemic and outbreak states can depend on covariates (Diebold et al., 1994). This allows for investigating the association of various factors with the probabilities of outbreak emergence and persistence and can aid in forecasting the outbreaks (Nunes et al., 2013). Additionally, Bayesian methods can be used to compute the posterior probability an outbreak is currently happening or will happen soon for the purpose of outbreak detection (Martínez-Beneito et al., 2008) and forecasting (Nunes et al., 2013), respectively. Outbreak/endemic Markov switching models have become increasingly popular in epidemiology, especially for outbreak detection (Unkel et al., 2012), with recent applications to influenza (Lytras et al., 2019), cutaneous leishmaniasis (Rahmanian et al., 2021) and salmonella (Zacher and Czogiel, 2022).

Despite their growing popularity, outbreak/endemic Markov switching models have not focused on the analysis of outbreaks in small areas, especially with many zeroes. Figure 4.2(b) shows weekly COVID-19 hospitalizations for the Hôpital Anna–Laberge in Châteauguay, a relatively small hospital. As illustrated in Figure 4.2(b), it may be more appropriate in smaller areas to describe switching between three periods: absence, endemic and outbreak. This has a strong epidemiological justification as it is well known that many infectious dis-

eases frequently go extinct in small communities (Bartlett, 1957; Keeling and Rohani, 2007). Ignoring long periods of disease absence would end up likely assigning too many zeroes to the endemic state, biasing its mean towards zero. This is undesirable as it could lead to the outbreak state being too dominant at medium counts, leading to false alarms (Rath et al., 2003). Another advantage of considering an absence state is that we can model the probabilities of disease extinction and reemergence, which are epidemiologically interesting (Douwes-Schultz and Schmidt, 2022).

There has also not been much work done on spatio-temporal outbreak/endemic Markov switching models, where the focus is on the analysis of outbreaks in several different but connected areas across time. This is a focus of our analysis as many important decisions are made at the individual hospital level and clearly, there is some, though not perfect, synchronization of outbreaks between hospital catchment areas, compare Figures 4.2(a) and 4.2(b) for example. Amorós et al. (2020) fit a spatio-temporal outbreak/endemic Markov switching model, but they only borrowed spatial strength in the observation component of the model. Individuals will mix between areas, causing outbreaks to spread geographically (Grenfell et al., 2001). Therefore, an outbreak should be more likely to emerge or persist in an area if there are outbreaks likely occurring in neighboring areas. One way to achieve this is to use a coupled Markov switching model (Pohle et al., 2021) where the states of neighbors are directly entered into the transition probabilities (Douwes-Schultz and Schmidt, 2022; Touloupou et al., 2020). For outbreak detection and forecasting, considering evidence of outbreaks in neighboring areas could provide an early warning at the very beginning of an outbreak when there is still uncertain evidence within the area. Heaton et al. (2012) did let the probabilities of outbreak emergence depend on outbreaks in neighboring areas, however, they did not allow the transition probabilities to depend on covariates, and they used an absorbing state model, meaning it is difficult to apply their model to time series that contain multiple outbreaks.

In our framework, we assume the disease switches between three states in each area: absence, endemic and outbreak. The epidemiological count in the absence state is assumed to always be 0, while it follows two autoregressive negative binomial processes in the outbreak and endemic states, distinguished by a lower level of transmission during the endemic period. Switching between the three states is governed by a first-order Markov chain where we focus on modeling the following four transition probabilities: absence to endemic (disease emergence), endemic to absence (disease extinction), endemic to outbreak (outbreak emergence) and outbreak to outbreak (outbreak persistence). Each transition probability can depend on covariates as well as a weighted sum of outbreak occurrence in neighboring areas. This allows us to investigate associations with important epidemiological transitions while also incorporating outbreak spread between areas. Nonhomogeneous Markov switching models can be sensitive to overfitting as the transition matrix may be non-persistent at some levels of the covariates, potentially leading to rapid switching between the states. Clearly, it is not realistic to rapidly switch between outbreak and endemic states, so we additionally introduce clone states (Kaufmann, 2018) into the model to enforce a minimum endemic and outbreak duration.

This paper is structured as follows. In Section 4.2 we introduce our proposed model, a three-state coupled Markov switching model. In Section 4.3 we describe Bayesian inference using data augmentation and make use of the individual forward filtering backward sampling algorithm of Touloupou et al. (2020). In Section 4.4 we evaluate outbreak detection and forecasting results from the model on simulated data where the exact start and end times of the outbreaks are known. In Section 4.5 we apply the model to COVID-19 outbreaks across Quebec based on admissions in the 30 largest hospitals. We close with a general discussion in Section 4.6.

4.2 A Three-state Coupled Markov Switching Model

Let y_{it} be an epidemiological count indicator, e.g., counts of hospitalizations, associated with area $i = 1, \ldots, N$ across $t = 1, \ldots, T$ time periods. Let $S_{it} \in \{1, 2, 3\}$ be an indicator for the epidemiological state of the disease, where $S_{it} = 1$ if the disease is absent in area i during time t, $S_{it} = 2$ if the disease is in an endemic state and $S_{it} = 3$ if the disease is in an outbreak state.

We assume that the epidemiological count indicator is always 0 when the disease is absent and follows two distinct autoregressive negative binomial processes in the endemic and outbreak states, that is,

$$y_{it} \mid S_{it}, y_{i(t-1)} \sim \begin{cases} 0, & \text{if } S_{it} = 1 \text{ (absence)} \\ NB(\lambda_{it}^{EN}, r^{EN}), & \text{if } S_{it} = 2 \text{ (endemic)} \\ NB(\lambda_{it}^{OB}, r^{OB}), & \text{if } S_{it} = 3 \text{ (outbreak)}, \end{cases} \tag{4.1}$$

where λ_{it}^{EN} and λ_{it}^{OB} are the means in the endemic and outbreak states respectively and r^{EN} and r^{OB} are the overdispersion parameters, so that, for example, $\text{Var}(y_{it}|S_{it} = 3, y_{i(t-1)}) = \lambda_{it}^{OB}(1+\lambda_{it}^{OB}/r^{OB})$. A zero count could be produced by all three states while a positive count can be produced by either the endemic or outbreak states, therefore, none of the states are observable and S_{it} is a latent variable. Intuitively, we often do not know if an outbreak is occurring and there could be zeroes during an endemic period, or early outbreak period, due to a failure to detect the disease or a lack of severe cases for hospitalizations and deaths. As justification for the negative binomial distribution in (4.1), over a Poisson distribution, we found there was a high amount of overdispersion in the Quebec hospitalizations, see Table 4.3.

As we are modeling an infectious disease we would expect the previous count $y_{i(t-1)}$ to affect the current expected count when the disease is present (Bauer and Wakefield, 2018).

Therefore, we use log-linear autoregressive models (Liboschik et al., 2017) for λ_{it}^{EN} and λ_{it}^{OB},

$$\begin{aligned}\log(\lambda_{it}^{EN}) &= \beta_{0i}^{EN} + \boldsymbol{x}_{it}^T\boldsymbol{\beta}^{EN} + \rho^{EN}\log(y_{i(t-1)}+1)\\ \log(\lambda_{it}^{OB}) &= \beta_{0i}^{OB} + \boldsymbol{x}_{it}^T\boldsymbol{\beta}^{OB} + \rho^{OB}\log(y_{i(t-1)}+1),\end{aligned} \tag{4.2}$$

where $\beta_{0i}^{EN} \sim N\left(\beta_0^{EN}, \left(\sigma^{EN}\right)^2\right)$ and $\beta_{0i}^{OB} \sim N\left(\beta_0^{OB}, \left(\sigma^{OB}\right)^2\right)$ are random intercepts meant to account for between area differences and $\boldsymbol{x}_{it}$ is a vector of space-time covariates that may affect transmission of the disease within the endemic and outbreak periods. We allow the covariate effects, $\boldsymbol{\beta}^{EN}$ and $\boldsymbol{\beta}^{OB}$, to be different as outbreaks can lead to behavioral changes in hosts (Verelst et al., 2016). We also considered spatially correlated random effects (Amorós et al., 2020) for the area specific intercepts in (4.2). However, we found for our motivating example that, possibly due to the amount of uncertainty about the underlying disease states, the spatial association of the intercepts could not be estimated precisely enough.

To model the switching between absence, endemic and outbreak periods we assume that S_{it} follows a three-state nonhomogeneous Markov chain within each area. In order to incorporate outbreak spread between areas we condition the transition matrix on $\boldsymbol{S}_{(-i)(t-1)} = (S_{1(t-1)}, \ldots, S_{(i-1)(t-1)}, S_{(i+1)(t-1)}, \ldots, S_{N(t-1)})^T$, the vector of all state indicators excluding area i at time $t-1$. We propose the following conditional transition matrix for the Markov chain, for $t = 2, \ldots, T$,

$$\Gamma\left(S_{it}|\boldsymbol{S}_{(-i)(t-1)}\right) =$$

State	S_{it}=1 (absence)	S_{it}=2 (endemic)	S_{it}=3 (outbreak)
$S_{i(t-1)}$=1 (absence)	$1-p12_{it}$	$p12_{it}$	0
$S_{i(t-1)}$=2 (endemic)	$p21_{it}$	$1-p21_{it}-p23_{it}$	$p23_{it}$
$S_{i(t-1)}$=3 (outbreak)	0	$1-p33_{it}$	$p33_{it}$

(4.3)

where $\Gamma\left(S_{it}|\boldsymbol{S}_{(-i)(t-1)}\right)_{lk} = P(S_{it}=k|S_{i(t-1)}=l, \boldsymbol{S}_{(-i)(t-1)})$ for $l,k = 1,2,3$ and we have the

following epidemiological interpretations of the transition probabilities,

$p12_{it} =$ probability of disease emergence, $\quad p21_{it} =$ probability of disease extinction,

$p23_{it} =$ probability of outbreak emergence, $\quad p33_{it} =$ probability of outbreak persistence.

From (4.3), we assume it is not possible to move from absence to outbreak and vice versa in a single time step. For a time step of one week, like in our application, this is reasonable. However, for longer time steps this assumption should be examined.

We assume each transition probability in (4.3) can depend on a p-dimensional vector of space-time covariates $\boldsymbol{z}_{it}$ as well as a weighted sum of outbreak occurrence in neighboring areas during the previous time period, to model outbreak spread between areas. Starting with the probabilities of disease emergence, $p12_{it}$, and outbreak persistence, $p33_{it}$, we let

$$\text{logit}(plk_{it}) = \alpha_{lk,0} + \boldsymbol{z}_{it}^T\boldsymbol{\alpha}_{lk} + \alpha_{lk,p+1}\sum_{j\in NE(i)}\omega_{ji}I[S_{j(t-1)} = 3], \tag{4.4}$$

for $lk = 12, 33$, where $I[\bullet]$ is an indicator function, $NE(i)$ is the set of all neighboring areas of area i and ω_{ji} is a fixed weight meant to reflect *a priori* knowledge of the level of influence area j has on area i. For example, one could use distance-based weights or weights based on trade between regions, see Schrödle et al. (2012). It is also possible to estimate connectivity in a coupled Markov switching model (Douwes-Schultz and Schmidt, 2022). However, we do not consider this here due to the complexity and number of transitions in our model, the model used in Douwes-Schultz and Schmidt (2022) only had one general presence state and one absence state. For the probabilities of disease extinction, $p21_{it}$, and outbreak emergence, $p23_{it}$, a multinomial logistic regression is needed so that the second row of (4.3) sums to 1, that is,

$$\log\left(\frac{plk_{it}}{1 - p21_{it} - p23_{it}}\right) = \alpha_{lk,0} + \boldsymbol{z}_{it}^T\boldsymbol{\alpha}_{lk} + \alpha_{lk,p+1}\sum_{j\in NE(i)}\omega_{ji}I[S_{j(t-1)} = 3], \tag{4.5}$$

for $lk = 21, 23$. Therefore, $\boldsymbol{\alpha}_{21}$ and $\boldsymbol{\alpha}_{23}$ in (4.5) represent the effects of the covariates on the relative odds of transitioning to the absence and outbreak states respectively compared to remaining in the endemic state.

Covariates will likely affect the transition probabilities in most cases. For example, dengue outbreaks are unlikely to occur during the winter when mosquito activity is low (Descloux et al., 2012), the introduction of new variants, such as Omicron, has likely played an important role in the development of new COVID-19 waves (Maslo et al., 2022) and disease extinction is less likely in smaller communities (Bartlett, 1957). As for the inclusion of neighboring outbreak indicators in (4.4)-(4.5), individuals will mix with those in other areas which can cause outbreaks to spread geographically, a phenomenon known as traveling waves (Grenfell et al., 2001). Therefore, an outbreak in a neighboring area should affect the transition probabilities due to either direct spread or because it indicates spread from a common source, such as a large city.

4.2.1 Identifiability constraints

Constraints are often placed on the parameters of a Bayesian mixture model to remove multimodality in the posterior distribution (Frühwirth-Schnatter, 2006), which often occurs genuinely and due to the invariance of the likelihood function to permutations in the state labeling of the parameters (Jasra et al., 2005). The constraints are commonly chosen based on knowledge of the process being modeled (Martínez-Beneito et al., 2008; Stoner and Economou, 2020). To help motivate constraints in our case, note that another way to express a log-linear model in (4.2) is as, for example, $\lambda_{it}^{OB} = \exp(\beta_{0i}^{OB} + \boldsymbol{x}_{it}^T \boldsymbol{\beta}^{OB}) \left(y_{i(t-1)} + 1\right)^{\rho^{OB}}$, which takes the form of a transmission rate multiplied by the previous counts. Disease transmission should always increase when moving from the endemic state to the outbreak state, especially if the model is to be used for issuing alarms for surveillance. Additionally, the autoregressive coefficients, ρ^{EN} and ρ^{OB}, control the speed the disease moves through the local population (Wakefield et al., 2019), which should be higher during the outbreak periods.

Therefore, we assume the following constraints,

$$\begin{aligned} &\beta_{0i}^{EN} + \boldsymbol{x}_{it}^T\boldsymbol{\beta}^{EN} + .01 < \beta_{0i}^{OB} + \boldsymbol{x}_{it}^T\boldsymbol{\beta}^{OB} \text{ for } i = 1, \ldots, N \text{ and } t = 2, \ldots, T, \text{ and,} \\ &\rho^{EN} + .05 < \rho^{OB}, \end{aligned} \tag{4.6}$$

meaning we are assuming at least a one percent increase in transmission when moving from the endemic state to the outbreak state (in Section 4.5.2 we show the results are not very sensitive to reasonably changing this one percent value). Initially, we had considered a simpler constraint on just the intercepts and the autoregressive coefficients, however, we found in a simulation study, in Section B.2 of the supplementary materials (SM), that the simpler constraint does not ensure consistent convergence of our Markov chain Monte Carlo (MCMC) algorithm. Constraining the entirety of the transmission rates, as in (4.6), greatly improved the convergence rate, did not introduce any significant bias into the inferential procedure and did not add a substantial amount of time to the model fitting.

4.2.2 Prior specification

For the count part of the model, we specified wide independent normal and gamma priors for most lower-level elements of $\boldsymbol{\beta}$. We used Unif$(0, 1)$ priors for ρ^{EN} and ρ^{OB} to meet the stability conditions of Liboschik et al. (2017) which ensures the counts are not expected to grow without bound. For r^{EN} and r^{OB} we used Unif$(0, 10)$ and Unif$(0, 50)$ priors respectively. An upper limit of 10 was chosen for r^{EN} as it is important for the endemic state distribution to have a long right tail to prevent frequent false alarms during outbreak detection (Rath et al., 2003). To impose the constraints in (4.6) we truncated the prior distribution of $\boldsymbol{\beta}$ (Kaufmann, 2018). Some shrinkage to the null is generally recommended for logistic and multinomial logistic regression parameters to avoid separation issues and reduce bias away from 0 (Bull et al., 2002). Therefore, following Gelman et al. (2008), we used Cauchy priors with scale $2.5/2 \cdot \text{sd}(z_{itq})$ for the effects of covariate z_{itq} on the transition probabilities in

(4.4)-(4.5). For the effects of neighboring outbreaks on the transition probabilities, i.e., $\alpha_{lk,p+1}$ for $lk = 12, 21, 23, 33$ in (4.4)-(4.5), we used more aggressive shrinkage and assigned $N(0, (.36/\max\{\omega_{ji}\}_{\substack{j,i=1 \\ j\neq i}}^{N})^2)$ priors. This states that with probability .95 *a priori* we believe that an outbreak occurring in a single neighboring area should not more than double or less than halve the odds, or relative odds, of any epidemiological transition, e.g., should not more than double the odds of an outbreak emerging relative to remaining in the endemic state (Wakefield, 2013). Our reasoning is that we do not want a single area to be given too much influence as we want the model to consider evidence of outbreaks across multiple neighboring areas. Additionally, while borrowing spatial strength can be important, we do not want it to overpower within-area information too strongly.

4.2.3 Clone states

As the transition probabilities in (4.3) depend both on covariates and latent neighboring states, the transition matrix may fluctuate between persistence and non-persistence. This could lead to time periods where there is rapid switching between the outbreak and endemic states, which is not realistic. Heaton et al. (2012) dealt with this by using an absorbing state model where the probability of outbreak persistence was fixed at 1, but that does not allow the analysis of multiple outbreaks or the study of outbreak persistence. Our solution is to introduce clone states into the model with determined transitions to enforce a minimum endemic and outbreak state duration, which is common in econometrics (Kaufmann, 2018). We can introduce a new latent state indicator $S_{it}^* \in \{1, 2, 3, 4, 5, 6, 7\}$ such that $S_{it} = 1$ if $S_{it}^* = 1$, $S_{it} = 2$ if $S_{it}^* \in \{2, 3\}$ and $S_{it} = 3$ if $S_{it}^* \in \{4, 5, 6, 7\}$, and with the following

conditional transition matrix, for $t = 2, \ldots, T$,

$$\Gamma\left(S^*_{it}|\boldsymbol{S}_{(-i)(t-1)}\right) =$$

State	S^*_{it}=1	S^*_{it}=2	S^*_{it}=3	S^*_{it}=4	S^*_{it}=5	S^*_{it}=6	S^*_{it}=7
$S^*_{i(t-1)}$=1, $S_{i(t-1)}$=1 **(absence)**	$1 - p12_{it}$	$p12_{it}$	0	0	0	0	0
$S^*_{i(t-1)}$=2, $S_{i(t-1)}$=2 **(endemic)**	0	0	1	0	0	0	0
$S^*_{i(t-1)}$=3, $S_{i(t-1)}$=2 **(endemic)**	$p21_{it}$	0	$1 - p21_{it} - p23_{it}$	$p23_{it}$	0	0	0
$S^*_{i(t-1)}$=4, $S_{i(t-1)}$=3 **(outbreak)**	0	0	0	0	1	0	0
$S^*_{i(t-1)}$=5, $S_{i(t-1)}$=3 **(outbreak)**	0	0	0	0	0	1	0
$S^*_{i(t-1)}$=6, $S_{i(t-1)}$=3 **(outbreak)**	0	0	0	0	0	0	1
$S^*_{i(t-1)}$=7, $S_{i(t-1)}$=3 **(outbreak)**	0	$1 - p33_{it}$	0	0	0	0	$p33_{it}$

(4.7)

where $\Gamma\left(S^*_{it}|\boldsymbol{S}_{(-i)(t-1)}\right)_{lk} = P(S^*_{it} = k|S^*_{i(t-1)} = l, \boldsymbol{S}_{(-i)(t-1)})$ for $l, k = 1, \ldots, 7$. The new transition matrix (4.7) will prevent rapid switching between endemic and outbreak periods by imposing a minimum endemic duration of 2 weeks and a minimum outbreak duration of 4 weeks for our motivating example. A COVID-19 outbreak that lasts less than 4 weeks in Quebec is likely either a false alarm or too small to be concerning. Also, there should be at least some time between outbreaks for the susceptible population to replenish (Keeling and Rohani, 2007). Clearly, the idea of (4.7) could be used to impose any arbitrary minimum state durations, and so we assume (4.7) throughout the rest of the paper w.l.o.g..

We will refer to the model defined by (4.1)-(4.5), with (4.3) replaced by (4.7), as the coupled Markov switching negative binomial model with 1 absence state, 2 endemic states and 4 outbreak states, i.e., the CMSNB(1,2,4) model. To finish model specification we also need to specify an initial state distribution for the Markov chain in each area, i.e., $p(S^*_{i1})$ for $i = 1, ..., N$, which we assume does not depend on any unknown parameters. Note that moving from (4.3) to (4.7) does not add any new parameters to the model, however, the

restrictive transition matrix does slightly complicate the inferential procedure, which will now be discussed.

4.3 Inferential Procedure

Let $\boldsymbol{S}^* = (S^*_{11}, \ldots, S^*_{1T}, \ldots, S^*_{N1}, \ldots, S^*_{NT})^T$ be the vector of all state indicators, let $\boldsymbol{y} = (y_{11}, \ldots, y_{1T}, \ldots, y_{N1}, \ldots, y_{NT})^T$ be the vector of all counts, let $\boldsymbol{\beta}$ be the vector of all model parameters in the count part of the model, i.e., parameters in (4.1)-(4.2), let $\boldsymbol{\theta}$ be the vector of all model parameters in the Markov chain part of the model, i.e., parameters in (4.4)-(4.5), and, finally, let $\boldsymbol{v} = (\boldsymbol{\beta}, \boldsymbol{\theta})^T$ be the vector of all model parameters. Then the likelihood of $\boldsymbol{v}$ given $\boldsymbol{y}$ and $\boldsymbol{S}^*$ is given by,

$$\mathrm{L}(\boldsymbol{y}, \boldsymbol{S}^* | \boldsymbol{v}) = \prod_{i=1}^{N} \prod_{t=2}^{T} p(y_{it} | S_{it}, y_{i(t-1)}, \boldsymbol{\beta}) \prod_{i=1}^{N} p(S^*_{i1}) \prod_{t=2}^{T} p(S^*_{it} | S^*_{i(t-1)}, \boldsymbol{S}_{(-i)(t-1)}, \boldsymbol{\theta}). \tag{4.8}$$

Recall from the previous section that $\boldsymbol{S}^*$ is not observed. It is not possible to marginalize out $\boldsymbol{S}^*$ from (4.8) as doing so would require matrix multiplication with a $7^N \times 7^N$ matrix (Douwes-Schultz and Schmidt, 2022). Additionally, we want to make inferences about $\boldsymbol{S}^*$ for the purpose of outbreak detection and forecasting as well as historical retrospection. Therefore, we estimate $\boldsymbol{S}^*$ along with $\boldsymbol{v}$ by sampling both from their joint posterior distribution which, from Bayes' theorem, is proportional to,

$$p(\boldsymbol{S}^*, \boldsymbol{v} | \boldsymbol{y}) \propto \mathrm{L}(\boldsymbol{y}, \boldsymbol{S}^* | \boldsymbol{v}) p(\boldsymbol{v}), \tag{4.9}$$

where $p(\boldsymbol{v})$ is the prior distribution of $\boldsymbol{v}$.

As the joint posterior (4.9) is not available in closed form, we resort to MCMC methods, in particular, we used a hybrid Gibbs sampling algorithm with some steps of the Metropolis–Hastings algorithm to sample from it. We sampled all elements of $\boldsymbol{v}$ without conjugate priors individually, using an adaptive random walk Metropolis step (Shaby and Wells,

2010). It is easy to implement the sampling of each element of $\boldsymbol{S}^*$ one-at-a-time from $p(S^*_{it}|\boldsymbol{y}, \boldsymbol{v}, \{S^*_{jl}\}_{jl\neq it})$ (Douwes-Schultz and Schmidt, 2022). However, we found that one-at-a-time sampling mixed so slowly that it is not usable with our model. To see this, consider the following hypothetical state sequence for S^*_{it}, 456723. Note it is not possible to sample any new single element of this sequence and in general we found that one-at-time sampling gets stuck in small regions of the parameter space. To avoid issues like this with one-at-a-time sampling when fitting Markov switching models, Chib (1996) proposed to sample all of $\boldsymbol{S}^*$ jointly from $p(\boldsymbol{S}^*|\boldsymbol{v}, \boldsymbol{y})$. However, this is not possible with our model as it would involve matrix multiplication with $7^N \times 7^N$ matrix (Douwes-Schultz and Schmidt, 2022). As an alternative, we can block sample $\boldsymbol{S}^*$ with each block containing all the state indicators in a single location (Touloupou et al., 2020). Let $\boldsymbol{S}^*_i = (S^*_{i1}, \ldots, S^*_{iT})^T$ denote the vector of all state indicators in area i and let $\boldsymbol{S}^*_{(-i)}$ be $\boldsymbol{S}^*$ with $\boldsymbol{S}^*_i$ removed. Then we can sample all of $\boldsymbol{S}^*_i$ jointly from its full conditional distribution, which is given by,

$$p(\boldsymbol{S}^*_i|\boldsymbol{v}, \boldsymbol{S}^*_{(-i)}, \boldsymbol{y}) = p(S^*_{iT}|\boldsymbol{S}^*_{(-i)}, \boldsymbol{y}, \boldsymbol{v}) \prod_{t=1}^{T-1} p(S^*_{it}|S^*_{i(t+1)}, \boldsymbol{S}^*_{(-i)(1:t+1)}, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}), \tag{4.10}$$

using an individual forward filter backward sampling (iFFBS) algorithm (Touloupou et al., 2020). More details are given in SM Section B.1. It is also possible to block sample $\boldsymbol{S}^*$ in multi-location blocks (Douwes-Schultz and Schmidt, 2022) but we do not consider this here as it does not scale well with large transition matrices.

Our hybrid Gibbs sampler was implemented using the R package Nimble (de Valpine et al., 2017). Nimble comes with built-in Metropolis–Hastings and categorical (equivalent to one-at-a-time sampling) samplers. The iFFBS samplers were implemented using Nimble's custom sampler feature. All Nimble R code, including for the custom iFFBS samplers, are provided on GitHub (`https://github.com/Dirk-Douwes-Schultz/CMSNB124_code`). Nimble was chosen as it is extremely fast (C++ compiled) and only requires the coding of new samplers. In the SM Section B.2, we provide a simulation study, which shows that our proposed

Gibbs sampler can recover the true parameters of a CMSNB(1,2,4) model that is specified like in our motivating example in Section 4.5.

4.3.1 Outbreak detection, forecasting and historical retrospection

Once a sample from the joint posterior (4.9) has been obtained, the posterior probability that the disease is currently in epidemiological state s, for $s = 1$ (absence), $s = 2$ (endemic) and $s = 3$ (outbreak), in area i can be approximated as $P(S_{iT} = s|\boldsymbol{y}) \approx \frac{1}{Q-M}\sum_{m=M+1}^{Q} I[S_{iT}^{[m]} = s]$, where the superscript $[m]$ denotes a draw from the posterior distribution of the variable, M is the size of the burn-in sample and Q is the total MCMC sample size. The posterior probability $P(S_{iT} = 3|\boldsymbol{y})$ represents the probability that an outbreak is currently happening in area i given all observed data and can be used for the purpose of outbreak detection (Martínez-Beneito et al., 2008). We recommend using $P(S_{iT} = 3|\boldsymbol{y})$ along with other, external, pieces of information to help determine if an outbreak is likely occurring. While $P(S_{iT} = 3|\boldsymbol{y})$ has no closed form, we can condition on the parameters and previous states to gain some intuition into how the model performs outbreak detection,

$$p(S_{iT}|S^*_{i(T-1)}, \boldsymbol{S}_{(-i)(T-1)}, \boldsymbol{v}, \boldsymbol{y}) \propto p(y_{iT}|S_{iT}, y_{i(T-1)}, \boldsymbol{\beta})p(S_{iT}|S^*_{i(T-1)}, \boldsymbol{S}_{(-i)(T-1)}, \boldsymbol{\theta}). \tag{4.11}$$

From (4.11), the model generally weighs two factors for outbreak detection, the relative likelihood that the outbreak state generated the observed count and the probability of entering the outbreak state at the current time $P(S_{iT} = 3|S^*_{i(T-1)}, \boldsymbol{S}_{(-i)(T-1)}, \boldsymbol{\theta})$, which recall may depend on outbreaks in neighboring areas and covariates, see (4.4)-(4.5) and (4.7).

As for forecasting, we used a simulation procedure to draw realizations from the posterior predictive distributions (Frühwirth-Schnatter, 2006). Algorithm 2 in the SM will obtain realizations from the posterior predictive distribution of the counts, $y^{[m]}_{i(T+k)} \sim p(y_{i(T+k)}|\boldsymbol{y})$, and the epidemiological state of the disease, $S^{[m]}_{i(T+k)} \sim p(S_{i(T+k)}|\boldsymbol{y})$, for $i = 1, \ldots, N$, $k =$

$1, \ldots, K$ and $m = M+1, \ldots, Q$. Then, the posterior probability that the disease will be in state s, k time steps from now, in area i, can be approximated as $P(S_{i(T+k)} = s|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} I[S_{i(T+k)}^{[m]} = s]$.

Finally, it is also important to examine the posterior probability that the disease was in state s during past time periods $P(S_{it} = s|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} I[S_{it}^{[m]} = s]$ for $t = 1, \ldots, T-1$, $i = 1, \ldots, N$ and $s = 1$ (absence), $s = 2$ (endemic) and $s = 3$ (outbreak) for several reasons. Firstly, it may be of historical interest to investigate the epidemiological history of the disease in various areas. Also, the estimation of $\boldsymbol{v}$ depends heavily on the estimates of the past states as indicated by the joint likelihood function (4.8). If the classification of past epidemiological periods is not sensible, this would cast doubt on the estimates of $\boldsymbol{v}$ and may point to model misspecification.

4.3.2 Model comparison

To compare competing models, we use the widely applicable information criterion (WAIC) (Gelman et al., 2014). For state-space models, the WAIC is more accurate when the latent states are marginalized (Auger-Méthé et al., 2021). For models without neighboring outbreak indicators in the transition probabilities, such as the Non-coupled Model in Section 4.5, we can use the forward filter (Frühwirth-Schnatter, 2006) to calculate the marginal density $p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})$, where $\boldsymbol{y}_{1:(t-1)} = (y_{11}, \ldots, y_{1(t-1)}, \ldots, y_{N1}, \ldots, y_{N(t-1)})^T$, and use it to calculate the WAIC. However, for models with neighboring outbreak indicators in the transition probabilities it is not computationally possible to completely marginalize $\boldsymbol{S}^*$ (Douwes-Schultz and Schmidt, 2022) and, therefore, we use the partially marginalized density $p(y_{it}|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})$ instead. See SM Section B.3.1 for more details. The model with the lowest WAIC is considered to have the best fit and as a rule of thumb, a difference of 5 or more in the WAIC is considered significant. A simulation study we conducted, in SM Section B.3.2, suggests that when there is no significant difference in the WAIC to prefer the less complex model and when there is a significant difference the WAIC almost always

chooses the correct model, between a spatial and non-spatial Markov switching model.

4.4 Simulation Study

In Sections 4.5.3 and 4.5.4 below we attempt to evaluate the retrospective and real-time state estimates of a CMSNB(1,2,4) model (see Section 4.3.1) fitted to the Quebec hospitalizations. In this case, it is difficult to exactly quantify the accuracy of the state estimates as the true underlying states are not known, and we can only evaluate the estimates visually. One solution is to design a simulation study where the exact start and end times of the outbreaks are pre-determined, although this does have the disadvantage of using simulated and not real data (Buckeridge, 2007).

We designed a simulation study with 30 areas divided into 5 clusters of 6 areas each. Each area alternated between a 15-week endemic period and a 15-week outbreak period for a total of 4 of each period, 120 weeks total. For every area in a cluster, the start of each outbreak was randomized to occur within the first 4 weeks of the corresponding outbreak period. Therefore, we did not explicitly model outbreak spread within a cluster, we just assumed the outbreaks occurred around the same time. We also assumed a 40% chance, taken roughly from our results in Section 4.5.2, that each endemic period contains a 7-week absence period inserted into the middle. The counts were simulated from a negative binomial distribution with overdispersion 10 and conditional means $e^{.1\text{beds}_i}(y_{i(t-1)})^{.5}$ (endemic), $e^{.75+.05\text{beds}_i}(y_{i(t-1)})^{.75}$ (outbreak) and $y_{it} = 0$ (absence). The covariate beds_i was taken from our real data example to produce some between-area heterogeneity in the transmission. These values are somewhat similar to those from Section 4.5.2 and produced realistic looking simulations, SM Figure B.4 shows some of the simulated time series. We only generated one simulation as each model must be fit 20 times to produce the needed state estimates, see below.

To fit to the simulated data, we considered a CMSNB(1,2,4) model with the count part correctly specified, including no random intercepts, and $\sum_{j\in NE(i)} I[S_{j(t-1)} = 3]$ included

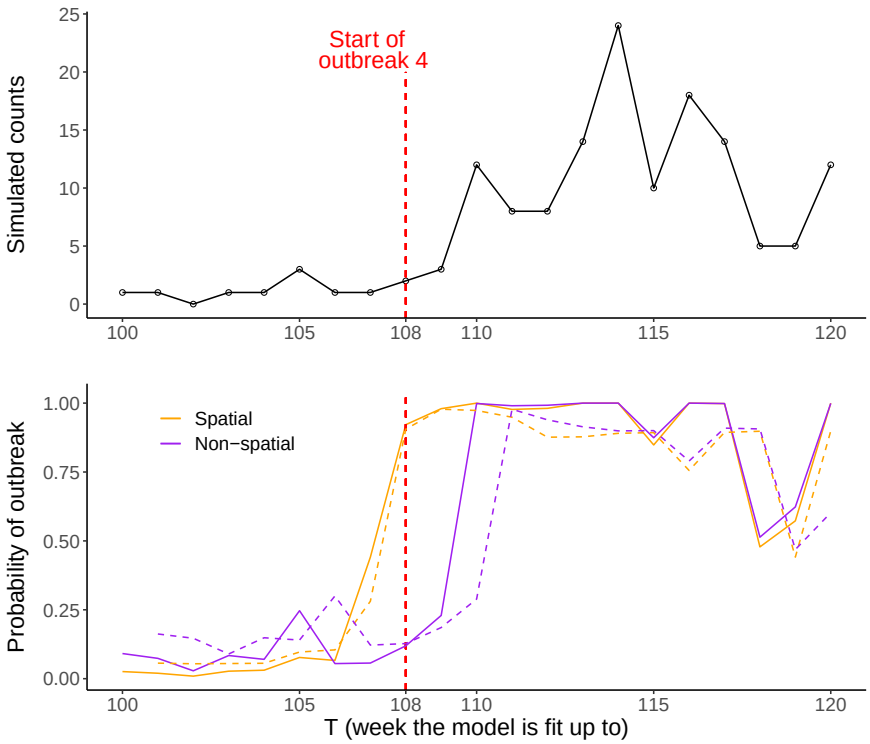


Figure 4.3: Top graph shows simulated counts from the last 20 weeks in one of the areas from the simulation study. Bottom graph solid lines show the real-time outbreak detection probabilities, $P(S_{iT} = 3|\boldsymbol{y})$, versus T. Bottom graph dashed lines show the one-week ahead outbreak forecasts from the previous week, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$, versus T. The red dotted line indicates the exact pre-determined start time of outbreak 4 in the area. The Spatial Model is in orange and the Non-spatial Model is in purple.

only in the relative odds of outbreak emergence, where $NE(i)$ contained all areas in the same cluster as area i. We will call this the Spatial Model, and for comparison purposes, we also considered a homogeneous non-spatial model without $\sum_{j \in NE(i)} I[S_{j(t-1)} = 3]$ in the transition probability. We focused on outbreak emergence in the simulation study, as it is the most important transition to capture in a surveillance setting. We evaluated three outbreak state estimates from the models: (1) retrospective probabilities, $P(S_{it} = 3|\boldsymbol{y})$ from fitting the models to the full simulated data set (2) real-time detection probabilities, $P(S_{iT} = 3|\boldsymbol{y})$ from fitting the models up to time T for $T = 100, \ldots, 120$ (20 separate fits) and (3) real-time one week ahead forecasts, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$ from fitting the models up to time $T-1$ for $T = 101, \ldots, 120$. We only evaluated the real-time state estimates on the last outbreak in each area to ensure stable model fitting. We evaluated the outbreak state estimates using

the area under the ROC curve (AUC), sensitivity, specificity and timeliness (Buckeridge, 2007), the last three calculated with a 50% threshold. Timeliness for a single outbreak is simply the number of weeks into the outbreak, starting at one, when the outbreak state estimate first rises above 50%, and the overall timeliness is the average timeliness across all outbreaks.

The results are summarized in Table B.4 of the SM. The Spatial Model was able to recover the true underlying outbreak states well retrospectively, AUC of .995, and during outbreak detection, AUC of .983, and one week ahead forecasting, AUC of .965. The Spatial Model improved, over the Non-spatial Model, every criterion for each type of outbreak state estimate, especially for real-time outbreak detection and forecasting. For instance, the Spatial Model improved the sensitivity of the real-time outbreak detection probabilities by 6% and reduced the timeliness of the one-week ahead outbreak forecasts by one week. Additionally, the Spatial Model showed a fair tradeoff between outbreak detection and forecasting, with the outbreak forecasts having two percent lower sensitivity but an improvement in timeliness of .43 weeks. In contrast, the forecasts from the Non-spatial Model reduced the sensitivity by 6.7% and only improved the timeliness by .06 weeks. This suggests outbreak forecasting should only be attempted by nonhomogeneous models. These points are further illustrated in Figure 4.3 which compares the real-time outbreak detection probabilities and the one-week ahead outbreak forecasts in one of the areas from the simulation study.

4.5 Application to COVID-19 Outbreaks Across Quebec

4.5.1 Model specification and fitting

We fitted a CMSNB(1,2,4) model to weekly COVID-19 hospitalizations across the 30 largest hospitals in Quebec ($i = 1, \ldots, N = 30$), in terms of overall COVID-19 admissions, between March 16th 2020 and May 9th 2022 ($t = 1, \ldots, T = 113$). Hospitalizations typically lag infections by 1–2 weeks (Ward and Johnsen, 2021), and there are a few days of reporting

delay. We did not include hospital-acquired infections so that the outbreak periods represent outbreaks in the catchment areas of the hospitals. A hospital's catchment area is typically difficult to exactly determine and can stretch disjointly across a broad geographical region (Gilmour, 2010). By examining the locations of a sample of patients from each hospital, we found that the catchment areas are mostly contained within, and sometimes adjacent to, the same county where the hospital is located. We assumed that the number of beds in a hospital is a reasonable proxy for the population size of the catchment area.

Increased mobility has been a major concern of the Quebec government during the epidemic due to a fear that it will lead to increased COVID-19 transmission, potentially overwhelming hospitals (Wilton, 2020). In general, epidemiologists have long theorized that mobility plays an important role in the development of recurring outbreaks (Soper, 1929). While quantifying mobility is challenging, Google mobility metrics (Google LLC, 2022) tend to give a reasonable approximation to more accurate, but only privately available, mobility measures based on close contact rates (Crawford et al., 2022). Google LLC (2022) provides daily metrics that measure mobility as a percent change in the number of visits to certain venues, based on personal electronic devices, from days of the same weekday in January 2020. We used the retail+recreation mobility metric which measures changes in visits to places like restaurants, cafes, shopping centers, museums, libraries, and movie theaters as such venues have been heavily targeted by the Quebec government (Gouvernement du Québec, 2022). In Quebec, Google mobility metrics are available at the county level, although the metrics are missing for around 10 percent of counties. Another major concern has been the introduction of new COVID-19 variants, especially the Alpha and Omicron variants (Olivier, 2021; Stevenson, 2021). New variants of COVID-19 can be more contagious and more resistant to existing vaccines compared to previous variants, although they can also be less deadly (Chenchula et al., 2022).

Based on the above discussion, we considered $\boldsymbol{x}_{it} = \boldsymbol{z}_{it} = (\text{beds}_i, \text{mobility}_{\text{county}(i)(t-4)}, \text{new\_variant}_t)^T$ as covariates possibly associated with transmission during the endemic and outbreak periods in Equation (4.2) and disease emergence, disease extinction, outbreak emergence and outbreak persistence in Equations (4.4)-(4.5). Here, beds_i is the number of beds in hospital i and $\text{mobility}_{\text{county}(i)(t-4)}$ is the retail+recreation Google mobility metric for the county where hospital i is located averaged across week $t-4$ (recall it is a daily metric). We lagged the mobility metrics by one month to account for delays between infection, hospitalization and reporting, as well as the time needed for the effects of a spike in mobility to trickle down into the general population. For counties with missing mobility metrics, we substituted the metrics from the nearest county with a similar urban makeup. The binary covariate new_variant_t was 1 if 3 months after the introduction of the Alpha variant in Quebec, introduced December 29, 2020, and the Omicron variant, introduced November 29, 2021, and 0 otherwise (NCCID, 2022). The introduction of the variants are marked in Figure 4.2(a) and we will in general refer to the outbreak that occurs in most hospitals around the introduction of Omicron as the Omicron outbreak. We included new_variant_t only in the relative odds of outbreak emergence and in the outbreak transmission rate to capture the short-term effects of the introduction of a new variant on outbreak emergence and transmission. We combined the Omicron and Alpha variants since there is likely not enough information in the data to estimate the effects of each separately. Additionally, having separate effects for the two variants can make it challenging to use the model in real-time, since we may not know if a future variant is more like Alpha or Omicron. Note, we did not have county-level information on the introduction of the new variants so new_variant_t is the same for all hospitals.

From a sample of patients from hospitals i and j grouped into $l = 1, \ldots, 91$ neighborhoods, which divide Quebec, we calculated the spatial weight ω_{ji}, in Equations (4.4)-(4.5), as

$$\omega_{ji} = \sum_{l=1}^{91} \sqrt{p_{jl} * p_{il}},$$

where, for example, p_{jl} is the proportion of sampled patients from hospital j who lived in neighborhood l, which is known as the Bhattacharyya coefficient (Bi et al., 2019). The Bhattacharyya coefficient must be between 0 and 1 and is a way to measure the amount of overlap between two categorical samples. Therefore, hospitals with more overlap in their catchment areas were given a larger weight. We took the 5 nearest neighbors of hospital i, in terms of the largest weights, to form the neighborhood set $NE(i)$. Note ω_{ji} can be seen as a distance-based decay weight as $\omega_{ji} = \exp(-d_{ji}^{BD})$ where d_{ji}^{BD} is the Bhattacharyya distance (Bi et al., 2019). To finish model specification, we used a uniform initial state distribution in each area.

We fit the CMSNB(1,2,4) model specified above to the Quebec hospitalization data using our proposed hybrid Gibbs sampler from Section 4.3. We ran the Gibbs sampler for 200,000 iterations on three chains with an initial burn-in of 50,000 iterations. All sampling was started from random values in the parameter space to avoid convergence to local modes. Convergence was checked using the Gelman–Rubin statistic (all estimated parameters$<$1.05), the minimum effective sample size ($>$1000) and by visually examining the traceplots (Plummer et al., 2006).

For comparison purposes we also fit a model without neighboring outbreak indicators in the transition probabilities, that is, with $\alpha_{lk,p+1} = 0$ for $lk = 12, 21, 23, 33$ in (4.4)-(4.5), which we will refer to as the Non-coupled Model. Additionally, we fit a two-state model without any absence or clone states, which we will refer to as the No Absence/Clone State Model. We chose these two models for comparison to examine the effects of our main contributions. We will sometimes refer to our CMSNB(1,2,4) model as the Full Coupled Model during comparison. Finally, we also compared to an endemic-epidemic (EE) model (Bracher and Held, 2022), which lies outside our Markov switching framework. The EE model is a state-of-the-art multivariate autoregressive count time series model which is popular for modeling spatio-temporal infectious disease counts (Bauer and Wakefield, 2018; Ssentongo et al., 2021)

Table 4.1: Shows the WAIC of the 4 considered models from Section 5.1 fitted to the Quebec hospitalizations. The best fitting model, the one with the lowest WAIC, is bolded.

Model	**WAIC**
Full Coupled	**17,516**
Non-coupled	17,639
No Absence/Clone	17,545
Endemic-epidemic	17,845

and is commonly used as a benchmark (Bauer et al., 2016; Stojanović et al., 2019). The EE model is usually used for forecasting and for investigating associations between covariates and disease transmission (Ssentongo et al., 2021). We compared to the latest iteration of the EE model (Bracher and Held, 2022) which allows for multiple temporal lags, whose order we chose using the WAIC. More details, including the parameter estimates, are given in SM Section B.4.

