Bayesian Hierarchical Modeling of COVID-19 Cases and Government Response in the United States

Lincole Jiang, Safiya Sirota, Anja Shahu, Bin Yang, Ting-Hsuan Chang

2023-05-05

Introduction

Since its outbreak on Dec. 1st, 2019, COVID-19 has had a profound impact on the world and caused significant disruptions in public health, economic stability, and social life. In the United States, due to the sporadic nature of public health administration, different states implemented a wide range of containment measures, economic support policies, and public health interventions in face of this unprecedented public health crisis. This project aims to investigate the state-level association of various government response measures, mobility changes in public spaces, and basic demographic landscape of each state with the spread of COVID-19 by using Bayesian hierarchical modeling techniques.

The dataset considered, `covid_working_data.csv`, is a comprehensive dataset multiple sources of information to facilitate the Bayesian hierarchical modeling of COVID-19 cases and government response, providing a good picture of the pandemic landscape in the U.S. in 2020. For each of the 50 U.S. states from the week of Jan. 26th, 2020 to Dec. 27th, 2020, it comprises the starting date of each week (`week_start`); the weekly data of the number of confirmed cases (`weekly_cases`); the weekly infection rate (`infection_rate`, calculated as the number of weekly cases divided by the total population of each state in 2019); the weekly average percentage change in mobility trends for retail and recreational places, parks, and transit stations compared to a baseline period (`retail_and_recreation_percent_change_from_baseline`, `parks_percent_change_from_baseline`, and `transit_stations_percent_change_from_baseline`, respectively); and the weekly average containment health index (`containment_index`, a composite measure of the stringency of COVID containment policies), economic support index (`economic_support_index`, measuring the extent of government financial support in response to COVID), stringency index (`stringency_index`, a composite measure of strictness of government policies including lockdowns, closures, travel restrictions, etc.), and government response index (`government_response_index`, average of the three government response indexes), each scaled from 0 to 100. In addition, state-level demographic information are also incorporated, including that of population (`pop2019`, total population of each state in 2019), land area in square miles (`LandArea`), population density (`population_density`, `pop2019/LandArea`), and the percentage of population aged 65 or older (`Percentage_over_65`). From preliminary exploratory analysis, we note an increasing trend in weekly cases and infection rate for each state over time (Figure 1), a slightly positive association between average weekly cases for each state and population among time-insensitive covariates (Figure 2), a positive correlation between government response indexes and log-infection rate (Figure 3), and less ostensible association between mobility changes and public spaces (Figure 4).

To summarize, the main objective of this project is to investigate how various state-level government response measures, percent mobility changes in public locations, and demographic landscape influence COVID-19 case count in the U.S. for each state during 2020. To this end, we (1) calculated the posterior distribution of model parameters for fixed effects coefficients, population-level variance, and residual variance based on the outlined Bayesian hierarchical model; (2) designed and implemented a Metropolis Hasting algorithm for the outlined Bayesian hierarchical model and monitored the convergence of MCMC chains using diagnostic plots and summary statistics; (3) computed posterior summaries and 95% credible intervals for the effect of government interventions, mobility changes, population density, and elderly percentage on the infection rate; and finally (4) interpreted the results in the context of improving public health outcomes and bettering policy-making process. Ultimately, the insights gained from this study is hoped to help inform future policy decisions and guide effective response strategies in ongoing and future public health emergencies.

Methods

Model

In this project, we use a Bayesian approach to estimate fixed effects of interest through the posterior likelihood of our model parameters. To find posterior estimates for each parameter, we are given some distributional assumptions. First, let's introduce the notation we will use to describe variables and parameters. Let Y_{ij} be the number of new infections in state i during week j. Let α be the intercept of our model. Let β be a vector of fixed effects coefficients for covariates $\mathbf{X_{ij}}$. Let γ be the fixed effect coefficient for the population density, P_{ij}. Let δ be the fixed effect for the percentage of elderly population, E_{ij}. Finally, let u_i be the state-level random effect and ϵ_{ij} be the residual error. Then our distributional assumptions are

$$\begin{aligned}
Y_{ij} &\sim Poisson(\lambda_{ij} n_{ij}) \\
log(\lambda_{ij}) &= \alpha + \beta \mathbf{X_{ij}^T} + \gamma P_{ij} + \delta E_{ij} + u_i + \epsilon_{ij} \\
\alpha,\ \beta,\ \gamma,\ \delta &\sim N(0,1) \\
u_i &\sim N(0, \sigma_u^2) \\
\epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \\
\sigma_\epsilon^2, \sigma_u^2 &\sim HN(0, 100).
\end{aligned}$$