Table 4.1 shows the WAIC for the 4 considered models. The Full Coupled Model has the lowest WAIC which lends support to the inclusion of neighboring outbreak indicators in the transition probabilities and the addition of the absence/clone states. The Markov switching models all have a lower WAIC compared to the Endemic-epidemic model, however, we would argue the main advantage of the Markov switching models is that they offer more interesting and richer interpretations. For instance, the Endemic-epidemic model concludes that mobility had a strong positive association with overall disease transmission, see SM Table B.3. From Section 4.5.2 below, the CMSNB(1,2,4) model broadly agrees with this conclusion but provides a deeper understanding of the effect; mobility mainly affects disease transmission during the outbreak and not endemic periods and mobility increases the risk of outbreaks emerging and persisting and reduces the chances of the disease going extinct in smaller areas. These richer covariate interpretations could help policymakers better identify useful interventions. Additionally, the Markov switching models can be utilized for retrospective and real-time state estimation/forecasting, as we illustrate in Sections 4.5.3 and 4.5.4.

4.5.2 Results

Table 4.2 gives the estimated parameters from the Markov chain part of the fitted CM-SNB(1,2,4) model. As relative odds ratios from multinomial logistic regression can be difficult to interpret, we also plot some of the transition probabilities versus some of the covariates in Figure 4.4. Smaller areas were associated with a higher probability of COVID-19 extinction, which follows well known theories from Bartlett (1957) that state population size is inversely related to the rate of extinction of an infectious disease. We did not find evidence that larger areas experience more frequent or longer outbreaks compared to smaller areas (although outbreaks in larger areas do tend to be more severe in terms of transmission, see Table 4.3). We found that mobility in retail and recreation venues had a positive association with both outbreak emergence and outbreak persistence and a negative association with disease extinction. For example, we estimated that a 20 percent (one standard deviation) increase in the number of visits to retail and recreation venues, from the baseline week, was associated with a 73 (8, 169) percent increase in the odds of an outbreak emerging relative to remaining in the endemic state. However, this result may not imply a causal relationship between mobility and outbreak emergence. Mobility could be high at the start of an outbreak because outbreaks typically begin after a long period of relative calm, so it is natural for people to start going out more to movie theaters, restaurants, etc. It is difficult to control for this in our framework, as we would have to account for the time since the last outbreak ended, making the model non-Markovian. Also, note the high amount of uncertainty around the odds ratios for mobility.

The introduction of the Alpha and Omicron variants had a very strong association with the emergence of new COVID-19 outbreaks. We estimated that the probability of an outbreak emerging, at average levels of the other covariates and assuming no neighboring outbreaks, was .02 (.01, .03) if no new variant had been introduced recently and .19 (.09, .32) if a new variant had been introduced in the last 3 months. Emerging variants can trigger new

Table 4.2: Posterior means and 95% posterior credible intervals (in parentheses) for the estimated parameters from the Markov chain part of the fitted CMSNB(1,2,4) model. The intercept row shows the transition probabilities at an average level of mobility and beds, no new variant and assuming no neighboring outbreaks. Other rows show the (relative) odds ratios of the corresponding covariate. Units are given in parentheses after the covariates, for reference, .56 is the average weight ω_{ji} between two neighboring areas. The units for beds and mobility are equal to one standard deviation. Odds ratios whose 95% posterior credible intervals do not contain 0 are bolded.

	Probability or (relative) Odds Ratio			
Covariate	**Disease Emergence**	**Disease Extinction**	**Outbreak Emergence**	**Outbreak Persistence**
Intercept	.33 (.15, .61)	.03 (.01, .06)	.02 (.01, .03)	.88 (.83, .93)
beds (100s)	1.71 (.82, 3.44)	**.44 (.21, .74)**	1.15 (.86, 1.49)	1.13 (.85, 1.48)
mobility (20%)	.77 (.38, 1.36)	**.53 (.24, .98)**	**1.73 (1.08, 2.69)**	**1.65 (1.18, 2.26)**
new variant	–	–	**16.22 (6.91, 33.14)**	–
weighted neighborhood outbreak sum (.56)	1.15 (.84, 1.53)	1.01 (.73, 1.35)	**1.98 (1.58, 2.46)**	**1.31 (1.14, 1.51)**

COVID-19 waves due to them potentially being more contagious and more resistant to existing vaccines compared to previous variants (Maslo et al., 2022). Finally, we found that the presence of outbreaks in neighboring areas was strongly associated with outbreak emergence and persistence. We estimated that an outbreak occurring in a single neighboring area of average connectivity was associated with a 98 (58, 146) percent increase in the odds of an outbreak emerging relative to remaining in the endemic state and a 31 (14, 51) percent increase in the odds of an outbreak persisting. This could be due to either direct disease spread from the neighboring area, the signaling of spread from a common source and, potentially, partly due to correlated missing covariates. Interestingly, we found no strong evidence of an association between neighboring outbreaks and disease emergence or disease extinction, despite there being strong epidemiological justification for a relationship

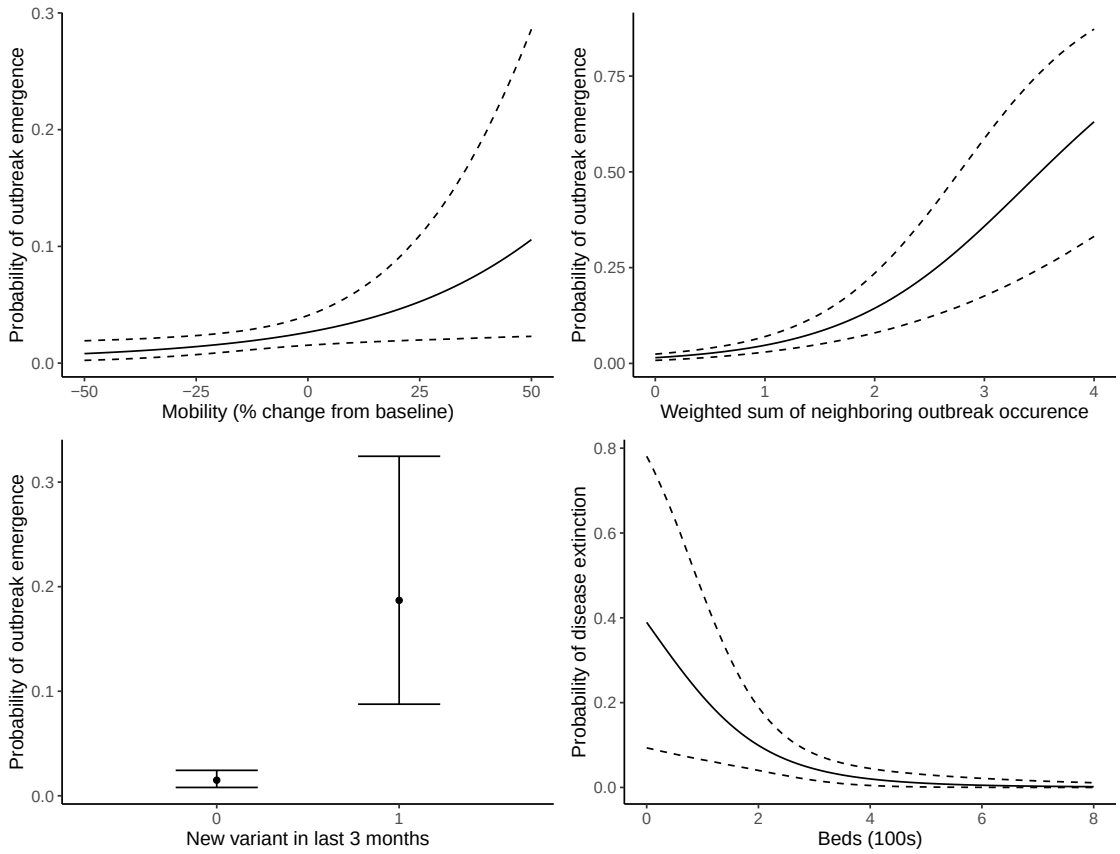


Figure 4.4: Posterior means (solid lines) and 95% posterior credible intervals (dashed lines) of some of the estimated transition probabilities versus some of the covariates. Other covariates were fixed at either their average values, for beds and mobility, or at 0, for new variant and the weighted neighborhood outbreak sum.

due to outbreak spread (Grenfell et al., 2001). Recall that the weighted sum of neighboring outbreak occurrence, $\sum_{j \in NE(i)} \omega_{ji} I[S_{j(t-1)} = 3]$ in (4.4)-(4.5), is a latent covariate and so there is likely not a lot of information about its effect on these rarer transitions (the disease only goes extinct in the smaller catchment areas).

Table 4.3 gives the estimated parameters from the count part of the model. Transmission during the outbreak periods was on average 121 (91, 157) percent higher than during the endemic periods. Mobility in retail and recreation venues and population size were both positively associated with transmission during the outbreak and endemic periods, and outbreaks amplified the effects of population size and mobility on transmission. Note, as we are modeling hospitalizations, an effect in terms of transmission in the hospitalizations could

Table 4.3: Posterior means and 95% posterior credible intervals (in parentheses) from the count part of the fitted CMSNB(1,2,4) model. The intercepts and covariate effects are exponentiated so that they represent rates and rate ratios. Rate ratios whose 95% posterior credible intervals do not contain 0 are bolded. The units for beds and mobility are equal to one standard deviation.

		Rate Ratios	
Covariate	**Parameter**	**Endemic**	**Outbreak**
Intercept of random intercepts	e^{β_0}	.99 (.85, 1.13)	2.18 (1.97, 2.41)
Std. dev of random intercepts	σ	.31 (.21, .44)	.13 (.1, .17)
beds (100s)	$e^{\beta_{\text{beds}}}$	**1.07** **(1.02, 1.11)**	**1.20** **(1.09, 1.32)**
mobility (20%)	$e^{\beta_{\text{mobility}}}$	**1.06** **(1, 1.13)**	**1.17** **(1.14, 1.20)**
new variant	$e^{\beta_{\text{new\_variant}}}$	–	1.01 (.97, 1.06)
autoregressive	ρ	.69 (.65, .72)	.75 (.72, .78)
overdispersion	r	5.50 (4.13, 7.37)	8.95 (7.89, 10.14)

reflect an effect on transmission in the actual cases and/or an effect on the severity of the disease. Mobility and beds should not affect disease severity; however, new variants often do (Chenchula et al., 2022). This could partly explain why we found the introduction of a new variant likely did not have a large effect on transmission in the hospitalizations, Omicron was much more transmissible person to person but less severe (NCCID, 2022). Also, we combined the Alpha and Omicron variants, and, as can be seen in Figure 4.2, Alpha does not appear to have had a large impact on transmission in the hospitalizations. Finally, for sensitivity analysis, we also fit the CMSNB(1,2,4) model using .05 and .1 in place of .01 in the constraint (4.6). The only posterior that changed noticeably was for the effect of mobility on transmission in the endemic period ($e^{\beta^{EN}_{\text{mobility}}}$) which increased from 1.06 (1, 1.13) to 1.08

(1.02, 1.15) going from .01 to .1 in the constraint.

4.5.3 Retrospective evaluation and comparison

As mentioned in Section 4.3.1 it is important to examine the retrospective state estimates of an epidemiological Markov switching model, that is, the posterior probability that the disease was in each state during each week of the study period, to ensure the models' estimates of the epidemiological history of the disease are sensible. Starting with the bottom graphs of Figures 4.5 (a) and (c), we compare the retrospective state estimates for Fleury Hospital in Montreal, one of the smaller hospitals, between models with, (a), and without, (c), an absence state, as it is a good example of how accounting for long periods of disease absence in smaller areas can be important. The top graphs show the posteriors of the endemic and outbreak state distributions, which can be helpful for better understanding the state estimates. To draw from the posterior of the state distributions we drew from $p(y_{it}|S_{it} = s, y_{i(t-1)}, \boldsymbol{\beta}^{[m]})$, see (4.1)-(4.2), for $m = M + 1, \ldots, Q$, $t = 1, \ldots, T$, $s = 2$ (endemic) and $s = 3$ (outbreak). The Full Coupled Model identifies 3, unlikely 4, outbreak periods for Fleury Hospital which are separated by a few endemic and absence periods. Long sequences of zeroes in Fleury Hospital are generally assigned to absence periods by the Full Coupled Model. In contrast, the No Absence/Clone State Model assigns the long strings of zeroes to endemic periods, bringing the endemic state distribution much closer to 0 compared to the Full Coupled Model. This causes some issues in the state estimation for the No Absence/Clone State Model, as the outbreak state becomes too dominant. Firstly, the No Absence/Clone State Model classifies a very small increase in hospitalizations, around week 56, as likely a short outbreak period, which is not realistic. Secondly, the No Absence/Clone State Model identifies a smaller outbreak, around week 81, before the final Omicron outbreak. In contrast, the Full Coupled Model only identifies the Omicron outbreak here, which appears more realistic as the number of hospitalizations in the few weeks before Omicron never becomes high enough to be very concerning and, by checking the other plots, there is no strong evidence of outbreaks in any

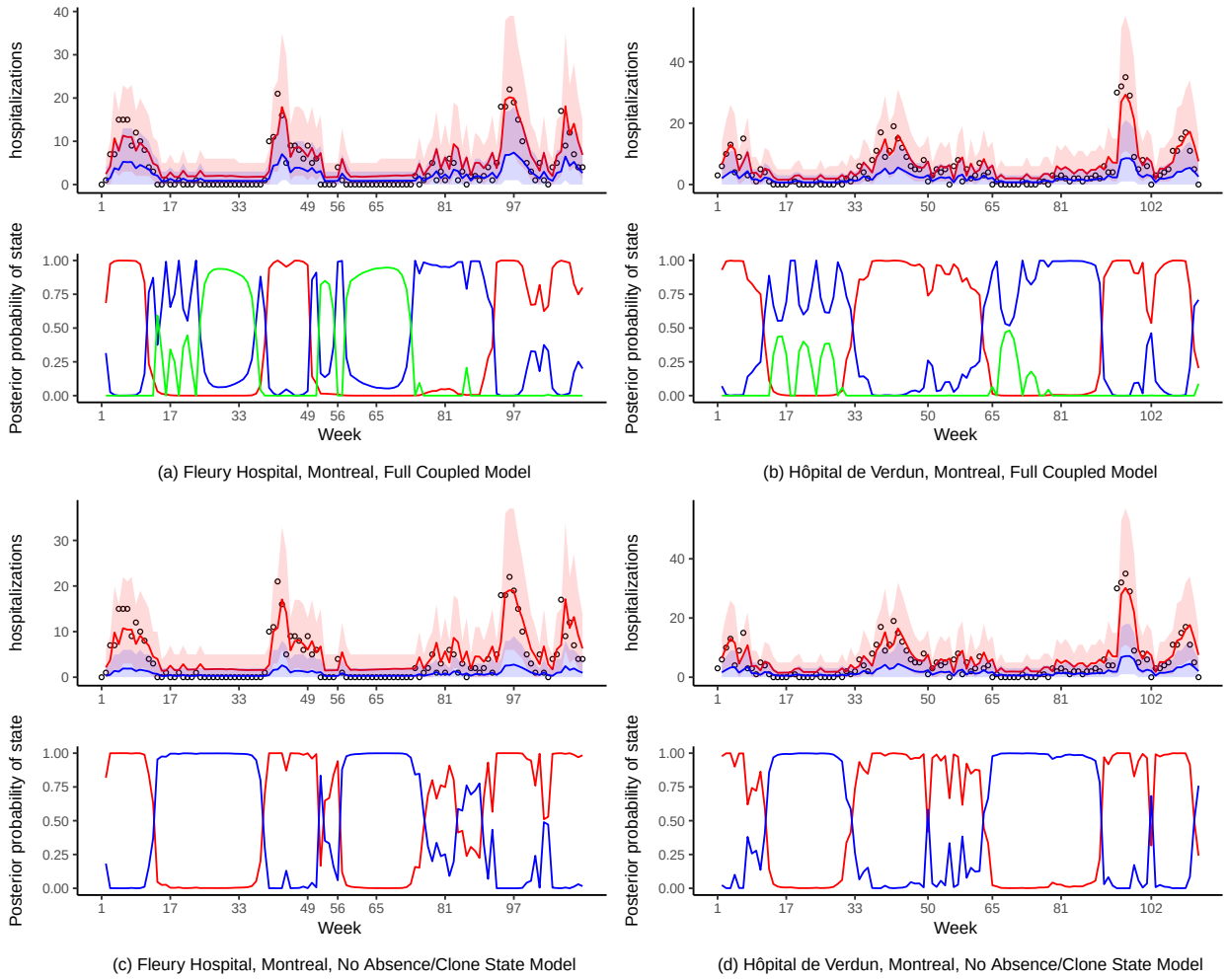


Figure 4.5: Top graphs show the posterior means (solid lines) and 95% posterior credible intervals (shaded areas) of the endemic state distribution (blue) and the outbreak state distribution (red) versus the observed hospitalizations (points). The bottom graphs show the posterior probability that the disease was in the absence state (green line), the endemic state (blue line) and the outbreak state (red line) during each week of the study period, that is, $P(S_{it} = s|\boldsymbol{y})$ for $t = 1, \ldots, T$ and $s = 1$ (absence) in green, $s = 2$ (endemic) in blue and $s = 3$ (outbreak) in red, see Section 4.3.1.

neighboring areas.

In Figures 4.5 (b) and (d) we compare the retrospective states estimates for Hôpital de Verdun between models with, (b), and without, (d), clone states. The No Absence/Clone State Model shows rapid switching between the outbreak and endemic states around weeks 50 and 102. During week 50, it is more realistic that the one-week dip in the hospitalizations is due to random variation, which is what the Full Coupled Model identifies, and that there

are not multiple outbreaks occurring. The Full Coupled Model does place some weight on an endemic period between weeks 102 and 103. A two-week dip into the endemic period seems more realistic than a one-week dip, however, we could prevent two-week dips by increasing the number of endemic clone states. For the Full Coupled Model, we looked through the plots in Figure 4.5 for each hospital and found the retrospective state estimates to be sensible.

4.5.4 Real-time evaluation and comparison

To evaluate the real-time performance of the models we fit the Full Coupled and Non-coupled models up to 6 weeks before the introduction of Omicron and then up to every week after, that is, we fit the models up to time T for $T = 84 =$ 2021-10-18, ..., $113 =$ 2022-05-09, producing 30 sets of posterior samples for each model. We did not include new_variant$_t$ as a covariate since many of the fitted models are fit before, or just after, Omicron and so there is likely not enough information to estimate the effect for many of the fits. Additionally, we found after discarding 27% of the data for the real-time evaluation that MCMC chains would sometimes get stuck in local modes in regions of the parameter space that only used the absence state and one of the count states. This is not an uncommon problem in complex Bayesian mixture modeling and, following the discussion and recommendations in Section 4.2.3 of Frühwirth-Schnatter (2006), we applied mild shrinkage to the transition probabilities by assigning $N(0, 2.5^2)$ priors to the intercepts in Equations (4.4)-(4.5), which stabilized model fitting. Figure 4.6 compares the posterior probabilities that an outbreak is currently happening (bottom graphs solid lines) between the Full Coupled Model, in orange, and the Non-coupled Model, in purple, for four hospitals. Recall from Section 4.3.1 that the posterior probability that an outbreak is currently happening is given by $P(S_{iT} = 3|\boldsymbol{y})$ and would be used for the purpose of outbreak detection in a real-time scenario. The Full Coupled Model gives an earlier warning of the Omicron outbreak in the first three hospitals, (a)-(c), but not in the final hospital, (d). In (d) the hospitalizations rise very rapidly and so there is a lot of within area information about the Omicron outbreak, making the coupling less useful

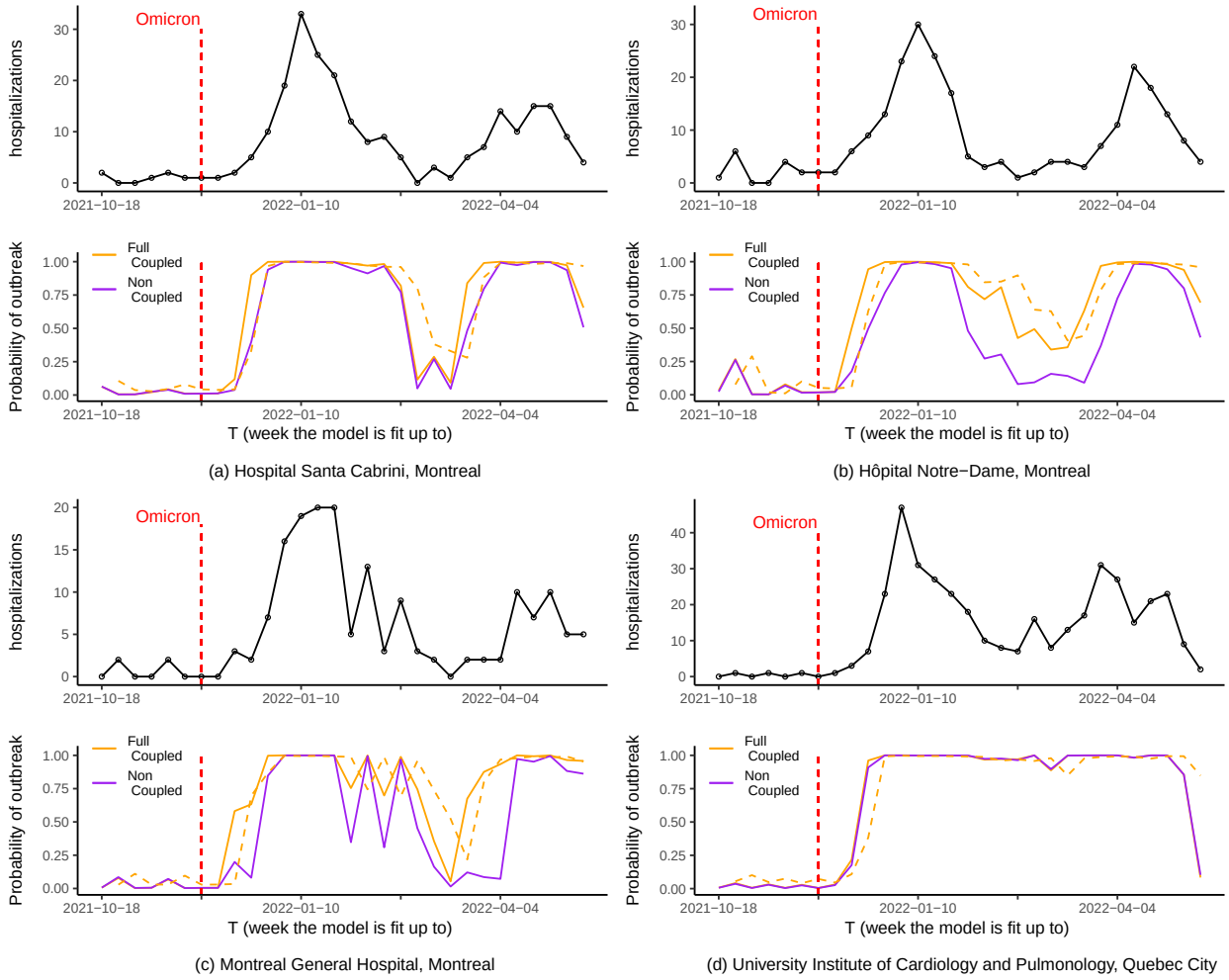


Figure 4.6: Top graphs show the observed hospitalizations for the last 30 weeks of the study period where we conducted the real-time evaluation. Bottom graphs solid lines show the posterior probabilities that an outbreak is currently happening, that is, $P(S_{iT} = 3|\boldsymbol{y})$, versus T. Bottom graphs dashed line shows the one-week ahead outbreak forecasts from the previous week, that is, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$, versus T. The Full Coupled Model is in orange and the Non-coupled Model is in purple. The dotted red lines are drawn at the introduction of the Omicron variant for all of Quebec.

for providing an early warning. SM Figure B.5 (a) shows that, when averaged across all hospitals, early on during the Omicron wave the Full Coupled Model gives a .15-.2 higher posterior probability that an outbreak is currently happening, an earlier warning compared to the Non-coupled Model, while it does not show an increased risk of an outbreak prior to Omicron. The Omicron wave was highly synchronized between the hospitals and, therefore, as can be seen from SM Figure B.5 (b), the probability of outbreak emergence would have

been higher for the Full Coupled Model during the start of the wave as evidence of outbreaks accumulated in neighboring hospitals. In general, SM Figure B.5 shows that the Full Coupled Model tends to enforce the status quo, making outbreak occurrence more (less) likely if there are (are not) outbreaks occurring in neighboring areas.

The dashed line in the bottom graphs of Figure 4.6 show the one-week ahead outbreak forecasts $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$, from the previous week, for the Full Coupled Model. Note that, we typically observe $P(S_{iT} = 3|\boldsymbol{y}) > P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)}) > P(S_{i(T-1)} = 3|\boldsymbol{y}_{1:(T-1)})$ around the start of the Omicron wave. We looked through the plots in Figure 4.6 for each hospital and found the real-time state estimates/forecasts to be sensible in most hospitals. We did find evidence of false alarms being thrown in 2/30 hospitals during the real-time evaluation, shown in the SM Section B.7.3. The false alarms appear to have been mostly caused by hospitalizations spiking during periods of high outbreak risk, meaning the model has no information that an outbreak starting should be unlikely to counter the false alarm. Still, the model corrected itself quickly, by the next week in most cases.

Finally, the map in SM Figure B.6 shows the likely state in the catchment area of each hospital during the last week of the study period, March 5th, 2022, from the Full Coupled Model with new variant as a covariate. According to the map, most catchment areas were in the outbreak state at this time, especially in/around Montreal.

4.6 Discussion

We have proposed a three-state coupled nonhomogeneous Markov switching model for the general analysis of spatio-temporal outbreak occurrence. The model can be used for investigating associations between various factors, including geographical outbreak spread, and outbreak emergence, outbreak persistence, disease extinction and disease emergence, as well as for detecting and forecasting the outbreaks. We made three main contributions to the existing (Amorós, 2017) endemic/outbreak Markov switching literature. Firstly, to account

for long periods of disease absence in smaller areas (Bartlett, 1957), we added an absence state to our model in addition to the more traditional endemic and outbreak states. Secondly, to incorporate geographical outbreak spread (Grenfell et al., 2001), we allowed the transition probabilities to depend on whether outbreaks were occurring in neighboring areas. Previous two-state approaches (Heaton et al., 2012) have allowed the probabilities of outbreak emergence to depend on neighboring outbreaks, but they did not allow outbreak emergence to depend on covariates, and they fixed the probability of outbreak persistence at one to prevent rapid switching between endemic and outbreak periods, meaning their method could not be applied to multiple outbreaks. Finally, to allow for the analysis of multiple outbreaks, we instead introduced clone states (Kaufmann, 2018) into the model, which prevents rapid switching between endemic and outbreak periods by enforcing a minimum endemic and outbreak duration.

We applied our model to the analysis of COVID-19 outbreaks across Quebec based on admissions in the 30 largest hospitals. We found that mobility in retail and recreation venues, the introduction of new variants and the occurrence of outbreaks in neighboring areas all had positive associations with the emergence or persistence of new COVID-19 outbreaks in Quebec. Additionally, mobility and population size had positive associations with transmission during the outbreak and endemic periods, and the effects were amplified during the outbreak periods. As for disease extinction and emergence, we only found evidence that the disease is more likely to go extinct in smaller areas and at lower levels of mobility. Disease extinction and emergence are rarer transitions, so more data needs to be collected to study them with high precision.

Regarding model performance, we found our model gave realistic estimates of past epidemiological states for the Quebec hospitalizations and appeared to perform outbreak detection/ forecasting well in a real-time evaluation of the Omicron wave, absent a small number of false alarms during the real-time evaluation. However, on the Quebec hospitalizations it

was difficult to exactly quantify the accuracy of the state estimates, e.g., calculate AUC, sensitivity, etc., as the true underlying states were not known. Therefore, we also conducted a simulation study with known outbreaks and found our model achieved good sensitivity, specificity and timeliness both retrospectively and during real-time outbreak detection and forecasting.

Regarding model comparison among different Markov switching models. In our simulation study, where we assumed outbreaks in neighboring areas occurred around the same time, including neighboring outbreak indicators in the transition probabilities greatly improved sensitivity, specificity and timeliness during real-time outbreak detection and forecasting. As for our analysis of the Quebec hospitalizations, model comparison using WAIC supported the inclusion of absence/clone states and the addition of neighboring outbreak indicators in the transition probabilities. We also found for the Quebec analysis that the addition of an absence state can be important in smaller areas for reducing bias in the endemic state distribution towards 0, a bias that we observed often leads to unrealistic state estimates. We further observed that clone states are important for preventing rapid switching between endemic and outbreak periods, making the state estimates more realistic. Finally, in a real-time evaluation we found that the incorporation of outbreak spread led to an earlier warning of the Omicron wave in Quebec as the wave was highly synchronized across the hospitals. Ultimately, our contributions show a lot of promise for improving state estimation both retrospectively and in real-time, especially when there are small areas and highly spatially synchronized outbreaks. Although we applied our model to spatio-temporal counts of hospitalizations, it could also be applied to spatio-temporal counts of disease cases, which are popular to model for outbreak occurrence analysis (Knorr-Held and Richardson, 2003; Watkins et al., 2009), or deaths.

Now we will address the question of how our approach fits into the broader popular literature on spatio-temporal infectious disease modeling, including compartmental models (Bauer and

Wakefield, 2018) and multivariate time series models (Ssentongo et al., 2021). One aspect of our approach that we believe is unique is the ability to quantify associations, through odds ratios, between covariates and certain epidemiological transitions, such as outbreak emergence or disease extinction. Also, our model can provide the probability that a spike in an epidemiological indicator is due to an actual outbreak developing as opposed to just random endemic variation, which is valuable for outbreak detection (Unkel et al., 2012). On the Quebec hospitalizations, we found our model fit better and had better interpretations compared to the Endemic-epidemic model (Bracher and Held, 2022), a popular alternative.

There are also some limitations with our approach. Conditional on the states, we only consider first-order autoregression, see the discussion at the end of SM Section B.4. Given weekly data and a disease with a short serial interval, such as COVID-19, this is likely not a major limitation (Bracher and Held, 2022). However, in some cases, especially with daily data, it can be important to account for higher-order temporal dependencies. As another limitation, we assume that the probability of an outbreak emerging does not depend on the amount of time since the last outbreak has ended. After an outbreak ends, the susceptible population should increase over time due to demographic changes, waning immunity and other factors, leading to an increased outbreak risk (Keeling and Rohani, 2007). In a hidden semi-Markov model (Langrock and Zucchini, 2011), the transition probabilities can depend on the amount of time that has been spent in the state, which can be pursued in further work. Finally, given the large number of transitions in our model, the number of covariate effects grows quickly with the number of covariates (4 times). Shrinkage could be applied to the covariate effects in the transition probabilities to eliminate many unimportant effects (Wang et al., 2023).

Acknowledgements

This research was part of a larger project of INESSS (Institut national d'excellence en santé et en services sociaux), whose objective was to produce bed occupancy projections for COVID-19 patients. The access to data was made possible through a tripartite agreement between the MSSS, the RAMQ and INESSS. This work is part of the PhD thesis of D. Douwes-Schultz under the supervision of A. M. Schmidt in the Graduate Program of Biostatistics at McGill University, Canada. Douwes-Schultz is grateful for financial support from IVADO and the Canada First Research Excellence Fund/Apogée (PhD Excellence Scholarship 2021-9070375349). Schmidt is grateful for financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant RGPIN-2017-04999). Shen is grateful for financial support from the Fonds de recherché du Québec–Santé (FRQS) (Doctoral training award 313602). Buckeridge is supported by a Canada Research Chair in Health Informatics and Data Science (950-232679. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

Chapter 5

Markov switching zero-inflated space-time multinomial models for comparing multiple infectious diseases

Preamble to Manuscript 3. In this third manuscript, we consider coupled Markov switching models where multiple diseases switch between periods of presence and absence in each area. Therefore, this paper can be seen as a multivariate extension of the first manuscript from Chapter 3. Since we are mainly interested in comparing the transmission dynamics of the diseases, we assume that the reported cases of the present diseases in an area jointly follow an autoregressive multinomial model. The multinomial model can be used to investigate associations between certain factors, such as temperature, and differences in the transmission intensity of the diseases. Like in the first manuscript, we assume the absent diseases always report 0 cases. Additionally, we assume that the probability a disease is present in an area can depend on the presence of the other diseases previously, to account for disease interactions (Sherlock et al., 2013).

Our approach incorporates coupled Markov switching into the zero-inflated multinomial

framework of Zeng et al. (2022). Zero-inflated multinomial models assume that certain categories (in our case diseases) can be absent from each observation, in which case their corresponding multinomial probabilities are fixed at zero (Tang and Chen, 2019; Zeng et al., 2022; Koslovsky, 2023). Existing zero-inflated multinomial models are not appropriate for spatio-temporal infectious disease counts, as they assume that the multinomial counts conditional on the present categories are independent between observations. They also assume that the presence of a category in one observation does not affect whether that category or other categories are present in other observations (independence of categorical presence). If a disease is present, it is more likely to be present again in the area (Coutinho et al., 2006) and it may also affect the presence of other diseases due to disease interactions (Sherlock et al., 2013). For our motivating example to dengue, Zika and chikungunya counts, we found that the model of Zeng et al. (2022) did not fit as well as our coupled Markov switching extension, see Table 5.1, and produced estimates that were not consistent with the epidemiology of the diseases, see the discussion at the beginning of Section 5.4.2. We also found that the model of Zeng et al. (2022) was not able to produce a complex spatio-temporal pattern of disease presence like with our coupled Markov switching model, see Figure 5.4.

Finally, as an alternative to our multinomial approach, we could assume the cases of the present diseases jointly follow some form of a multivariate Poisson distribution (Jack et al., 2019). This would be more appropriate if we were interested in the overall effects of a covariate on transmission and not just the relative effects, see the discussion in Section C.1.2 of the supplementary materials. Zero-inflated spatio-temporal multivariate Poisson models have only been considered by Rotejanaprasert et al. (2021) and Pavani and Moraga (2022). These models, which relied primarily on random effects to capture correlations, have several important limitations such as not allowing the probabilities of disease presence to change across time. A random effect approach will likely become intractable due to the need to account for correlations across three dimensions in both the count and zero-inflated components of the model. Our approach is much more computationally feasible and could

be easily extended to the multivariate Poisson case, see Section C.1.2.

In conclusion, by extending the models from the first manuscript to the multivariate setting and by extending zero-inflated multinomial models to the space-time setting, this manuscript represents an important contribution to the area of multivariate spatio-temporal infectious disease modeling. Additionally, the multinomial approaches explored here could be extended to the multivariate Poisson case to address important gaps in that literature. Indeed, very little attention has been given to zero-inflated multivariate spatio-temporal disease modeling, despite the importance of comparing disease dynamics in a space-time setting (Freitas et al., 2019).

Markov switching zero-inflated space-time multinomial models for comparing multiple infectious diseases

Dirk Douwes-Schultz[1], Alexandra M. Schmidt[1].

[1]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

Abstract

The modeling of zero-inflated multinomial data across space and time is challenging due to the need to account for correlations across space, time and category in both the count and zero-inflated components of the model. Despite spatio-temporal counts often containing many zeroes, zero-inflated multinomial models for space-time data have not been considered. Here, we present a computationally feasible approach to the problem based on an autoregressive Markov switching framework. We are interested in comparing the transmission dynamics of several co-circulating infectious diseases where some can be absent for long periods. We first assume there is a baseline disease that is well-established and always present in the region. The other diseases switch between periods of presence and absence in each area through a series of coupled Markov chains that account for long periods of disease absence, disease interactions and disease spread from neighboring areas. Since we are only interested in comparing the diseases, we assume the cases of the present diseases in an area jointly follow an autoregressive multinomial model. We use the multinomial model to investigate whether there are associations between certain factors, such as temperature, and differences in the transmission intensity of the diseases. Inference is performed using efficient Bayesian Markov chain Monte Carlo methods based on jointly sampling all presence indicators. We apply the model to spatio-temporal counts of dengue, Zika and chikungunya cases in Rio de Janeiro during the first triple epidemic there.

5.1 Introduction

Comparing the transmission dynamics of several co-circulating infectious diseases is an important problem in epidemiology. In this paper, we are specifically concerned with comparing the transmission dynamics of the vector-borne diseases Zika, chikungunya and dengue. Since 2015 several countries in Latin America have experienced triple epidemics of these diseases, placing an immense burden on the local populations (Rodriguez-Morales et al., 2016; Bisanzio et al., 2018; Queiroz and Medronho, 2022). Despite being transmitted by the same vector, the *Aedes aegypti*, several studies have found spatio-temporal differences in clusters of the three diseases (Freitas et al., 2019; Kazazian et al., 2020). Investigating how these differences may be related to covariates could help more efficiently distribute medical resources, as health interventions can vary between the diseases (Freitas et al., 2019). Schmidt et al. (2022) attempted this using a multinomial model but they only investigated how the relative distribution of the diseases varied spatially. There is also an interest in seeing how the transmission of the three diseases may differ by spatio-temporal factors. For instance, laboratory studies have shown that the transmission of Zika may be more sensitive to temperature compared to dengue, which is important as it implies Zika will not be able to spread as effectively into North America and the highlands (Tesla et al., 2018).

As a motivating example, we analyze bi-weekly reported case counts of Zika, chikungunya, and dengue in the 160 neighborhoods of Rio de Janeiro, Brazil, between 2015-2016, during the first triple epidemic there. Since we are only interested in comparing the diseases, we condition on the total number of cases in each area and bi-week and use a multinomial distribution to model their relative allocations (Dabney and Wakefield, 2005; Dreassi, 2007; Schmidt et al., 2022). Several authors have proposed modeling multinomial data across time, space, or both by adding correlated random effects to multinomial link functions, such as the log relative odds (Pereira et al., 2018; Sosa et al., 2023). Proposed distributions for the random effects include conditional autoregressive distributions (Schmidt et al., 2022),

multivariate logit-beta distributions (Bradley et al., 2019) and multivariate normal random walks (Cargnoni et al., 1997). Autoregressive approaches (Loredo-Osti and Sutradhar, 2012) have also been popular, where the current multinomial probabilities in an area are assumed to depend on past observations in the area (Fokianos and Kedem, 2003) and neighboring areas (Tepe and Guldmann, 2020). Since new infectious disease cases are transmitted from previous cases, conditioning on past observations is often a natural way to model infectious disease counts across time and space (Bauer and Wakefield, 2018). Therefore, we primarily take an autoregressive-driven approach in this paper. In particular, we assume that if conditions are more favorable for one disease over the others, then the share of that disease will grow over time. We represent the relative favorability of the diseases through a series of autoregressive parameters that are linked to covariates. This allows us to investigate whether a covariate is associated with conditions that are more favorable toward the transmission of one of the diseases. As we will show, our approach is epidemiologically motivated and can be derived from a series of Reed-Frost chain binomial susceptible-infectious-recovered (SIR) models (Abbey, 1952; Vynnycky, 2010; Bauer and Wakefield, 2018).

A potential issue with applying a multinomial model to multivariate spatio-temporal infectious disease counts is that many infectious diseases, especially vector-borne diseases, can be absent for long periods of time in an area (Bartlett, 1957; Coutinho et al., 2006; Adams and Boots, 2010). This behavior is illustrated in Figure 5.1 which shows bi-weekly reported case counts of dengue, Zika and chikungunya in the western Rio neighborhood of Praça Seca, between 2015-2016. While dengue cases are reported consistently throughout the two years, very few Zika cases and no chikungunya cases are reported in the first year. These patterns are typical of most neighborhoods. If a disease is absent in an area then there should be no chance of cases being reported. However, a multinomial model will always assign a positive probability to cases being reported for any of the diseases. This may lead to many more zeroes in the observed counts than can be realistically produced by a fitted multinomial model (Koslovsky, 2023). Also, we do not want to compare the transmission dynamics of

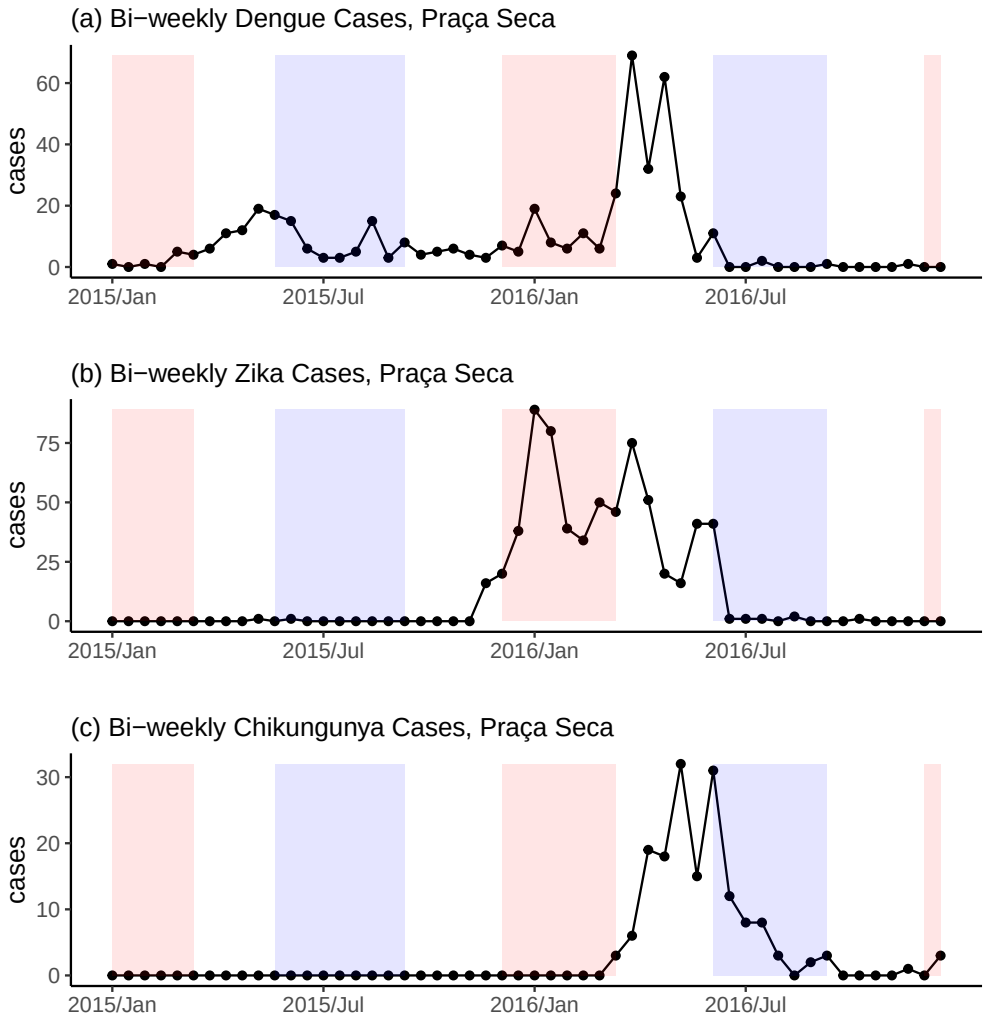


Figure 5.1: Bi-weekly reported case counts of dengue (a), Zika (b), and chikungunya (c) in Praça Seca, a west neighborhood of Rio de Janeiro, Brazil, between 2015-2016. Summer/winter seasons highlighted in red/light blue.

a present disease and an absent one as it would likely overestimate the favorability of the present disease.

To account for long periods of disease absence, we will place our proposed autoregressive multinomial model within a zero-inflated multinomial modeling framework (Zeng et al., 2022). Zero-inflated multinomial models allow certain categories, in our case diseases, to be absent from each observation (Diallo et al., 2018). If a category is absent from an observation its corresponding multinomial probability is fixed at 0 (Tang and Chen, 2019). The

presence of a potentially absent category is treated as random and generated through an independent Bernoulli process (Koslovsky, 2023). Tang and Chen (2019), Zeng et al. (2022) and Koslovsky (2023) have proposed zero-inflated multinomial models that allow all categories to be either present or absent from each observation. These proposals mainly differ in how the multinomial distributions are parametrized and whether covariates can affect the probability of a category being present (Tang and Chen, 2019; Koslovsky, 2023). We will largely follow Zeng et al. (2022) as they use a log link for the relative odds which matches our proposed autoregressive multinomial model.