The first assumption in this hierarchical Bayesian framework (that $Y_{ij} \sim P(\lambda_{ij} n_{ij})$) adds an extra layer of distributional assumptions. We attempted to use this information to find posterior estimates for the fixed effects, however we ran into convergence issues in our MCMC algorithm. Therefore, for practicality, we simplified the problem by ignoring this assumption, essentially removing a layer from the hierarchical framework. Although this was ultimately a pragmatic decision in order to produce results we could interpret, we have some justifications for this. First, we are asked in this project to analyze the effect of our variables on the infection rate λ_{ij}, rather than the count Y_{ij}. So our outcome of interest is not Y_{ij}. Second, removing the top layer allows us to have a normal likelihood and priors, leading to posterior distribution that is easy to work with. Lastly, we are confident that, given our assumption, $log(\lambda_{ij})$ can be fit using a linear mixed effects model, particularly because the equation for $log(\lambda_{ij})$ includes a random error term ϵ_{ij}. On the other hand, we are not sure how to deal with this error term in a Poisson mixed effects model, as it is not usually defined in that setting.

Given these justifications, we end up with the following likelihood calculations. First, given the fixed effects and the distributional assumptions on u_i and ϵ_{ij}, we can see that

$$\log(\lambda_{ij}) \sim N(\mu = \alpha + \beta \mathbf{X_{ij}^T} + \gamma P_{ij} + \delta E_{ij}, \sigma^2 = \sigma_u^2 + \sigma_\epsilon^2).$$

Therefore the log-likelihood is

$$l(data|\lambda_{ij}) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(log(\lambda_{ij}) - \mu)^2}{2\sigma^2}.$$

It is straightforward to show from the distributional assumptions that the log-prior distribution is proportional to

$$-\frac{1}{200}(\alpha^2 + sum(\beta^{\mathbf{2}}) + \gamma^2 + \delta^2 + \sigma_u^2 + \sigma_\epsilon^2).$$

.

Finally, we can add together the log-likelihood and log-prior distributions to get the log-posterior:

$$\log(posterior) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(log(\lambda_{ij}) - \mu)^2}{2\sigma^2} - \frac{1}{200}(\alpha^2 + sum(\beta^{\mathbf{2}}) + \gamma^2 + \delta^2 + \sigma_u^2 + \sigma_\epsilon^2),$$

where $\mu = \alpha + \beta \mathbf{X_{ij}^t} + \gamma P_{ij} + \delta E_{ij}$ and $\sigma^2 = \sigma_u^2 + \sigma_\epsilon^2$.

Therefore we are essentially fitting a linear random effects model with a random intercept for state, with log infection rate as our outcome. Ultimately, we are only interested in fixed effects. Along with population density and elderly population, the covariates we include as a part of $\mathbf{X_{ij}}$ are retail and recreation utilization percent change from baseline, parks utilization percent change from baseline, transit stations utilization percent change from baseline, government response index, containment index, economic support index, stringency index, and the week of the year as a continuous variable. Now that our likelihood and model are clear, we will describe our alogirthm for finding posterior estimates.

Algorithm

We use a component-wise Metropolis Hastings (MH) algorithm that updates the p parameters at the kth iteration one at a time. In contrast to the Gibbs sampler, the MH algorithm uses the full joint distribution to generate potential values, allowing us to avoid computing the full conditional posteriors for each of the parameters.