However, the framework of Zeng et al. (2022) is not entirely appropriate for our application to multivariate spatio-temporal infectious disease counts. Firstly, past approaches have ignored that all categories cannot be absent from an observation if the total count is not equal to 0. This may have been overlooked as existing zero-inflated multinomial models have been applied to data with many categories, usually greater than 50 (Zeng et al., 2022). However, in our motivating example, we have only 3 diseases. Another issue is that past approaches have made two independence assumptions that are inappropriate for spatio-temporal infectious disease counts. Firstly, they have assumed that the multinomial counts are independent conditional on the present categories of each observation. Infectious disease counts usually exhibit a high amount of correlation across space and time due to disease transmission (Bauer and Wakefield, 2018). Secondly, past approaches assumed that if a category is present in one observation it does not affect whether that category or other categories are present in other observations (independence of categorical presence). If an infectious disease is present in an area then it is more likely to be present in the next time period, as disease presence and absence are usually persistent (Coutinho et al., 2006). Also, the presence of one disease may affect whether other diseases are present due to potential disease interactions (Paul et al., 2008). Finally, the diseases will spread between neighboring areas (Grenfell et al., 2001), meaning the status of the disease in an area will affect its presence in neighboring areas.

To address the above issues, we first choose a baseline disease that is well-established in the region and assume it is always present. This avoids all diseases being absent in an area at once which would be impossible if the total number of cases were not 0. We assume all other diseases switch back and forth between periods of presence and absence in each area through a series of Markov chains. The Markov chains account for long periods of disease absence, disease interactions and covariates, including the prevalence of the diseases in neighboring areas. For the diseases present in an area together, we assume that the distribution of their cases conditional on the total follows our proposed autoregressive multinomial model. The absent diseases always report 0 cases. The rest of this paper is organized as follows. Section 5.2 introduces our proposed autoregressive multinomial model for comparing the transmission dynamics of multiple infectious diseases in a space-time setting. Section 5.2.1 extends the model to incorporate zero inflation and deal with long periods of disease absence. Section 5.3 details our Bayesian inferential procedure which utilizes the forward filtering backward sampling (FFBS) algorithm of Chib (1996) for efficient inference. In Section 5.4, we apply our model to cases of dengue, Zika and chikungunya in Rio de Janeiro during the first triple epidemic there. We investigate how factors like temperature were related to differences in the transmission intensity of the diseases. We close with a general discussion in Section 5.5.

5.2 An Autoregressive Multinomial Model for Comparing Transmission Dynamics

Assume we have reported case counts for K different infectious diseases across $i = 1, \ldots, N$ areas and $t = 1, \ldots, T$ time periods. Let $\boldsymbol{y}_{it} = (y_{1it}, \ldots, y_{Kit})^T$ represent the vector of case counts for all diseases in area i during the time interval $(t-1, t]$, where y_{kit} represents the reported case counts for disease k. We will begin by presenting our model without zero inflation and then extend to the zero-inflated case in Section 5.2.1. Since we are mainly

interested in comparing the diseases, we condition on the total number of cases and model $\boldsymbol{y}_{it}$ using a multinomial distribution,

$$\boldsymbol{y}_{it}|\text{total}_{it}, \boldsymbol{y}_{t-1} \sim \text{Multinom}(\boldsymbol{\pi}_{it}, \text{total}_{it}), \tag{5.1}$$

where $\text{total}_{it} = \sum_{k=1}^{K} y_{kit}$ and $\boldsymbol{y}_{t-1} = (\boldsymbol{y}_{1(t-1)}, \ldots, \boldsymbol{y}_{N(t-1)})^T$ is the vector of all case counts reported in the previous time, giving the model an autoregressive structure. In (5.1), $\boldsymbol{\pi}_{it} = (\pi_{1it}, \ldots, \pi_{Kit})^T$ represents the expected relative distribution of the disease counts at time t in area i given the previous observed counts $\boldsymbol{y}_{t-1}$. If conditions in $(t-1, t]$ do not favor one disease over the others, then the relative distribution of the diseases should remain the same. Otherwise, if conditions in $(t-1, t]$ favor some diseases over others then the share of the favored diseases should grow relative to those less favored. Therefore, we model the relative odds as,

$$\frac{\pi_{kit}}{\pi_{1it}} = \lambda^*_{kit} = \lambda_{kit} \frac{\left(y_{ki(t-1)} + 1\right)^{\zeta_k}}{\left(y_{1i(t-1)} + 1\right)^{\zeta_1}}, \tag{5.2}$$

for $k = 2, \ldots, K$, where we assume disease 1 is the baseline disease and we add 1 to the top and bottom to avoid dividing by zero, adding such constants is common in autoregressive count models (Liboschik et al., 2017; Fritz and Kauermann, 2022). The parameters ζ_k and ζ_1, assumed to be between 0 and 1, are meant to account for nonhomogeneous mixing (Wakefield et al., 2019). That is, as cases increase there tends to be a dampening effect on transmission as individuals avoid infection and governments implement interventions. In (5.2), $\lambda_{kit} > 0$, which we link to covariates below, is a parameter that represents the favorability of conditions for the transmission of disease k relative to the baseline disease during $(t-1, t]$. For instance, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then we would expect the share of disease k to grow (shrink) relative to the baseline disease in $(t-1, t]$, after adjusting for nonhomogeneous mixing with ζ_k and ζ_1. If $\zeta_k \approx \zeta_1 << 1$, like for our motivating example in Section 5.4, then this means that if $\lambda_{kit} > 1$ the share of disease k will grow initially from an equal share. However, as disease k

becomes more dominant the share will be pulled back towards equality as individuals avoid infection from the dominant disease.

We link λ_{kit} to covariates and random effects using a log-linear model,

$$\log(\lambda_{kit}) = \alpha_{0ki} + \boldsymbol{x}_{kit}^T \boldsymbol{\alpha}_k + \phi_{kit}, \tag{5.3}$$

where $\alpha_{0ki} \sim N(\alpha_{0k}, \sigma_k^2)$ is a normal random intercept meant to account for between area differences and $\boldsymbol{x}_{kit}$ is a vector of covariates that might affect the favorability of disease k relative to the baseline disease. We model the random effects ϕ_{kit} using a multivariate normal distribution (Xia et al., 2013; Zeng et al., 2022),

$$\boldsymbol{\phi}_{it} = \begin{pmatrix} \phi_{2it} \\ \vdots \\ \phi_{Kit} \end{pmatrix} \sim \text{MVN}_{K-1}(\mathbf{0}, \Sigma), \tag{5.4}$$

where Σ is a $K-1$ by $K-1$ variance-covariance matrix. The main purpose of these random effects is to account for overdispersion which, like with Poisson counts, is a very common issue when modeling multinomial counts (Xia et al., 2013). As we show using a simulation study in the Supplementary Materials (SM) Section C.2, the variance-covariance matrix Σ needs to be interpreted with caution. In the simulation study we focused on the case of $K = 3$ where we have $\boldsymbol{\phi}_{it} = (\phi_{2it}, \phi_{3it})^T$. We found that the correlation between ϕ_{2it} and ϕ_{3it} needs to be around .5 or greater for the correlation between the individual counts, marginalizing the random effects and conditional on their total, to be larger than what is produced by the standard multinomial distribution.

Now we will discuss the interpretation of the covariate effects in (5.3). Broadly, if $\alpha_{kj} > 0$ ($\alpha_{kj} < 0$) then as x_{kitj} increases, conditions will become more (less) favorable for the transmission of disease k relative to the baseline disease. If $\alpha_{kj} = 0$ then x_{kitj} does not affect the favorability of disease k. Note, we can express the log relative odds in (5.2)

as $\log(\pi_{kit}/\pi_{1it}) = \alpha_{0ki} + \boldsymbol{x}_{it}^T\boldsymbol{\alpha}_k + \phi_{kit} + \zeta_k \log(y_{ki(t-1)} + 1) - \zeta_1 \log(y_{1i(t-1)} + 1)$. Therefore, $e^{\alpha_{kj}}$ is the ratio of π_{kit}/π_{1it} (relative odds ratio), conditional on the other covariates, random effects and previous distribution of the diseases, and the ratio of λ_{kit} (rate ratio), from Equations (5.2)-(5.3), conditional on the other covariates and random effects, corresponding to a unit increase in x_{kitj}. There is also an epidemiological interpretation of the model parameters. As we show in SM Section C.1, if all diseases follow a Reed-Frost chain binomial SIR model (Abbey, 1952; Vynnycky, 2010; Bauer and Wakefield, 2018) with roughly the same serial interval, the time between successive generations of cases, then we can derive Equations (5.1)-(5.4). Under this derivation, λ_{kit} in (5.2) represents the ratio of the effective reproduction number of disease k and the baseline disease. The effective reproduction number is an important measure of transmission in epidemiology that gives the average number of new infections produced by a single infectious individual in the current population before they recover (Vynnycky, 2010). The covariate effect α_{kj} then represents the difference between the effect of covariate x_{kitj} on the effective reproduction number of disease k and the baseline disease, see SM Section C.1 for more details.

We recommend treating the Reed-Frost derivation with some caution as the Reed-Frost model, although commonly used (Wakefield et al., 2019), makes assumptions that are not appropriate for many diseases (Abbey, 1952), and the model can be sensitive to underreporting and reporting delay when it comes to estimating the effective reproduction number (Bracher and Held, 2021; Quick et al., 2021). Regardless, the Reed-Frost derivation does reveal an important source of potential confounding due to differences in the susceptible populations, i.e., the population of individuals who are not immune to the disease. From the Reed-Frost derivation, see SM Equation (C.8), ideally, we would add $\log(\delta_{ki(t-1)}/\delta_{1i(t-1)})$ to $\log(\lambda_{kit})$ in Equation (5.3), where $\delta_{ki(t-1)}$ is the size of the susceptible population for disease k in area i at time $t-1$. This makes sense intuitively, if a disease has a larger susceptible population compared to another then transmission for that disease will naturally be favored. The effect of any covariates correlated with the ratio of the susceptible populations could

then be confounded. We conduct sensitivity analysis to attempt to control for this in Section 5.4.4.

Finally, we note that the number of cases in neighboring areas may affect the favorability of the diseases as individuals will mix with those in neighboring areas (Grenfell et al., 2001). Therefore, we allow functions of disease cases in neighboring areas, such as the prevalence across neighboring areas, to be added to the covariate vector $\boldsymbol{x}_{kit}$. Also, we allow previous counts of other diseases, e.g., $\log(y_{ji(t-1)}+1)$ for $j \neq k$, to be added to $\boldsymbol{x}_{kit}$ to account for potential disease interactions (Freitas et al., 2019), see Section 5.4 for an example. We will call the model defined by (5.1)-(5.4) the autoregressive multinomial (ARMN) model. Note the multinomial probabilities are completely defined by (5.2)-(5.4) as $\pi_{1it} = \frac{1}{1+\sum_{j=2}^{K}\lambda^*_{jit}}$ and $\pi_{kit} = \frac{\lambda^*_{kit}}{1+\sum_{j=2}^{K}\lambda^*_{jit}}$ for $k = 2, \ldots, K$.

5.2.1 Incorporating zero-inflation

As detailed in the introduction, if some of the diseases are absent in an area then the ARMN model is not appropriate as it will always assign a positive probability to cases being reported. In this subsection, we will keep the definitions from above and extend the ARMN model to deal with long periods of disease absence by adapting the proposal of Zeng et al. (2022) to our space-time setting. Let S_{kit}, for $k = 1, \ldots, K$, be an indicator for the presence of disease k, so that $S_{kit} = 1$ if disease k was present in area i during time t and $S_{kit} = 0$ if disease k was absent. Under the zero-inflated multinomial framework of Zeng et al. (2022), the multinomial probabilities are given by,

$$\begin{aligned} \pi_{1it} &= \frac{S_{1it}}{S_{1it} + \sum_{j=2}^{K} S_{jit}\lambda^*_{jit}}, \\ \pi_{kit} &= \frac{S_{kit}\lambda^*_{kit}}{S_{1it} + \sum_{j=2}^{K} S_{jit}\lambda^*_{jit}}, \end{aligned} \tag{5.5}$$

for $k = 2, \ldots, K$. An immediate issue with (5.5) is that we cannot have $S_{1it} = \cdots = S_{Kit} = 0$ as we would divide by 0. Even if we were to define $\boldsymbol{\pi}_{it} = \mathbf{0}$ if $S_{1it} = \cdots = S_{Kit} = 0$, we would have the issue that this implies $\text{total}_{it} = 0$. A multinomial model conditions on the total count and does not generate it. There is nothing in the model of Zeng et al. (2022) to prevent $S_{1it} = \cdots = S_{Kit} = 0$. This may have been overlooked by them as they dealt with the case of $K \geq 50$ and, therefore, it is very unlikely $S_{1it} = \cdots = S_{Kit} = 0$. However, in our motivating example, we have $K = 3$.

There are a few possible solutions to prevent $S_{1it} = \cdots = S_{Kit} = 0$ including truncating the joint distribution of $\boldsymbol{S}_{it} = (S_{1it}, \ldots, S_{Kit})^T$. As it makes sense in many epidemiological applications, we will assume that the baseline disease is well-established in the region and is always present. That is, we assume $S_{1it} = 1$ for all i and t. For instance, in our motivating example, dengue has been circulating in Rio since 1986 and is well-established there (Teixeira et al., 2009). Also, most neighborhoods in Rio reported dengue cases fairly consistently between 2015-2016, see Figure 5.1 for an example. Therefore, we will take dengue as the baseline disease. We replace Equation (5.5) then by,

$$\begin{aligned} \pi_{1it} &= \frac{1}{1 + \sum_{j=2}^{K} S_{jit}\lambda^*_{jit}}, \\ \pi_{kit} &= \frac{S_{kit}\lambda^*_{kit}}{1 + \sum_{j=2}^{K} S_{jit}\lambda^*_{jit}}, \end{aligned} \tag{5.6}$$

for $k = 2, \ldots, K$.

The presence of a disease in an area should depend on whether it was present previously, as disease presence and absence are often highly persistent (Coutinho et al., 2006), and may also depend on whether other diseases were present previously, as the diseases can interact with one another (Paul et al., 2008; Sherlock et al., 2013). Also, the presence of a disease might depend on environmental or socioeconomic factors, like temperature and water supply in the case of vector-borne diseases (Schmidt et al., 2011; Xu et al., 2017). Therefore, we

model S_{kit} as,

$$S_{kit}|\boldsymbol{S}_{i(t-1)}, \boldsymbol{y}_{t-1} \sim \text{Bern}(p_{kit})$$

$$\text{logit}(p_{kit}) = \eta_{0k} + \boldsymbol{z}_{kit}^T\boldsymbol{\eta}_k + \underbrace{\rho_k^{AR} S_{ki(t-1)}}_{\substack{\text{autoregressive} \\ \text{effect}}} + \underbrace{\sum_{j>1, j\neq k} \rho_{jk}^{DI} S_{ji(t-1)}}_{\substack{\text{disease interaction} \\ \text{effect}}}, \tag{5.7}$$

where $\boldsymbol{z}_{kit}$ is a vector of space-time covariates that might affect the presence of disease k. As the disease may spread from neighboring areas (Grenfell et al., 2001), we allow functions of disease cases in neighboring areas to be added to $\boldsymbol{z}_{kit}$, hence the dependence on $\boldsymbol{y}_{t-1}$ in (5.7). For instance, the prevalence of the disease in neighboring areas could be added to $\boldsymbol{z}_{kit}$, see Section 5.4. Note, as ρ_k^{AR} in Equation (5.7) increases, the probability of getting a consecutive period of disease presence or absence approaches 1. Therefore, (5.7) can account for long consecutive periods of disease presence and absence which are often observed for infectious diseases, see Figure 5.1, Coutinho et al. (2006) and Adams and Boots (2010).

We will call the ARMN model modified by Equations (5.6)-(5.7) the Markov switching zero-inflated autoregressive multinomial (MS-ZIARMN) model. There are some properties of the MS-ZIARMN model that are worth mentioning. Firstly, from Equation (5.6), $\pi_{1it} + \cdots + \pi_{Kit} = 1$ for any possible values of the parameters and presence indicators. Also note that if diseases k and j are present in area i during time t we have $\pi_{kit}/\pi_{jit} = \lambda^*_{kit}/\lambda^*_{jit}$, which is the same as in the ARMN model described in Equations (5.2)-(5.4). That is, the MS-ZIARMN model preserves the relative distribution of the present diseases from the ARMN model. Although not explicitly mentioned by them, this is a convenient property of Zeng et al. (2022), in general, their framework preserves the relative distribution of the present categories from the non-zero-inflated model. Finally, if $K = 2$ and we assume Equation (5.6), where $\log(\lambda^*_{2it}) = \text{logit}(\pi_{2it}) = \alpha_{02i} + \boldsymbol{x}_{2it}^T\boldsymbol{\alpha}_2$ and $\text{logit}(p_{2it}) = \eta_{02} + \boldsymbol{z}_{2it}^T\boldsymbol{\eta}_2$, we get the well known and popular zero-inflated binomial (ZIB) model of Hall (2000). Therefore, our model without space-time effects could be seen as a multinomial extension of Hall (2000).

As we will show in the next section, the MS-ZIARMN model is a mixture of different multinomial distributions. More specifically, it is a Markov mixture or Markov switching model, meaning we can perform efficient Bayesian inference (Frühwirth-Schnatter, 2006).

5.3 Inferential Procedure

In this section, we will implicitly assume the number of diseases K is small enough for matrix multiplication with a 2^K by 2^K matrix to be computationally feasible. This should be the case in most epidemiological applications, for instance, in our motivating example we compare $K = 3$ diseases. Our inferential strategy revolves around reparametrizing the MS-ZIARMN model as a Markov switching model (Frühwirth-Schnatter, 2006), which allows for very efficient Bayesian inference using the forward filtering backward sampling (FFBS) algorithm (Chib, 1996). To simplify the explanations we will assume, without loss of generality, that $K = 3$.

In the case of $K = 3$, we have $\boldsymbol{S}_{it} = (S_{2it}, S_{3it})^T$ as S_{1it} is always equal to 1. Let S^*_{it} be an indicator for the possible values of $\boldsymbol{S}_{it}$, so that $S^*_{it} = 1$ if $\boldsymbol{S}_{it} = (1,1)^T$, $S^*_{it} = 2$ if $\boldsymbol{S}_{it} = (0,1)^T$, $S^*_{it} = 3$ if $\boldsymbol{S}_{it} = (1,0)^T$ and $S^*_{it} = 4$ if $\boldsymbol{S}_{it} = (0,0)^T$. Then we have that, from Equation (5.6),

$$\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it}, \text{total}_{it} \sim \begin{cases} \text{Multinom}\left(\left(\frac{1}{1+\lambda^*_{2it}+\lambda^*_{3it}}, \frac{\lambda^*_{2it}}{1+\lambda^*_{2it}+\lambda^*_{3it}}, \frac{\lambda^*_{3it}}{1+\lambda^*_{2it}+\lambda^*_{3it}}\right)^T, \text{total}_{it}\right), & \text{if } S^*_{it} = 1 \\ \text{Multinom}\left(\left(\frac{1}{1+\lambda^*_{3it}}, 0, \frac{\lambda^*_{3it}}{1+\lambda^*_{3it}}\right)^T, \text{total}_{it}\right), & \text{if } S^*_{it} = 2 \\ \text{Multinom}\left(\left(\frac{1}{1+\lambda^*_{2it}}, \frac{\lambda^*_{2it}}{1+\lambda^*_{2it}}, 0\right)^T, \text{total}_{it}\right), & \text{if } S^*_{it} = 3 \\ \text{Multinom}\left((1,0,0)^T, \text{total}_{it}\right), & \text{if } S^*_{it} = 4, \end{cases} \tag{5.8}$$

meaning, for $K = 3$, the MS-ZIARMN model is a mixture of 4 different multinomial dis-

tributions where the multinomial probabilities are fixed at 0 for the diseases absent in a component.

Now note that, from Equation (5.7), as $\boldsymbol{S}_{it}$ only depends on $\boldsymbol{S}_{i(t-1)}$, $\boldsymbol{y}_{t-1}$ and covariates, S^*_{it} follows a first order nonhomogeneous Markov chain. We will denote the transition matrix of S^*_{it} as $\Gamma(S^*_{it}|\boldsymbol{y}_{t-1})$, where $\Gamma(S^*_{it}|\boldsymbol{y}_{t-1})_{jk} = P(S^*_{it} = k|S^*_{i(t-1)} = j, \boldsymbol{y}_{t-1})$ for $j, k = 1, 2, 3, 4$, $i = 1, \ldots, N$ and $t = 2, \ldots, T$. As S^*_{it} is an indicator for $\boldsymbol{S}_{it}$, we can derive the transition matrix by calculating,

$$P(\boldsymbol{S}_{it} = \boldsymbol{s}_{it}|\boldsymbol{S}_{i(t-1)} = \boldsymbol{s}_{i(t-1)}, \boldsymbol{y}_{t-1}) = \prod_{k=2}^{3} P(S_{kit} = s_{kit}|S_{2i(t-1)} = s_{2i(t-1)}, S_{3i(t-1)} = s_{3i(t-1)}, \boldsymbol{y}_{t-1}),$$

for $\boldsymbol{s}_{i(t-1)}, \boldsymbol{s}_{it} = (1,1), (0,1), (1,0), (0,0)$, using Equation (5.7). Finally, as S^*_{i2} depends on S^*_{i1} we require an initial distribution for the Markov chain, that is, $P(S^*_{i1} = j)$ for $j = 1, 2, 3, 4$ and $i = 1, \ldots, N$. The initial distribution can be set by the modeler based on how likely they believe the diseases to be present at the beginning of the study period.

Given $p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it}, \text{total}_{it})$ and $\Gamma(S^*_{it}|\boldsymbol{y}_{t-1})$ for $i = 1, \ldots N$ and $t = 2, \ldots T$, and $p(S^*_{i1})$ for $i = 1, \ldots, N$, we completely define a Markov switching model (Frühwirth-Schnatter, 2006). A Markov switching model assumes a time series can be described by several submodels, often called states or regimes, where switching between submodels is governed by a first-order Markov chain (Hamilton, 1989). In our case, the submodels are given by the 4 multinomial distributions in (5.8) and we switch between these submodels through the transition matrix $\Gamma(S^*_{it}|\boldsymbol{y}_{t-1})$. Typically, S^*_{it} is called the state indicator.

Let $\boldsymbol{S}^*_t = (S^*_{1t}, \ldots, S^*_{Nt})^T$ be the vector of state indicators for time t, let $\boldsymbol{S}^* = (\boldsymbol{S}^*_1, \ldots, \boldsymbol{S}^*_T)^T$ be the vector of all state indicators, let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)^T$ be the vector of all observations, let $\boldsymbol{\beta}$ be the vector of all parameters in the multinomial part of the model, i.e., parameters in (5.2)-(5.4), let $\boldsymbol{\theta}$ be the vector of all parameters in the Markov chain part of the model, i.e., parameters in (5.7), and, finally, let $\boldsymbol{v} = (\boldsymbol{\beta}, \boldsymbol{\theta})^T$ be the vector of all model parameters.

Then the likelihood of $\boldsymbol{v}$ given $\boldsymbol{S}^*$ and $\boldsymbol{y}$ is given by,

$$p(\boldsymbol{S}^*, \boldsymbol{y}|\boldsymbol{v}) = \prod_{i=1}^{N}\prod_{t=2}^{T} p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it}, \text{total}_{it}, \boldsymbol{\beta}) \prod_{i=1}^{N} p(S^*_{i1}) \prod_{t=2}^{T} p(S^*_{it}|S^*_{i(t-1)}, \boldsymbol{y}_{t-1}, \boldsymbol{\theta}). \tag{5.9}$$

Note, from Equation (5.8), S^*_{it} is only fully known when $y_{1it}, y_{2it}, y_{3it} > 0$, in which case we must have $S^*_{it} = 1$. Therefore, the inferential procedure must deal with $\boldsymbol{S}^*$ in some way. It is possible, assuming K is not too large, to use the forward filter (Hamilton, 1989) to completely marginalize $\boldsymbol{S}^*$ from (5.9) and calculate $p(\boldsymbol{y}|\boldsymbol{v})$, see SM Section C.3. However, we want to make inferences about $\boldsymbol{S}^*$ to investigate when the model believes the diseases were present or absent for model checking, see Section 5.4.3. Therefore, we estimate $\boldsymbol{S}^*$ along with $\boldsymbol{v}$ by sampling both from their joint posterior distribution which, from Bayes' theorem, is proportional to,

$$p(\boldsymbol{v}, \boldsymbol{S}^*|\boldsymbol{y}) \propto p(\boldsymbol{S}^*, \boldsymbol{y}|\boldsymbol{v})p(\boldsymbol{v}), \tag{5.10}$$

where $p(\boldsymbol{v})$ is the prior distribution of $\boldsymbol{v}$. We specified wide independent gamma and normal priors for most lower-level elements of $\boldsymbol{v}$. We specified a conjugate inverse-Wishart prior for Σ with K degrees of freedom and identity matrix, as it implies the correlations in Σ have marginal uniform prior distributions (Gelman et al., 2013).

As the joint posterior (5.10) is not available in closed form, we resorted to Markov chain Monte Carlo (MCMC) methods, in particular, we used a hybrid Gibbs sampling algorithm with some steps of the Metropolis–Hastings algorithm to sample from it. We sampled each $\boldsymbol{\phi}_{it}$ in Equations (5.3)-(5.4) jointly using an adaptive blocked random walk metropolis step (Roberts and Sahu, 1997; Shaby and Wells, 2010). For our motivating example, we sampled some of the regression coefficients in $\boldsymbol{v}$ jointly using an automated factor slice sampler (Tibbits et al., 2014), as they mixed slowly and exhibited high posterior correlations. All other elements of $\boldsymbol{v}$ without conjugate priors were sampled individually using an adaptive random

walk metropolis step (Shaby and Wells, 2010). We sampled all of $\boldsymbol{S}^*$ jointly from $p(\boldsymbol{S}^*|\boldsymbol{v}, \boldsymbol{y})$ using the FFBS algorithm for Markov switching models (Chib, 1996). See SM Section C.3 for more details. We found our MCMC algorithm mixed much faster when jointly sampling $\boldsymbol{S}^*$ as opposed to sampling each presence indicator individually. This is typically the case for Markov switching models (Frühwirth-Schnatter, 2006) and was one of our main motivations for reparametrizing the MS-ZIARMN model as a Markov switching model.

Our MCMC algorithm was implemented using the R package Nimble (de Valpine et al., 2017). We implemented the FFBS algorithm using Nimble's custom sampler feature. All other MCMC samplers we used, which are mentioned above, are built into Nimble.

5.3.1 Fitted values

Comparing the fit of a time series model to the observed data is an important part of model checking in time series analysis (Knorr-Held and Richardson, 2003; Liboschik et al., 2017; Chen et al., 2019). We will let $\boldsymbol{y}'_{it}$ denote a fitted value. We assume that $\boldsymbol{y}'_{it}$ is a new count from the same MS-ZIARMN model (5.8) that is assumed to have conditionally generated $\boldsymbol{y}_{it}$, with the same past counts, parameter values, covariates and present diseases, except we assume a new value for the random effect $\boldsymbol{\phi}'_{it}$ corresponding to $\boldsymbol{y}'_{it}$. That is, we assume $\boldsymbol{y}'_{it}$ comes from the same space and time as $\boldsymbol{y}_{it}$ and from the same assumed model (5.8), except we posit a scenario where the unknown residual space-time factors influencing $\boldsymbol{y}'_{it}$ could be different. We assume a new value for the random effect ϕ_{kit} as they represent a residual for each observation and, therefore, it would be misleading to add them to the fit of the model.

To be more concrete, let $\boldsymbol{v}'$ represent $\boldsymbol{v}$ with Σ and $\boldsymbol{\phi}_{it}$, for all i and t, removed. We assume that $\boldsymbol{y}'_{it}|\boldsymbol{y}_{t-1}, S^*_{it}, \text{total}_{it}, \boldsymbol{v}', \boldsymbol{\phi}'_{it}$ follows the mixture distribution in (5.8) with λ^*_{kit} replaced by

$\lambda^{*'}_{kit}$ where,

$$\log(\lambda^{*'}_{kit}) = \alpha_{0ki} + \boldsymbol{x}_{it}^T\boldsymbol{\alpha}_k + \phi^{'}_{kit} + \zeta_k \log(y_{ki(t-1)} + 1) - \zeta_1 \log(y_{1i(t-1)} + 1).$$

We assume $\boldsymbol{\phi}'_{it}$ is a new value for the random effect corresponding to $\boldsymbol{y}'_{it}$ representing potentially different residual space-time factors. Therefore, $\boldsymbol{\phi}'_{it}$ does not feature in the likelihood and only depends on $\boldsymbol{y}$ through Σ, that is,

$$\boldsymbol{\phi}'_{it}|\Sigma, \boldsymbol{y} = \boldsymbol{\phi}'_{it}|\Sigma \sim \text{MVN}_{K-1}(\boldsymbol{0}, \Sigma).$$

The posterior distribution of $\boldsymbol{y}'_{it}$ is then given by,

$$p(\boldsymbol{y}'_{it}|\boldsymbol{y}) = \int_{\boldsymbol{v}',\Sigma,\boldsymbol{\phi}'_{it}} \sum_{S^*_{it}} p(\boldsymbol{y}'_{it}|\boldsymbol{y}_{t-1}, S^*_{it}, \text{total}_{it}, \boldsymbol{v}', \boldsymbol{\phi}'_{it}) p(\boldsymbol{\phi}'_{it}|\Sigma) p(\boldsymbol{v}', \Sigma, S^*_{it}|\boldsymbol{y})\, d\boldsymbol{v}' d\Sigma d\boldsymbol{\phi}'_{it}. \quad (5.11)$$

Let the superscript $[m]$, for $m = M+1, \ldots Q$, where M is the size of the burn-in sample and Q is the total MCMC sample size, denote a draw from the posterior distribution of a variable. From (5.11), to draw $\boldsymbol{y}^{'[m]}_{it} \sim p(\boldsymbol{y}'_{it}|\boldsymbol{y})$ we can first draw $\boldsymbol{\phi}^{'[m]}_{it}$ from $p(\boldsymbol{\phi}'_{it}|\Sigma = \Sigma^{[m]})$. Then we can draw $\boldsymbol{y}^{'[m]}_{it}$ from $p(\boldsymbol{y}'_{it}|\boldsymbol{y}_{t-1}, S^*_{it} = S^{*[m]}_{it}, \text{total}_{it}, \boldsymbol{v}' = \boldsymbol{v}^{'[m]}, \boldsymbol{\phi}'_{it} = \boldsymbol{\phi}^{'[m]}_{it})$. To help assess the fit of the model, we then compare the posterior mean and credible interval of $y^{'}_{kit}$ with the observed value y_{kit}.

The posterior probability that disease k was present in area i during time t is given by $P(S_{kit} = 1|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} S^{[m]}_{kit}$. It is important to view plots of $P(S_{kit} = 1|\boldsymbol{y})$ versus t for various diseases k and areas i for model checking, see Section 5.4.3. We cannot compare $P(S_{kit} = 1|\boldsymbol{y})$ to the true value of S_{kit} when $y_{kit} = 0$, since it is unknown. We usually do not know if the diseases were circulating whenever 0 cases were reported as they could have been circulating undetected. However, we can still check to make sure the model's estimates of when the diseases were present are reasonable and correspond to our general knowledge

of the epidemiological situation. Note, when $y_{kit} > 0$, $P(S_{kit} = 1|\boldsymbol{y})$ is not informative of the model fit as we will always have $P(S_{kit} = 1|\boldsymbol{y}) = 1$ if $y_{kit} > 0$, see Equation (5.8).

5.4 Application to Counts of Dengue, Zika and Chikungunya in Rio

Between 2015-2016 Rio de Janeiro experienced its first triple epidemic of dengue, Zika and chikungunya. Using space-time cluster analysis, Freitas et al. (2019) found that Zika and dengue clusters were much more likely in the western region of Rio (there were no chikungunya clusters found in the west), while other regions contained many clusters of each disease though they often differed in time. However, they did not investigate how these differences in clustering might depend on covariates or other factors. Schmidt et al. (2022), using a spatial multinomial regression model, found that cases of chikungunya were more likely in urban areas and Zika cases were less likely in population-dense areas, compared to dengue cases. A spatial analysis cannot investigate how space-time factors might have affected differences in the transmission intensity of the diseases, which may be important in this example, see below. Therefore, we apply our MS-ZIARMN model to cases of the three diseases, $K = 3$, across the 160 neighborhoods of Rio de Janeiro, $N = 160$, for each bi-week between 2015-2016, $T = 52$. Our model seems appropriate for this example as there appear to be long periods of disease absence in both the time series of chikungunya and Zika cases, see Figure 5.1 for example (overall 65% of Zika cases are equal to 0 and 75% of chikungunya cases are equal to 0).

5.4.1 Model specification and fitting

We modeled the Rio data with a bi-weekly time step as this roughly corresponds to the serial interval of the three diseases (Majumder et al., 2016; Riou et al., 2017). As mentioned in Section 5.2.1, we took dengue as the baseline disease (k=1) since it is well-established in the

city and cases are reported consistently throughout the study period. We then considered $k = 2$ to represent Zika and $k = 3$ to represent chikungunya.

For $\boldsymbol{x}_{kit}$ in (5.3), covariates that may affect the favorability of Zika or chikungunya transmission, relative to dengue transmission, we first took all spatial covariates considered in Schmidt et al. (2022): the percentage of neighborhood i covered in green area, verde_i; the social development index of neighborhood i, SDI_i; and the population density of neighborhood i, popdens_i. We additionally considered the percentage of neighborhood i that was occupied by favelas (slums), favela_i. These spatial covariates were chosen as the transmission of arboviral diseases can be significantly influenced by socioeconomic factors such as water supply and urbanicity (Schmidt et al., 2011), see Schmidt et al. (2022) for more details. As for space-time covariates, we first took the average weekly maximum temperature in neighborhood i during bi-week t, temp_{it}. Temperature is an important factor to consider since if the transmission of Zika or chikungunya is more sensitive (less sensitive) to temperature compared to dengue then their range of spread will be smaller (bigger) (Tesla et al., 2018; Mercier et al., 2022). If there are many more cases of Zika or chikungunya in neighboring areas, compared to dengue, then we would expect the share of those diseases to grow all else equal, due to between-area mixing (Stoddard et al., 2013). Therefore, we considered the previous prevalence of disease k and disease 1 (dengue) across neighboring areas, $\log\left(\frac{\sum_{j\in NE(i)} y_{kj(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$ and $\log\left(\frac{\sum_{j\in NE(i)} y_{1j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$, in $\boldsymbol{x}_{kit}$, where $NE(i)$ is the set of neighboring areas of neighborhood i and pop_j is the population of neighborhood j. We considered two areas neighbors if they shared a border. To simplify the model, we assumed the effect of neighboring dengue prevalence was the same on both $\log(\pi_{2it}/\pi_{1it})$ and $\log(\pi_{3it}/\pi_{1it})$, like with the within-area autoregression. Finally, it was speculated in Freitas et al. (2019) that Zika circulation could have been inhibiting chikungunya transmission. To account for these kinds of disease interactions, we considered $\log(y_{3i(t-1)} + 1)$ in $\boldsymbol{x}_{2it}$ and $\log(y_{2i(t-1)} + 1)$ in $\boldsymbol{x}_{3it}$. That is, chikungunya cases were allowed to affect the favorability of Zika transmission relative to dengue transmission and Zika cases were allowed to affect the

favorability of chikungunya transmission relative to dengue transmission. To summarize, for covariates potentially related to the favorability of Zika transmission compared to dengue transmission, we considered,

$$\boldsymbol{x}_{2it} = \Bigg(\text{verde}_i, \text{SDI}_i, \text{popdens}_i, \text{favela}_i, \text{temp}_{it}, \log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right), \log\left(\frac{\sum_{j\in NE(i)} y_{1j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right), \log(y_{3i(t-1)} + 1)\Bigg)^T,$$

and similar for chikungunya.

As for covariates potentially associated with the presence of Zika or chikungunya, $\boldsymbol{z}_{kit}$ in (5.7), we considered mostly the same factors included in $\boldsymbol{x}_{kit}$. An exception is that we do not include the cases of the other disease as the Markov chain defined in (5.7) already accounts for disease interactions through ρ_{jk}^{DI}. Also, since the presence is not defined relatively, we do not include neighboring dengue prevalence. Therefore, for covariates potentially associated with the presence of Zika, we considered,

$$\boldsymbol{z}_{2it} = \left(\text{verde}_i, \text{SDI}_i, \text{popdens}_i, \text{favela}_i, \text{temp}_{it}, \log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)\right)^T,$$

and similar for chikungunya.

We also need to specify the initial state distributions, see Section 5.3. As there were no chikungunya cases reported at all in the first year, we assumed chikungunya had only a 5 percent chance of being present at the start of 2015. For Zika, there are a few, but not many, cases reported through most of 2015 and, therefore, we assumed a 10 percent chance of initial presence. We fitted the MS-ZIARMN model specified above to the Rio data using our proposed Gibbs sampler from Section 5.3. We ran the Gibbs sampler for 250,000 iterations on three chains with an initial burn-in of 50,0000 iterations. All sampling was started from random values in the parameter space to avoid convergence to local modes. Convergence was

checked using the Gelman-Rubin statistic (all estimated parameters <1.05), the minimum effective sample size (>1000) and by visually examining the traceplots (Plummer et al., 2006).

For comparison purposes, we also considered a model without zero inflation to ensure there is justification for treating the many zeroes in the data, as hypothesized. That is, the ARMN model described in Equations (5.1)-(5.4) with the same covariates $\boldsymbol{x}_{2it}$ and $\boldsymbol{x}_{3it}$ as the MS-ZIARMN model. We also considered a zero-inflated model without the coupled Markov chains, that is, with $\rho_k^{AR} = 0$ and $\rho_{jk}^{DI} = 0$ for all k and $j > 1$, $j \neq k$ in Equation (5.7), which we will call the ZIARMN model. Without the coupled Markov chains, our model is a finite mixture and not a Markov mixture model, and the inference is greatly simplified, e.g., there is no need to use the FFBS algorithm. Therefore, we want to make sure this component is important. Finally, we compared with the zero-inflated multinomial model of Zeng et al. (2022) as our approach is built on their framework. Zeng et al. (2022) assumed that the probability a category is absent from an observation only depended on the category and, following this, we let $\text{logit}(p_{kit}) = \eta_{0k}$ for $k = 2, 3$ in Equation (5.7). We kept the rest of the model the same as the MS-ZIARMN model. Note, Zeng et al. (2022) did not consider any covariates or space-time structure in the multinomial probabilities of the present categories. However, this is inappropriate for our example and did not fit the data well (results not shown). Also, Zeng et al. (2022) did not assume the baseline category was always present, however, this appears to be necessary when modeling a small number of categories, see Section 5.2.1.

Table 5.1 shows the widely applicable information criterion (WAIC) (Gelman et al., 2014) of the 4 considered models. We calculated the WAIC by marginalizing out the state indicators $\boldsymbol{S}^*$ as recommended by Auger-Méthé et al. (2021), see SM Section C.4 for more details. Note, the model with the smallest WAIC is considered to have the best fit and as a rule of thumb a difference of 5 or more in the WAIC is considered significant. Table 5.1 shows it is not only

Table 5.1: Shows the WAIC of the 4 considered models from Section 5.4.1 fitted to the Rio data. The best fitting model, the one with the lowest WAIC, is bolded.

Model	WAIC
MS-ZIARMN	**22,091**
ARMN	23,974
ZIARMN	22,756
Zeng (2022)	23,986

important to account for zero inflation, but also covariates and correlations across space, time and disease in the zero-inflated process (i.e., the process of disease presence). Indeed, the model of Zeng et al. (2022) had a worse fit compared to the ARMN model (without zero inflation), despite the other zero-inflated models having a much better fit.

5.4.2 Results

Table 5.2 shows the estimated coefficients from the multinomial part of the fitted MS-ZIARMN model, that is the intercepts and covariate effects from Equation (5.3). From the intercept row, at average values of the covariates, Zika transmission was favored over dengue and chikungunya transmission. That is, under average conditions the share of Zika cases tended to grow over time in an area (at least from an equal share, see the nuances in Section 5.2). It is interesting to compare these estimated intercepts with those from the ARMN and Zeng (2022) models in SM Table C.1. Unlike the MS-ZIARMN model, the Zeng (2022) and ARMN models both estimated that on average Zika and chikungunya transmission was much less intense in Rio compared to dengue transmission. This does not correspond to our knowledge of the epidemiology of these diseases and is likely due to a failure to properly account for the many zeroes. Zika is generally considered more transmisable due to *Ae. aegypti* transmitting Zika at a higher rate (Freitas et al., 2019). This further illustrates it is important to account for zero inflation and to model zero inflation properly (e.g., account for correlations).

Table 5.2: Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted MS-ZIARMN model. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Section 5.4.1 for an explanation of the covariates.

	Relative Odds Ratio or Rate Ratio	
Covariates	**Zika-dengue**	**chik.-dengue**
Intercept	1.14 (1.03, 1.26)	1.02 (.91, 1.13)
verde_i	1.02 (.95, 1.09)	.92 (.85, 1)
SDI_i	1.07 (.99, 1.15)	1.01 (.92, 1.11)
popdens_i	1.02 (.94, 1.11)	1.06 (.96, 1.16)
favela_i	.98 (.91, 1.05)	.94 (.87, 1.03)
temp_{it}	**1.14 (1.09, 1.20)**	**.85 (.80, .90)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{1j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**.70 (.67, .74)**	**.70 (.67, .74)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**1.59 (1.48, 1.70)**	–
$\log\left(\frac{\sum_{j\in NE(i)} y_{3j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	–	**1.43 (1.36, 1.50)**
$\log(y_{2i(t-1)} + 1)$	–	**.90 (.85, .96)**
$\log(y_{3i(t-1)} + 1)$	.98 (.95, 1.02)	–

Interestingly, unlike in Schmidt et al. (2022), we did not find strong evidence that any of the spatial covariates were important for explaining differences in the transmission of the diseases. We found that Zika transmission was favored at warmer temperatures and chikungunya transmission was favored at colder temperatures, compared to dengue transmission. Both of these findings have been observed in laboratory studies (Tesla et al., 2018; Mercier et al., 2022). Indeed, a concern of chikungunya is that it could spread into colder areas of Europe unreachable by dengue (Mercier et al., 2022). To better quantify the temperature effects, we plot λ_{kit}, from Equation (5.2), versus temperature for $k = 2$ (Zika) and $k = 3$ (chikungunya)

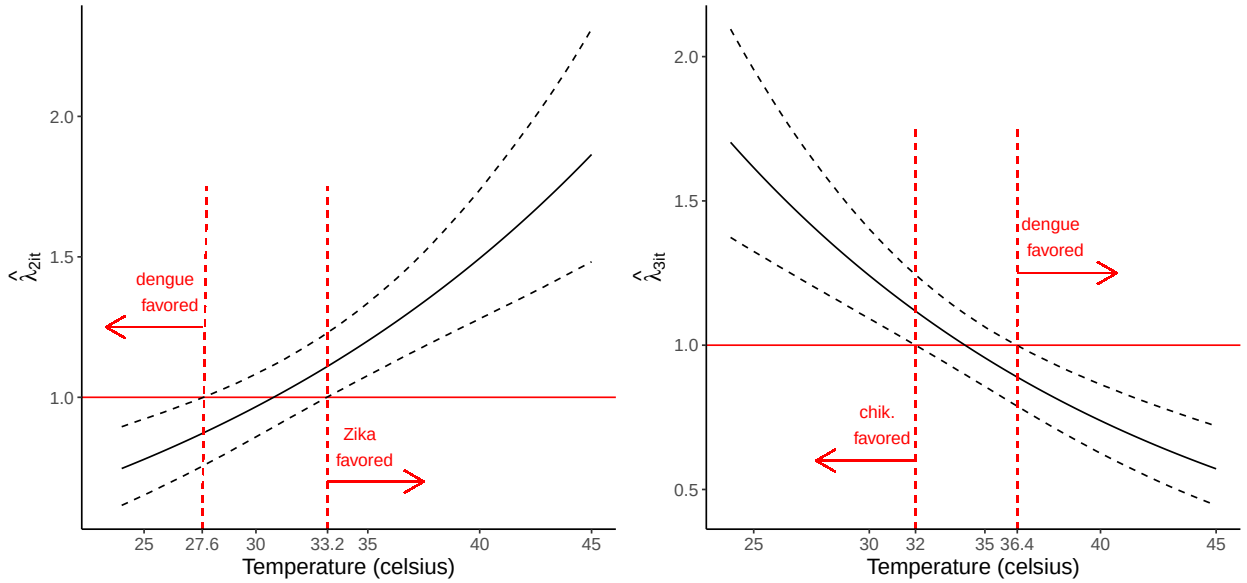


Figure 5.2: Shows the posterior means (black solid line) and 95% posterior credible intervals (black dashed lines) of λ_{kit} versus temperature for $k = 2$ (Zika) (left) and $k = 3$ (chikungunya) (right). We assumed an average area, other covariates fixed at their average values and a null space-time random effect. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. The red horizontal dashed lines are drawn at temperatures below and above which λ_{kit} is significantly different from 1.

in Figure 5.2. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the transmission of disease k is favored (disfavored) over dengue transmission, and the share of disease k relative to dengue will tend to grow (shrink) over time. From the figure, we have strong evidence, under otherwise average conditions, that Zika transmission was favored above 33.2 degrees and chikungunya transmission was favored below 32 degrees, compared to dengue transmission.

The neighboring prevalence effects are by far the strongest in Table 5.2 for explaining differences in the relative distribution of the diseases. For instance, we estimated that a standard deviation increase in the log of the prevalence of Zika across neighboring areas during the previous week was associated with a 59 (48, 70) percent increase in the odds of a new case being Zika relative to dengue. As *Aedes aegypti* primarily bite during the day, its arboviruses can spread rapidly between areas (Stoddard et al., 2013). We found that areas with more Zika cases were less favorable towards the transmission of chikungunya compared to dengue.

That is, we estimated that a standard deviation increase in the log of the previous number of Zika cases in an area was associated with a 10 (4, 15) percent decrease in the odds of a new case being chikungunya relative to dengue. This does not necessarily mean that Zika transmission inhibited chikungunya transmission in a non-relative sense. It could be that Zika transmission encouraged both dengue and chikungunya transmission but the effect on dengue transmission was larger. Therefore, this result needs to be interpreted with some caution.

We estimated ζ_1, ζ_2 and ζ_3, from Equation (5.2), as .44 (.38, .49), .36 (.31, .32) and .43 (.36, .51) respectively, indicating a high degree of nonhomogeneous mixing. That is, the model predicts that if one of the diseases becomes very dominant in a neighborhood the relative distribution of the diseases will move towards a uniform distribution, perhaps because individuals avoid infection from the dominant disease. However, as avoiding one of the diseases should help avoid the others, as they are transmitted by the same vector, ζ_k could also be mainly reflecting differences in disease immunity, which are hard to account for explicitly, see Section 5.4.4. From Equations (5.3)-(5.4), we estimated the standard deviation of ϕ_{2it} as .75 (.71, .79), of ϕ_{3it} as .80 (.75, .86) and the correlation between ϕ_{2it} and ϕ_{3it} as .59 (.52, .65). This implies that the correlation between the disease cases, marginalizing the random effects and conditional on their total, is about what we would expect without the random effects, see the simulation study in SM Section C.2. That is, the random effects do not induce any excess correlation between the disease counts. However, they do induce a large amount of overdispersion, see Figure 5.4.

Table 5.3 shows the estimated parameters from the Markov chain component of the model, that is, from Equation (5.7). The Zika column shows the ratio of the odds of Zika presence corresponding to a unit increase in the covariate. However, the intercept for chikungunya is highly negative making the odds ratios difficult to interpret and, therefore, we did not transform the chikungunya estimates. From the intercept row, chikungunya was very unlikely

Table 5.3: Posterior means and 95% posterior credible intervals (in parentheses) for the estimated parameters from the Markov chain part of the fitted MS-ZIARMN model. The intercept row shows the probability of Zika presence and the logit of the probability of chikungunya presence, assuming no previous disease presence in the area, no cases in neighboring areas and at average values of the other covariates. The Zika column shows the ratio of $p_{2it}/(1-p_{2it})$ (odds of Zika presence) corresponding to a unit increase in the covariate. The chik. column shows the untransformed estimates (not exponentiated, see text). All covariates are standardized except the neighboring prevalences which are given in units of 1 case per 25,000 (average combined size of neighboring areas) at small prevalences.

	Odds Ratio	**Untransformed**
Covariate	**Zika presence**	**chik. presence**
Intercept	.1 (.08, .12)	-9.06 (-11.97, -7)
verde_i	1.14 (.98, 1.31)	08 (-.20 .37)
SDI_i	1.11 (.96, 1.29)	.12 (-.18, .42)
popdens_i	1 (.85, 1.19)	.02 (-.33, .37)
favela_i	1.03 (.89, 1.19)	0 (-.3, .29)
temp_{it}	**1.44 (1.24, 1.66)**	**1.11 (.6, 1.7)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j}+1\right)$	**1.70 (1.55, 1.87)**	–
$\log\left(\frac{\sum_{j\in NE(i)} y_{3j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j}+1\right)$	–	**1.09 (.85, 1.35)**
$S_{2i(t-1)}$	**6.41 (3.95, 9.94)**	**5.12 (3.2, 8)**
$S_{3i(t-1)}$	**3.46 (2.42, 4.98)**	**11.55 (8.13, 16.12)**

to emerge in an area under average conditions. Indeed, there are only two effects large enough from Table 5.3 to cause chikungunya emergence: (1) Zika emerged and then chikungunya emerged in the following summer when temperatures were high, and (2) chikungunya spread from a neighboring area where it was detected. In contrast, the emergence of Zika was much more random, occurring with a 10 (8, 12) percent chance bi-weekly under average conditions. Once chikungunya did emerge it was likely to stay in the area (probability of .88 (.63, .99) assuming no neighboring cases, no Zika presence in the previous bi-week and otherwise average conditions). Both Zika and chikungunya were more likely to be present at high

temperatures and when there were cases in neighboring areas. The effects of neighboring prevalence are especially large, again illustrating the disease's high rate of geographical spread. Zika presence was an indicator for chikungunya presence and chikungunya presence was an indicator for Zika presence. This could be unrelated to disease interactions and due to unmeasured confounding as the diseases share the same vector.

5.4.3 Model fit

Firstly, we want to investigate how the favorability of the diseases differed across Rio. The maps in Figure 5.3 show the posterior mean of $\bar{\lambda}_{ki} = \left(1/\sum_{t=2}^{T} S_{kit}\right) \sum_{t=2}^{T} \lambda_{kit} S_{kit}$ (average value of λ_{kit} when disease k was present in neighborhood i) for $k = 2$ (Zika) (top map) and $k = 3$ (chik.) (bottom map). Note, if $\bar{\lambda}_{ki} > 1$ then the transmission of disease k was favored on average over dengue when it was present in the neighborhood. From the bottom map, chikungunya transmission was disfavored in the west of the city when it was present and the share of chikungunya cases tended to decline there. This was partly due to temperatures being 2-3 degrees higher on average in the west, see SM Figure C.2, and chikungunya transmission being favored at colder temperatures relative to the other two diseases, see Table 5.2. Also, there were many Zika cases in the west of the city, and chikungunya transmission was disfavored in areas with many Zika cases. Zika transmission was favored throughout most of the city when it was present. This can be partly explained by the inherent favorability of Zika, i.e., the intercept in Table 5.2, and temperatures typically being higher in Rio when Zika was present (the temperature covariate was centered using the overall average temperature) and Zika transmission being favored at higher temperatures.

Figure 5.4 shows fitted values and posterior probabilities of disease presence from three models. By comparing panels (a) and (b), the random effects ϕ_{kit}, from Equations (5.3)-(5.4), are an important component of the model and induce a large amount of necessary overdispersion. From panel (c), the model of Zeng et al. (2022), which does not account for correlations in disease presence across space, time and disease, estimates chikungunya

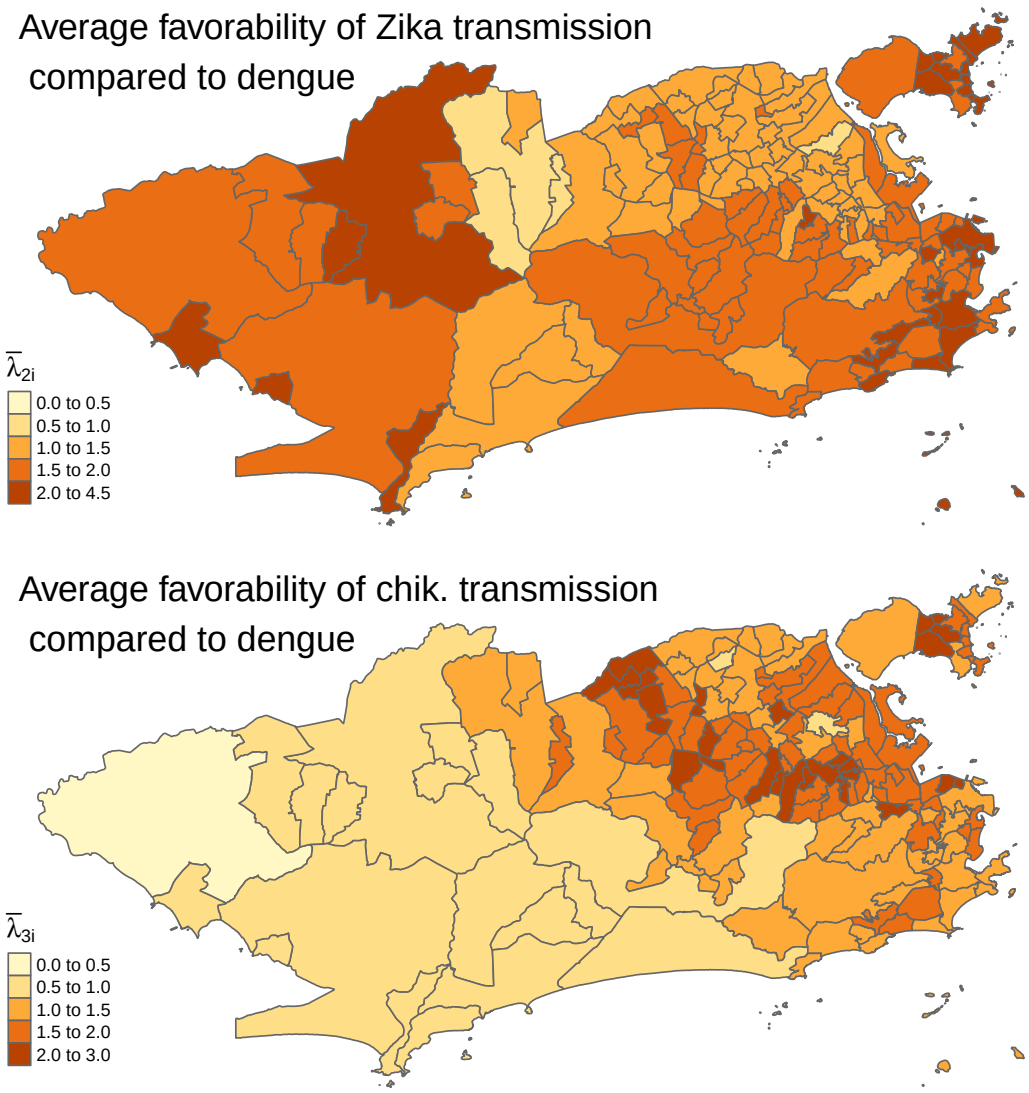


Figure 5.3: Shows the posterior mean of $\bar{\lambda}_{ki} = \left(1/\sum_{t=2}^{T} S_{kit}\right) \sum_{t=2}^{T} \lambda_{kit} S_{kit}$ for $k = 2$ (Zika) (top map) and $k = 3$ (chik.) (bottom map). If $\bar{\lambda}_{ki} > 1$ ($\bar{\lambda}_{ki} < 1$) then the transmission of disease k was favored (disfavored) on average over dengue when it was present in the neighborhood, and the share of disease k tended to grow (shrink) relative to dengue in the neighborhood.

as always present in Catumbi, except for the initial state (this is also true for Zika and in all other areas). Therefore, the model of Zeng et al. (2022) is too restrictive to produce the likely complex space-time distribution of disease presence in the city. Comparing panels (a) and (d) shows the MS-ZIARMN model believes the presence of chikungunya was much more persistent compared to Zika, which follows Table 5.3 (from the last two rows). One

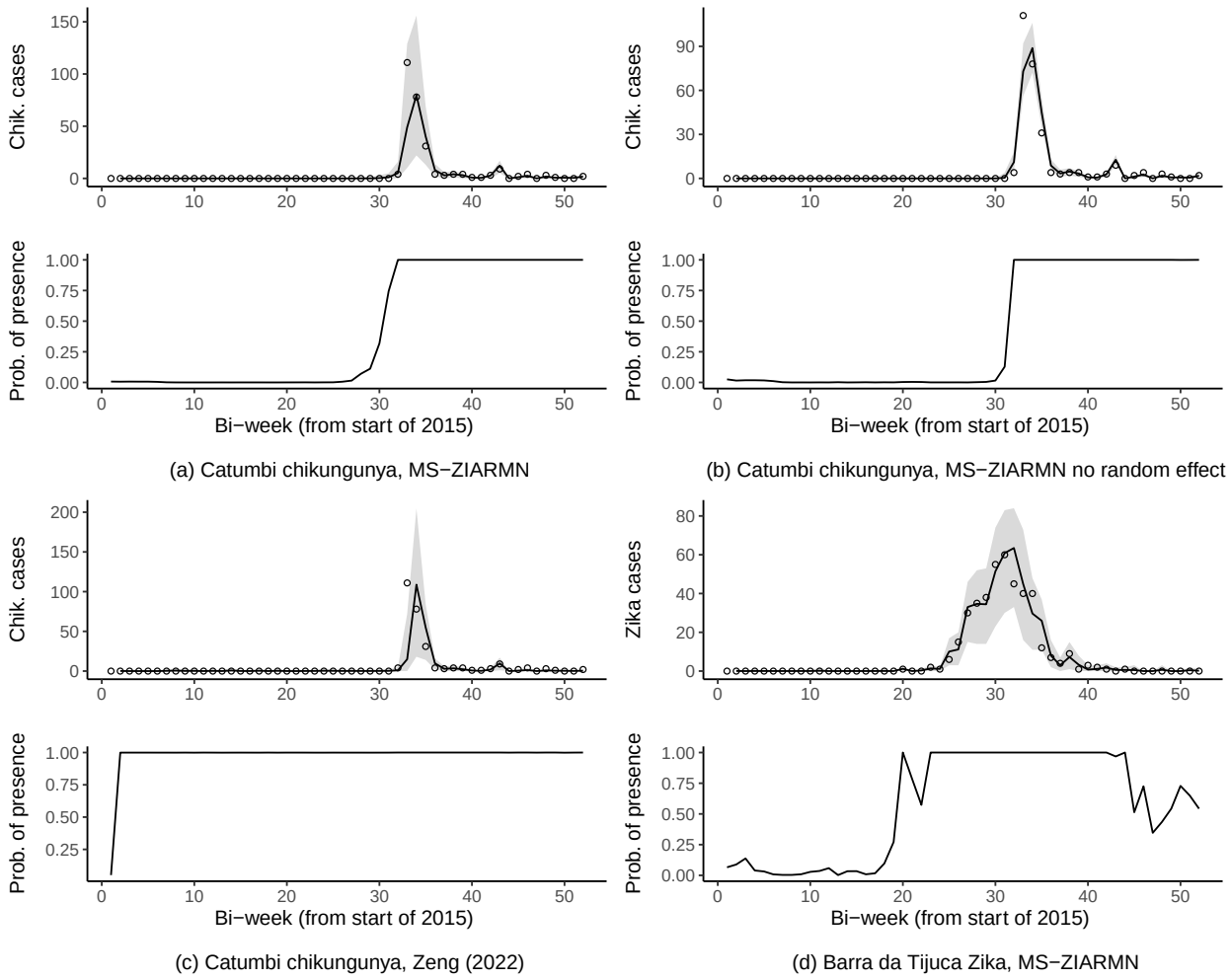


Figure 5.4: Top graphs show the posterior means (solid lines) and 95% posterior credible intervals (shaded areas) of the fitted values y'_{kit}, see Section 5.3.1. The bottom graphs show the posterior probability the disease was present in each bi-week, $P(S_{kit} = 1|\boldsymbol{y})$ versus t. "MS-ZIARMN no random effect" refers to the MS-ZIARMN model fit without including ϕ_{kit} from Equations (5.3)-(5.4).

way to think of $P(S_{kit} = 1|\boldsymbol{y})$ (from the bottom graphs) is as a weight for the observation's contribution to the likelihood of the multinomial parameters of the model. For example, from panel (a), the MS-ZIARMN model does not use any chikungunya observations before around week 28 in Catumbi to inform the multinomial estimates in Table 5.2. In contrast, the model of Zeng et al. (2022), from panel (c), will use all observations to inform the multinomial model which likely causes bias due to the large number of 0s, see the discussion at the beginning of Section 5.4.2.

Finally, we will discuss some nuances in interpreting $P(S_{kit} = 1|\boldsymbol{y})$ as a probability of disease presence. The model takes disease absence to mean there was no chance of cases being reported. However, if the diseases are present in very small amounts, the chance of cases being reported could be close enough to 0 that it is indistinguishable. Indeed, if we replaced the 0s in Equation (5.8) by say .005, and normalized the probabilities, we would likely get similar parameter estimates. Therefore, disease presence should be interpreted closer to "a feasible chance of cases being reported" than "the presence of at least one infection".

5.4.4 Sensitivity analysis to control for differences in the susceptible populations

As mentioned in Section 5.2, differences in the susceptible populations of the diseases can be an important source of unmeasured confounding. If one of the diseases has a much larger susceptible population then its transmission will be favored. It is difficult to directly measure susceptible populations of vector-borne diseases (also most diseases) due to underreporting, different strains, cross-immunity, waning immunity and demographic changes (Reich et al., 2013; Freitas et al., 2019; Aogo et al., 2023). As a proxy for differences in the susceptible populations, we considered the differences in cumulative incidence per person. That is, we refit the MS-ZIARMN model including $\left(\sum_{j=1}^{t-1} y_{kij} - \sum_{j=1}^{t-1} y_{1ij}\right)/\text{pop}_i$ in $\boldsymbol{x}_{kit}$ for $k = 2$ (Zika) and k=3 (chik.). The multinomial estimates are given in SM Table C.3 and none are substantially different from those in Table 5.2. The effects of the differences in cumulative incidence were not significant and the WAIC of the model increased by more than 5 indicating a worse fit. That is, if one of the diseases accumulated many more cases compared to the others, there was likely not a large effect on the favorability of its transmission. One explanation is that the parameters ζ_k in Equation (5.2) are already mostly accounting for this, as the effect of nonhomogeneous mixing on transmission is similar to the effect of the depletion of the susceptible population.

5.5 Discussion

We have proposed a zero-inflated spatio-temporal multinomial model for comparing the transmission dynamics of multiple co-circulating infectious diseases. Our approach can account for long periods of disease absence while investigating how certain factors are related to differences in the transmission intensity of the diseases. Past zero-inflated multinomial models (Tang and Chen, 2019; Zeng et al., 2022; Koslovsky, 2023) have made independence assumptions that are inappropriate for spatio-temporal infectious disease counts, such as assuming categorical presence was independent between observations. We accounted for correlations across space, time and disease, in both the multinomial and zero-inflated components, through a combination of autoregression, regressing on past observations, and by assuming the presence of the diseases followed a series of coupled Markov chains. This approach allowed for efficient and computationally feasible Bayesian inference using the FFBS algorithm.

We assumed that the cases of the diseases present in an area jointly followed a multinomial model, conditioning on the total cases. It is more standard in the disease mapping literature to jointly model multiple disease counts with some form of a multivariate Poisson distribution (Jack et al., 2019). There are advantages to both approaches and we give a more detailed comparison of the two (without model fitting) in SM Section C.1.2. If one is only interested in comparing the transmission of the diseases, like here, then a multinomial model eliminates many nuisance parameters and any shared space-time factors. A multivariate Poisson approach could estimate the overall effects of a covariate on transmission, not just the relative effects, and does not need to assume one of the diseases is always present. To the best of our knowledge, only Pavani and Moraga (2022) and Rotejanaprasert et al. (2021) have considered zero-inflated spatio-temporal multivariate Poisson models. Both relied on random effects to capture correlations in the data. However, these proposals have some important limitations including not considering any space-time or space-time-disease inter-

actions and not allowing the probabilities of disease presence to change across time. These modeling assumptions are inappropriate for our motivating example, and likely others, see Figure 5.4 for instance. An approach based on random effects could become intractable as one would need to consider random effects correlated across three dimensions, space, time and disease, in both the count and zero-inflated components of the model. Our approach seems much more computationally feasible and could be easily extended to the multivariate Poisson case, see SM Section C.1.2. We will focus on these extensions and comparisons with the multinomial model in future work.

In our application, we found that Zika generally had more intense transmission compared to dengue and chikungunya, but was also not able to transmit as well at a lower temperature. This was not a major factor in tropical Rio, see Figure 5.3, but could be important in North America and Europe. In contrast, chikungunya transmission was relatively higher at lower temperatures meaning it might fare better in colder regions compared to Zika and dengue. Laboratory studies have come to similar conclusions (Tesla et al., 2018; Mercier et al., 2022). Alternative models that did not account for zero inflation or did not model correlations in disease presence (Zeng et al., 2022) did not fit as well and produced estimates inconsistent with our knowledge about the epidemiology of the diseases.

There are also some important limitations to our work. Being an observational study, unmeasured confounding could be an important issue, especially concerning differences in the susceptible populations, which are unobservable. We tried using differences in cumulative incidence as a proxy for differences in the susceptible populations in a sensitivity analysis, and our results did not change. However, differences in cumulative incidence could be an imperfect proxy due to, for example, changes in reporting rates across space and time. Another limitation is that climate covariates, such as temperature, can have lagged and non-linear effects on the transmission of arboviral diseases (Lowe et al., 2018). As our multinomial model estimates differences in effects on transmission, our approach might be somewhat

robust to non-linearity if the non-linear patterns are similar between diseases. It would be challenging to incorporate non-linear and lagged effects into our already complex modeling scheme. Shrinkage methods could help deal with the many parameters that would need to be added to the model by eliminating unimportant effects (Wang et al., 2023).

Acknowledgements

This work is part of the PhD thesis of D. Douwes-Schultz under the supervision of A. M. Schmidt in the Graduate Program of Biostatistics at McGill University, Canada. Douwes-Schultz is grateful for financial support from IVADO and the Canada First Research Excellence Fund/Apogée (PhD Excellence Scholarship 2021-9070375349). Schmidt is grateful for financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant RGPIN-2017-04999). This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

Chapter 6

Conclusion

6.1 Summary

In this thesis, we proposed several novel coupled Markov switching models to account for, study, detect, and forecast abrupt shifts in the behavior of spatio-temporal infectious disease counts, such as due to an outbreak emerging or a disease going extinct. We assumed the shifts were caused by the disease switching between different epidemiological periods or states. We considered shifts due to disease (re)emergence (absence to presence), disease extinction (presence to absence), outbreak emergence (endemic to outbreak) and outbreak end (outbreak to endemic) (i.e., the inverse of outbreak persistence). Other epidemiological transitions could also be analyzed, e.g., outbreak plateau to decline (Lytras et al., 2019), by adding more states to the model (Cliff et al., 1987).

We made three important main contributions to the existing epidemiological Markov switching literature (see Amorós (2017) for a recent review and Cliff et al. (1987)) in a space-time setting. Firstly, we considered a disease absence state in our models to account for the long strings of zeroes common in spatio-temporal infectious disease counts (Arab, 2015). This also allowed us to study the reemergence and persistence of the disease. For instance, in the first

manuscript, we investigated how temperature, socioeconomic factors, and the presence of the disease in neighboring areas were related to the reemergence and persistence of dengue fever in Rio neighborhoods. Secondly, we allowed the probability of transitioning between states (i.e., the probability of an abrupt shift in behavior) to depend on covariates **and** the states in neighboring areas, to account for geographical disease spread. For instance, in the second manuscript, we let the probability of an outbreak emerging in an area depend on mobility in retail and recreation venues and whether outbreaks were occurring in neighboring areas. Finally, in the third manuscript, we considered a model where multiple (existing approaches have only focused on a single disease) interacting diseases switch between different epidemiological periods across several areas. We used the model to investigate important questions (Tesla et al., 2018; Freitas et al., 2019) about how certain factors, such as temperature, were related to differences in the transmission intensity of Zika, dengue and chikungunya in Rio de Janeiro. We also make many important contributions to the broader area (outside just a Markov switching framework) of spatio-temporal infectious disease modeling, see the preambles and below.

Spatio-temporal infectious disease modeling is mostly dominated by approaches outside a Markov switching framework, such as spatio-temporal autoregressive models (Bracher and Held, 2022) and epidemiological compartmental models (Crawford et al., 2022). Therefore, it is worth discussing why, in general, our Markov switching models are useful in epidemiological modeling. Firstly, we can study the transitions between the disease states and how they depend on certain factors. For instance, in the second manuscript, we investigated the relationship between the emergence of COVID-19 outbreaks across Quebec and factors such as mobility and geographical disease spread. We found that the introduction of new variants and outbreak spread from neighboring areas were important contributors to the emergence of the outbreaks. This indicates that carefully monitoring the introduction of new variants and preventing cross-infection between areas could be effective interventions for preventing future COVID-19 outbreaks from emerging. In general, gaining a better understanding of how the

outbreaks of an infectious disease develop can help give early warnings or suggest preventive measures (Descloux et al., 2012; Lu et al., 2014). Therefore, we would argue that being able to study important epidemiological transitions, like outbreak emergence, is a very useful feature of our models. While spatio-temporal autoregressive models and compartmental models can investigate the relationship between a covariate and disease transmission (Bauer and Wakefield, 2018; Ssentongo et al., 2021), such a relationship does not necessarily imply an association with outbreak emergence, as explained in the introduction in Chapter 1. Additionally, past epidemiological Markov switching models have allowed the probability of outbreak emergence to depend on covariates (Nunes et al., 2013) or whether outbreaks were occurring in neighboring areas (Heaton et al., 2012), but not both like with our approach. As mentioned, we found that both covariates, like mobility and the introduction of new variants, and the occurrence of outbreaks in neighboring areas were important for explaining the emergence of COVID-19 outbreaks across Quebec.

Another way that our models can be useful in epidemiology is that they can approximate the posterior probability that an observation, including future observations, came, or will come, from any of the disease states. As illustrated in the second manuscript, this allows us to provide the posterior probability that an outbreak is currently happening or will happen in the next week for each area. If one of these probabilities rises above .5 then an alarm could be issued as in Martínez-Beneito et al. (2008). Issuing an alarm early during an outbreak is very important, as the rapid rise in cases can overwhelm the health care system (Shingler and Hendry, 2022). In the second manuscript, our model was able to provide early real-time warnings of outbreaks for both simulated data and COVID-19 hospitalizations across Quebec. This was especially true when including neighboring outbreak indicators in the transition probabilities. See Sections 4.4 and 4.5.4. Heaton et al. (2012) also allowed the probability of outbreak emergence to depend on whether outbreaks were occurring in neighboring areas. However, their model can only be applied to data with at most one outbreak in each area, and they did not allow the probability of outbreak emergence to

depend on covariates. Therefore, their model is not widely applicable. For instance, for our application to COVID-19 hospitalizations across Quebec, there were multiple outbreaks per area, and several covariates, such as mobility and the introduction of new variants, were important for explaining the emergence of the outbreaks.

Finally, the state switching in our models can account for behavior that other statistical models might struggle to capture. For instance, negative binomial and multinomial models can often not account for a large number of zeroes in univariate or multivariate counts (Arab, 2015; Koslovsky, 2023). Long strings of zeroes are a common feature of infectious disease time series (Chen et al., 2019). As explained in Section 3.2, switching with the absence state in our models can produce many consecutive zeroes interspersed between periods of high or low disease activity. Finite mixture zero-inflated models (Hoef and Jansen, 2007), which are popular for dealing with excess zeroes in spatio-temporal infectious disease counts (Arab, 2015), also switch with an absence state. However, the switching is not governed by a Markov chain but by a series of conditionally independent Bernoulli random variables. As extensively detailed in the first manuscript, our approach has several advantages including allowing for each covariate to have a separate effect on the emergence and persistence of the disease, which is often epidemiologically justified, and allowing the effects of disease spread between areas to depend on space-time factors. Model comparison using WAIC supported Markov switching models that switched with an absence state in all three manuscripts; compared to finite mixture zero-inflated models, Markov switching models that switched between only count states, and autoregressive negative binomial and multinomial models without state switching. See Tables 3.1, 4.1 and 5.1. Note that past epidemiological Markov switching models did not include an absence state and could struggle with capturing many zeroes in the counts (Zou et al., 2014). Markov switching models outside of epidemiology have considered switching with a degenerate zero distribution (Wang, 2001; Malyshkina and Mannering, 2010). However, these models did not allow the transition probabilities to depend on the state in neighboring areas and, therefore, are not appropriate in an epidemiological setting

where we expect the disease to spread between areas. We will expand on some of the points from the last three paragraphs in the next section, where we provide specific references from the spatio-temporal epidemiological modeling literature concerning potential applications of our models.

6.2 Potential Applications

There are several useful potential applications of this thesis to analyzing spatio-temporal infectious disease counts. Firstly, our models can be used when there is an interest in investigating associations between certain factors, including disease spread, and important epidemiological transitions, such as outbreak emergence or disease extinction. Studying these transitions is often of considerable interest in epidemiology. For instance, Walter et al. (2016) was concerned with how covariates, such as distance to a major water body, and neighboring disease intensity were related to the reemergence (absence to presence) and persistence (presence to presence) of Lyme disease in Connecticut towns. However, they assumed the disease was present whenever cases were reported and absent when no cases were reported. The disease could have been present when no cases were reported if it went undetected. The models explored in the first manuscript account for imperfect detection by allowing zero reported cases to arise when the disease is present. This feature is shared with finite mixture zero-inflated models (Vergne et al., 2016), however, they do not model the reemergence and persistence of the disease separately (which Walter et al. (2016) was interested in). Descloux et al. (2012) was interested in how climate factors, such as temperature, were related to the emergence of dengue outbreaks in New Caledonia. However, they assumed an outbreak had occurred whenever cases rose above either the 50th or 67th percentile and then fit a support vector machine to the classified data. The models explored in the second manuscript will account for the structure of the time series (i.e., the probabilities of transitioning in and out of a certain state) when segmenting the

counts into outbreak and endemic periods. Additionally, we avoid a two-step procedure of first classifying the data as either endemic or outbreak and then fitting a separate model to the classifications. Lim et al. (2020), investigating questions similar to Descloux et al. (2012), also used a two-step procedure by fitting a logistic regression to the state estimates of a homogeneous endemic/outbreak Markov switching model. In the second manuscript, we showed it is possible to do this kind of analysis in a single step using a nonhomogeneous Markov switching model. Nunes et al. (2013) also used a nonhomogeneous Markov switching model to investigate associations between covariates and outbreak emergence, however, only in a purely temporal setting. In a space-time setting, we can additionally investigate how outbreaks in neighboring areas affect outbreak emergence and we can account for the long strings of zeroes common in spatio-temporal counts.

The model explored in the second manuscript can also be used for outbreak detection and forecasting. Detecting an outbreak early is paramount to having an effective response (Buckeridge, 2007). Markov switching models have been utilized for outbreak detection since the late 1990s (Le Strat and Carrat, 1999). However, unlike most previous approaches, our model considers whether outbreaks are occurring in neighboring areas. In the second manuscript, we showed, using both simulated and real data, that including neighboring outbreak indicators in the transition probabilities greatly improves the accuracy and timeliness of outbreak detection and forecasting. Heaton et al. (2012) also allowed the probability of outbreak emergence to depend on whether outbreaks were occurring in neighboring areas. However, their model can only be applied to data with at most one outbreak in each area and they did not allow covariates to affect the probability of outbreak emergence.

Finally, our models can be used to deal with the problem of excess zeroes in spatio-temporal infectious disease counts. That is, there are often many more zeroes in these counts than can be realistically produced by popular spatio-temporal Poisson and negative binomial regression models (Arab, 2015). For a single disease, usually finite mixture zero-inflated models

are used to deal with excess zeroes in spatio-temporal infectious disease counts (Fernandes et al., 2009; Young et al., 2020). However, as detailed in the first manuscript, our zero-state coupled Markov mixture approach has several advantages over these models including being able to easily account for many consecutive zeroes; allowing for each covariate, and disease spread, to have a separate effect on the reemergence (absence to presence) and persistence (presence to presence) of the disease, which is often epidemiologically justified (Walter et al., 2016); and allowing the effects of disease spread between neighboring areas to depend on space-time factors, such as the population of the areas (gravity effect (Tuite et al., 2011)). Modeling cases of multiple interacting diseases across space and time when there are many zeroes has received very little attention. For instance, zero-inflated multinomial models (Zeng et al., 2022) have not been considered at all in a dependent data setting. The models we explored in the third manuscript can be used to investigate differences in the transmission intensity of several diseases across space and time while accounting for long periods of disease absence.

We would also argue for our models to be used for just the general analysis (Wakefield et al., 2019) of spatio-temporal infectious disease counts. Indeed, our approaches have important interpretational advantages over many popular alternative models (outside a Markov switching framework) for these counts. For example, multivariate autoregressive time series models, like the Endemic-epidemic (EE) model (Bracher and Held, 2022) we compared to in Section 4.5.1 of the second manuscript, and epidemiological compartmental models (Bauer and Wakefield, 2018). Both of these approaches are often used to investigate associations between covariates and overall disease transmission, usually represented as the expected number of new cases per each recent previous case (Bauer and Wakefield, 2018; Ssentongo et al., 2021). The difference with our models is that we can estimate the effect of a covariate on disease transmission in the different periods and on the epidemiological transitions, such as outbreak emergence or disease extinction (see Section 4.5.1 for an example). We would argue breaking down the covariate effects like this is useful for policymakers in many instances.

For example, knowing a covariate only affects transmission during one of the periods could help better plan suitable interventions. Also, understanding factors associated with outbreak emergence may help with preparing a response (Descloux et al., 2012), and understanding factors associated with disease extinction and reemergence could help eliminate and prevent the spread of a disease (Walter et al., 2016).

In conclusion, there are many useful potential applications of this thesis to the field of spatio-temporal infectious disease modeling. However, there are also important limitations to our approaches which we will now discuss along with some future avenues of research.

6.3 Limitations and Avenues for Future Work

The discussions of the individual manuscripts list several important limitations of those works. This section is not meant to be a repetition of those discussions. Instead, we will focus on new limitations that were not mentioned there.

Firstly, all our models take a long time to fit, around 8 to 12 hours. This is because we use computationally intensive Markov chain Monte Carlo (MCMC) methods to draw from the posterior distributions. Such methods are standard for conducting Bayesian inferences with coupled and non-coupled Markov switching models (Chib, 1996; Touloupou et al., 2020), as the posterior is not available in closed form. Recently, Chen et al. (2023) proposed using a variational Bayes (VB) method to fit nonhomogeneous Markov switching models. This method is much faster than MCMC, however, it only samples from an approximation of the posterior distribution. This approximation needs to make assumptions about the structure of the posterior such as which parameters are *a-posteriori* dependent. Also, they use the forward filter (Frühwirth-Schnatter, 2006) which cannot be run with coupled Markov switching models that have many interacting chains, like our models from the first two manuscripts (see Section 3.3 of the first manuscript). Still, it may be possible to modify the methods of Chen et al. (2023) to be appropriate for these coupled Markov switching models.

This, and pursuing other potential methods of fast inference, is an important avenue of future research. Faster methods of inference would allow us to fit our models to much larger data sets and enable more convenient online outbreak detection and forecasting.

A challenge in modeling infectious disease spread is that neighborhood structures can change over time. For instance, during the COVID-19 epidemic, the Quebec government restricted travel to different areas at different times (Sherwin, 2020). In the first manuscript, we allow the effects of disease spread between areas to change across space and time with covariates, see the coupling parameters in Equation (3.6). Therefore, if the changes in neighborhood structure can be explained by observable factors they can be accounted for by that model. For instance, if we know that the border between two areas is locked down at a certain point, we could include a binary indicator that is 0 before the lockdown and 1 after in Equation (3.6). In future work, we could extend the coupling parameters from the first manuscript to the models in the second and third manuscripts. This was not considered in the thesis as the models from the second and third manuscripts are more complex. The models from the second manuscript switch between a zero-state and two count states, as opposed to switching between a zero-state and a single count state in the first manuscript, and the models from the third manuscript are multivariate. If the changes in neighborhood structure cannot all be accounted for by covariates, then future work could incorporate splines or temporal random effects, like random walks (Cargnoni et al., 1997), into Equation (3.6) to capture unobserved heterogeneity.

Conditional on the non-absence states, we regressed on past observations in the area and neighboring areas, known as autoregression, to capture correlations in the counts across space and time within an epidemiological period. Autoregression is popular for capturing spatio-temporal correlations in infectious disease data (Bauer and Wakefield, 2018; Bracher and Held, 2022). Another popular approach for capturing these correlations is to add random effects that are correlated across space, time, or both, such as conditional autoregressive

effects (Amorós et al., 2020) or random walks (Knorr-Held and Richardson, 2003), to the model's link functions. We preferred using autoregression for a few reasons. Firstly, it is epidemiologically motivated. As shown in Bauer and Wakefield (2018), the negative binomial autoregressive models we use in the first two manuscripts can be motivated from discrete-time susceptible-infectious-recovered (SIR) models, see also Equations (C.1)-(C.2). An advantage of these models is that they allow aggregate level disease transmission to be estimated, that is, the expected number of new cases per each recent previous case (Bauer et al., 2016; Ssentongo et al., 2021). Similarly, in Section C.1, we showed how our autoregressive multinomial model from the third manuscript can be derived from a multivariate discrete-time SIR model. Additionally, correlated random effects can mix very slowly in MCMC algorithms, often requiring specialized block samplers to be developed (Knorr-Held and Richardson, 2003). Therefore, it would not be straightforward to implement these effects into our models, especially given the already large number of parameters and states that we consider. However, if these computational challenges can be overcome, then an interesting future extension of our work would be to consider both autoregression and correlated random effects. For instance, the endemic and outbreak intercepts in Equation (4.2) could follow normal autoregressive random effects (Otting et al., 2020), which would allow for correlated unobserved heterogeneity in outbreak and endemic disease transmission over time. Indeed, a limitation of our models is that within the endemic and outbreak periods, disease transmission can only change according to covariates. However, there may be factors not accounted for by the covariates that affect disease transmission within these periods, such as government interventions or migration from high-risk areas. Finally, we note that we use a negative binomial distribution to account for overdispersion conditional on the non-absence states in the first two manuscripts, as autoregression does not induce a large amount of overdispersion. This is standard for autoregressive count models (Dunbar and Held, 2020). In contrast, count models with many random effects often assume a Poisson distribution as the random effects will induce a large amount of overdispersion (Lawson and

Kim, 2022).

An important property of the multinomial distribution is that the categories must be mutually exclusive. In the context of the third manuscript, this means that we cannot incorporate individuals who are diagnosed with more than one of the diseases, dengue, Zika or chikungunya, at the same time. This should not be a major limitation in that manuscript, as co-infection with these diseases rarely occurs (Schmidt et al., 2022). However, it may be an important issue in other applications, like with tuberculosis (TB) and HIV (Pawlowski et al., 2012). If one is interested in, for example, the odds of a new case being HIV-TB co-infection relative to infection with just HIV or TB, then HIV-TB co-infection could be added as an extra category. This and other possible solutions, depending on the study's objectives, could be explored in future work.

A departure of the third manuscript from the first is that we did not assume that the presence of a disease in an area depends on the presence of the disease previously in neighboring areas, only on the number of cases in neighboring areas (i.e., the Markov chains were coupled between diseases in an area but not across areas). We felt this was a necessary simplification of the models explored in the first manuscript given the complex nature of modeling zero-inflated multivariate spatio-temporal disease counts. However, it does have the limitation that a disease may have been present in a neighboring area and spread from there even if no cases were reported (e.g., if the disease was undetected). In future work, we could consider the influence of neighboring disease presence on the transition probabilities in Equation (5.7) using ideas similar to the first and second manuscripts.

There are other interesting extensions of the third manuscript that we could explore. For instance, we could consider multiple diseases switching between absence, endemic and outbreak states in each area instead of just presence and absence states. This would allow us to investigate how outbreaks of one disease are related to outbreaks of the others. While interesting, such models could be quite complex. Each disease would have two states that require

parametrization (the absence state has no parameters) and the 3 by 3 transition matrix for each disease could be coupled across diseases and perhaps areas. Shrinkage methods (Wang et al., 2023) may help reduce the complexity.

Finally, our models could be extended to areas outside of epidemiology. For example, Williams et al. (2024) recently proposed an autoregressive Markov switching model for weekly counts of battle deaths across several countries. They assumed each country could be in one of three conflict states: non-violent, stable violence and intensified violence. As speculated by them, if a country is in a violent state it may indicate that nearby countries are more likely to become violent. Our methods could be used to add indicators for violent states in neighboring areas to the transition probabilities.

In conclusion, there are several interesting extensions of our work. As we have detailed in the preambles and Section 6.1, this thesis represents a useful contribution to the area of spatio-temporal infectious disease modeling. As such, we hope the ideas that we explored here are built on in the future.

Appendices

APPENDIX A

Appendix to Manuscript 1

A.1 The Blocked Forward Filtering Backward Sampling (bFFBS) Algorithm

Borrowing the notation from the main text, sampling $\boldsymbol{S}$ in the Gibbs sampler involves the following steps. First we initialize $\boldsymbol{S}^{[1]}$ by setting $S_{it}^{[1]} = 0$ whenever $y_{it} = 0$, and we set $S_{it}^{[m]} = 1$ whenever $y_{it} > 0$ for all m, since only the count process can produce a positive count. Then the following steps are repeated for $m = 2, ..., Q$, where Q is the total number of iterations,

1. Sample $\boldsymbol{v}^{[m]}$ from $p(\boldsymbol{v}|\boldsymbol{S}^{[m-1]}, \boldsymbol{y})$
2. Sample $\boldsymbol{S}_{(\boldsymbol{b}_k)}^{[m]}$ from $p(\boldsymbol{S}_{(\boldsymbol{b}_k)}|\boldsymbol{S}_{(\boldsymbol{b}_1)}^{[m]}, ..., \boldsymbol{S}_{(\boldsymbol{b}_{k-1})}^{[m]}, \boldsymbol{S}_{(\boldsymbol{b}_{k+1})}^{[m-1]}, ..., \boldsymbol{S}_{(\boldsymbol{b}_B)}^{[m-1]}, \boldsymbol{v}^{[m]}, \boldsymbol{y})$ for $k = 1, ..., B$.

Step 1 is broken up into (mostly) Metropolis-Hastings steps as described in the main text. In step 2, we only sample unknown (i.e. when $y_{it} = 0$) state indicators in the block. Here we provide the algorithms for sampling from $p(\boldsymbol{S}_{(\boldsymbol{b}_k)}|\boldsymbol{S}_{(-\boldsymbol{b}_k)}, \boldsymbol{v}, \boldsymbol{y})$ needed for step 2. Touloupou et al. (2020) derived the algorithm for $n_k = 1$ for all k, and it is straightforward to generalize it to arbitrary n_k.