For our algorithm, we first define some notation:

- N_{iter} is the number of iterations that the algorithm was run for.
- $\theta_i^{(k)}$ and $\theta_i^{(k+1)}$ are the current and updated values, respectively, for the ith parameter at the kth iteration.
- $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_p^{(k)})$ is the vector of parameters $\boldsymbol{\theta}$ at the start of the kth iteration.
- $\boldsymbol{\theta}_{-i}^{(k)} = (\theta_1^{(k+1)}, \theta_2^{(k+1)}, \ldots, \theta_{i-1}^{(k+1)}, \theta_{i+1}^{(k)}, \ldots \theta_p^{(k)})$ is the current vector of parameters $\boldsymbol{\theta}$ at the kth iteration, excluding the ith parameter.
- θ_i^* is the proposal value for the ith parameter.
- $Q_i(\theta^*|\theta_i^{(k)})$ is the proposal distribution for the ith parameter. More generally, the proposal distribution can also depend on $\boldsymbol{\theta}_{-i}^{(k)}$, but for our algorithm, our choice of proposal distribution depends only on $\theta_i^{(k)}$.

The algorithm is given as follows:

1. Choose starting values $\boldsymbol{\theta}^{(0)}$.
2. For the kth iteration where $k = 1, \ldots, N_{iter}$:
    1. For the ith parameter where $i = 1, \ldots, p$:
        1. Generate θ_i^* from $Q_i(\theta^*|\theta_i^{(k)})$.

2. Calculate the MH ratio given by:

$$\begin{aligned} r_i^{(k)} &= \frac{\pi(\theta_i^*, \boldsymbol{\theta}_{-i}^{(k)}|\boldsymbol{X})}{\pi(\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)}|\boldsymbol{X})} \frac{Q_i(\theta_i^{(k)}|\theta_i^*)}{Q_i(\theta_i^*|\theta_i^{(k)})} \\ &= \frac{f(\theta_i^*, \boldsymbol{\theta}_{-i}^{(k)}, \boldsymbol{X})}{f(\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)}, \boldsymbol{X})} \frac{Q_i(\theta_i^{(k)}|\theta_i^*)}{Q_i(\theta_i^*|\theta_i^{(k)})} \\ &= \frac{f(\theta_i^*, \boldsymbol{\theta}_{-i}^{(k)}, \boldsymbol{X})}{f(\theta_i^{(k)}, \boldsymbol{\theta}_{-i}^{(k)}, \boldsymbol{X})} \end{aligned} \tag{1}$$

3. Calculate acceptance probability $\alpha_i(\theta_i^{(k)}, \theta_i^*) = \min\left(1, r_i^{(k)}\right)$.
4. Draw $U \sim Unif(0,1)$. If $U < \alpha_i(\theta_i^{(k)}, \theta_i^*)$, then set $\theta_i^{(k+1)} = \theta_i^*$. Otherwise, set $\theta_i^{(k+1)} = \theta_i^{(k)}$.

For each $Q_i(\theta^*|\theta_i^{(k)})$, we use a uniform distribution. More specifically, θ_i^* is generated as $\theta_i^* = \theta_i^{(k)} + Unif(-a_i, a_i)$, where a_i represents the chosen window length for the ith parameter. We tune a_i for each of the parameters by trying different a_i until we find ones that allow us to accept about 30-60% of N_{iter} iterations for a parameter. We keep track of the number of acceptances by computing the number of unique $\theta_i^{(k)}$ across all N_{iter} iterations. Tuning a_i is important because: 1) a_i being too large means the proposed moves will be too large and unlikely to be accepted, taking the chain a long time to sample the entire parameter space (i.e. the chain is sampling a lot of values outside of the support of the posterior distribution); 2) a_i being too small means that the proposed moves will be too small and accepted too often, taking the chain a long time to move around the parameter space.

In equation (1), the second equality holds since $\pi(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{X})}{m(\boldsymbol{X})}$, so the $m(\boldsymbol{X})$'s cancel out. Additionally, the third equality holds since θ_i^* is drawn from $Unif(\theta_i^{(k)} - a_i, \theta_i^{(k)} + a_i)$. Since the uniform distribution is a symmetric distribution, the Q_i's cancel out.