Note that,

$$p(\boldsymbol{S}_{(\boldsymbol{b}_k)}|\boldsymbol{S}_{(-\boldsymbol{b}_k)},\boldsymbol{y},\boldsymbol{v}) = p(\boldsymbol{S}_{(\boldsymbol{b}_k)T}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:T)},\boldsymbol{y},\boldsymbol{v}) \times \prod_{t=0}^{T-1} p(\boldsymbol{S}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(\boldsymbol{b}_k)t+1},\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)},\boldsymbol{y}^{(t)},\boldsymbol{v}). \tag{A.1}$$

and that,

$$p(\boldsymbol{S}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(\boldsymbol{b}_k)t+1},\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)},\boldsymbol{y}^{(t)},\boldsymbol{v}) \propto p(\boldsymbol{S}_{(\boldsymbol{b}_k)t+1}|\boldsymbol{S}_{(\boldsymbol{b}_k)t},\boldsymbol{S}_{(-\boldsymbol{b}_k)t},\boldsymbol{y}^{(t)},\boldsymbol{v}) \times p(\boldsymbol{S}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)},\boldsymbol{y}^{(t)},\boldsymbol{v}).$$

Now, $P(\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)},\boldsymbol{y}^{(t)},\boldsymbol{v})$ for $t = 0,...,T$ and $\boldsymbol{s}_{(\boldsymbol{b}_k)t} \in \{0,1\}^{n_k}$ (i.e. the presence/absence status of all locations in the block) are known as the filtered probabilities. These are calculated using the forward part of the bFFBS algorithm, starting with $t = 0$, we have,

$$P(\boldsymbol{S}_{(\boldsymbol{b}_k)0} = \boldsymbol{s}_{(\boldsymbol{b}_k)0}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:1)},\boldsymbol{v}) \propto \prod_{j\in(-\boldsymbol{b}_k)} p(S_{j1}|\boldsymbol{S}_{(\boldsymbol{b}_k)0} = \boldsymbol{s}_{(\boldsymbol{b}_k)0},\boldsymbol{S}_{(-\boldsymbol{b}_k)0},\boldsymbol{v}) \prod_{i\in\boldsymbol{b}_k} P(S_{i0} = s_{i0}(\boldsymbol{s}_{(\boldsymbol{b}_k)0})),$$

where $s_{i0}(\boldsymbol{s}_{(\boldsymbol{b}_k)0})$ is the state indicator for location i in $\boldsymbol{s}_{(\boldsymbol{b}_k)0}$. Now for $t = 1,...,T$ the predictive probabilities are first calculated and then used to calculate the filtered probabilities. The predictive probability is given by,

$$P(\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t)},\boldsymbol{y}^{(t-1)},\boldsymbol{v}) = \sum_{\boldsymbol{s}'_{(\boldsymbol{b}_k)t-1}\in\{0,1\}^{n_k}} P(\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(\boldsymbol{b}_k)(t-1)} = \boldsymbol{s}'_{(\boldsymbol{b}_k)t-1},\boldsymbol{S}_{(-\boldsymbol{b}_k)(t-1)},\boldsymbol{y}^{(t-1)},\boldsymbol{v}) \times P(\boldsymbol{S}_{(\boldsymbol{b}_k)(t-1)} = \boldsymbol{s}'_{(\boldsymbol{b}_k)t-1}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t)},\boldsymbol{y}^{(t-1)},\boldsymbol{v}), \tag{A.2}$$

which is most efficiently calculated by multiplying the $2^{n_k} \times 2^{n_k}$ conditional transition matrix of $\boldsymbol{S}_{(\boldsymbol{b}_k)t}$, $\Gamma(\boldsymbol{S}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(t-1)},\boldsymbol{y}^{(t-1)})$, transposed, by the vector of previous filtered probabili-

ties. Note that an element of $\Gamma(\boldsymbol{S}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(t-1)}, \boldsymbol{y}^{(t-1)})$ is given by,

$$P(\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(\boldsymbol{b}_k)(t-1)} = \boldsymbol{s}'_{(\boldsymbol{b}_k)t-1}, \boldsymbol{S}_{(-\boldsymbol{b}_k)(t-1)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{v}) =$$
$$\prod_{i \in \boldsymbol{b}_k} P(S_{it} = s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t})|\boldsymbol{S}_{(\boldsymbol{b}_k)(t-1)} = \boldsymbol{s}'_{(\boldsymbol{b}_k)t-1}, \boldsymbol{S}_{(-\boldsymbol{b}_k)(t-1)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{v}),$$

where $s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t})$ is the state indicator for location i in $\boldsymbol{s}_{(\boldsymbol{b}_k)t}$, which can be calculated from the individual area conditional transition matrices defined in equation (3.3) of the main text. The predictive probabilities are then used to calculate the filtered probabilities,

$$\begin{aligned} &P(\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)}, \boldsymbol{y}^{(t)}, \boldsymbol{v}) \propto \\ &P(\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}|\boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{v}) \prod_{i \in \boldsymbol{b}_k} p(y_{it}|S_{it} = s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t}), \boldsymbol{y}^{(t-1)}, \boldsymbol{v}) \\ &\times \prod_{j \in (-\boldsymbol{b}_k)} p(S_{j(t+1)}|\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}, \boldsymbol{S}_{(-\boldsymbol{b}_k)t}, \boldsymbol{y}^{(t)}, \boldsymbol{v}), \end{aligned} \tag{A.3}$$

which involves the predictive probability. In (A.3), if $y_{it} > 0$ and $s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t}) = 0$, then $p(y_{it}|S_{it} = s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t}), \boldsymbol{y}^{(t-1)}, \boldsymbol{v}) = 0$. This means the entire filtered probability can be set to 0 and, therefore, these conditions should be checked first to avoid unnecessary calculations. In the extreme case where $y_{it} > 0$ for all $i \in \boldsymbol{b}_k$, then t can be skipped altogether as the entire vector of filtered probabilities is then known. Also, if $y_{it} > 0$ but $s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t}) = 1$ then $p(y_{it}|S_{it} = s_{it}(\boldsymbol{s}_{(\boldsymbol{b}_k)t}), \boldsymbol{y}^{(t-1)}, \boldsymbol{v})$ can be set to 1, since it will cancel in the calculation when normalizing. For $t = T$, there is no $\prod_{j \in (-\boldsymbol{b}_k)} p(S_{j(t+1)}|\boldsymbol{S}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}, \boldsymbol{S}_{(-\boldsymbol{b}_k)t}, \boldsymbol{y}^{(t)}, \boldsymbol{v})$ term included.

After all filtered probabilities have been calculated then the backward sampling step is performed. Starting at $T = t$, $\boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)T}$ is sampled categorically from the final filtered probabilities.

Then, for $t = T-1, ..., 0$ we sample $\boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)t}$ from,

$$
\begin{aligned}
&P(\boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t} | \boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)(t+1)}, \boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)}, \boldsymbol{y}^{(t)}, \boldsymbol{v}) \propto \\
&p(\boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)(t+1)} | \boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t}, \boldsymbol{S}_{(-\boldsymbol{b}_k)t}, \boldsymbol{y}^{(t)}, \boldsymbol{v}) \\
&\times P(\boldsymbol{S}^{[m]}_{(\boldsymbol{b}_k)t} = \boldsymbol{s}_{(\boldsymbol{b}_k)t} | \boldsymbol{S}_{(-\boldsymbol{b}_k)(0:t+1)}, \boldsymbol{y}^{(t)}, \boldsymbol{v}).
\end{aligned}
\tag{A.4}
$$

The first probability in (A.4) comes from $\Gamma(\boldsymbol{S}_{(\boldsymbol{b}_k)(t+1)} | \boldsymbol{S}_{(-\boldsymbol{b}_k)t}, \boldsymbol{y}^{(t)})$, the second probability is the filtered probability. Note that when $y_{it} > 0$ the algorithm will always sample $S^{[m]}_{it} = 1$ and, therefore, these can just be set ahead of running the MCMC as mentioned.

A.1.1 Validating the algorithm

To validate the bFFBS algorithm, we compared the posteriors produced by the bFFBS2 sampler, described in the main text, with those produced by the binary sampler. These were compared on the dengue data, whose fitting is described in the main text. The binary sampler is validated in the simulation study in Section A.2 of this appendix. Therefore, the bFFBS2 sampler should produce the same posterior distributions as the binary sampler.

In Figure A.1, we show a plot comparing the posterior means of $\boldsymbol{S}$ from the two samplers. Both samplers produce the same posterior means for $\boldsymbol{S}$ within reasonable Monte Carlo error. Additionally, we compared the posterior means and 95% posterior credible intervals for all elements of $\boldsymbol{v}$ (not shown). There was no meaningful difference in the posteriors of $\boldsymbol{v}$ produced by the two samplers.

A.2 Simulation Study

We designed a simulation study to ensure the proposed hybrid Gibbs sampler can recover the true parameters of the ZS-CMSNB model. In particular, we are interested in whether we can estimate heterogeneous effects of disease spread between areas and separate covariate

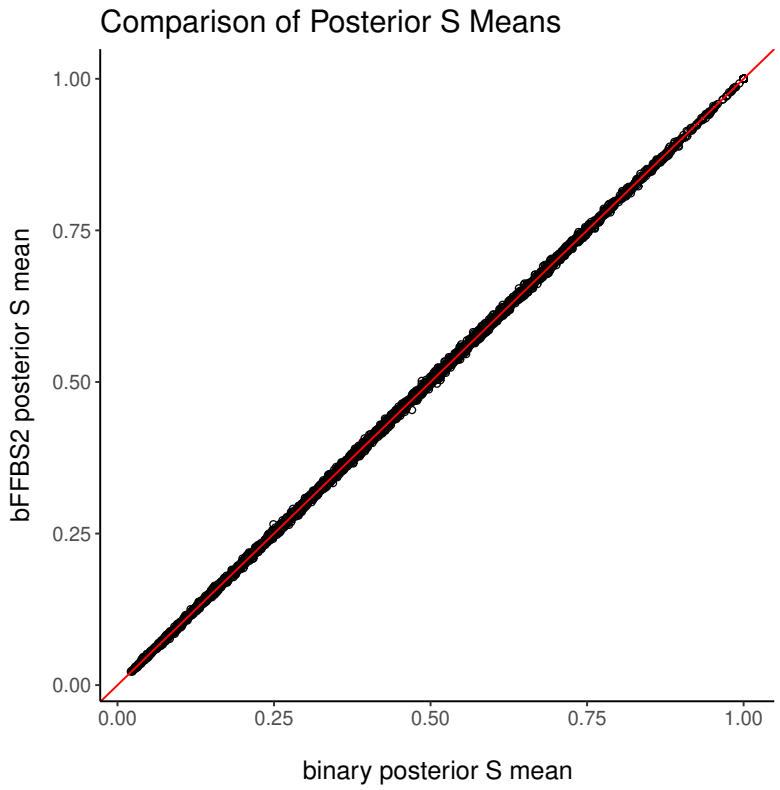


Figure A.1: Comparison of the posterior means of $\boldsymbol{S}$ produced by the binary and bFFBS2 samplers.

effects for the reemergence and persistence of the disease. To simulate heterogeneous effects of between area disease spread, we assume there is a 30 percent chance that any two neighboring areas are separated by some barrier to disease spread that reduces the effect of disease spread between the two areas by 60 percent. This corresponds very similarly to the situation in Smith et al. (2002) who looked at whether rivers reduce the effect of rabies spread between two neighboring areas. To simulate separate covariate effects we assume that temperature has half the effect on the persistence. We generated data from a ZS-CMSNB model with the following specifications for λ_{it}, r_{it}, $\boldsymbol{z}_{it}$, $\boldsymbol{z}_{01,ijt}^{(c)}$ and $\boldsymbol{z}_{11,ijt}^{(c)}$ (see equations (3.1)-(3.6) of the

main text),

$$\begin{aligned}
\log(\lambda_{it}) &= \beta_0 + \beta_1 \text{Temp}_{t-1} + \beta_2 \text{HDI}_i, \\
r_{it} &= r \\
\boldsymbol{z}_{it} &= \left[\text{Temp}_{t-1}\right]^T \\
\boldsymbol{z}^{(c)}_{01,ijt} &= \left[\text{barr}_{ij}\right]^T \\
\boldsymbol{z}^{(c)}_{11,ijt} &= \emptyset,
\end{aligned} \tag{A.5}$$

where $\boldsymbol{v} = \left[\beta_0, \beta_1, \beta_2, r, \zeta_0, \zeta_1, \zeta_0^{(c)}, \zeta_1^{(c)}, \eta_0, \eta_1, \eta_0^{(c)}\right]^T$ $= [1, .1, 10, 1.5, -1.5, .1, .25, -.15, 1.5, .05, .1]^T$, Temp_{t-1} and HDI_i are taken from the motivating example in the main text, and $\text{barr}_{ij} = 1$ if locations i and j are separated by a barrier, for $i = 1, ..., 160$ and $t = 1, ..., 84$ like the motivating example. The neighborhood structure is that of Rio like the motivating example of the main text. Additionally we assumed $p(S_{i0}) \sim Bern(.5)$ for $i = 1, ..., N$. The simulations are set up so that the effects of temperature and between area disease spread on the reemergence and persistence of the disease are somewhat similar to our motivating example. Additionally, 50 percent of the counts are 0 and 50 percent of the zeroes come from the Markov chain, which represents a good amount of missing information. We fit the ZS-CMSNB model (correctly specified) to 100 replications of (A.5). We ran 80,000 iterations, with a burn-in of 30,000, of the Gibbs sampler on 3 chains started randomly from different points. For each replication, convergence of each parameter was checked using the effective sample size (>1000) and the Gelman-Rubin statistic (<1.05) (Plummer et al., 2006). Additionally, as many zeroes often lead to model instability and identifiability issues in complex zero-inflated models (Agarwal et al., 2002), we also ran a set of simulations at 80 percent zeroes by changing β_0 from 1 to -1.

The sampling distributions of the posterior means from each set of simulations can be seen in Figure A.2. Figure A.2 shows that while the precision of the posterior means at 80 percent zeroes is much less then at 50 percent zeroes, the posterior means still have minimum bias in

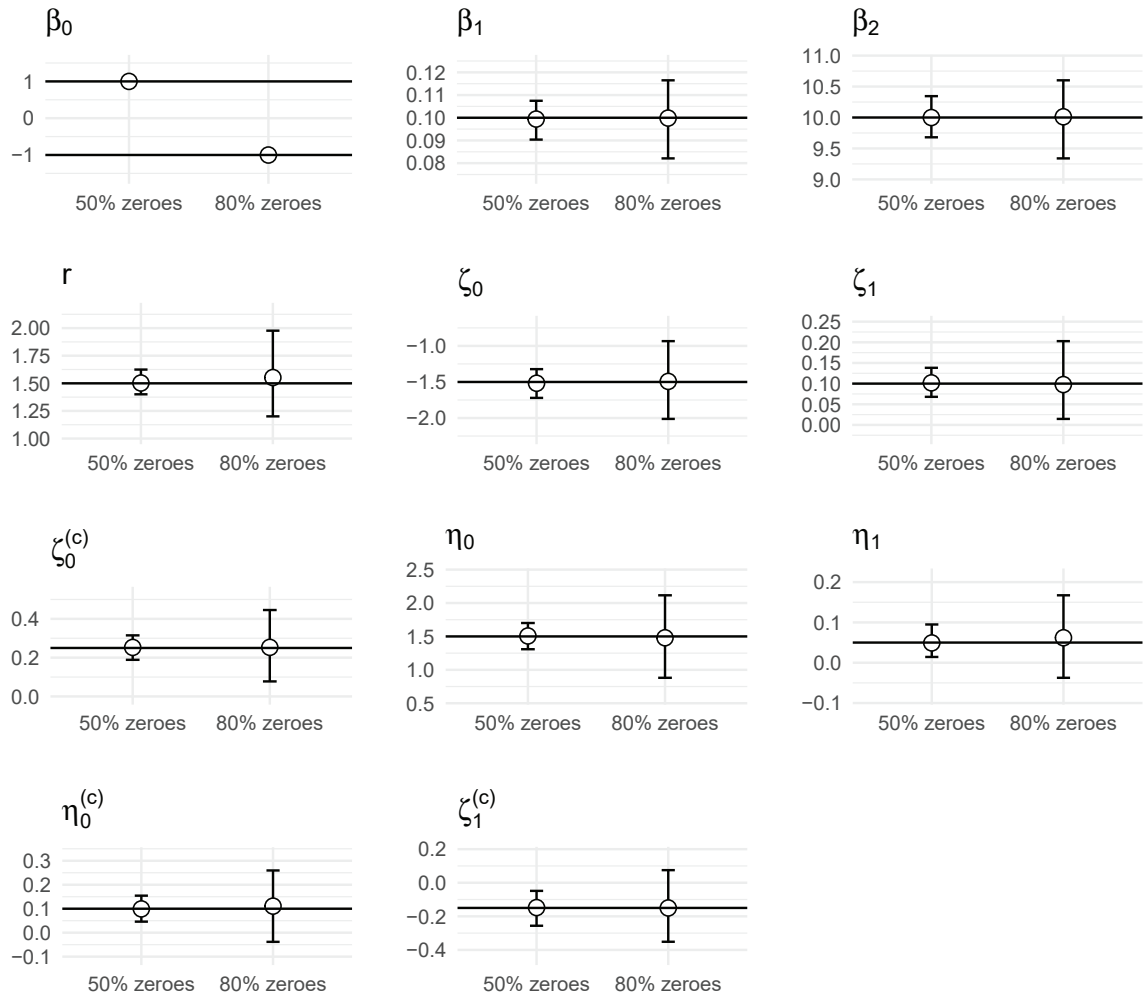


Figure A.2: Averages and 95% quantiles for the sampling distributions (100 replications) of the posterior means from a simulation study with 50 percent and 80 percent 0s. Horizontal lines drawn at true parameter values.

either scenario. Additionally, the average coverage of the 95% posterior credible intervals at 80 percent zeroes was 94.3% (min=92.7%, max=95.3%) and at 50 percent zeroes was 94.4% (min=91.4%, max=95.9%). However, we did notice some convergence issues at 80 percent 0s that were not present at 50 percent 0s. We had to run the model for 5 times as long at 80 percent zeroes and around 5 percent of the simulations did not converge, while all simulations converged at 50 percent zeroes.

In conclusion, our Gibbs sampler applied to the ZS-CMSNB model is able to estimate hetero-

geneous between area effects of disease spread and separate covariate effects with minimum bias and good coverage. However, at a large number of 0s (80 percent) the precision of the estimates are quite low and there can be rare convergence issues.

A.3 Monte Carlo Approximations for the Posterior Predictive Distributions

We are interested in the K step ahead posterior predictive distribution of both disease presence, $p(S_{i(T+K)}|\boldsymbol{y})$, and the cases counts, $p(y_{i(T+K)}|\boldsymbol{y})$, for $K = 1, 2, ...$ and $i = 1, ..., N$. We assume only dependence on the previous times counts for simplicity in notation, then we have,

$$\begin{aligned} p(S_{i(T+K)}|\boldsymbol{y}) = \int & p(S_{i(T+K)}|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-1}, \boldsymbol{\theta}) \\ & \times p(\boldsymbol{y}_{T+K-1}|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-2}, \boldsymbol{\beta}) p(\boldsymbol{S}_{T+K-1}|\boldsymbol{S}_{T+K-2}, \boldsymbol{y}_{T+K-2}, \boldsymbol{\theta}) \\ & \ldots \times p(\boldsymbol{y}_{T+1}|\boldsymbol{S}_{T+1}, \boldsymbol{y}_T, \boldsymbol{\beta}) p(\boldsymbol{S}_{T+1}|\boldsymbol{S}_T, \boldsymbol{y}_T, \boldsymbol{\theta}) \\ & \times p(\boldsymbol{v}, \boldsymbol{S}_T|\boldsymbol{y})\, d\boldsymbol{y}_{T+K-1} d\boldsymbol{S}_{T+K-1} \ldots d\boldsymbol{y}_{T+1} d\boldsymbol{S}_{T+1} d\boldsymbol{S}_T d\boldsymbol{\beta} d\boldsymbol{\theta}, \end{aligned} \tag{A.6}$$

and,

$$
\begin{aligned}
p(y_{i(T+K)}|\boldsymbol{y}) = \int & p(y_{i(T+K)}|S_{i(T+K)}, \boldsymbol{y}_{T+K-1}, \boldsymbol{\beta}) p(S_{i(T+K)}|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-1}, \boldsymbol{\theta}) \\
& \times p(\boldsymbol{y}_{T+K-1}|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-2}, \boldsymbol{\beta}) p(\boldsymbol{S}_{T+K-1}|\boldsymbol{S}_{T+K-2}, \boldsymbol{y}_{T+K-2}, \boldsymbol{\theta}) \\
& \ldots \times p(\boldsymbol{y}_{T+1}|\boldsymbol{S}_{T+1}, \boldsymbol{y}_T, \boldsymbol{\beta}) p(\boldsymbol{S}_{T+1}|\boldsymbol{S}_T, \boldsymbol{y}_T, \boldsymbol{\theta}) \\
& \times p(\boldsymbol{v}, \boldsymbol{S}_T|\boldsymbol{y})\, dS_{i(T+K)} d\boldsymbol{y}_{T+K-1} d\boldsymbol{S}_{T+K-1} \ldots d\boldsymbol{y}_{T+1} d\boldsymbol{S}_{T+1} d\boldsymbol{S}_T d\boldsymbol{\beta} d\boldsymbol{\theta} \\
= \int & \left[p(y_{i(T+K)}|S_{i(T+K)} = 1, \boldsymbol{y}_{T+K-1}, \boldsymbol{\beta}) P(S_{i(T+K)} = 1|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-1}, \boldsymbol{\theta}) \right. \\
& \left. + I\left[y_{i(T+K)} = 0\right] \left(1 - P(S_{i(T+K)} = 1|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-1}, \boldsymbol{\theta})\right)\right] \\
& \times p(\boldsymbol{y}_{T+K-1}|\boldsymbol{S}_{T+K-1}, \boldsymbol{y}_{T+K-2}, \boldsymbol{\beta}) p(\boldsymbol{S}_{T+K-1}|\boldsymbol{S}_{T+K-2}, \boldsymbol{y}_{T+K-2}, \boldsymbol{\theta}) \\
& \ldots \times p(\boldsymbol{y}_{T+1}|\boldsymbol{S}_{T+1}, \boldsymbol{y}_T, \boldsymbol{\beta}) p(\boldsymbol{S}_{T+1}|\boldsymbol{S}_T, \boldsymbol{y}_T, \boldsymbol{\theta}) \\
& \times p(\boldsymbol{v}, \boldsymbol{S}_T|\boldsymbol{y})\, d\boldsymbol{y}_{T+K-1} d\boldsymbol{S}_{T+K-1} \ldots d\boldsymbol{y}_{T+1} d\boldsymbol{S}_{T+1} d\boldsymbol{S}_T d\boldsymbol{\beta} d\boldsymbol{\theta}.
\end{aligned}
\tag{A.7}
$$

The above integrals can be approximated through Monte Carlo integration,

$$
p(S_{i(T+K)}|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} p(S_{i(T+K)}|\boldsymbol{S}_{T+K-1}^{[m]}, \boldsymbol{y}_{T+K-1}^{[m]}, \boldsymbol{\theta}^{[m]}), \tag{A.8}
$$

and,

$$
\begin{aligned}
p(y_{i(T+K)}|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} & \left[p(y_{i(T+K)}|S_{i(T+K)} = 1, \boldsymbol{y}_{T+K-1}^{[m]}, \boldsymbol{\beta}^{[m]}) \right. \\
& \times P(S_{i(T+K)} = 1|\boldsymbol{S}_{T+K-1}^{[m]}, \boldsymbol{y}_{T+K-1}^{[m]}, \boldsymbol{\theta}^{[m]}) \\
& \left. + I\left[y_{i(T+K)} = 0\right] \left(1 - P(S_{i(T+K)} = 1|\boldsymbol{S}_{T+K-1}^{[m]}, \boldsymbol{y}_{T+K-1}^{[m]}, \boldsymbol{\theta}^{[m]}\right)\right],
\end{aligned}
\tag{A.9}
$$

where the superscript $[m]$ denotes a draw from the posterior distribution of the variable ($\boldsymbol{y}_t^{[m]} = \boldsymbol{y}_t$ if $t \leq T$), M is the size of the burn-in sample, Q is the total MCMC sample size and $I[\cdot]$ is an indicator function.

Substituting $K = 1$ into (A.8) approximates the one step ahead posterior predictive distribution of disease presence in location i,

$$P(S_{i(T+1)} = 1|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} P(S_{i(T+1)} = 1|S_{iT}^{[m]}, \boldsymbol{S}_{(-i)T}^{[m]}, \boldsymbol{y}_T, \boldsymbol{\theta}^{[m]}), \tag{A.10}$$

for $i = 1, ..., N$. We can similarly approximate the one step ahead posterior predictive distribution of the reported cases in location i,

$$\begin{aligned} p(y_{i(T+1)}|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} \big[& p(y_{i(T+1)}|S_{i(T+1)} = 1, \boldsymbol{y}_T, \boldsymbol{\beta}^{[m]}) \\ & \times P(S_{i(T+1)} = 1|S_{iT}^{[m]}, \boldsymbol{S}_{(-i)T}^{[m]}, \boldsymbol{y}_T, \boldsymbol{\theta}^{[m]}) \\ & + I\left[y_{i(T+1)} = 0\right] \left(1 - P(S_{i(T+1)} = 1|S_{iT}^{[m]}, \boldsymbol{S}_{(-i)T}^{[m]}, \boldsymbol{y}_T, \boldsymbol{\theta}^{[m]})\right) \big], \end{aligned} \tag{A.11}$$

for $i = 1, ..., N$, where $I[\cdot]$ is an indicator function. Therefore, the one step ahead posterior predictive distribution of the counts is zero-inflated, where the mixing probability depends on past histories of the states and counts. We can expand (A.10) to gain a better understanding of this one step ahead prediction for the risk of disease presence in location i,

$$\begin{aligned} & P(S_{i(T+1)} = 1|S_{iT}^{[m]}, \boldsymbol{S}_{(-i)(T)}^{[m]}, \boldsymbol{y}_T, \boldsymbol{\theta}^{[m]}) = p01_{i(T+1)}^{[m]}(1 - S_{iT}^{[m]}) + p11_{i(T+1)}^{[m]} S_{iT}^{[m]}, \text{ where,} \\ & \text{logit}(p01_{i(T+1)}^{[m]}) = \zeta_0^{[m]} + \boldsymbol{z}_{i(T+1)}^T \boldsymbol{\zeta}^{[m]} + \sum_{j \in NE(i)} \phi_{01,j\to i}^{T\to(T+1)\,[m]} S_{jT}^{[m]}, \\ & \text{logit}(p11_{i(T+1)}^{[m]}) = \eta_0^{[m]} + \boldsymbol{z}_{i(T+1)}^T \boldsymbol{\eta}^{[m]} + \sum_{j \in NE(i)} \phi_{11,j\to i}^{T\to(T+1)\,[m]} S_{jT}^{[m]}. \end{aligned} \tag{A.12}$$

Even if for $j \in NE(i)$ $y_{jT} = 0$, $S_{jT}^{[m]}$ could be 1 for many m if there is a high chance the disease went undetected in neighbor j. Similarly, if $y_{iT} = 0$, $S_{iT}^{[m]}$ could be 1 for many m if there is a high chance the disease is undetected in location i. Therefore, this warning system accounts for the fact that the disease may already have been circulating undetected in location i for some time. It also considers the possible risk of spread from neighboring

Algorithm 1: Posterior Predictive Simulation

for m *in* $M+1:Q$ **do**
  for k *in* $1:K$ **do**
    for i *in* $1:N$ **do**
      1. Draw $S^{[m]}_{i(T+k)}$ from $p(S_{i(T+k)}|\boldsymbol{S}^{[m]}_{T+k-1}, \boldsymbol{y}^{[m]}_{T+k-1}, \boldsymbol{\theta}^{[m]})$, where $\boldsymbol{y}^{[m]}_{T} = \boldsymbol{y}_T$,

         $p(S_{i(T+k)}|\boldsymbol{S}^{[m]}_{T+k-1}, \boldsymbol{y}^{[m]}_{T+k-1}, \boldsymbol{\theta}^{[m]})$=Bern$\left(\pi^{[m]}_{i(T+k)}\right)$, where,

         $\pi^{[m]}_{i(T+k)} = p01^{[m]}_{i(T+k)}(1 - S^{[m]}_{i(T+k-1)}) + p11^{[m]}_{i(T+k)}S^{[m]}_{i(T+k-1)}$.

      2. Draw $y^{[m]}_{i(T+k)}$ from $p(y_{i(T+k)}|S^{[m]}_{i(T+k)}, \boldsymbol{y}^{[m]}_{T+k-1}, \boldsymbol{\beta}^{[m]})$, where $\boldsymbol{y}^{[m]}_{T} = \boldsymbol{y}_T$,

         $p(y_{i(T+k)}|S^{[m]}_{i(T+k)}, \boldsymbol{y}^{[m]}_{T+k-1}, \boldsymbol{\beta}^{[m]}) = NB(S^{[m]}_{i(T+k)}\lambda^{[m]}_{i(T+k)}, r^{[m]}_{i(T+k)})$.

    end
  end
end

locations with 0 reported cases that have a high chance of the disease being undetected. More traditional warning systems, based on autoregressive models (Chan et al., 2015), cannot account for these important risks. This makes our warning system particularly useful in cities like Rio where underreporting is a major issue, and the disease may circulate and spread unnoticed.

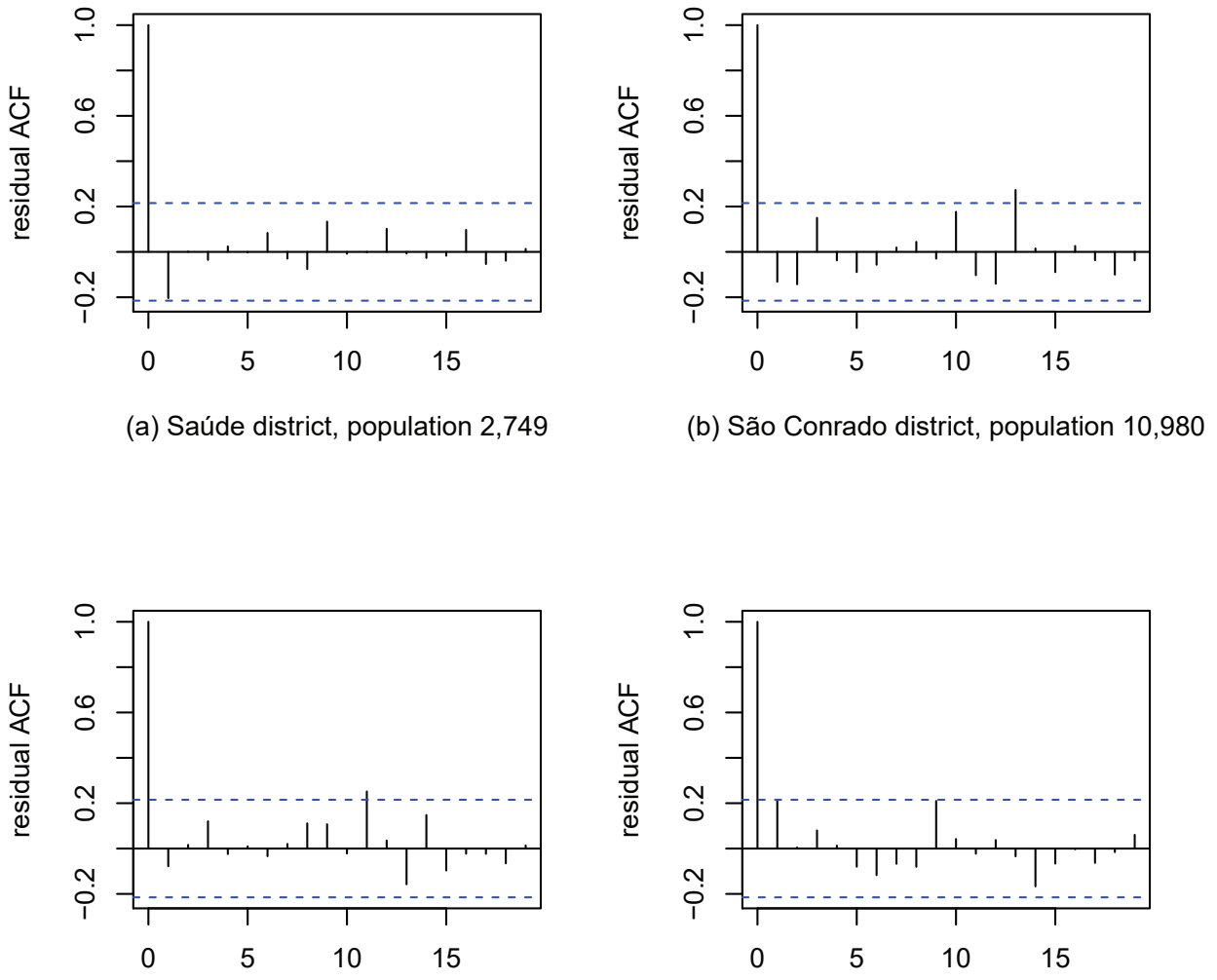
For arbitrary K step ahead temporal predictions, we use a simulation procedure (Frühwirth-Schnatter, 2006) to draw samples from the posterior predictive distributions. Algorithm 1 will obtain realizations from the posterior predictive distribution of the cases, $y^{[m]}_{i(T+k)} \sim p(y_{i(T+k)}|\boldsymbol{y})$, and the presence of the disease, $S^{[m]}_{i(T+k)} \sim p(S_{i(T+k)}|\boldsymbol{y})$, for $i = 1, ..., N$, $k = 1, ..., K$ and $m = M+1, ..., Q$. However, as $S^{[m]}_{i(T+k)}$ can only take two values, 0 or 1, it is difficult to interpret the uncertainty around this prediction for the presence of the disease. Therefore, instead of using summaries of $S^{[m]}_{i(T+k)}$ we use summaries of $P(S_{i(T+k)} = 1|S^{[m]}_{i(T+k-1)}, \boldsymbol{S}^{[m]}_{(-i)(T+k-1)}, \boldsymbol{y}^{[m]}_{T+k-1}, \boldsymbol{\theta}^{[m]})$.

A.4 Pearson Residuals

A common model diagnostic in time series analysis is the autocorrelation function (ACF) of the Pearson residuals (Bracher and Held, 2022). Following Section 3.3.2, we define the Pearson residual in area i for our model as,

$$\text{Pres}_{it} = \frac{y_{it} - E[y_{it}|\boldsymbol{y}^{(t-1)}, \boldsymbol{S}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{v}]}{\sqrt{Var[y_{it}|\boldsymbol{y}^{(t-1)}, \boldsymbol{S}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{v}]}}.$$

We estimate $E[y_{it}|\boldsymbol{y}^{(t-1)}, \boldsymbol{S}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{v}]$ and $Var[y_{it}|\boldsymbol{y}^{(t-1)}, \boldsymbol{S}_{(-\boldsymbol{b}_{NE(i)})(0:t)}, \boldsymbol{v}]$ using the sample mean and variance of $y_{it}^{c1*[m]}$. Figure A.3 (below) shows the ACF of the Pearson residuals for the 4 districts from Figure 3.6 of the main text. There are not many significant autocorrelations. Overall, the figure suggests we are capturing well the structure present in the data.



(a) Saúde district, population 2,749

(b) São Conrado district, population 10,980

(c) Mangueira district, population 17,835

(d) Engenho Novo district, population 42,172

Figure A.3: ACF of the Pearson residuals for the 4 districts from Figure 3.6 of the main text.

APPENDIX B

Appendix to Manuscript 2

B.1 The Individual Forward Filtering Backward Sampling (iFFBS) Algorithm

In this section, we describe how $\boldsymbol{S}^*$ is sampled in our hybrid Gibbs sampling algorithm. We will borrow all notation from the main text. First, to obtain valid initial values for the hidden Markov chain we sample $\boldsymbol{S}_i^{*[1]}$ from a 6 state Markov chain with transition Matrix given by Equation (4.7) in the main text without the absence state. The transition probabilities are fixed at .8 for remaining in a state and .2 for transitioning out of a state since we expect persistence. We do not include the absence state in the initial values since it might cause the initial joint likelihood function, Equation (4.8) in the main text, to be evaluated at 0 as the absence state cannot produce a positive count. After initialization, the following steps are repeated for $m = 2, \ldots, Q$, where Q is the total number of iterations for the Gibbs sampler,

1. Sample $\boldsymbol{v}^{[m]}$ from $p(\boldsymbol{v}|\boldsymbol{S}^{*[m-1]}, \boldsymbol{y})$
2. Sample $\boldsymbol{S}_i^{*[m]}$ from $p(\boldsymbol{S}_i^*|\boldsymbol{S}_1^{*[m]}, \ldots, \boldsymbol{S}_{i-1}^{*[m]}, \boldsymbol{S}_{i+1}^{*[m-1]}, ..., \boldsymbol{S}_N^{*[m-1]}, \boldsymbol{v}^{[m]}, \boldsymbol{y})$ for $i = 1, ..., N$.

As mentioned in the main text, in the first step elements of $\boldsymbol{v}$ without conjugate priors are sampled individually using an adaptive random walk Metropolis step (Shaby and Wells, 2010). Here we will provide the individual forward filtering backward sampling (iFFBS) algorithm for sampling from $p(\boldsymbol{S}_i^*|\boldsymbol{S}_{(-i)}^*, \boldsymbol{v}, \boldsymbol{y})$ needed for step 2. The algorithm was originally proposed by Touloupou et al. (2020). In this section we will sometimes use the subscript $t_1{:}t_2$ to denote a temporally indexed vector subsetted to the interval t_1 to t_2, e.g., $\boldsymbol{y}_{i(1:t)} = (y_{i1}, \ldots, y_{it})^T$.

First note that,

$$p(\boldsymbol{S}_i^*|\boldsymbol{S}_{(-i)}^*, \boldsymbol{v}, \boldsymbol{y}) = p(S_{iT}^*|\boldsymbol{S}_{(-i)}^*, \boldsymbol{y}, \boldsymbol{v}) \prod_{t=1}^{T-1} p(S_{it}^*|S_{i(t+1)}^*, \boldsymbol{S}_{(-i)(1:t+1)}^*, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}), \tag{B.1}$$

and that, from Bayes' Theorem,

$$p(S_{it}^*|S_{i(t+1)}^*, \boldsymbol{S}_{(-i)(1:t+1)}^*, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}) \propto p(S_{i(t+1)}^*|S_{it}^*, \boldsymbol{S}_{(-i)t}, \boldsymbol{\theta}) p(S_{it}^*|\boldsymbol{S}_{(-i)(1:t+1)}^*, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}). \tag{B.2}$$

The density $p(S_{i(t+1)}^*|S_{it}^*, \boldsymbol{S}_{(-i)t}, \boldsymbol{\theta})$ in (B.2) is simply the appropriate transition probability of the Markov chain in area i, which can be obtained from Equation (4.7) of the main text. Therefore, if we can calculate $P(S_{it}^* = s^*|\boldsymbol{S}_{(-i)(1:t+1)}^*, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v})$ for $s^* = 1, \ldots, 7$ and $t = 1, \ldots, T$, called the filtered probabilities, then $\boldsymbol{S}_i^*$ can be sampled backward using Equations (B.1) and (B.2).

Starting with $t = 1$ we have that,

$$\begin{aligned} p(S_{i1}^*|\boldsymbol{S}_{(-i)(1:2)}^*, y_{i1}, \boldsymbol{v}) &\propto p(S_{i1}^*|y_{i1}) p(\boldsymbol{S}_{(-i)2}^*|S_{i1}^*, \boldsymbol{S}_{(-i)1}^*, \boldsymbol{\theta}) \\ &\propto p(S_{i1}^*|y_{i1}) \prod_{j\,:\,i \in NE(j)} p(S_{j2}^*|S_{j1}^*, \boldsymbol{S}_{(-j)1}, \boldsymbol{\theta}). \end{aligned} \tag{B.3}$$

Here $p(S_{i1}^*|y_{i1})$ is the initial state distribution, which is fixed by the modeler, and $p(S_{j2}^*|S_{j1}^*, \boldsymbol{S}_{(-j)1}, \boldsymbol{\theta})$ is a transition probability of the Markov chain in area j. Note that

$\prod_{j\,:\,i\,\in\,NE(j)} p(S^*_{j2}|S^*_{j1}, \boldsymbol{S}_{(-j)1}, \boldsymbol{\theta})$ only depends on whether area i is in the outbreak state or not so only 2 values need to be calculated. Also note that since S^*_{i1} can only take seven values it is straightforward to derive the filtered probabilities using Equation (B.3),

$$P(S^*_{i1} = s^*|\boldsymbol{S}^*_{(-i)(1:2)}, y_{i1}, \boldsymbol{v}) = \frac{P(S^*_{i1} = s^*|y_{i1}) \prod\limits_{\substack{j\,:\,i\,\in\,NE(j)\\ S^*_{i1}=s^*}} p(S^*_{j2}|S^*_{j1}, \boldsymbol{S}_{(-j)1}, \boldsymbol{\theta})}{\sum_{k=1}^{7} P(S^*_{i1} = k|y_{i1}) \prod\limits_{\substack{j\,:\,i\,\in\,NE(j)\\ S^*_{i1}=k}} p(S^*_{j2}|S^*_{j1}, \boldsymbol{S}_{(-j)1}, \boldsymbol{\theta})},$$

for $s^* = 1, \ldots, 7$.

For $t = 2, \ldots, T-1$ we have that,

$$p(S^*_{it}|\boldsymbol{S}^*_{(-i)(1:t+1)}, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}) \propto p(y_{it}|S_{it}, y_{i(t-1)}, \boldsymbol{\beta}) p(S^*_{it}|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}) \times \prod_{j\,:\,i\,\in\,NE(j)} p(S^*_{j(t+1)}|S^*_{jt}, \boldsymbol{S}_{(-j)t}, \boldsymbol{\theta}).$$

Here $p(y_{it}|S_{it}, y_{i(t-1)}, \boldsymbol{\beta})$ is given by Equation (4.1) of the main text. Note that,

$$P(S^*_{it} = s^*|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}) = \sum_{k=1}^{7} P(S^*_{it} = s^*|S^*_{i(t-1)} = k, \boldsymbol{S}_{(-i)(t-1)}, \boldsymbol{\theta}) P(S^*_{i(t-1)} = k|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}),$$

where $P(S^*_{i(t-1)} = k|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v})$ is the previous filtered probability. It then follows that,

$$P(S^*_{it} = s^*|\boldsymbol{S}^*_{(-i)(1:t+1)}, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}) = \frac{p(y_{it}|S^*_{it} = s^*, y_{i(t-1)}, \boldsymbol{\beta}) P(S^*_{it} = s^*|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}) \prod\limits_{\substack{j\,:\,i\,\in\,NE(j)\\ S^*_{it}=s^*}} p(S^*_{j(t+1)}|S^*_{jt}, \boldsymbol{S}_{(-j)t}, \boldsymbol{\theta})}{\sum_{k=1}^{7} p(y_{it}|S^*_{it} = k, y_{i(t-1)}, \boldsymbol{\beta}) P(S^*_{it} = k|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}) \prod\limits_{\substack{j\,:\,i\,\in\,NE(j)\\ S^*_{it}=k}} p(S^*_{j(t+1)}|S^*_{jt}, \boldsymbol{S}_{(-j)t}, \boldsymbol{\theta})}, \tag{B.4}$$

for $s^* = 1, \ldots, 7$.

The logic for $t = T$ is similar but there is no forward product term,

$$P(S^*_{iT} = s^* | \boldsymbol{S}^*_{(-i)}, \boldsymbol{y}, \boldsymbol{v}) = \frac{p(y_{iT} | S^*_{iT} = s^*, y_{i(T-1)}, \boldsymbol{\beta}) P(S^*_{iT} = s^* | \boldsymbol{S}^*_{(-i)}, \boldsymbol{y}_{i(1:T-1)}, \boldsymbol{v})}{\sum_{k=1}^{7} p(y_{iT} | S^*_{iT} = k, y_{i(T-1)}, \boldsymbol{\beta}) P(S^*_{iT} = k | \boldsymbol{S}^*_{(-i)}, \boldsymbol{y}_{i(1:T-1)}, \boldsymbol{v})},$$

for $s^* = 1, \ldots, 7$.

Once the filtered probabilities have been calculated $\boldsymbol{S}^*_i$ can be sampled backward using Equations (B.1) and (B.2). Firstly, $S^{*[m]}_{iT}$ is drawn from $p(S^*_{iT} | \boldsymbol{S}^*_{(-i)}, \boldsymbol{y}, \boldsymbol{v})$. Then, for $t = T-1, \ldots, 1$, $S^{*[m]}_{it}$ is drawn from the density defined by,

$$P(S^*_{it} = s^* | S^*_{i(t+1)} = S^{*[m]}_{i(t+1)}, \boldsymbol{S}_{(-i)(1:t+1)}, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v}) =$$

$$\frac{P(S^*_{i(t+1)} = S^{*[m]}_{i(t+1)} | S^*_{it} = s^*, \boldsymbol{S}_{(-i)t}, \boldsymbol{\theta}) P(S^*_{it} = s^* | \boldsymbol{S}^*_{(-i)(1:t+1)}, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v})}{\sum_{k=1}^{7} P(S^*_{i(t+1)} = S^{*[m]}_{i(t+1)} | S^*_{it} = k, \boldsymbol{S}_{(-i)t}, \boldsymbol{\theta}) P(S^*_{it} = k | \boldsymbol{S}^*_{(-i)(1:t+1)}, \boldsymbol{y}_{i(1:t)}, \boldsymbol{v})},$$

for $s^* = 1, \ldots, 7$.

As mentioned in the main text, all Nimble code for the iFFBS samplers is provided on GitHub (https://github.com/Dirk-Douwes-Schultz/CMSNB124_code). Note that the only calculations that separate the iFFBS sampler from a traditional FFBS sampler for Markov switching models (Chib, 1996; Frühwirth-Schnatter, 2006) are the forward product terms $\prod_{j\,:\,i \in NE(j)} p(S^*_{j(t+1)} | S^*_{jt}, \boldsymbol{S}_{(-j)t}, \boldsymbol{\theta})$ which are needed to account for between chain dependencies. Therefore, if an area is not a neighbor of any other area the iFFBS sampler reduces to the FFBS sampler which, being computationally simpler, should be used instead. As such, in our code, we assign all areas that are not neighbors of any other areas FFBS samplers, which we also custom code and provide on GitHub.

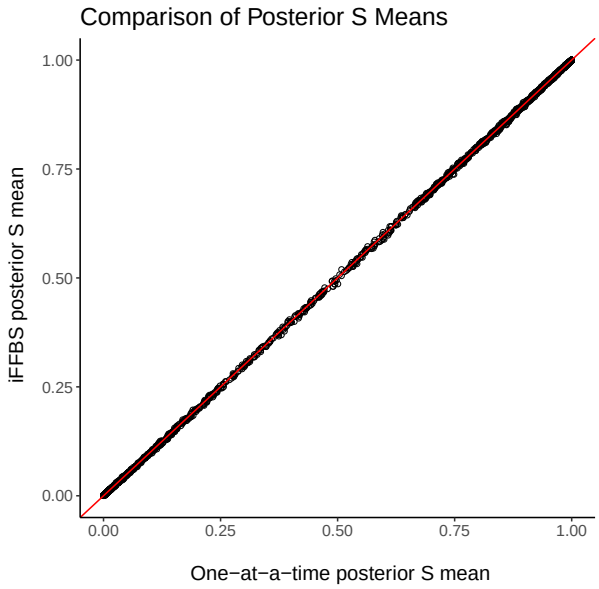


Figure B.1: Comparison of the posterior means of $\boldsymbol{S}$ produced by the one-at-a-time and iFFBS samplers when applied to the No Absence/Clone State Model from Section 4.5 of the main text. $S_{it} = 0$ indicates the endemic state and $S_{it} = 1$ indicates the outbreak state.

B.1.1 Validating the algorithm

One way to validate a Markov chain Monte Carlo (MCMC) sampler is to compare the posterior distributions produced by the sampler to those produced by a more established or simpler sampler. Different MCMC sampling algorithms should return the same posterior distributions. Figure B.1 compares the posteriors means of $\boldsymbol{S}$ produced by the one-at-a-time, described in Section 4.3 of the main text, and iFFBS samplers. These were compared on the COVID-19 hospitalization data from the main text with the No Absence/Clone State Model specified in Section 4.5.1 of the main text. We compared the samplers using the No Absence/Clone State Model as the one-at-a-time samplers do not converge when clone states are present. As can be seen from the figure, both samplers produce the same posterior means for $\boldsymbol{S}$ within reasonable Monte Carlo error. Additionally, we compared the posterior means and 95% posterior credible intervals for all elements of $\boldsymbol{v}$ (not shown) and there were no meaningful differences between the two samplers.

B.2 Simulation Study to Assess Parameter Recovery

We designed a simulation study to ensure our hybrid Gibbs sampling algorithm could recover the true parameters of the CMSNB(1,2,4) model. We simulated data from a slightly simplified version of the CMSNB(1,2,4) model specified in Section 4.5.1 of the main text. More specifically, we removed the random intercepts, we removed some insignificant effects from the Markov chain and the outbreak and endemic transmission rates, and we assumed that the overdispersion in the hospitalizations was the same during the endemic and outbreak periods. This was done to reduce the number of parameters in the model as we need to run many simulations each of which is computationally costly. We generated data from the following CMSNB(1,2,4) model,

$$
\begin{aligned}
\log(\lambda_{it}^{EN}) &= \beta_0^{EN} + \beta_{\text{beds}}^{EN}\text{beds}_i + \beta_{\text{mob}}^{EN}\text{mobility}_{\text{county}(i)(t-4)} + \rho^{EN}\log(y_{i(t-1)}+1) \\
\log(\lambda_{it}^{OB}) &= \beta_0^{OB} + \beta_{\text{beds}}^{OB}\text{beds}_i + \beta_{\text{mob}}^{OB}\text{mobility}_{\text{county}(i)(t-4)} + \rho^{OB}\log(y_{i(t-1)}+1) \\
r^{EN} &= r^{OB} = r \\
\text{logit}(p12_{it}) &= \alpha_{12,0} + \alpha_{12,\text{beds}}\text{beds}_i \\
\log\left(\frac{p21_{it}}{1-p21_{it}-p23_{it}}\right) &= \alpha_{21,0} + \alpha_{21,\text{beds}}\text{beds}_i + \alpha_{21,\text{mob}}\text{mobility}_{\text{county}(i)(t-4)} \\
\log\left(\frac{p23_{it}}{1-p21_{it}-p23_{it}}\right) &= \alpha_{23,0} + \alpha_{23,\text{mobi}}\text{mobility}_{\text{county}(i)(t-4)} + \alpha_{23,\text{newv}}\text{new\_variant}_t \\
&\quad + \alpha_{23,\text{spat}}\sum_{j\in NE(i)}\omega_{ji}I[S_{j(t-1)}=3] \\
\text{logit}(p33_{it}) &= \alpha_{33,0} + \alpha_{33,\text{mobi}}\text{mobility}_{\text{county}(i)(t-4)} \\
&\quad + \alpha_{33,\text{spat}}\sum_{j\in NE(i)}\omega_{ji}I[S_{j(t-1)}=3],
\end{aligned}
\tag{B.5}
$$

for $i = 1,\ldots,30$ and $t = 2,\ldots,113$, and with the following true parameter values $\boldsymbol{v} = (\beta_0^{EN}, \beta_{\text{beds}}^{EN}, \beta_{\text{mob}}^{EN}, \rho^{EN}, \beta_0^{OB}, \beta_{\text{beds}}^{OB}, \beta_{\text{mob}}^{OB}, \rho^{OB}, r, \alpha_{12,0}, \alpha_{12,\text{beds}}, \alpha_{21,0}, \alpha_{21,\text{beds}}, \alpha_{21,\text{mob}}, \alpha_{23,0}, \alpha_{23,\text{mobi}}, \alpha_{23,\text{newv}}, \alpha_{23,\text{spat}}, \alpha_{33,0}, \alpha_{33,\text{mobi}}, \alpha_{33,\text{spat}})^T = (0, .17, .003, .65, .78, , 06, .007, .75, 10, -.76, .45, -3.6,$

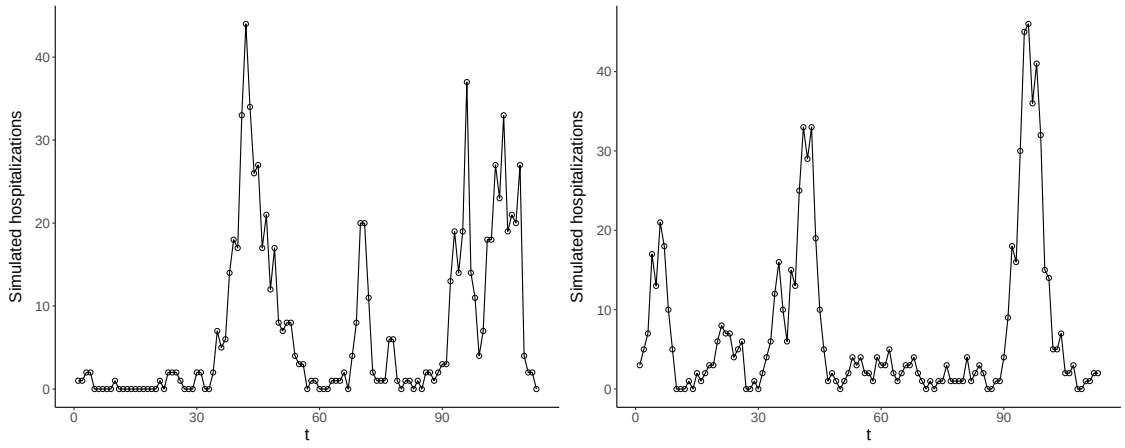


Figure B.2: Shows simulated hospitalizations in 2 areas from a single replication of (B.5).

$-.9, -.035, -4.15, .025, 2.5, 1.15, 2, .025, .45)^T$. The true parameter values were chosen to be similar to those estimated in our motivating example. In (B.5) beds_i, $\text{mobility}_{\text{county}(i)(t-4)}$, new_variant_t, $NE(i)$ and ω_{ji} are all the same as in our motivating example. Finally, we assumed a uniform initial state distribution for the Markov chain in each area. Figure B.2 shows the simulated hospitalizations in 2 areas from a single replication of (B.5) and they appear somewhat realistic.

We considered two sets of constraints for fitting the CMSNB(1,2,4) model to simulations of (B.5). Firstly, we considered a constraint on just the intercepts and the autoregressive coefficients, $\beta_0^{EN} + .1 < \beta_0^{OB}$ and $\rho^{EN} + .05 < \rho^{OB}$, which we will refer to as the weak constraints. We also considered constraining the entirety of the transmission rates,

$$\beta_0^{EN} + \beta_{\text{beds}}^{EN}\text{beds}_i + \beta_{\text{mob}}^{EN}\text{mobility}_{\text{county}(i)(t-4)} + .01 < \beta_0^{OB} + \beta_{\text{beds}}^{OB}\text{beds}_i + \beta_{\text{mob}}^{OB}\text{mobility}_{\text{county}(i)(t-4)}$$

for $i = 1, \ldots, 30$ and $t = 2, \ldots, 113$, and $\rho^{EN} + .05 < \rho^{OB}$, which we will refer to as the strong constraints. We chose minimum differences of .1 for the weak constraints on the intercepts and .01 for the strong constraints on the entirety of the transmission rates as the strong constraints constrain the minimum difference in transmission while the weak constraints constrain the average difference in transmission (we center all covariates) which should be

a larger difference. With our hybrid Gibbs sampling algorithm we fit the CMSNB(1,2,4) model (correctly specified) to 250 replications of (B.5) using the strong constraints and to another 700 replications of (B.5) (due to the low convergence rate, see below) using the weak constraints. We mostly assumed the same prior distribution for $\boldsymbol{v}$ as specified in Section 4.2.2 of the main text. The only exception is that we used wider priors for $\alpha_{23,\text{spat}}$ and $\alpha_{33,\text{spat}}$ as our goal was not to shrink these effects but to recover the true parameter values. We ran our Gibbs sampler for 200,000 iterations on 3 chains, started from random values in the parameter space, with an initial burn-in of 50,000 iterations. For each replication convergence of the Gibbs sampler was checked using the minimum effective sample size (>1000) and the maximum Gelman-Rubin statistic (<1.05) (Plummer et al., 2006).

When using the weak constraints, our Gibbs sampler passed the convergence checks for only $284/700 = 40.6\%$ of the replications. When using the strong constraints, the convergence rate was much higher with our Gibbs sampler converging for $237/250 = 94.8\%$ of the replications. Figure B.3 gives the sample mean and 95% quantile of the posterior medians for the 237/250 replications that converged using the strong constraints and the 284/700 replications that converged using the weak constraints. Figure B.3 shows that when our Gibbs sampler converges, using either the weak or the strong constraints, the posterior distributions are centered close to the true parameter values on average. Additionally, the average coverage of the 95% credible intervals was .95 with a minimum coverage of .91 and a maximum coverage of .98, when using either constraint, showing good coverage of the true parameter values.

In conclusion, regardless of whether the strong or weak constraints were used our hybrid Gibbs sampling algorithm was able to recover the true parameter values of the CMSNB(1,2,4) model well when it converged. However, the convergence rate was low when using the weak constraints. The strong constraints greatly improved the convergence rate and, from Figure B.3, did not introduce any significant bias into the inferential procedure. Additionally, check-

ing the strong constraints did not add a substantial amount of time to the model fitting (less than 1 hour). Therefore, we would recommend constraining the entirety of the transmission rates in practice and we do so throughout the main manuscript. We believe these constraints are reasonable in most applications since transmission should always increase when moving from the endemic state to the outbreak state.

We found in our simulation study that even when the strong constraints are used our Gibbs sampler can still run into rare convergence issues, around 5% of the time. By examining a few of the non-converged MCMC samples, when the strong constraint was used, we found the convergence issues were caused by genuine multimodality (the presence of multiple non-symmetric modes) in the posterior distribution, a common occurrence in Bayesian mixture modeling (Stephens, 2000; Jasra et al., 2005). Multimodality in the posterior can be an important issue in Bayesian inference as most standard MCMC sampling algorithms, such as Metropolis-Hastings, do not mix well between the modes and, therefore, may not explore all important regions of the parameter space (Yao et al., 2022). We cannot guarantee the absence of genuine multimodality in the 237/250 or 284/700 replications where our Gibbs sampler passed convergence checks, as all chains could have been stuck in the same mode (although we do start the chains from random values in the parameter space making this less likely). However, when the Gibbs sampler passed convergence checks in our simulation study, we were able to recover the true parameters well. This implies that if there were extra modes in those posteriors they were too minor to affect inference significantly. To double-check for genuine multimodality in our motivating example in Section 4.5 of the main text, we ran an additional 9 chains for the Full Coupled Model, started from random values in the parameter space, and they all converged to the same mode.

In all non-converged MCMC samples that we checked, when using the strong constraints, there was a mode centered close to the true parameter values and an extra mode centered on a submodel of the CMSNB(1,2,4) model that only used the absence state and the endemic state

(based on the locations of the different MCMC chains). That is, the posterior probability of outbreak emergence and persistence, at average covariate values, was essentially 0 for the extra mode. It is not uncommon for the likelihood of a complex mixture model to contain multiple local modes in regions of the parameter space where some of the states have close to 0 probability of being visited, see the discussion in Section 4.2.3 of Frühwirth-Schnatter (2006). (Note our strong constraints would not prevent this kind of multimodality as if the outbreak state is never visited then the transmission rate in the outbreak state does not contribute to the likelihood.) If such multimodality is encountered, it is recommended by Frühwirth-Schnatter (2006) to bound the posterior away from these regions. For instance, in our model, one could assign priors to the intercepts $\alpha_{lk,0}$ for $lk = 12, 21, 23, 33$, in Equations (4.4)-(4.5) of the main text, that shrink them towards 0 so that all states are likely to have a non-negligible chance of being visited. Another possible solution is to treat it as a model comparison problem, fitting the reduced model, that only uses an absence and one count state, and discarding the chains from the full model associated with the reduced model if it has a higher widely applicable information criterion (WAIC) (the WAIC is discussed in Section B.3 below). Other possible solutions to consider if multimodality is encountered include stacking (Yao et al., 2022) (weighting the chains associated with different modes based on some comparison criteria) and tempered MCMC (Jasra et al., 2005) (a type of MCMC algorithm that is often successful in sampling multimodal posteriors). It is difficult to test any of these possible solutions formally as, when using the strong constraints, we only ran into convergence issues 5% of the time, and so many simulations would have to be run to see if they could recover the true parameters despite the non-convergence of our Gibbs sampler.

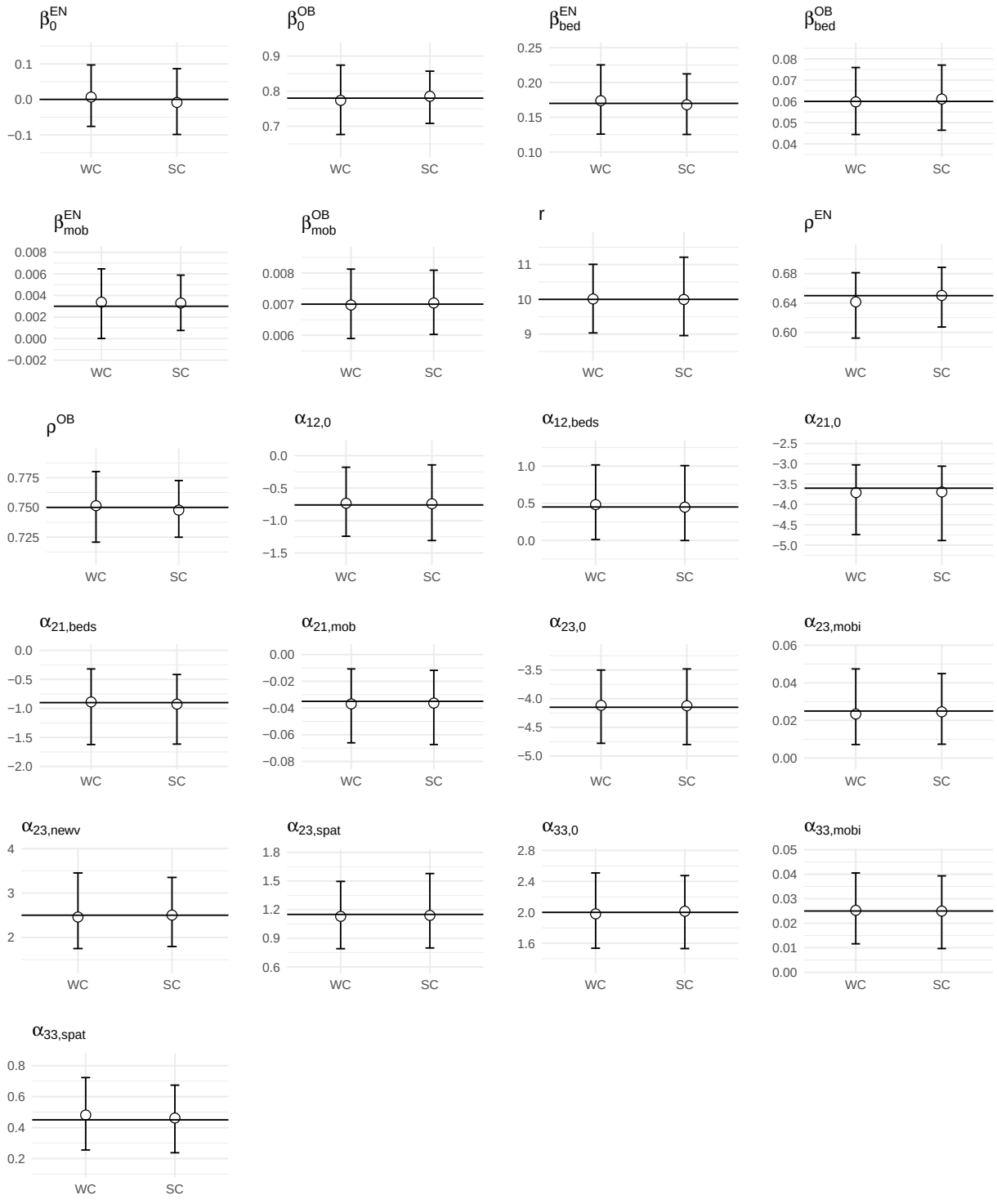


Figure B.3: Shows the sample mean (circles) and 95% quantile (caps) of the posterior medians from fitting 237 replications of (B.5) using the strong constraints (SC) (see text around figure) and an additional 284 replications of (B.5) using the weak constraints (WC), with our hybrid Gibbs sampling algorithm. The horizontal lines are drawn at the true parameter values.

B.3 Widely Applicable Information Criterion (WAIC)

B.3.1 Formulation

As mentioned in the main text, the WAIC for a state-space model is more accurate when the latent states are marginalized (Auger-Méthé et al., 2021). Starting with models that do not contain neighboring outbreak indicators in the transition probabilities, such as the Non-coupled Model from Section 4.5 of the main text, we can use the forward filtering part of the FFBS algorithm (Frühwirth-Schnatter, 2006) to calculate the marginalized density $p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})$, where $\boldsymbol{y}_{1:(t-1)}$ is the vector of all counts in all areas through $t-1$. Then, following Gelman et al. (2014), the WAIC can be calculated as,

$$\begin{aligned}
\text{lpdd} &= \sum_{i=1}^{N}\sum_{t=2}^{T} \log\left(\frac{1}{Q-M}\sum_{m=M+1}^{Q} p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]})\right), \\
\text{pwaic} &= \sum_{i=1}^{N}\sum_{t=2}^{T} Var_{m=M+1}^{Q} \log\left(p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]})\right), \\
\text{WAIC} &= -2(\text{lpdd} - \text{pwaic}),
\end{aligned} \tag{B.6}$$

where the superscript $[m]$ denotes a draw from the posterior of the variable and Var denotes the sample variance.

For models with neighboring outbreak indicators in the transition probabilities, such as the Full Coupled Model from Section 4.5 of the main text, we cannot completely marginalize $\boldsymbol{S}^*$ and calculate $p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})$ (Douwes-Schultz and Schmidt, 2022). As an alternative we could condition on $\boldsymbol{S}^*$ and use $p(y_{it}|\boldsymbol{S}^{*[m]}, \boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]}) = p(y_{it}|S_{it}^{[m]}, y_{i(t-1)}, \boldsymbol{\beta}^{[m]})$, which is given by Equation (4.1) in the main text, in place of $p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]})$ in Equation (B.6). For state-space models this is sometimes called the conditional WAIC, however, it has been shown to be inaccurate (Auger-Méthé et al., 2021). As a compromise we marginalize as much of $\boldsymbol{S}^*$ as is computationally possible from $p(y_{it}|\boldsymbol{S}^*, \boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})$. Note, from Equation (B.4) in Section B.1 above, it is possible to use the forward filtering part of the iFFBS algorithm to

calculate the partially marginalized density,

$$
\begin{aligned}
p(y_{it}|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}) &= p(y_{it}|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}) \\
&= \sum_{k=1}^{7} p(y_{it}|S^*_{it} = k, y_{i(t-1)}, \boldsymbol{v}) p(S^*_{it} = k|\boldsymbol{S}^*_{(-i)(1:t)}, \boldsymbol{y}_{i(1:t-1)}, \boldsymbol{v}),
\end{aligned}
$$

meaning we can marginalize all of $\boldsymbol{S}^*$ that is from the area and all future states in neighboring areas. Therefore, we use $p(y_{it}|\boldsymbol{S}^{*[m]}_{(-i)(1:t)}, \boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]})$ in place of $p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]})$ in Equation (B.6). In the next subsection, we show that the WAIC calculated in this manner can distinguish the correct model.

B.3.2 Simulation study

We designed a simulation study to ensure the WAIC, as formulated in the previous subsection, can choose the true data generating model. We focused on a comparison between models with/without neighboring outbreak indicators in the transition probabilities as this is one of the most important comparisons in the main text.

Firstly, we considered a CMSNB(1,2,4) model without neighboring outbreak indicators in the Markov chain,

$$
\begin{aligned}
\log(\lambda^{EN}_{it}) &= \beta^{EN}_0 + \beta^{EN}_{\text{beds}}\text{beds}_i + \beta^{EN}_{\text{mob}}\text{mobility}_{\text{county}(i)(t-4)} + \rho^{EN}\log(y_{i(t-1)}+1) \\
\log(\lambda^{OB}_{it}) &= \beta^{OB}_0 + \beta^{OB}_{\text{beds}}\text{beds}_i + \beta^{OB}_{\text{mob}}\text{mobility}_{\text{county}(i)(t-4)} + \rho^{OB}\log(y_{i(t-1)}+1) \\
r^{EN} &= r^{OB} = r \\
\text{logit}(p12_{it}) &= \alpha_{12,0} + \alpha_{12,\text{beds}}\text{beds}_i \\
\log\left(\frac{p21_{it}}{1-p21_{it}-p23_{it}}\right) &= \alpha_{21,0} + \alpha_{21,\text{beds}}\text{beds}_i \\
\log\left(\frac{p23_{it}}{1-p21_{it}-p23_{it}}\right) &= \alpha_{23,0} + \alpha_{23,\text{mobi}}\text{mobility}_{\text{county}(i)(t-4)} + \alpha_{23,\text{newv}}\text{new\_variant}_t \\
\text{logit}(p33_{it}) &= \alpha_{33,0} + \alpha_{33,\text{mobi}}\text{mobility}_{\text{county}(i)(t-4)},
\end{aligned}
\tag{B.7}
$$

for $i = 1, \ldots, 30$ and $t = 2, \ldots, 113$, and with the following true parameter values $\boldsymbol{v} = (\beta_0^{EN}, \beta_{\text{beds}}^{EN}, \beta_{\text{mob}}^{EN}, \rho^{EN}, \beta_0^{OB}, \beta_{\text{beds}}^{OB}, \beta_{\text{mob}}^{OB}, \rho^{OB}, r, \alpha_{12,0}, \alpha_{12,\text{beds}}, \alpha_{21,0}, \alpha_{21,\text{beds}}, \alpha_{23,0}, \alpha_{23,\text{mobi}}, \alpha_{23,\text{newv}}, \alpha_{33,0}, \alpha_{33,\text{mobi}})^T = (0, .1, 0, .5, .75, .05, .007, .75, 10, -1, .5, -3, -1, -3.5, .04, 1, 2.5, .02)^T$. We will refer to the model defined by Equation (B.7) as the Non-spatial Model.

We also considered a spatial extension of the Non-spatial Model with neighboring outbreak indicators in each transition probability,

$$
\begin{aligned}
\log(\lambda_{it}^{EN}) &= \beta_0^{EN} + \beta_{\text{beds}}^{EN}\text{beds}_i + \beta_{\text{mob}}^{EN}\text{mobility}_{\text{county}(i)(t-4)} + \rho^{EN}\log(y_{i(t-1)} + 1) \\
\log(\lambda_{it}^{OB}) &= \beta_0^{OB} + \beta_{\text{beds}}^{OB}\text{beds}_i + \beta_{\text{mob}}^{OB}\text{mobility}_{\text{county}(i)(t-4)} + \rho^{OB}\log(y_{i(t-1)} + 1) \\
r^{EN} &= r^{OB} = r \\
\text{logit}(p12_{it}) &= \alpha_{12,0} + \alpha_{12,\text{beds}}\text{beds}_i + \alpha_{12,\text{spat}}\sum_{j \in NE(i)} \omega_{ji} I[S_{j(t-1)} = 3] \\
\log\left(\frac{p21_{it}}{1 - p21_{it} - p23_{it}}\right) &= \alpha_{21,0} + \alpha_{21,\text{beds}}\text{beds}_i + \alpha_{21,\text{spat}}\sum_{j \in NE(i)} \omega_{ji} I[S_{j(t-1)} = 3] \\
\log\left(\frac{p23_{it}}{1 - p21_{it} - p23_{it}}\right) &= \alpha_{23,0} + \alpha_{23,\text{mobi}}\text{mobility}_{\text{county}(i)(t-4)} + \alpha_{23,\text{newv}}\text{new\_variant}_t \\
&\quad + \alpha_{23,\text{spat}}\sum_{j \in NE(i)} \omega_{ji} I[S_{j(t-1)} = 3] \\
\text{logit}(p33_{it}) &= \alpha_{33,0} + \alpha_{33,\text{mobi}}\text{mobility}_{\text{county}(i)(t-4)} + \alpha_{33,\text{spat}}\sum_{j \in NE(i)} \omega_{ji} I[S_{j(t-1)} = 3],
\end{aligned}
\tag{B.8}
$$

for $i = 1, \ldots, 30$ and $t = 2, \ldots, 113$, and with the following true parameter values $\boldsymbol{v} = (\beta_0^{EN}, \beta_{\text{beds}}^{EN}, \beta_{\text{mob}}^{EN}, \rho^{EN}, \beta_0^{OB}, \beta_{\text{beds}}^{OB}, \beta_{\text{mob}}^{OB}, \rho^{OB}, r, \alpha_{12,0}, \alpha_{12,\text{beds}}, \alpha_{12,\text{spat}}, \alpha_{21,0}, \alpha_{21,\text{beds}}, \alpha_{21,\text{spat}}, \alpha_{23,0}, \alpha_{23,\text{mobi}}, \alpha_{23,\text{newv}}, \alpha_{23,\text{spat}}, \alpha_{33,0}, \alpha_{33,\text{mobi}}, \alpha_{33,\text{spat}})^T = (0, .1, 0, .5, .75, .05, .007, .75, 10, -1, .5, .25, -3, .-1, -.25, -4, .04, 1, 1.2, 2, .02, .5)^T$. We will refer to the model defined by Equation (B.8) as the Spatial Model.

Like in Section B.2 above we simplified the models slightly compared to the main manuscript, such as by removing the random intercepts, to reduce the computational burden of running

Table B.1: Shows the results of model comparison using the WAIC for 24 replications of the Non-spatial Model and 24 replications of the Spatial Model. Each row represents the true model from which the 24 replications were produced.

True Model	**Non-spatial Model Chosen**	**Spatial Model Chosen**	**No Significant Difference ($\|\Delta\text{WAIC}\| < 5$)**
Non-spatial	13	1	10
Spatial	0	24	0

many simulations. The true parameter values were chosen to be like those estimated in our motivating example. In (B.7) and (B.8) beds_i, $\text{mobility}_{\text{county}(i)(t-4)}$, new_variant_t, $NE(i)$ and ω_{ji} are all the same as in our motivating example. Finally, we assumed a uniform initial state distribution for the Markov chain in each area.

We simulated 25 replications of the Non-spatial Model and fit the Spatial and Non-spatial Model to each replication. Then we simulated 25 replications of the Spatial Model and again fit the Spatial and Non-spatial Model to each replication. This resulted in 100 models being fit in total. Table B.1 gives the results of the model comparison using WAIC across all replications. There are 24 replications in each row as 2 of the models did not converge and were removed from the table, this is in line with the convergence rate reported in Section B.2 above. From Table B.1, when the Non-spatial Model generated the data the WAIC was split between choosing the correct Non-spatial Model and showing no significant difference between the models, and the WAIC only once chose the incorrect Spatial Model. For the replication where the WAIC chose the wrong model one of the spatial effects was highly significant, which can occur through random chance even when there is no real effect. When the Spatial Model generated the data the WAIC always chose the correct Spatial Model. These results suggest that when the WAIC shows no significant difference to prefer the less complex model and that the WAIC rarely selects the incorrect model when there is a significant difference.

A limitation is that we only compared a spatial and non-spatial model. The simulations are very computationally costly as one needs to run 2 models for every replication and there are multiple replications per model. Comparison between spatial and non-spatial models is arguably the most important comparison problem in the main text so we focused on it here.

B.4 Endemic-epidemic Model

Following Bracher and Held (2022), the Endemic-epidemic (EE) framework assumes the counts follow a multivariate autoregressive negative binomial model,

$$y_{it}|\boldsymbol{y}_{1:(t-1)} \sim NB(\lambda_{it}, \psi),$$

where $\boldsymbol{y}_{1:(t-1)}$ is the vector of all counts in all areas through $t-1$. The conditional mean λ_{it} is decomposed as,

$$\lambda_{it} = \lambda_{it}^{EP} \sum_{d=1}^{p} \sum_{j \in NE(i)} \lfloor u_d \rfloor \lfloor \omega_{ji} \rfloor y_{j,t-d} + \lambda_{it}^{BL}. \tag{B.9}$$

Here we will assume $NE(i)$ is the same as in Section 4.5.1 of the main text but also includes area i. The first term on the right-hand side of (B.9) accounts for expected new incidence due to local and neighboring disease transmission. The effects of covariates on transmission can be captured by the epidemic component λ_{it}^{EP} which is modeled in a log-linear fashion, in our case we assume,

$$\log(\lambda_{it}^{EP}) = \gamma_{0i}^{EP} + \gamma_1^{EP}\text{beds}_i + \gamma_2^{EP}\text{mobility}_{\text{county}(i)(t-4)} + \gamma_3^{EP}\text{newv}_t,$$

where $\gamma_{0i}^{EP} \sim N(\gamma_0^{EP}, \left(\sigma^{EP}\right)^2)$ is a normal random intercept. In (B.9) the spatial relationship between areas is described by the weights $\omega_{ji} > 0$ which are normalized $\lfloor \omega_{ji} \rfloor =$

$\omega_{ji} / \sum_{h:j \in NE(h)} \omega_{jh}$. The spatial weights are often assumed to follow a function, with unknown parameters, that decays with some measure of distance between areas (Bracher and Held, 2022). Following the main text Section 4.5.1, we assume $\omega_{ji} = \exp(-\phi d_{ji}^{BD})$, where $\phi > 0$ is decay parameter to be estimated and given a $U(0, 1000)$ prior (note that $d_{ii}^{BD} = 0$).

In (B.9) the relationship in the counts across time is described by the weights $u_d > 0$ which are normalized $\lfloor u_d \rfloor = u_d / \sum_{g=1}^{p} u_g$ for identifiability (Bracher and Held, 2022). We use a geometric specification (Bracher and Held, 2022) for the weights, $u_d = (1 - \kappa)^{d-1}\kappa$, where $0 < \kappa < 1$ is to be estimated and given a $U(0, 1)$ prior. Geometric weights decay over time. This is a reasonable assumption in our application as the serial interval, i.e., the time between the appearance of symptoms in successive generations, of COVID-19 is likely less than a week (Nishiura et al., 2020) implying most of the weight should be placed on the initial lag (Bracher and Held, 2022).

In (B.9) λ_{it}^{BL} is the baseline component, sometimes known as the endemic component (Bracher and Held, 2022) (we use baseline to avoid confusion with λ_{it}^{EN} in the main text), which captures contributions to incidence not directly related to local and neighboring disease transmission. For example, the baseline component might help capture imported cases from outside the study area. Like the epidemic component, the baseline component follows a log-linear model, in our case we assume,

$$\log(\lambda_{it}^{BL}) = \gamma_{0i}^{BL} + \gamma_1^{BL}\text{beds}_i + \gamma_2^{BL}\text{mobility}_{\text{county}(i)(t-4)} + \gamma_3^{BL}\text{newv}_t,$$

where $\gamma_{0i}^{BL} \sim N(\gamma_0^{BL}, \left(\sigma^{BL}\right)^2)$ is a normal random intercept. We found models with covariates in the baseline component fit better, according to the WAIC, compared to models without covariates.

To fit the EE models, we used standard MCMC methods in Nimble with flat priors, the code is available on GitHub `https://github.com/Dirk-Douwes-Schultz/CMSNB124_code`.

Table B.2: Shows the WAIC of Endemic-epidemic models with different maximum lags p. The best fitting model, the one with the lowest WAIC, is bolded.

Maximum Lag (p)	**WAIC**
1	17,884
2	**17,845**
3	17,848

We used the WAIC to choose the maximum temporal lag p and to compare the EE model to the Markov switching models from the main text (see Section 4.5.1 in the main text). To compare models using the WAIC they must be fit to the same data. The Markov switching models conditioned on the first observation and were fit to y_{it} for $i = 1, \ldots, 30$ and $t = 2, \ldots, 113$. Therefore, to remain consistent we introduced times $t = 0$ and $t = -1$ and fit the EE models across $t = 2, \ldots, 113$ conditioning on $t = 1, 0, -1$ (depending on p). Then to calculate the WAIC for the EE models we used Equation B.6 in Section B.3.1 above with $p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}^{[m]})$ substituted with $p(y_{it}|\boldsymbol{y}_{(-1):(t-1)}, \boldsymbol{v}^{[m]}) = NB(y_{it}|\lambda_{it}^{[m]}, \psi^{[m]})$, where $NB(y|\lambda, \psi)$ denotes a negative binomial density evaluated at y with mean λ and overdispersion ψ. Before time $t = -1$ there were no COVID-19 hospitalizations, so we only considered EE models with p up to 3. This should not be a major limitation as, from Table B.2, the WAIC increased going from $p = 2$ to $p = 3$. From Table B.2, we decided on an EE model with $p = 2$ and we will only discuss the results from that model in this section.

We estimated the decay parameter for the spatial weights with distance, ϕ, as 79.47 (67.13, 93.81) (posterior mean and 95% posterior credible interval), which is difficult to interpret given the normalization of the weights. From examining the posteriors of the normalized weights $\lfloor \omega_{ji} \rfloor$, for most areas 85-90 percent of the weight was placed on the home area ($\lfloor \omega_{ii} \rfloor$) and the remaining 10-15 percent was largely distributed over the closest 1-2 neighbors. This implies most transmission occurred locally and from close neighbors. We estimated κ as .85 (.80, .90) and 87 (83, 91) percent of the temporal weight was placed on the first lag ($\lfloor u_1 \rfloor$) and 13 (9, 17) percent was placed on the second lag ($\lfloor u_2 \rfloor$). A large portion of the temporal

Table B.3: Posterior means and 95% posterior credible intervals (in parenthesis) from the baseline and epidemic components of the fitted Endemic-epidemic model. The intercepts and covariate effects are exponentiated so that they represent rates and rate ratios. Rate ratios whose 95% posterior credible intervals do not contain 0 are bolded. The units for beds and mobility are equal to one standard deviation.

		Rate Ratios	
Covariate	**Parameter**	**Baseline**	**Epidemic**
Intercept of random intercepts	e^{γ_0}	.84 (.71, .99)	.93 (.88, .99)
Std. dev of random intercepts	σ	.66 (.42, .99)	.13 (.09, .17)
beds (100s)	e^{γ_1}	**1.18** (1.01, 1.36)	1.02 (.98, 1.06)
mobility (20%)	e^{γ_2}	**1.25 (1.05, 1.49)**	**1.14 (1.11, 1.18)**
new variant	e^{γ_3}	**1.46 (1.20, 1.75)**	**1.06 (1.01, 1.11)**

weight was placed on the previous week which is sensible considering the short serial interval of COVID-19.

Table B.3 shows the estimates from the baseline and epidemic components. The "Epidemic" column of the table shows the effect of each covariate on local and neighboring disease transmission. Mobility had a strong positive association with local and neighboring transmission while new variant and beds mainly affected the disease process through the baseline component. Note, as we are modeling hospitalizations, an effect in terms of transmission in the hospitalizations could reflect an effect on transmission in the actual cases and/or an effect on the severity of the disease. Mobility and beds should not affect disease severity; however, new variants often do (Chenchula et al., 2022). Therefore, the effects of new variant need to be interpreted with caution. It is interesting to compare the estimates in Table B.3 to those in Tables 4.2 and 4.3 in the main text. Both models largely agree that all three covariates are important for the disease process. However, we would argue the CMNSB(1,2,4) model

offers a deeper understanding of the covariate effects as it breaks down the effect of each covariate into its effect on transmission during the endemic and outbreak periods and on the epidemiological transitions. By breaking down the effects it could make it easier for policy makers to implement suitable interventions.

There are also some advantages of the EE models. The EE model only took one hour to fit while the CMSNB(1,2,4) model took eight hours. Also, the EE model can be fit using standard MCMC methods while the CMSNB(1,2,4) model requires the coding of custom iFFBS samplers. The CMSNB(1,2,4) model additionally requires implementing constraints to ensure consistent convergence. Finally, the EE models allow one to better understand the temporal and spatial structure of the disease counts by estimating the weights u_d and ω_{ji}. It is difficult to implement spatial and temporal weighting of the disease counts in the CMSNB(1,2,4) model due to the complexity of those models. Appropriate constraints would have to be considered as well as how the weight structure might switch between the states.

B.5 Temporal Predictions

In this section, we are interested in the posterior predictive distribution of the counts $p(y_{i(T+k)}|\boldsymbol{y})$ and the epidemiological state of the disease $p(S^*_{i(T+k)}|\boldsymbol{y})$ for $i = 1, \ldots, N$ and

$k = 1, \ldots, K$. The posterior predictive distribution of the counts is given by,

$$\begin{aligned} p(y_{i(T+k)}|\boldsymbol{y}) = \int & p(y_{i(T+k)}|S_{i(T+k)}, y_{i(T+k-1)}, \boldsymbol{\beta}) p(S^*_{i(T+k)}|S^*_{i(T+k-1)}, \boldsymbol{S}_{(-i)(T+k-1)}, \boldsymbol{\theta}) \\ & \times \prod_{j=1}^{N} p(y_{j(T+k-1)}|S_{j(T+k-1)}, y_{j(T+k-2)}, \boldsymbol{\beta}) p(S^*_{j(T+k-1)}|S^*_{j(T+k-2)}, \boldsymbol{S}_{(-j)(T+k-2)}, \boldsymbol{\theta}) \\ & \ldots \times \prod_{j=1}^{N} p(y_{j(T+1)}|S_{j(T+1)}, y_{jT}, \boldsymbol{\beta}) p(S^*_{j(T+1)}|S^*_{jT}, \boldsymbol{S}_{(-j)T}, \boldsymbol{\theta}) \\ & \times p(\boldsymbol{S}^*_{(1:N)T}, \boldsymbol{v}|\boldsymbol{y})\, dS^*_{i(T+k)} d\boldsymbol{y}_{(1:N)(T+1:T+k-1)} d\boldsymbol{S}^*_{(1:N)(T:T+k-1)} d\boldsymbol{\beta} d\boldsymbol{\theta}, \end{aligned} \tag{B.10}$$

and the posterior predictive distribution for the state of the disease in area i is given by,

$$\begin{aligned} p(S^*_{i(T+k)}|\boldsymbol{y}) = \int & p(S^*_{i(T+k)}|S^*_{i(T+k-1)}, \boldsymbol{S}_{(-i)(T+k-1)}, \boldsymbol{\theta}) \\ & \times \prod_{j=1}^{N} p(S^*_{j(T+k-1)}|S^*_{j(T+k-2)}, \boldsymbol{S}_{(-j)(T+k-2)}, \boldsymbol{\theta}) \\ & \ldots \times \prod_{j=1}^{N} p(S^*_{j(T+1)}|S^*_{jT}, \boldsymbol{S}_{(-j)T}, \boldsymbol{\theta}) \\ & \times p(\boldsymbol{S}^*_{(1:N)T}, \boldsymbol{v}|\boldsymbol{y})\, d\boldsymbol{S}^*_{(1:N)(T:T+k-1)} d\boldsymbol{\beta} d\boldsymbol{\theta}. \end{aligned} \tag{B.11}$$

The integrals (B.10) and (B.11) are intractable, however, they can be approximated by Monte Carlo integration,

$$p(y_{i(T+k)}|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} p(y_{i(T+k)}|S^{*[m]}_{i(T+k)}, y^{[m]}_{i(T+k-1)}, \boldsymbol{\beta}^{[m]}), \tag{B.12}$$

and

$$p(S^*_{i(T+k)}|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} p(S^*_{i(T+k)}|S^{*[m]}_{i(T+k-1)}, \boldsymbol{S}^{[m]}_{(-i)(T+k-1)}, \boldsymbol{\theta}^{[m]}), \tag{B.13}$$

where, like in the main text, the superscript $[m]$ denotes a draw from the posterior of the

Algorithm 2: Posterior Predictive Simulation

for m *in* $M+1:Q$ **do**
  for k *in* $1:K$ **do**
    for i *in* $1:N$ **do**
      1. Draw $S^{*[m]}_{i(T+k)}$ from $p(S^{*}_{i(T+k)}|S^{*[m]}_{i(T+k-1)}, \boldsymbol{S}^{[m]}_{(-i)(T+k-1)}, \boldsymbol{\theta}^{[m]})$.
      2. Draw $y^{[m]}_{i(T+k)}$ from $p(y_{i(T+k)}|S^{*[m]}_{i(T+k)}, y^{[m]}_{i(T+k-1)}, \boldsymbol{\beta}^{[m]})$, where $y^{[m]}_{iT} = y_{iT}$.
    end
  end
end

variable.