To allow the results to converge more quickly and save computation time, we choose the $\boldsymbol{\theta}^{(0)}$ at the beginning of the algorithm based on results from model fitting using the frequentist approach (under the assumption that the Bayesian and frequentist approaches will yield similar results). More specifically, we fit a linear mixed effect model (LMM) with random intercepts on log(infection rate) for the ith state during the jth week (see Table 1 for results). The LMM is given as follows:

$$log(\lambda_{ij}) = \alpha + \boldsymbol{\beta}\boldsymbol{x}_{ij}^T + \gamma P_{ij} + \delta E_{ij} + u_i + e_{ij}$$

where α is the fixed intercept, $\boldsymbol{\beta}$ is the vector of fixed-effect coefficients associated with the vector of covariates $\boldsymbol{x}_{ij}^T$, γ is the fixed-effect coefficient associated with population density P_{ij}, δ is the fixed-effect coefficient associated with the perentage of elderly population E_{ij}, $u_i \sim N(0, \sigma_u^2)$ is the state-specific random intercept, and $e_{ij} \sim N(0, \sigma_e^2)$ is the residual error term.

Results

Convergence Diagnostics

We present the convergence diagnostics and summaries of the posterior distribution of the MCMC algorithm. Specifically, we focus on the coefficients for the following covariates: effect of government interventions, mobility changes, population density and elderly percentage. We ran the aforementioned MCMC algorithm with 200,000 iterations and first monitor the convergence results of the MCMC samples. Before we formally present the model diagnostics, we note that there are two common issues in MCMC samples. One common problem is that a large portion of the sample is drawn from distributions that are significantly different from the target. Another common issue is that the effective size of the sample is too small. If the above major issues are absent, then the following hypotheses hold: The majority of the observations in the MCMC sample have been drawn from distributions that are very similar to the target distribution and the effective size of the sample is not too small. These hypotheses imply that the empirical distribution of any large chunk of the sample is a good approximation of the target distribution.

To test for this implication, we use the tools of diagnostic plots including trace plot, density plot and autocorrelation function (ACF) plot[1]. A trace plot is a plot of the chain values against iteration number. It can be used to visually inspect the chain for convergence and mixing properties. If the trace plot shows a random walk with no discernible pattern, the chain is likely to be well-mixed and converged. If there are any trends or patterns in the plot, further investigation may be required. A density plot is a plot of the estimated probability density function of the chain values. It can be used to assess the shape and distribution of the chain, and to identify any potential problems such as multimodality or skewness. An auto-correlation plot is a plot of the auto-correlation function of the chain values against lag. It can be used to assess the degree of correlation between values in the chain at different lags, and to identify any potential problems such as long-term dependence or lack of mixing.

We first examine the trace plot and density plot for the coefficients of interest. As shown in figure 5-8, MCMC samples converged well for the coefficient for mobility change and population density. For government response index, we observe a lot of serial correlation between successive draws. The chain is very slow in exploring the sample space. There seems to be few independent observations in our sample. The effective size of our sample is too small. Multimodality is also detected in the density plot. For elderly population percentage, we observe moderate correlation between successive draws and the effective size of the sample is also too small.

In addition, we create ACF plot to identify any pattern of correlation with respect to different values of lags. As shown in figure 9, we observe that for population density and mobility change, the autocorrelation is large at short lags, but then goes to zero pretty quickly. For elderly population percentage, government response, the autocorrelation is large at short lags, but it also dies out very slowly indicating that the effective sample size is small.

Finally, we present a discussion for other diagnostics methods. We could further compute effective sample size based on some exiting packages. We could also create multiple chains with different starting values and then compare and monitor the convergence of different chains using Gelman-Rubin Statistics (R-hat). To further improve the algorithms, we could use multiple-try metropolis[2] to address the dimensionality issues. Hamiltonian (or hybrid) Monte Carlo (HMC)[3] can be implemented to avoid random walks. And finally, we can also consider using a hybrid approach where we can use empirical bayes to estimate some hyperparameters when n is large.

Posterior Summaries

Table 2 shows the posterior means and 95% credible intervals of the parameters for mobility changes, government response, population density, and percentage of elderly population. The posterior mean is calculated by taking the mean of the MCMC chain after burn-in (i.e., last 50,000 states of the chain), and the 95% credible interval is the 2.5% and 97.5% percentiles of the chain after burn-in. Because the Government Response Index measures the overall government response to the pandemic – including containment measures, economic support, and stringency of public health measures – we only present the posterior results of the parameter for Government Response Index (β_4) here. Based on the posterior summeries, changes in mobility trends (compared to baseline) for retail and recreation were associated with a slight decrease in the expected log infection rate, whereas changes in mobility trends for parks and transit stations were associated with an increase in the expected log infection rate. A 1-unit increase in the weekly average Government Response Index was associated with a 4.95 unit decrease in the expected log infection rate, adjusting for the other covariates. A 1% increase in percentage of elderly population was associated with a 12.74 unit decrease in the expected log infection rate, adjusting for the other covariates. The credible interval with respect to population density suggests that population density had no significant impact on the infection rate.