We can use a simulation procedure to draw realizations from the posterior predictive distributions (Frühwirth-Schnatter, 2006). Algorithm 2 will obtain realizations from the posterior predictive distribution of the counts, $y^{[m]}_{i(T+k)} \sim p(y_{i(T+k)}|\boldsymbol{y})$, and the epidemiological state of the disease, $S^{*[m]}_{i(T+k)} \sim p(S^{*}_{i(T+k)}|\boldsymbol{y})$, for $i = 1, \ldots, N$, $k = 1, \ldots, K$ and $m = M+1, \ldots, Q$. Then the realizations from the posterior predictive distributions can be substituted into Equations (B.12) and (B.13) if the posterior predictive distributions themselves need to be calculated. Although, as $S^{*}_{i(T+k)}$ can only take 7 values it is easier to approximate $p(S^{*}_{i(T+k)}|\boldsymbol{y})$ with the frequency distribution of $S^{*[m]}_{i(T+k)}$. Note that in Algorithm 2, the first step draws a new state for the disease from the Markov chain in each area and then step 2 draws new counts conditional on the new states.

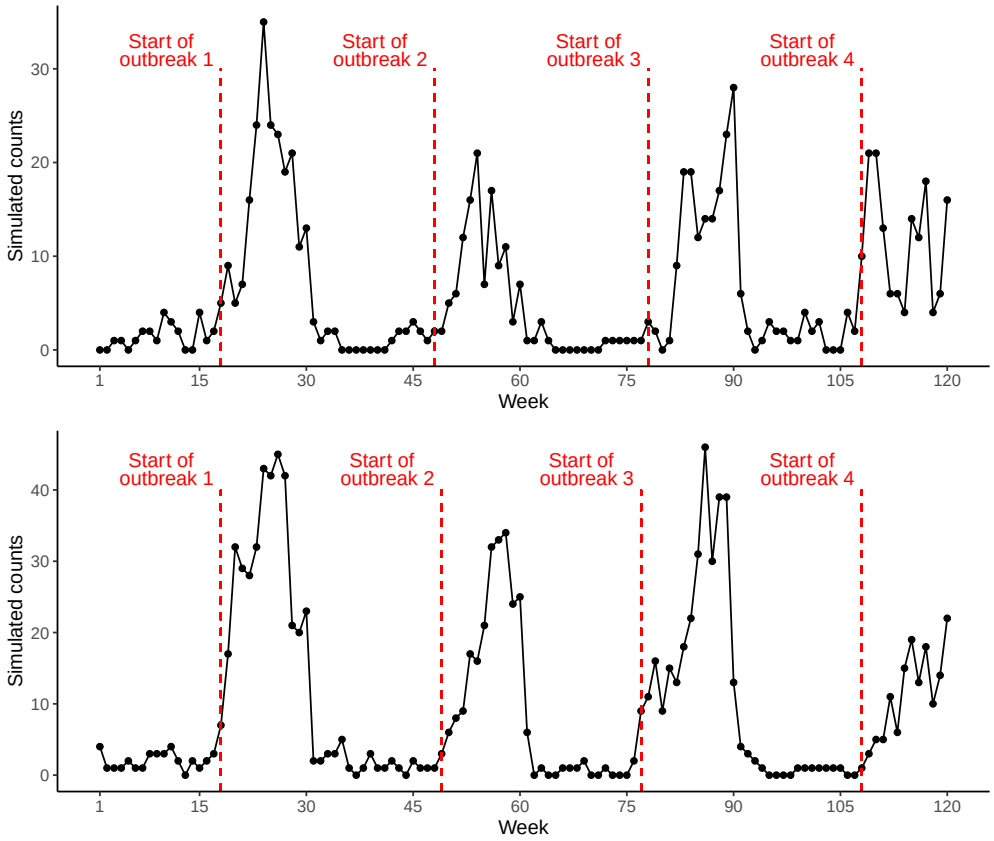


Figure B.4: Shows simulated counts in 2 areas from the simulation study described in Section 4.4 of the main text. The red dotted lines indicate the exact pre-determined start time of each outbreak in the area.

B.6 Further Results from the Simulation Study to Quantify and Compare State Estimation (Section 4.4 of the Main Text)

B.6.1 Simulated counts from 2 areas (SM Figure B.4)

Figure B.4 shows simulated counts in 2 areas from the simulation study described in Section 4.4 of the main text.

Table B.4: Shows the area under the ROC curve (AUC), sensitivity, specificity and timeliness (last 3 calculated with a 50 percent threshold) of three outbreak state estimates: (a) retrospective probabilities, $P(S_{it} = 3|\boldsymbol{y})$ from fitting the models to the full simulated data set (b) real-time detection probabilities, $P(S_{iT} = 3|\boldsymbol{y})$ from fitting the models up to time T for $T = 100, \ldots, 120$ (20 separate fits) and (c) real-time one week ahead forecasts, $P(S_{iT} = 3|\boldsymbol{y}_{1:(T-1)})$ from fitting the models up to time $T-1$ for $T = 101, \ldots, 120$. All outbreak state estimates were evaluated on the simulated data set described in Section 4.4 of the main text. The best criteria for each outbreak state estimate are bolded.

	(a) Retrospective		**(b) Real-time Detection**		**(c) Real-time Forecast**	
Criterion	**Spatial**	**Non-spatial**	**Spatial**	**Non-spatial**	**Spatial**	**Non-spatial**
AUC	**.995**	.992	**.983**	.960	**.965**	.919
Sensitivity	**.944**	.935	**.894**	.835	**.874**	.768
Specificity	**.981**	.967	**.982**	.964	**.985**	.964
Timeliness	**1.48**	1.62	**1.90**	2.53	**1.47**	2.47

B.6.2 Primary results from the simulation study (SM Table B.4)

Table B.4 summarizes the primary results of the simulation study described in Section 4.4 of the main text.

B.7 Further Results for the Application to COVID-19 Outbreaks Across Quebec

B.7.1 Difference in the posterior probability that an outbreak is currently happening between the Full Coupled and Non-coupled models averaged across all hospitals (SM Figure B.5)

Figure B.5 shows the difference in the posterior probability that an outbreak is currently happening between the Full Coupled and Non-coupled models averaged across all hospitals and was discussed in Section 4.5.4 of the main text.

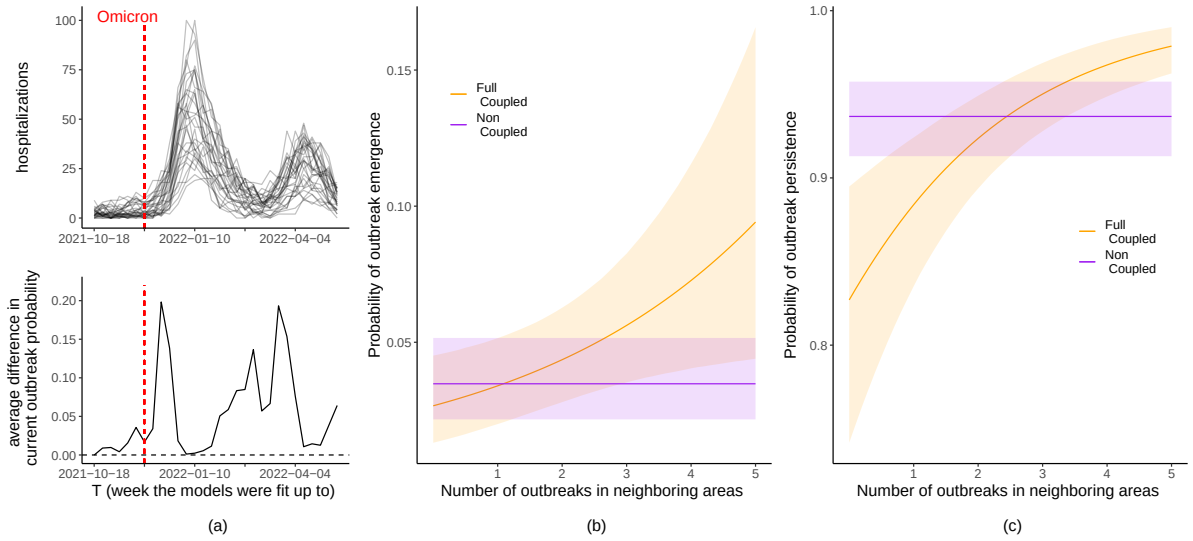


Figure B.5: (a)(top) Each line gives the number of hospitalizations in one of the 30 hospitals included in the study. (a)(bottom) shows the difference in the posterior probability that an outbreak is currently happening between the Full Coupled and Non-coupled models averaged across all hospitals, that is, $1/30 \sum_{i=1}^{30} P(S_{iT} = 3|\boldsymbol{y}, \text{Full Coupled Model}) - P(S_{iT} = 3|\boldsymbol{y}, \text{Non-coupled Model})$, versus T. (b) and (c) show posterior means (solid lines) and 95% posterior credible intervals (shaded areas) of the probabilities of outbreak emergence, (b), and outbreak persistence, (c), versus the number of outbreaks in neighboring areas assuming average connectivity and other covariates fixed at their average values. For (b) and (c), the Full Coupled Model is in orange, the Non-coupled Model is in purple, and the models were fit up to $T = 84$.

B.7.2 Map of the likely state in each catchment area during the last week of the study period (SM Figure B.6)

Figure B.6 shows a map of the likely state in each catchment area during the last week of the study period and was discussed in Section 4.5.4 of the main text.

B.7.3 Analysis of real-time false alarms

Figure B.7, similar in structure to Figure 4.6 of the main text, shows the posterior probabilities that an outbreak is currently happening (bottom graphs) for the Full Coupled Model in the two hospitals where we found evidence of false alarms during our real-time evaluation. Recall in the real-time evaluation the model was fit up to time T for $T = 84 =$ 2021-10-18, ..., $113 =$ 2022-05-09 and then the summaries in Figure B.7 were calculated for

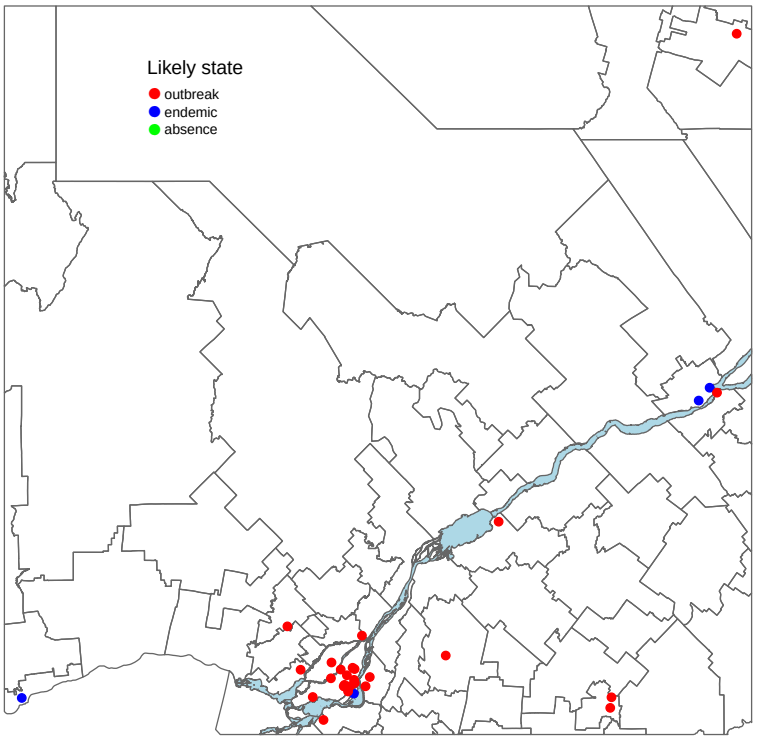


Figure B.6: A map of the part of Quebec where the 30 hospitals (points) included in the study are located. Borders separate counties. The color of the points represents the likely state in the catchment area of the hospital during the last week of the study period, that is, red if $P(S_{iT} = 3|\boldsymbol{y}) > .5$, blue if $P(S_{iT} = 2|\boldsymbol{y}) > .5$ and green if $P(S_{iT} = 1|\boldsymbol{y}) > .5$, where $T = 113 =$ 2022-05-09. From the Full Coupled Model with new variant as a covariate.

each T. We decided there was evidence of a false alarm being triggered at the time T if $P(S_{iT} = 3|\boldsymbol{y}) > .5$ and it did not appear that an outbreak had occurred at the time T in hindsight. As can be seen in Figure B.7, in both hospitals there is a sharp increase in $P(S_{iT} = 3|\boldsymbol{y})$ that appears to not be associated with any outbreaks. The model corrects itself quickly, however, with $P(S_{iT} = 3|\boldsymbol{y})$ typically declining sharply 1-2 weeks after the false alarm. Figure B.8, following the same graphical structure as Figure 4.5 in the main text, shows the retrospective state estimates in the same two hospitals as Figure B.7. Note, we did not include new variant as a covariate in the model used to produce Figure B.8 so that the model would be the same as the one used in the real-time evaluation. The false alarms do not show up in the retrospective state estimates, again showing that the model

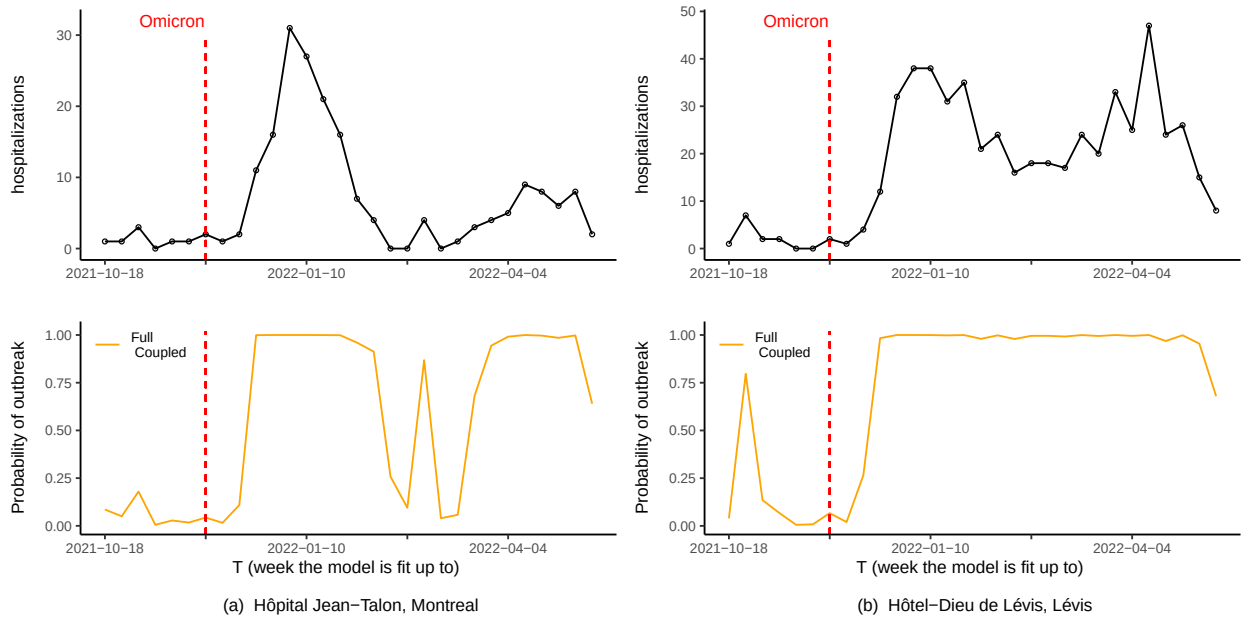


Figure B.7: Top graphs show the hospitalizations for the last 30 weeks of the study period where we conducted the real-time evaluation. Bottom graphs: solid lines show the posterior probabilities that an outbreak is currently happening, that is, $P(S_{iT} = 3|\boldsymbol{y})$, versus T. The dotted red lines are drawn at the introduction of the Omicron variant for all of Quebec. Shows results from the Full Coupled Model for the two hospitals where we found evidence of false alarms during the real-time evaluation.

corrects itself after gathering further data. We found no evidence in any areas of false alarms in the retrospective state estimates.

As to the cause of the false alarms in Figure B.7, the risk of an outbreak was high in both hospitals when the false alarms were triggered. In Hôpital Jean-Talon mobility was high and in Hôpital Hôtel-Dieu de Lévis there was strong evidence of outbreaks in several neighboring areas. Compared to the Non-coupled Model (not shown), the same false alarm was triggered in Hôpital Hôtel-Dieu de Lévis but not in Hôpital Jean-Talon. This is likely because outbreak risk was elevated in Hôpital Jean-Talon, when the false alarm was triggered, due to evidence of outbreaks in neighboring areas. In this case, a single false alarm seems worth the tradeoff for the earlier warning of the Omicron outbreak across the hospitals.

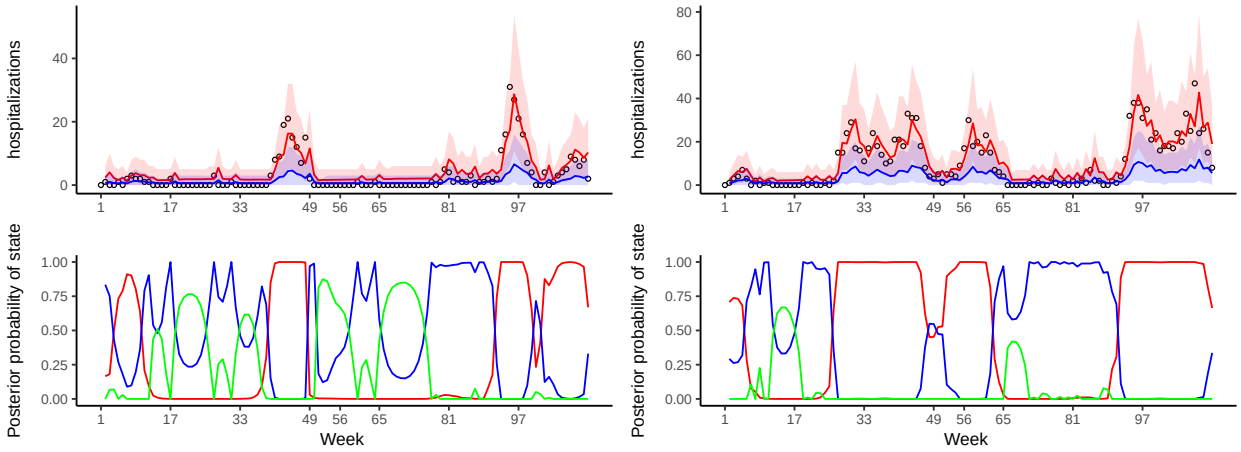


Figure B.8: Shows the retrospective state estimates from the Full Coupled Model, without new variant as a covariate, for the two hospitals where we found evidence of false alarms during the real-time evaluation. Follows the same graphical structure as Figure 4.5 from the main text.

APPENDIX C

Appendix to Manuscript 3

C.1 Reed-Frost Derivation of the ARMN Model

In this section, we will derive the ARMN model, see Section 5.2 of the main manuscript, from a multivariate Reed-Frost model. The ARMN model stands independently based on the justifications given in Section 5.2. However, the Reed-Frost derivation gives a more epidemiological interpretation for some of the parameters and reveals important sources of potential confounding.

The Reed-Frost model is a discrete-time susceptible-infectious-recovered (SIR) model originally developed by Lowell J. Reed and Wade Hampton Frost in a series of lectures at John Hopkins University in the late 1920s (Abbey, 1952). The model takes a time step of one serial interval, the time between successive cases, and assumes that an infected individual does not become infectious till the next time step and then afterward is immune. See Abbey (1952) and Chapter 6 of Vynnycky (2010) for more details. Following the autoregressive Poisson derivation of Bauer and Wakefield (2018), if disease k follows a Reed-Frost model in

area i then the number of new infections in the interval $(t-1, t]$ is given by,

$$\begin{aligned} y_{kit}|y_{ki(t-1)} &\sim \text{Pois}(\Phi_{kit}) \\ \Phi_{kit} &= \frac{\delta_{ki(t-1)}}{\text{pop}_i} R_{kit} y_{ki(t-1)}, \end{aligned} \tag{C.1}$$

where $\delta_{ki(t-1)}$ is the size of the susceptible population, the population not immune to the disease, for disease k at time $t-1$ in area i, and pop_i is the total population of area i. The parameter R_{kit} represents the time-varying basic reproduction number of disease k. That is, the average number of new infections in $(t-1, t]$ resulting from a single infectious individual if the population were fully susceptible, i.e., if $\delta_{ki(t-1)} = \text{pop}_i$. The parameter $R^*_{kit} = \frac{\delta_{ki(t-1)}}{\text{pop}_i} R_{kit}$ represents the effective reproduction number, sometimes called the net reproduction number, i.e., the average number of new infections resulting from a single infectious individual in the current, not fully susceptible, population. Both the basic reproduction and effective reproduction numbers are considered important measures of disease transmission in epidemiology (Vynnycky, 2010).

We make small adjustments to (C.1). Following Finkenstadt and Grenfell (2000), we raise $y_{ki(t-1)}$ to the power of $0 < \zeta_k < 1$ to account for nonhomogeneous mixing. We also add 1 to $y_{ki(t-1)}$ to avoid the absorbing 0 state of the Poisson autoregressive process in (C.1). It was shown by Fokianos and Tjøstheim (2011) that model results are not sensitive to the constant added to $y_{ki(t-1)}$. Therefore, we considered,

$$\Phi_{kit} = \frac{\delta_{ki(t-1)}}{\text{pop}_i} R_{kit} \left(y_{ki(t-1)} + 1\right)^{\zeta_k}, \tag{C.2}$$

in place of Φ_{kit} in Equation (C.1).

We use correlated log-linear models for the basic reproduction numbers,

$$\log(R_{kit}) = \beta_{0ki} + \boldsymbol{x}_{it}^T \boldsymbol{\beta}_k + \psi_{kit} + b_{it}, \tag{C.3}$$

where $\beta_{0ki} \sim N(\beta_{0k}, \sigma^2_{\beta,k})$ is a normal random intercept meant to account for between area differences and $\boldsymbol{x}_{it}$ is a vector of space-time covariates that may affect the transmission of the diseases. If we know a covariate x_{itl} does not affect the transmission of disease k, then we can fix $\beta_{kl} = 0$. To account for overdispersion and correlation between the reproduction numbers, we model the random effects ψ_{kit} using a multivariate normal distribution,

$$\boldsymbol{\psi}_{it} = \begin{pmatrix} \psi_{1it} \\ \vdots \\ \psi_{Kit} \end{pmatrix} \sim \text{MVN}_K(\mathbf{0}, \Sigma_{RF}), \tag{C.4}$$

where Σ_{RF} is a K by K variance-covariance matrix. In (C.3), b_{it} represents shared space-time factors. This may be particularly relevant for arboviruses that share the same vector as in our motivating example. We do not specify any model for b_{it} as the distribution of the disease counts conditional on their total does not depend on it, see below.

Equations (C.1)-(C.4) define a multivariate Reed-Frost model for the diseases. Following the well-known relationship between the multinomial and Poisson distributions we have that,

$$\boldsymbol{y}_{it} | \text{total}_{it}, \boldsymbol{y}_{t-1} \sim \text{Multinom}(\boldsymbol{\pi}_{it}, \text{total}_{it}), \tag{C.5}$$

where,

$$\pi_{kit} = \frac{\Phi_{kit}}{\sum_{j=1}^{K} \Phi_{jit}}. \tag{C.6}$$

The relative odds, relative to disease 1, are then given by,

$$\frac{\pi_{kit}}{\pi_{1it}} = \frac{\Phi_{kit}}{\Phi_{1it}} = \frac{R^*_{kit}}{R^*_{1it}} \frac{\left(y_{ki(t-1)} + 1\right)^{\zeta_k}}{\left(y_{1i(t-1)} + 1\right)^{\zeta_1}}, \tag{C.7}$$

where,

$$\log\left(\frac{R^*_{kit}}{R^*_{1it}}\right) = (\beta_{0ki} - \beta_{01i}) + \boldsymbol{x}_{it}^T(\boldsymbol{\beta}_k - \boldsymbol{\beta}_1) + (\psi_{kit} - \psi_{1it}) + (b_{it} - b_{it}) + \log\left(\frac{\delta_{ki(t-1)}}{\delta_{1i(t-1)}}\right). \tag{C.8}$$

Equations (C.5)-(C.8) define an ARMN model where,

- $\lambda_{kit} = \dfrac{R^*_{kit}}{R^*_{1it}}$; the ratio of the effective reproduction number of disease k and disease 1
- $\boldsymbol{x}_{kit} = \{x_{itl} \in \boldsymbol{x}_{it} \ : \ \beta_{kl} \neq 0 \text{ or } \beta_{1l} \neq 0\}$; the set of covariates that affect either the transmission of disease k or the transmission of disease 1.
- If $x_{itl} \in \boldsymbol{x}_{kit}$ let l_k represent the index of x_{itl} in $\boldsymbol{x}_{kit}$, then $\alpha_{kl_k} = \beta_{kl} - \beta_{1l}$; the difference between the effect of covariate x_{itl} on the effective reproduction number of disease k and disease 1
- $\alpha_{0ki} = \beta_{0ki} - \beta_{01i}$, $\alpha_{0k} = \beta_{0k} - \beta_{1k}$ and $\sigma_k^2 = \sigma_{\beta,k}^2 + \sigma_{\beta,1}^2$
- $\phi_{kit} = \psi_{kit} - \psi_{1it}$ and $\Sigma_{kj} = \Sigma_{RF,kj} - \Sigma_{RF,k1} - \Sigma_{RF,j1} + \Sigma_{RF,11}$
- $\log\left(\dfrac{\delta_{ki(t-1)}}{\delta_{1i(t-1)}}\right)$ is added as an offset to $\log(\lambda_{kit})$.

In the ARMN model the shared factors b_{it} are eliminated. Since we do not typically observe $\log\left(\frac{\delta_{ki(t-1)}}{\delta_{1i(t-1)}}\right)$ it is an important source of potential confounding. We discuss this issue extensively in Sections 5.2 and 5.4.4 of the main manuscript. Also, as we are modeling reported cases, there could be confounding due to changes in relative reporting rates. For example, consider a spatial covariate associated with areas more likely to report disease k compared to disease 1. In that case, it would seem like there was an increase in $\frac{R^*_{kit}}{R^*_{1it}}$ associated with the covariate even if there was no change in the true ratio.

C.1.1 Extensions to account for disease interactions and geographical spread

The multivariate Reed-Frost model defined by Equations (C.1)-(C.4) does not account for disease interactions, beyond the random effects ψ_{kit}, and disease spread between areas. To further account for disease interactions we can add $\log(y_{ji(t-1)} + 1)$ to $\boldsymbol{x}_{it}$ and assume its corresponding coefficient in $\boldsymbol{\beta}_j$ is equal to 0. That is, we can assume previous cases of disease j affect the basic reproduction number of the other diseases. This is justified in many instances, for example, it has been shown that mosquitoes infected by both Zika and chikungunya transmit Zika at a higher rate (Göertz et al., 2017). In the multinomial form of the model, the above is equivalent to adding $\log(y_{ji(t-1)} + 1)$ to $\boldsymbol{x}_{kit}$ for $j \neq k$ and $j \neq 1$, which we do for Zika and chikungunya in Section 5.4 of the main manuscript. Note any interactions with the baseline disease would be absorbed by ζ_1 or ζ_k in the multinomial representation of the model. A similar strategy for accounting for disease interactions in a multivariate Poisson model was used by Paul et al. (2008).

Several approaches have been proposed for extending the Reed-Frost model to deal with disease spread between areas (Wakefield et al., 2019). Bauer and Wakefield (2018), following Held et al. (2005), added a sum of weighted previous cases in neighboring areas to the conditional means of the Reed-Frost model. That is, in the context of our model, $\Phi'_{kit} = \Phi_{kit} + \lambda_k^{NE} \sum_{j \in NE(i)} \omega_{ji} y_{kj(t-1)}$, where Φ'_{kit} is the conditional mean adjusted for geographical disease spread, $\lambda_k^{NE} > 0$ is an unknown spatial effect and ω_{ji} are fixed weights representing the influence area j has on area i. However, an additive adjustment would be awkward in the multinomial form of the model (C.7). We can also consider a multiplicative adjustment, similar to Lawson and Kim (2022),

$$\Phi'_{kit} = \Phi_{kit} \left(\sum_{j \in NE(i)} \omega_{ji} y_{kj(t-1)} + 1 \right)^{\beta_k^{NE}},$$

where β_k^{NE} is an unknown spatial effect, which can be negative. This is equivalent to adding $\log(\sum_{j\in NE(i)} \omega_{ji} y_{kj(t-1)} + 1)$ to $\boldsymbol{x}_{it}$ and setting its coefficient in $\boldsymbol{\beta}_m$ for $m \neq k$ to 0. Unlike the additive adjustment, the multiplicative adjustment is convenient in the multinomial form of the model since it is the same as adding $\log(\sum_{j\in NE(i)} \omega_{ji} y_{kj(t-1)} + 1)$ and $\log(\sum_{j\in NE(i)} \omega_{ji} y_{1j(t-1)} + 1)$ to $\boldsymbol{x}_{kit}$. We do this in Section 5.4 of the main manuscript where we consider the weights $\omega_{ji} = 1/\sum_{m\in NE(i)} pop_m$, which gives the prevalence of the diseases across neighboring areas.

C.1.2 The difference between multinomial and multivariate Poisson approaches

We can use the above derivation of the ARMN model from the multivariate Reed-Frost model to better understand the advantages of both approaches. Firstly, the ARMN model has $(K-1)/K$ times as many parameters, not counting the shared factors b_{it}, compared to the multivariate Reed-Frost model. For instance, in the case of $K = 3$ diseases, the ARMN model has 66% as many parameters, which is a substantial reduction in model complexity. The ARMN model also eliminates all shared factors b_{it}. This may be especially relevant for arboviruses that share the same vector, like in our motivating example.

An immediate advantage of the multivariate Reed-Frost approach is that it allows for estimating the effect of a covariate on the transmission of any of the diseases. The ARMN model can only estimate the differences in the effects, which the multivariate Reed-Frost model could also provide. The multivariate Reed-Frost model could be extended to deal with zero inflation and long periods of disease absence in a similar way to the MS-ZIARMN model. Following Section 5.2.1 of the main manuscript, we could replace Φ_{kit} by $\Phi_{kit} S_{kit}$ and model $\boldsymbol{S}_{it}$ as following a coupled Markov chain. In this case, we would not have to assume S_{1it} is always equal to one. That is, we would not have to assume one of the diseases was always present like with the MS-ZIARMN model.

In conclusion, there seem to be advantages to both approaches. The multinomial approach is likely more appropriate when one is mainly interested in differences between the transmission of the diseases and we can assume one of the diseases is always present, like in our motivating example. It goes beyond the scope of this paper to do a comparison based on model fitting given the complexity of the multivariate Reed-Frost model and any zero-inflated extensions. This will be the subject of further work.

C.2 Simulation Study to Investigate Correlations Induced by the Random Effects

We designed a simulation study to better understand the correlations between the disease counts, conditional on their total, induced by the random effects ϕ_{kit} in Equations (5.3)-(5.4) of the main manuscript. We simulated from the following multinomial distribution with multivariate normal random effects added to the log relative odds,

$$\begin{aligned}
(y_1, y_2, y_3)^T \,|\, \text{total} &\sim \text{Multinom}(\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)^T, \text{total} = y_1 + y_2 + y_3) \\
\log\left(\frac{\pi_3}{\pi_1}\right) &= \alpha_{03} + \phi_3 \\
\log\left(\frac{\pi_2}{\pi_1}\right) &= \alpha_{02} + \phi_2 \\
(\phi_2, \phi_3)^T &\sim MVN_2\left((0,0)^T, \begin{pmatrix} \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}\right).
\end{aligned} \tag{C.9}$$

We fixed $\alpha_{02} = \log(1.14)$, $\alpha_{03} = 0$, $\sigma_2 = .75$ and $\sigma_3 = .8$ based on the estimates from our motivating example in Section 5.4.2 of the main manuscript. We assumed total = 10 as this was the average sum of dengue, Zika and chikungunya cases in a neighborhood.

Figure C.1 shows the correlation between y_2 and y_3, conditional on the total count and marginalizing the random effects, versus ρ based on simulations of (C.9) (to the best of our

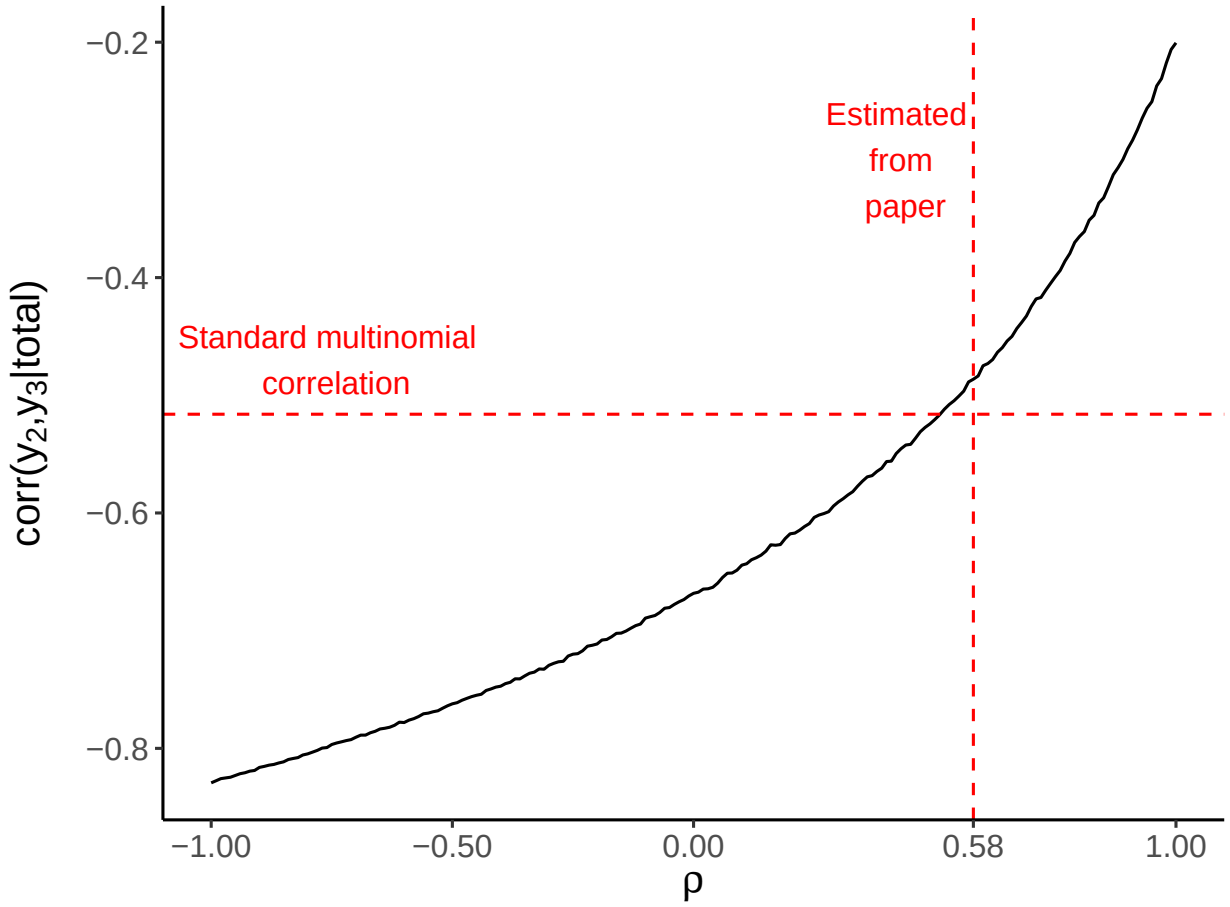


Figure C.1: The black solid line shows corr$(y_2, y_3|\text{total})$ versus ρ, marginalizing the random effects, based on simulations of (C.9). The red horizontal dashed line is drawn at corr$(y_2, y_3|\text{total})$ assuming the random effects are fixed at 0, i.e., $\phi_2 = \phi_3 = 0$ in (C.9). The red vertical dashed line is drawn at the value of ρ estimated in Section 5.4 of the main text.

knowledge there is no closed form solution). If there were no random effects then,

$$\text{corr}(y_2, y_3|\text{total}) = \frac{-\text{total}\left(\frac{e^{\alpha_{02}}}{1+e^{\alpha_{02}}+e^{\alpha_{03}}}\right)\left(\frac{e^{\alpha_{03}}}{1+e^{\alpha_{02}}+e^{\alpha_{03}}}\right)}{\sqrt{\text{total}\left(\frac{e^{\alpha_{02}}}{1+e^{\alpha_{02}}+e^{\alpha_{03}}}\right)\left(\frac{1+e^{\alpha_{03}}}{1+e^{\alpha_{02}}+e^{\alpha_{03}}}\right)}\sqrt{\text{total}\left(\frac{e^{\alpha_{03}}}{1+e^{\alpha_{02}}+e^{\alpha_{03}}}\right)\left(\frac{1+e^{\alpha_{02}}}{1+e^{\alpha_{02}}+e^{\alpha_{03}}}\right)}},$$

which is given by the horizontal red dashed line in Figure C.1. The correlation between the random effects ρ needs to be greater than around .5 for the marginalized correlation between the counts to be greater than that produced by the standard multinomial distribution (i.e., without the random effects). This is an interesting and somewhat unintuitive result, as

we would expect $\rho = 0$ to be a more likely threshold for this. In Section 5.4 of the main manuscript, we estimated ρ as .58 (.52, .65), and, therefore, the random effects did not likely induce any significant correlations between the counts in our motivating example.

From Figure C.1, even if ρ is close to 1 the marginal correlation between the counts, conditional on the total, is negative. We do not see negative correlations as a limitation of the model. The correlation is conditional on the total count and, therefore, if one of the counts increases another must necessarily decrease. Regardless, based on Figure C.1, adding multivariate normal random effects to the log relative odds leads to a more flexible correlation structure among the disease counts compared to the standard multinomial distribution. We tried different values of α_{02} and α_{03} in (C.9) and got similar results. Also, plotting $\text{corr}(y_1, y_3|\text{total})$ and $\text{corr}(y_1, y_2|\text{total})$ versus ρ (not shown) led to the same conclusions.

C.3 Forward Filtering Backward Sampling (FFBS) Algorithm

In this Section, we provide the FFBS algorithm (Chib, 1996) for sampling from $p(\boldsymbol{S}^*|\boldsymbol{y}, \boldsymbol{v})$ needed for the hybrid Gibbs sampler described in Section 5.3 of the main text. Like in Section 5.3 of the main manuscript, we will assume $K = 3$, which leads to 4 possible states for S^*_{it}. We will use the subscript $t_1 : t_2$ to denote a temporally indexed vector subsetted to the interval t_1 to t_2, e.g., $\boldsymbol{S}^*_{i(t_1:t_2)} = (S^*_{it_1}, \ldots, S^*_{it_2})^T$. We will also use the subscript $(-i)$ to denote the vector with the ith element removed.

First note that $p(\boldsymbol{S}^*|\boldsymbol{y}, \boldsymbol{v}) \propto p(\boldsymbol{S}^*, \boldsymbol{y}|\boldsymbol{v})$ which, from Equation (5.9) of the main text, factors into functions involving only $\boldsymbol{S}^*_{i(1:T)}$ for each i. Therefore, $\boldsymbol{S}^*_{1(1:T)}, \ldots, \boldsymbol{S}^*_{N(1:T)}$ are mutually independent give $\boldsymbol{y}$ and $\boldsymbol{v}$. This means to sample from $p(\boldsymbol{S}^*|\boldsymbol{y}, \boldsymbol{v})$ we can sample from $p(\boldsymbol{S}^*_{i(1:T)}|\boldsymbol{y}, \boldsymbol{v})$ for each i. A similar argument can be used to show that $\boldsymbol{S}^*_{1(1:t)}, \ldots, \boldsymbol{S}^*_{N(1:t)}$ are mutually independent conditional on either $\boldsymbol{y}_{1:t}$ and $\boldsymbol{v}$ or $\boldsymbol{y}_{1:(t-1)}$ and $\boldsymbol{v}$.

The initial forward filtering part of the FFBS algorithm starts at $t = 2$ and goes recursively to $t = T$. For each t, we first calculate the predictive probabilities and then the filtered probabilities. The predictive probabilities are given by,

$$P(S^*_{it} = l|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}) = \sum_{k=1}^{4} P(S^*_{it} = l|S^*_{i(t-1)} = k, \boldsymbol{y}_{t-1}, \boldsymbol{v}) P(S^*_{i(t-1)} = k|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}), \quad \text{(C.10)}$$

for $l = 1, \ldots 4$. For $t = 2$, $P(S^*_{i(t-1)} = k|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}) = P(S^*_{i1} = k|\boldsymbol{v})$, which is the initial state distribution. Recall that the modeler sets the initial state distribution based on how likely they believe the diseases to be present at the beginning of the study. The probability $P(S^*_{it} = l|S^*_{i(t-1)} = k, \boldsymbol{y}_{t-1}, \boldsymbol{v})$ is the probability of the Markov chain in area i transitioning from state k at time $t-1$ to state l at time t, i.e., $\Gamma(S^*_{it}|\boldsymbol{y}_{t-1})_{kl}$, see Section 5.3 of the main text. The filtered probabilities are given by,

$$\begin{aligned}
p(S^*_{it}|\boldsymbol{y}_{1:t}, \boldsymbol{v}) &= p(S^*_{it}|\boldsymbol{S}^*_{(-i)t}, \boldsymbol{y}_{1:t}, \boldsymbol{v}) \qquad \text{(by mutual independence)} \\
&\propto p(\boldsymbol{y}_t|\boldsymbol{S}^*_t, \boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}) p(\boldsymbol{S}^*_t|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}) \\
&= \prod_{j=1}^{N} p(y_{jt}|S^*_{jt}, \boldsymbol{y}_{1:(t-1)}, \text{total}_{jt}, \boldsymbol{\beta}) p(S^*_{jt}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}) \qquad \text{(by mutual independence)} \\
&\propto p(y_{it}|S^*_{it}, \boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{\beta}) p(S^*_{it}|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}),
\end{aligned}$$

implying,

$$P(S^*_{it} = l|\boldsymbol{y}_{1:t}, \boldsymbol{v}) = \frac{p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it} = l, \text{total}_{it}, \boldsymbol{\beta}) P(S^*_{it} = l|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})}{p(\boldsymbol{y}_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v})},$$

where,

$$p(\boldsymbol{y}_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v}) = \sum_{k=1}^{4} p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it} = k, \text{total}_{it}, \boldsymbol{\beta}) P(S^*_{it} = k|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}). \quad \text{(C.11)}$$

The distribution $p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it} = l, \text{total}_{it}, \boldsymbol{\beta})$ is given by Equation (5.8) of the main text. The probability $P(S^*_{it} = l|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v})$ is the predictive probability for time t calculated in the

previous step.

Once the filtered probabilities have been calculated for $t = 2, \ldots, T$, the backward sampling part of the FFBS algorithm is performed. We have that,

$$p(\boldsymbol{S}^*_{i(1:T)}|\boldsymbol{y}, \boldsymbol{v}) = \left[\prod_{t=1}^{T-1} p(S^*_{it}|S^*_{i(t+1)}, \boldsymbol{y}_{1:t}, \boldsymbol{v})\right] p(S^*_{iT}|\boldsymbol{y}, \boldsymbol{v}).$$

Also, note that,

$$p(S^*_{it}|S^*_{i(t+1)}, \boldsymbol{y}_{1:t}, \boldsymbol{v}) \propto p(S^*_{i(t+1)}|S^*_{it}, \boldsymbol{y}_{1:t}, \boldsymbol{v}) p(S^*_{it}|, \boldsymbol{y}_{1:t}, \boldsymbol{v}),$$

implying,

$$P(S^*_{it} = l|S^*_{i(t+1)} = j, \boldsymbol{y}_{1:t}, \boldsymbol{v}) = \frac{P(S^*_{i(t+1)} = j|S^*_{it} = l, \boldsymbol{y}_{1:t}, \boldsymbol{v}) P(S^*_{it} = l|, \boldsymbol{y}_{1:t}, \boldsymbol{v})}{\sum_{k=1}^{4} P(S^*_{i(t+1)} = j|S^*_{it} = k, \boldsymbol{y}_{1:t}, \boldsymbol{v}) P(S^*_{it} = k|, \boldsymbol{y}_{1:t}, \boldsymbol{v})}, \tag{C.12}$$

for $l = 1, \ldots, 4$. The probability $P(S^*_{i(t+1)} = j|S^*_{it} = l, \boldsymbol{y}_{1:t}, \boldsymbol{v})$ is given by $\Gamma(S^*_{i(t+1)}|\boldsymbol{y}_t)_{lj}$. The probability $P(S^*_{it} = l|, \boldsymbol{y}_{1:t}, \boldsymbol{v})$ is the filtered probability for time t.

Therefore, to sample $\boldsymbol{S}^{*[m]}_{i(1:T)} \sim p(\boldsymbol{S}^*_{i(1:T)}|\boldsymbol{y}, \boldsymbol{v})$ we first sample $S^{*[m]}_{iT} \sim p(S^*_{iT}|\boldsymbol{y}, \boldsymbol{v})$ using the final filtered probabilities. Then we work backwards sampling $S^{*[m]}_{it} \sim p(S^*_{it}|S^*_{i(t+1)} = S^{*[m]}_{i(t+1)}, \boldsymbol{y}_{1:t}, \boldsymbol{v})$ using the probabilities $P(S^*_{it} = l|S^*_{i(t+1)} = S^{*[m]}_{i(t+1)}, \boldsymbol{y}_{1:t}, \boldsymbol{v})$ for $l = 1, \ldots, 4$, from (C.12), for $t = T - 1, \ldots, 1$.

Finally, we note that the likelihood of $\boldsymbol{v}$ given $\boldsymbol{y}$, conditioning on the totals, is provided by,

$$p(\boldsymbol{y}|\boldsymbol{v}) = \prod_{i=1}^{N} \prod_{t=2}^{T} p(\boldsymbol{y}_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v}),$$

which can be calculated using the forward filter, see Equation (C.11). Therefore, we could base inference about $\boldsymbol{v}$ on the marginal, concerning the states, likelihood. However, as

explained in the main text, we are interested in making inferences about the unknown states as they represent when the model believes the diseases were present. The marginal likelihood contribution $p(\boldsymbol{y}_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v})$ is used, however, to calculate the WAIC as we explain in the next section.

C.4 Widely Applicable Information Criterion (WAIC)

The WAIC for a state space model is more accurate when the latent state indicators are marginalized (Auger-Méthé et al., 2021). We can use the forward filtering part of the FFBS algorithm to marginalize S^*_{it} from $p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it}, \text{total}_{it}, \boldsymbol{\beta})$,

$$p(\boldsymbol{y}_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v}) = \sum_{k=1}^{4} p(\boldsymbol{y}_{it}|\boldsymbol{y}_{t-1}, S^*_{it} = k, \text{total}_{it}, \boldsymbol{\beta}) P(S^*_{it} = k|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{v}),$$

see Equation (C.11) above. Then, following Gelman et al. (2014), the WAIC can be calculated as,

$$\begin{aligned}
\text{lpdd} &= \sum_{i=1}^{N}\sum_{t=2}^{T} \log\left(\frac{1}{Q-M}\sum_{m=M+1}^{Q} p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v}^{[m]})\right), \\
\text{pwaic} &= \sum_{i=1}^{N}\sum_{t=2}^{T} Var_{m=M+1}^{Q} \log\left(p(y_{it}|\boldsymbol{y}_{1:(t-1)}, \text{total}_{it}, \boldsymbol{v}^{[m]})\right), \\
\text{WAIC} &= -2(\text{lpdd} - \text{pwaic}),
\end{aligned} \tag{C.13}$$

where the superscript $[m]$ denotes a draw from the posterior of the variable and Var denotes the sample variance.

C.5 Estimates From the Comparison Models

This section provides the estimated coefficients from the multinomial part of the comparison models described in Section 5.4 of the main text. The tables here mirror Table 5.2 from

the main text. The estimates for the ARMN and Zeng (2022) models are given in Table C.1 and those for the ZIARMN model in Table C.2. We combined the ARMN and Zeng (2022) models as the Zeng (2022) model estimated that the diseases were always present, meaning the multinomial parameter estimates from the two models were nearly identical. The estimates for the sensitivity analysis model from Section 5.4.4 of the main text are given in Table C.3.

Table C.1: Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted Zeng (2022) and ARMN models. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Section 5.4.1 of the main text for an explanation of the covariates.

	Relative Odds Ratio or Rate Ratio	
Covariates	**Zika-dengue**	**chik.-dengue**
Intercept	.41 (.38, .45)	.19 (.17, .22)
verde_i	.98 (.92, 1.05)	**.90 (.83, .98)**
SDI_i	1.03 (.96, 1.10)	1 (.92, 1.09)
popdens_i	1 (.94, 1.09)	.97 (.88, 1.06)
favela_i	.96 (.90, 1.03)	.95 (.87, 1.03)
temp_{it}	**1.10 (1.04, 1.16)**	**.81 (.76, .86)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{1j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**.56 (.53, .59)**	**.56 (.53, .59)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**2.56 (2.37, 2.76)**	–
$\log\left(\frac{\sum_{j\in NE(i)} y_{3j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	–	**2.13 (2.01, 2.25)**
$\log(y_{2i(t-1)} + 1)$	–	**1.45 (1.36, 1.55)**
$\log(y_{3i(t-1)} + 1)$	**1.30 (1.25, 1.36)**	–

Table C.2: Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted ZIAMRN model. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Section 5.4.1 of the main text for an explanation of the covariates.

	Relative Odds Ratio or Rate Ratio	
Covariates	**Zika-dengue**	**chik.-dengue**
Intercept	1.19 (1.07, 1.31)	.99 (.89, 1.10)
verde_i	1 (.94, 1.07)	**.91 (.84, .99)**
SDI_i	1.07 (.99, 1.15)	1 (.91, 1.10)
popdens_i	1.02 (.94, 1.10)	1.05 (.96, 1.16)
favela_i	.97 (.91, 1.04)	.93 (.86, 1.01)
temp_{it}	**1.17 (1.11, 1.24)**	**.86 (.81, .92)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{1j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**.68 (.64, .71)**	**.68 (.64, .71)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**1.53 (1.44, 1.65)**	–
$\log\left(\frac{\sum_{j\in NE(i)} y_{3j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	–	**1.43 (1.36, 1.51)**
$\log(y_{2i(t-1)} + 1)$	–	1.01 (.96, 1.08)
$\log(y_{3i(t-1)} + 1)$	1.03 (.99, 1.07)	–

Table C.3: Posterior means and 95% posterior credible intervals (in parentheses) for the estimated coefficients from the multinomial part of the fitted sensitivity analysis model from Section 5.4.4 of the main text. The intercept row shows λ_{kit} for $k = 2$ (Zika) and $k = 3$ (chik.) in a typical area at average values of the covariates. Recall, if $\lambda_{kit} > 1$ ($\lambda_{kit} < 1$) then the share of disease k relative to dengue will tend to grow (shrink) over time. Other rows show the ratio of π_{kit}/π_{1it} (relative odds ratio) or the ratio of λ_{kit} (rate ratio) (both are the same, see Section 5.2) corresponding to a unit increase in the covariate. All covariates are standardized. Significant effects are bolded. See Sections 5.4.1 and 5.4.4 of the main text for an explanation of the covariates.

	Relative Odds Ratio or Rate Ratio	
Covariates	**Zika-dengue**	**chik.-dengue**
Intercept	1.13 (1.02, 1.26)	1 (.9, 1.11)
verde_i	1.02 (.94, 1.1)	.92 (.85, 1)
SDI_i	1.06 (.97, 1.15)	1.02 (.92, 1.12)
popdens_i	1.02 (.93, 1.12)	1.06 (.96, 1.17)
favela_i	.97 (.89, 1.04)	.94 (.86, 1.03)
temp_{it}	**1.14 (1.08, 1.20)**	**.85 (.80, .90)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{1j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**.7 (.67, .74)**	**.7 (.67, .74)**
$\log\left(\frac{\sum_{j\in NE(i)} y_{2j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	**1.59 (1.48, 1.70)**	–
$\log\left(\frac{\sum_{j\in NE(i)} y_{3j(t-1)}}{\sum_{j\in NE(i)} \text{pop}_j} + 1\right)$	–	**1.43 (1.36, 1.51)**
$\log(y_{2i(t-1)} + 1)$	–	**.9 (.84, .96)**
$\log(y_{3i(t-1)} + 1)$	.98 (.94, 1.02)	–
$\left(\sum_{j=1}^{t-1} y_{2ij} - \sum_{j=1}^{t-1} y_{1ij}\right)/\text{pop}_i$	.94 (.87, 1)	–
$\left(\sum_{j=1}^{t-1} y_{3ij} - \sum_{j=1}^{t-1} y_{1ij}\right)/\text{pop}_i$	–	.99 (.94, 1.05)

C.6 Map of Temperature in Rio

Figure C.2 shows a map of the average weekly maximum temperature across Rio between 2015-2016. Temperatures tend to be higher in the west of the city and lower downtown.

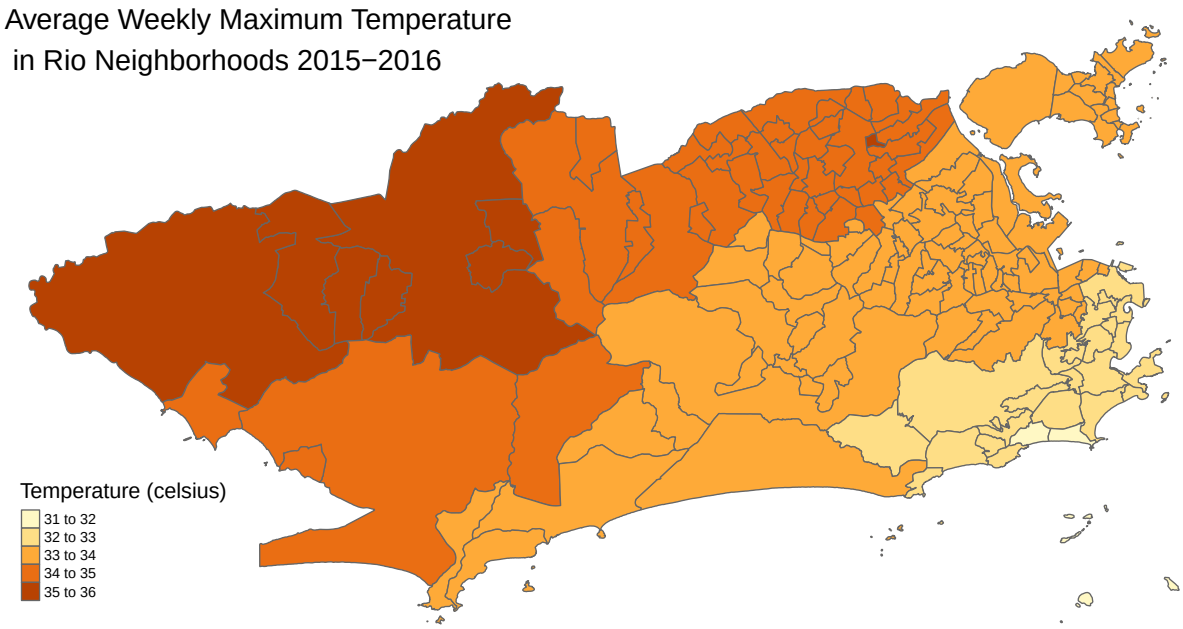


Figure C.2: Shows the average weekly maximum temperature across Rio neighborhoods between 2015-2016.

References

Abbey, H. (1952) An examination of the Reed-Frost theory of epidemics. *Human Biology*, **24**, 201–233.

Achterberg, M. A., Prasse, B., Ma, L., Trajanovski, S., Kitsak, M. and Van Mieghem, P. (2022) Comparing the accuracy of several network-based COVID-19 prediction algorithms. *International Journal of Forecasting*, **38**, 489–504.

Adams, B. and Boots, M. (2010) How important is vertical transmission in mosquitoes for the persistence of dengue? Insights from a mathematical model. *Epidemics*, **2**, 1–10.

Agarwal, D. K., Gelfand, A. E. and Citron-Pousty, S. (2002) Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, **9**, 341–355.

Agudze, K. M., Billio, M., Casarin, R. and Ravazzolo, F. (2022) Markov switching panel with endogenous synchronization effects. *Journal of Econometrics*, **230**, 281–298.

Aktekin, T. and Musal, M. (2015) Analysis of income inequality measures on human immunodeficiency virus mortality: A spatiotemporal Bayesian perspective. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178**, 383–403.

Amorós, R. (2017) *Bayesian Temporal and Spatio-Temporal Markov Switching Models for the Detection of Influenza Outbreaks*. Ph.D. thesis, Universitat de València.

Amorós, R., Conesa, D., López-Quílez, A. and Martinez-Beneito, M.-A. (2020) A spatio-

temporal hierarchical Markov switching model for the early detection of influenza outbreaks. *Stochastic Environmental Research and Risk Assessment*, **34**, 275–292.

Ansari, H., Mansournia, M., Izadi, S., Zeinali, M., Mahmoodi, M. and Holakouie-Naieni, K. (2015) Predicting CCHF incidence and its related factors using time-series analysis in the southeast of Iran: Comparison of SARIMA and Markov switching models. *Epidemiology & Infection*, **143**, 839–850.

Aogo, R. A., Zambrana, J. V., Sanchez, N., Ojeda, S., Kuan, G., Balmaseda, A., Gordon, A., Harris, E. and Katzelnick, L. C. (2023) Effects of boosting and waning in highly exposed populations on dengue epidemic dynamics. *Science Translational Medicine*, **15**, eadi1734.

Arab, A. (2015) Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health*, **12**, 10536–10548.

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A. A., Leos-Barajas, V., Mills Flemming, J., Nielsen, A., Petris, G. and Thomas, L. (2021) A guide to state–space modeling of ecological time series. *Ecological Monographs*, **91**, e01470.

Bartlett, M. S. (1957) Measles periodicity and community size. *Journal of the Royal Statistical Society: Series A (General)*, **120**, 48–60.

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**, 1379–1398.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z. and Wang, Y. (2016) Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in Medicine*, **35**, 1848–1865.

Baum, L. E. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**, 1554–1563.

Bi, S., Broggi, M. and Beer, M. (2019) The role of the Bhattacharyya distance in stochastic model updating. *Mechanical Systems and Signal Processing*, **117**, 437–452.

Billio, M., Casarin, R., Ravazzolo, F. and Van Dijk, H. K. (2016) Interconnections between Eurozone and US booms and busts using a Bayesian panel Markov-switching VAR model. *Journal of Applied Econometrics*, **31**, 1352–1370.

Bisanzio, D., Dzul-Manzanilla, F., Gomez-Dantés, H., Pavia-Ruz, N., Hladish, T. J., Lenhart, A., Palacio-Vargas, J., Roldan, J. F. G., Correa-Morales, F., Sánchez-Tejeda, G., Morales, P. K., Manrique-Saide, P., Longini, I. M., Halloran, M. E. and Vazquez-Prokopec, G. M. (2018) Spatio-temporal coherence of dengue, chikungunya and Zika outbreaks in Merida, Mexico. *PLOS Neglected Tropical Diseases*, **12**, e0006298.

Bjørnstad, O. N. (2023) *Epidemics: Models and Data Using R*. Use R! Cham: Springer International Publishing.

Bracher, J. and Held, L. (2021) A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *Biometrics*, **77**, 1202–1214.

— (2022) Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting*, **38**, 1221–1233.

Bradley, J. R., Wikle, C. K. and Holan, S. H. (2019) Spatio-temporal models for big multinomial data using the conditional multivariate logit-beta distribution. *Journal of Time Series Analysis*, **40**, 363–382.

Brand, M., Oliver, N. and Pentland, A. (1997) Coupled hidden Markov models for complex action recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 994–999.

Buckeridge, D. L. (2007) Outbreak detection through automated surveillance: A review of the determinants of detection. *Journal of Biomedical Informatics*, **40**, 370–379.

Bull, S. B., Mak, C. and Greenwood, C. M. T. (2002) A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis*, **39**, 57–74.

Cargnoni, C., Müller, P. and West, M. (1997) Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association*, **92**, 640–647.

Chan, T.-C., Hu, T.-H. and Hwang, J.-S. (2015) Daily forecast of dengue fever incidents for urban villages in a city. *International Journal of Health Geographics*, **14**, 9.

Chen, C. W. S., Khamthong, K. and Lee, S. (2019) Markov switching integer-valued generalized auto-regressive conditional heteroscedastic models for dengue counts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**, 963–983.

Chen, S.-C. and Hsieh, M.-H. (2012) Modeling the transmission dynamics of dengue fever: Implications of temperature effects. *Science of the Total Environment*, **431**, 385–391.

Chen, X., Li, Y., Feng, X. and Chang, J. T. (2023) Variational Bayesian analysis of non-homogeneous hidden Markov models with long and ultralong sequences. *The Annals of Applied Statistics*, **17**, 1615–1640.

Chenchula, S., Karunakaran, P., Sharma, S. and Chavan, M. (2022) Current evidence on efficacy of COVID-19 booster dose vaccination against the Omicron variant: A systematic review. *Journal of Medical Virology*, **94**, 2969–2976.

Chib, S. (1996) Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79–97.

Cliff, A. D., Haggett, P. and Ord, J. K. (1987) *Spatial Aspects of Influenza Epidemics*. London: Routledge Kegan & Paul.

Conesa, D., Martínez-Beneito, M., Amorós, R. and López-Quílez, A. (2015) Bayesian hierarchical Poisson models with a hidden Markov structure for the detection of influenza epidemic outbreaks. *Statistical Methods in Medical Research*, **24**, 206–223.

Counotte, M. J., Kim, C. R., Wang, J., Bernstein, K., Deal, C. D., Broutet, N. J. N. and Low, N. (2018) Sexual transmission of Zika virus and other flaviviruses: A living systematic review. *PLoS Medicine*, **15**, e1002611.

Coutinho, F. A. B., Burattinia, M. N., Lopeza, L. F. and Massada, E. (2006) Threshold conditions for a non-autonomous epidemic system describing the population dynamics of dengue. *Bulletin of Mathematical Biology*, **68**, 2263–2282.

Crawford, F. W., Jones, S. A., Cartter, M., Dean, S. G., Warren, J. L., Li, Z. R., Barbieri, J., Campbell, J., Kenney, P., Valleau, T. and Morozova, O. (2022) Impact of close interpersonal contact on COVID-19 incidence: Evidence from 1 year of mobile device data. *Science Advances*, **8**, eabi5499.

Dabney, A. R. and Wakefield, J. C. (2005) Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, **14**, 83–112.

Descloux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., Guillaumot, L., Teurlai, M., Gourinat, A.-C., Benzler, J., Pfannstiel, A., Grangeon, J.-P., Degallier, N. and Lamballerie, X. D. (2012) Climate–based models for understanding and forecasting dengue epidemics. *PLoS Neglected Tropical Diseases*, **6**, e1470.

Diallo, A. O., Diop, A. and Dupuy, J.-F. (2018) Analysis of multinomial counts with joint zero-inflation, with an application to health economics. *Journal of Statistical Planning and Inference*, **194**, 85–105.

Diebold, F., Lee, J. and Weinbach, G. (1994) Regime switching with time–varying transition probabilities. In *Nonstationary Time Series Analysis and Cointegration* (ed. C. Hargreaves), 283–302. Oxford: Oxford University Press.

Douwes-Schultz, D. and Schmidt, A. M. (2022) Zero-state coupled Markov switching count models for spatio-temporal infectious disease spread. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**, 589–612.

Dreassi, E. (2007) Polytomous disease mapping to detect uncommon risk factors for related diseases. *Biometrical Journal*, **49**, 520–529.

Dunbar, M. B.-N. and Held, L. (2020) Endemic-epidemic framework used in COVID-19 modelling : [Discussion on the paper by Nunes, Caetano, Antunes and Dias]. *REVSTAT-Statistical Journal*, **18**, 565–574.

Fernandes, M. V. M., Schmidt, A. M. and Migon, H. S. (2009) Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, **9**, 3–25.

Finkenstadt, B. F. and Grenfell, B. T. (2000) Time series modelling of childhood diseases: A dynamical systems approach. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **49**, 187–205.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C. and Bhatt, S. (2020) Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, **584**, 257–261.

Fokianos, K. and Kedem, B. (2003) Regression theory for categorical time series. *Statistical Science*, **18**, 357–376.

Fokianos, K. and Tjøstheim, D. (2011) Log-linear Poisson autoregression. *Journal of Multivariate Analysis*, **102**, 563–578.

Freitas, L. P., Cruz, O. G., Lowe, R. and Sá Carvalho, M. (2019) Space–time dynamics of a triple epidemic: dengue, chikungunya and Zika clusters in the city of Rio de Janeiro. *Proceedings of the Royal Society B: Biological Sciences*, **286**, 20191867.

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**, 400–424.

Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York: Springer-Verlag.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*. Boca Raton: Chapman and Hall/CRC, 3rd edn.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**, 1360–1383.

Ghosal, S., Lau, T. S., Gaskins, J. and Kong, M. (2020) A hierarchical mixed effect hurdle model for spatiotemporal count data and its application to identifying factors impacting health professional shortages. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **69**, 1121–1144.

Gilmour, S. J. (2010) Identification of hospital catchment areas using clustering: An example from the NHS. *Health Services Research*, **45**, 497–513.

Giorgi, E., Schlüter, D. K. and Diggle, P. J. (2018) Bivariate geostatistical modelling of the relationship between Loa loa prevalence and intensity of infection. *Environmetrics*, **29**, e2447.

Goldfeld, S. M. and Quandt, R. E. (1973) A Markov model for switching regressions. *Journal of Econometrics*, **1**, 3–15.

Google LLC (2022) COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility?hl=en.

Gouvernement du Québec (2022) Measures adopted by orders in council and ministerial orders in the context of the COVID-19 pandemic. https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/measures-orders-in-council-ministerial-orders.

Greene, W. H. (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *SSRN Scholarly Paper ID 1293115*, Social Science Research Network, Rochester, NY.

Grenfell, B. T., Bjørnstad, O. N. and Kappey, J. (2001) Travelling waves and spatial hierarchies in measles epidemics. *Nature*, **414**, 716–723.

Göertz, G. P., Vogels, C. B. F., Geertsema, C., Koenraadt, C. J. M. and Pijlman, G. P. (2017) Mosquito co-infection with Zika and chikungunya virus allows simultaneous transmission without affecting vector competence of Aedes aegypti. *PLoS Neglected Tropical Diseases*, **11**, e0005654.

Hall, D. B. (2000) Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**, 1030–1039.

Hamilton, J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

— (1993) 9 Estimation, inference and forecasting of time series subject to changes in regime. In *Handbook of Statistics*, vol. 11 of *Econometrics*, 231–260. Elsevier.

Harrington, L. C., Scott, T. W., Lerdthusnee, K., Coleman, R. C., Costero, A., Clark, G. G., Jones, J. J., Kitthawee, S., Kittayapong, P., Sithiprasasna, R. et al. (2005) Dispersal of the dengue vector Aedes aegypti within and between rural communities. *The American Journal of Tropical Medicine and Hygiene*, **72**, 209–220.

Heaton, M. J., Banks, D. L., Zou, J., Karr, A. F., Datta, G., Lynch, J. and Vera, F. (2012) A spatio-temporal absorbing state model for disease and syndromic surveillance. *Statistics in Medicine*, **31**, 2123–2136.

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

Hoef, J. M. V. and Jansen, J. K. (2007) Space—time zero-inflated count models of harbor seals. *Environmetrics*, **18**, 697–712.

Hooten, M. B. and Wikle, C. K. (2010) Statistical agent-based models for discrete spatio-temporal systems. *Journal of the American Statistical Association*, **105**, 236–248.

INSPQ (2022) COVID-19 data in Quebec [online]. https://www.inspq.qc.ca/covid-19/donnees.

Jack, E., Lee, D. and Dean, N. (2019) Estimating the changing nature of Scotland's health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, **182**, 1061–1080.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20**, 50–67.

Joe, H. and Zhu, R. (2005) Generalized Poisson distribution: The property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal*, **47**, 219–229.

Kaufmann, S. (2018) Hidden Markov models in time series, with applications in economics. In *Handbook of Mixture Analysis* (eds. S. Frühwirth-Schnatter, G. Celeux and C. P. Robert). New York: Chapman and Hall/CRC.

Kazazian, L., Neto, A. S. L., Sousa, G. S., Nascimento, O. J. d. and Castro, M. C. (2020) Spatiotemporal transmission dynamics of co-circulating dengue, Zika, and chikungunya viruses in Fortaleza, Brazil: 2011–2017. *PLoS Neglected Tropical Diseases*, **14**, e0008760.

Keeling, M. J. and Rohani, P. (2007) *Modeling Infectious Diseases in Humans and Animals*. Princeton: Princeton University Press, illustrated edn.

Keeling, M. J., Woolhouse, M. E., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J. and Grenfell, B. T. (2001) Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, **294**, 813–817.

Knorr-Held, L. and Richardson, S. (2003) A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 169–183.

Koslovsky, M. D. (2023) A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data. *Biometrics*, **79**, 3239–3251.

Laframboise, K. (2020) Quebec coronavirus cases soar past 4,000 as medical equipment shortage looms. *Global News*. URL: https://globalnews.ca/news/6755471/quebec-coronavirus-march-31/.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Langrock, R. and Zucchini, W. (2011) Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, **55**, 715–724.

Lawson, A. B. and Kim, J. (2022) Bayesian space-time SIR modeling of Covid-19 in two US states during the 2020–2021 pandemic. *PLoS One*, **17**, e0278515.

Le Strat, Y. and Carrat, F. (1999) Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, **18**, 3463–3478.

Legault, F. and Blais, M. (2020) Conférence de presse de M. François Legault, premier ministre et Mme Marguerite Blais, ministre responsable des aînés et des proches aidants, Quebec City. URL: http://www.assnat.qc.ca/fr/actualites-salle-presse/conferences-points-presse/ConferencePointPresse-60025.html.

Liboschik, T., Fokianos, K. and Fried, R. (2017) Tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, **82**, 1–51.

Lim, J. T., Dickens, B. S., Haoyang, S., Ching, N. L. and Cook, A. R. (2020) Inference on dengue epidemics with Bayesian regime switching models. *PLoS Computational Biology*, **16**, e1007839.

Loredo-Osti, J. C. and Sutradhar, B. C. (2012) Estimation of regression and dynamic dependence paremeters for non-stationary multinomial time series. *Journal of Time Series Analysis*, **33**, 458–467.

Lowe, R., Gasparrini, A., Meerbeeck, C. J. V., Lippi, C. A., Mahon, R., Trotman, A. R., Rollock, L., Hinds, A. Q. J., Ryan, S. J. and Stewart-Ibarra, A. M. (2018) Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. *PLoS Medicine*, **15**, e1002613.

Lu, G., Zhou, S., Horstick, O., Wang, X., Liu, Y. and Müller, O. (2014) Malaria outbreaks in China (1990–2013): A systematic review. *Malaria Journal*, **13**, 269.

Lu, H.-M., Zeng, D. and Chen, H. (2010) Prospective infectious disease outbreak detection using Markov switching models. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 565–577.

Lytras, T., Gkolfinopoulou, K., Bonovas, S. and Nunes, B. (2019) FluHMM: A simple and flexible Bayesian algorithm for sentinel influenza surveillance and outbreak detection. *Statistical Methods in Medical Research*, **28**, 1826–1840.

Majumder, M., Cohn, E., Fish, D. and Brownstein, J. (2016) Estimating a feasible serial interval range for Zika fever. *Bulletin of the World Health Organization*.

Malyshkina, N. V. and Mannering, F. L. (2010) Zero-state Markov switching count-data models: An empirical assessment. *Accident Analysis & Prevention*, **42**, 122–130.

Marquardt, W. H. (ed.) (2004) *Biology of Disease Vectors*. Burlington, MA: Academic Press, 2nd edn.

Martínez-Beneito, M. A., Conesa, D., López-Quílez, A. and López-Maside, A. (2008) Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, **27**, 4455–4468.

Maslo, C., Friedland, R., Toubkin, M., Laubscher, A., Akaloo, T. and Kama, B. (2022) Characteristics and outcomes of hospitalized patients in South Africa during the COVID-19 Omicron wave compared with previous waves. *JAMA*, **327**, 583–584.

Mercier, A., Obadia, T., Carraretto, D., Velo, E., Gabiane, G., Bino, S., Vazeille, M., Gasperi, G., Dauga, C., Malacrida, A. R., Reiter, P. and Failloux, A.-B. (2022) Impact of temperature on dengue and chikungunya transmission by the mosquito Aedes albopictus. *Scientific Reports*, **12**, 6973.

Mizumoto, K. and Chowell, G. (2020) Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. *Infectious Disease Modelling*, **5**, 264–270.

NCCID (2022) Updates on COVID-19 variants of concern (VOC). https://nccid.ca/covid-19-variants/.

Nishiura, H., Linton, N. M. and Akhmetzhanov, A. R. (2020) Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*, **93**, 284–286.

Nunes, B., Natário, I. and Carvalho, M. L. (2013) Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Statistics in Medicine*, **32**, 2643–2660.

Okano, J. T., Sharp, K., Valdano, E., Palk, L. and Blower, S. (2020) HIV transmission and source–sink dynamics in sub-Saharan Africa. *The Lancet HIV*, **7**, e209–e214.

Olivier, A. (2021) Quebec imposes lockdown for 3 cities as COVID-19 cases rise ahead of Easter long weekend. *Global News*. URL: https://globalnews.ca/news/7731357/quebec-imposes-lockdown-for-3-cities-as-covid-19-cases-rise-ahead-of-easter-long-weekend/.

Otranto, E. (2005) The multi-chain Markov switching model. *Journal of Forecasting*, **24**, 523–537.

Otting, M., Langrock, R., Deutscher, C. and Leos-Barajas, V. (2020) The hot hand in professional darts. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, **183**, 565–580.

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

Pavani, J. and Moraga, P. (2022) A Bayesian joint spatio-temporal model for multiple mosquito-borne diseases. In *New Frontiers in Bayesian Statistics* (eds. R. Argiento,

F. Camerlenghi and S. Paganin), Springer Proceedings in Mathematics & Statistics, 69–77. Cham: Springer International Publishing.

Pawlowski, A., Jansson, M., Sköld, M., Rottenberg, M. E. and Källenius, G. (2012) Tuberculosis and HIV co-infection. *PLoS Pathogens*, **8**, e1002464.

Pereira, S., Turkman, F. and Correia, L. (2018) Spatio-temporal analysis of regional unemployment rates: A comparison of model based approaches. *Revstat - Statistical Journal*, **16**, 515–536.

Phillips, K. L. (1991) A two-country model of stochastic output with changes in regime. *Journal of International Economics*, **31**, 121–142.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.

Pohle, J., Langrock, R., van der Schaar, M., King, R. and Jensen, F. H. (2021) A primer on coupled state-switching models for multiple interacting time series. *Statistical Modelling*, **21**, 264–285.

Poritz, A. (1982) Linear predictive hidden Markov models and the speech signal. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, 1291–1294.

Queiroz, E. R. d. S. and Medronho, R. d. A. (2022) Overlap between dengue, Zika and chikungunya hotspots in the city of Rio de Janeiro. *PLoS One*, **17**, e0273980.

Quick, C., Dey, R. and Lin, X. (2021) Regression models for understanding COVID-19 epidemic dynamics with incomplete data. *Journal of the American Statistical Association*, **116**, 1561–1577.

Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

Rahimi, I., Chen, F. and Gandomi, A. H. (2023) A review on COVID-19 forecasting models. *Neural Computing and Applications*, **35**, 23671–23681.

Rahmanian, V., Bokaie, S., Haghdoost, A. and Barouni, M. (2021) Predicting cutaneous leishmaniasis using SARIMA and Markov switching models in Isfahan, Iran: A time-series study. *Asian Pacific Journal of Tropical Medicine*, **14**, 83.

Rath, T., Carreras, M. and Sebastiani, P. (2003) Automated detection of influenza epidemics with hidden Markov models. In *Advances in Intelligent Data Analysis V. IDA 2003* (eds. M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse and C. Borgelt), vol. 2810 of *Lecture Notes in Computer Science*, 521–532. Berlin, Heidelberg: Springer.

Reich, B. J. and Ghosh, S. K. (2019) *Bayesian Statistical Methods*. Chapman & Hall.

Reich, N. G., Shrestha, S., King, A. A., Rohani, P., Lessler, J., Kalayanarooj, S., Yoon, I.-K., Gibbons, R. V., Burke, D. S. and Cummings, D. A. T. (2013) Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *Journal of The Royal Society Interface*, **10**, 20130414.

Riou, J., Poletto, C. and Boëlle, P.-Y. (2017) A comparative analysis of chikungunya and Zika transmission. *Epidemics*, **19**, 43–52.

Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, **59**, 291–317.

Rodriguez-Morales, A. J., Villamil-Gómez, W. E. and Franco-Paredes, C. (2016) The arboviral burden of disease caused by co-circulation and co-infection of dengue, chikungunya and Zika in the Americas. *Travel Medicine and Infectious Disease*, **14**, 177–179.

Rotejanaprasert, C., Lee, D., Ekapirat, N., Sudathip, P. and Maude, R. J. (2021) Spatiotem-

poral distributed lag modelling of multiple Plasmodium species in a malaria elimination setting. *Statistical Methods in Medical Research*, **30**, 22–34.

Schmidt, A. M., Freitas, L. P., Cruz, O. G. and Carvalho, M. S. (2022) A Poisson-multinomial spatial model for simultaneous outbreaks with application to arboviral diseases. *Statistical Methods in Medical Research*, **31**, 1590–1602.

Schmidt, W.-P., Suzuki, M., Thiem, V. D., White, R. G., Tsuzuki, A., Yoshida, L.-M., Yanai, H., Haque, U., Tho, L. H., Anh, D. D. and Ariyoshi, K. (2011) Population density, water supply, and the risk of dengue fever in vietnam: Cohort study and spatial analysis. *PLOS Medicine*, **8**, e1001082.

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

Scott, S. L. (2002) Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, **97**, 337–351.

Shaby, B. A. and Wells, M. T. (2010) Exploring an adaptive Metropolis algorithm.

Sherlock, C., Xifara, T., Telfer, S. and Begon, M. (2013) A coupled hidden Markov model for disease interactions. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **62**, 609–627.

Sherwin, C. (2020) Life in the red zone: Here's what you can and can't do. *CTV News*. URL: https://montreal.ctvnews.ca/life-in-the-red-zone-here-s-what-you-can-and-can-t-do-1.5125093.

Shingler, B. and Hendry, L. (2022) Overwhelmed by COVID, Quebec hospitals face tough choices in scaling back surgeries. *CBC News*. URL: https://www.cbc.ca/news/canada/montreal/surgeries-covid-quebec-1.6307895.

Smith, D. L., Lucey, B., Waller, L. A., Childs, J. E. and Real, L. A. (2002) Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *Proceedings of the National Academy of Sciences*, **99**, 3668–3672.

Soper, H. E. (1929) The interpretation of periodicity in disease prevalence. *Journal of the Royal Statistical Society*, **92**, 34–73.

Sosa, J., Briz-Redon, A., Flores, M., Abril, M. and Mateu, J. (2023) A spatio-temporal multinomial model of firearm death in Ecuador. *Spatial Statistics*, **54**, 100738.

Souza, R. C. (1982) Forecasting the progress of epidemics by means of a Bayesian-entropy framework. *Environment and Planning A: Economy and Space*, **14**, 49–60.

Spezia, L. (2006) Bayesian analysis of non-homogeneous hidden Markov models. *Journal of Statistical Computation and Simulation*, **76**, 713–725.

Ssentongo, P., Fronterre, C., Geronimo, A., Greybush, S. J., Mbabazi, P. K., Muvawala, J., Nahalamba, S. B., Omadi, P. O., Opar, B. T., Sinnar, S. A., Wang, Y., Whalen, A. J., Held, L., Jewell, C., Muwanguzi, A. J., Greatrex, H., Norton, M. M., Diggle, P. J. and Schiff, S. J. (2021) Pan-African evolution of within- and between-country COVID-19 dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2026664118.

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 795–809.

Stevenson, V. (2021) Quebec shuts down schools, bars, gyms and more as COVID-19 case counts soar. *CBC News*. URL: https://www.cbc.ca/news/canada/montreal/new-closures-covid-19-quebec-1.6292622.

Stoddard, S. T., Forshey, B. M., Morrison, A. C., Paz-Soldan, V. A., Vazquez-Prokopec, G. M., Astete, H., Reiner, R. C., Vilcarromero, S., Elder, J. P., Halsey, E. S., Kochel,

T. J., Kitron, U. and Scott, T. W. (2013) House-to-house human movement drives dengue virus transmission. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 994–999.

Stoddard, S. T., Morrison, A. C., Vazquez-Prokopec, G. M., Paz Soldan, V., Kochel, T. J., Kitron, U., Elder, J. P. and Scott, T. W. (2009) The role of human movement in the transmission of vector-borne pathogens. *PLoS Neglected Tropical Diseases*, **3**, e481.

Stojanović, O., Leugering, J., Pipa, G., Ghozzi, S. and Ullrich, A. (2019) A Bayesian Monte Carlo approach for predicting the spread of infectious diseases. *PLoS One*, **14**, e0225838.

Stoner, O. and Economou, T. (2020) An advanced hidden Markov model for hourly rainfall time series. *Computational Statistics & Data Analysis*, **152**, 107045.

Tang, Z.-Z. and Chen, G. (2019) Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, **20**, 698–713.

Teixeira, M. G., Costa, M. d. C. N., Barreto, F. and Barreto, M. L. (2009) Dengue: twenty-five years since reemergence in Brazil. *Cadernos de Saúde Pública*, **25**, S7–S18.

Tepe, E. and Guldmann, J.-M. (2020) Spatio-temporal multinomial autologistic modeling of land-use change: A parcel-level approach. *Environment and Planning B: Urban Analytics and City Science*, **47**, 473–488.

Tesla, B., Demakovsky, L. R., Mordecai, E. A., Ryan, S. J., Bonds, M. H., Ngonghala, C. N., Brindley, M. A. and Murdock, C. C. (2018) Temperature drives Zika virus transmission: evidence from empirical and mathematical models. *Proceedings of the Royal Society B: Biological Sciences*, **285**, 20180795.

Tibbits, M. M., Groendyke, C., Haran, M. and Liechty, J. C. (2014) Automated factor slice sampling. *Journal of Computational and Graphical Statistics*, **23**, 543–563.

Torabi, M. (2017) Zero-inflated spatio-temporal models for disease mapping. *Biometrical Journal*, **59**, 430–444.

Touloupou, P., Finkenstädt, B. and Spencer, S. E. F. (2020) Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *Journal of Computational and Graphical Statistics*, **29**, 238–249.

Tuite, A. R., Tien, J., Eisenberg, M., Earn, D. J., Ma, J. and Fisman, D. N. (2011) Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Annals of Internal Medicine*, **154**, 593–601.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C. and Andrews, N. (2012) Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **175**, 49–82.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T. and Bodik, R. (2017) Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, **26**, 403–413.

Verelst, F., Willem, L. and Beutels, P. (2016) Behavioural change models for infectious disease transmission: A systematic review (2010–2015). *Journal of the Royal Society Interface*, **13**, 20160820.

Vergne, T., Korennoy, F., Combelles, L., Gogin, A. and Pfeiffer, D. U. (2016) Modelling African swine fever presence and reported abundance in the Russian Federation using national surveillance data from 2007 to 2014. *Spatial and Spatio-temporal Epidemiology*, **19**, 70–77.

Vynnycky, E. (2010) *An Introduction to Infectious Disease Modelling*. New York: Oxford University Press, USA, 1st edn.

Wakefield, J. (2013) *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. New York, NY: Springer.

Wakefield, J., Dong, T. Q. and Minin, V. N. (2019) Spatio-temporal analysis of surveillance data. In *Handbook of Infectious Disease Data Analysis* (eds. L. Held, N. Hens, P. O'Neil and J. Wallinga), 455–475. Boca Raton: Chapman and Hall/CRC.

Walter, K. S., Pepin, K. M., Webb, C. T., Gaff, H. D., Krause, P. J., Pitzer, V. E. and Diuk-Wasser, M. A. (2016) Invasion of two tick-borne diseases across New England: Harnessing human surveillance data to capture underlying ecological invasion processes. *Proceedings of the Royal Society B: Biological Sciences*, **283**, 20160834.

Wang, E., Chiang, S., Haneef, Z., Rao, V., Moss, R. and Vannucci, M. (2023) Bayesian non-homogeneous hidden Markov model with variable selection for investigating drivers of seizure risk cycling. *The Annals of Applied Statistics*, **17**, 333–356.

Wang, P. (2001) Markov zero-inflated Poisson regression models for a time series of counts with excess zeros. *Journal of Applied Statistics*, **28**, 623–632.

Wang, P. and Puterman, M. L. (1999) Markov Poisson regression models for discrete time series. Part 1: Methodology. *Journal of Applied Statistics*, **26**, 855–869.

Wangdi, K., Clements, A. C. A., Du, T. and Nery, S. V. (2018) Spatial and temporal patterns of dengue infections in Timor-Leste, 2005–2013. *Parasites & Vectors*, **11**, 1–9.

Ward, T. and Johnsen, A. (2021) Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLoS One*, **16**, e0257978.

Watkins, R. E., Eagleson, S., Veenendaal, B., Wright, G. and Plant, A. J. (2009) Disease surveillance using a hidden Markov model. *BMC Medical Informatics and Decision Making*, **9**, 1–12.

Williams, J. P., Hermansen, G. H., Strand, H., Clayton, G. and Nygård, H. M. (2024) Bayesian hidden Markov models for latent variable labeling assignments in conflict research: application to the role ceasefires play in conflict dynamics. *Annals of Applied Statistics*. In press.

Wilton, K. (2020) COVID-19 updates March 15: Quebec now has 39 cases, orders further closures. *Montreal Gazette*. URL: https://montrealgazette.com/news/local-news/covid-19-live-updates-quebec-catholics-invited-to-celebrate-mass-online.

Xia, F., Chen, J., Fung, W. K. and Li, H. (2013) A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, **69**, 1053–1063.

Xu, L., Stige, L. C., Chan, K.-S., Zhou, J., Yang, J., Sang, S., Wang, M., Yang, Z., Yan, Z., Jiang, T., Lu, L., Yue, Y., Liu, X., Lin, H., Xu, J., Liu, Q. and Stenseth, N. C. (2017) Climate variation drives dengue dynamics. *Proceedings of the National Academy of Sciences*, **114**, 113–118.

Yang, M., Zamba, G. K. D. and Cavanaugh, J. E. (2013) Markov regression models for count time series with excess zeros: A partial likelihood approach. *Statistical Methodology*, **14**, 26–38.

Yao, Y., Vehtari, A. and Gelman, A. (2022) Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *Journal of Machine Learning Research*, **23**, 1–45.

Young, D. S., Roemmele, E. S. and Shi, X. (2020) Zero-inflated modeling part II: Zero-inflated models for complex data structures. *WIREs Computational Statistics*, **14**, e1540.

Zacher, B. and Czogiel, I. (2022) Supervised learning using routine surveillance data improves outbreak detection of salmonella and campylobacter infections in Germany. *PLoS One*, **17**, e0267510.

Zeger, S. L. and Qaqish, B. (1988) Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, **44**, 1019–1031.

Zeng, Y., Pang, D., Zhao, H. and Wang, T. (2022) A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, **118**, 2356—-2369.

Zou, J., Karr, A. F., Datta, G., Lynch, J. and Grannis, S. (2014) A Bayesian spatio–temporal approach for real–time detection of disease outbreaks: a case study. *BMC Medical Informatics and Decision Making*, **14**, 108.