Conclusion

Our posterior results suggest that the overall government response to the pandemic may be crucial for slowing down the spread of COVID-19. Larger mobility changes in retail and recreation may also slow down the spread of COVID-19. However, the temporal order of changes in mobility trends and infection rate is unclear given the available data. A next step for future studies is thus to evaluate the lagged relationship between mobility trends and infection rate. In terms of demographics, higher percentage of elderly population appeared to be associated with lower infection rate, which may be due to other factors (e.g., less interactions among older adults). We therefore suggest further investigation on the causal mechanisms underlying the spread of COVID-19 among the younger population. Finally, population density did not appear to have a significant impact on the infection rate. There have been mixed findings in the literature regarding the importance of population density on the spread of COVID-19. Although a higher population density indicates a higher chance of being in contact with a positive case, larger cities also tend to have more resources to combat the pandemic than rural areas. More in-depth research on these relationships is thus needed to determine the important predictors for the spread of COVID-19.

References

1. Taboga M. Markov chain monte carlo (MCMC) diagnostics. *Lectures on probability theory and mathematical statistics*. Published online 2021. https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo-diagnostics
2. Liu JS, Liang F, Wong WH. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*. 2000;95(449):121-134.
3. Hoffman MD, Gelman A, et al. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J Mach Learn Res*. 2014;15(1):1593-1623.

Tables and Figures

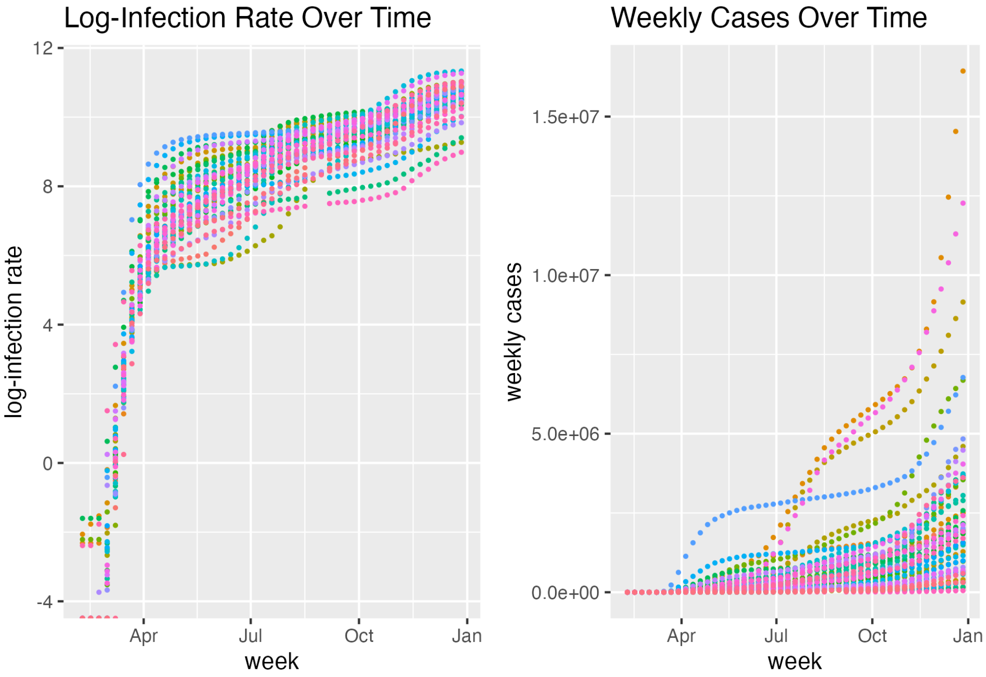

Figure 1: Weekly Cases and Log-Infection Rate Over Time

Table 1: Summary of LMM fit on log(infection rate)

Coefficient	Estimate
(Intercept)	-0.601
Retail/rec. % change	-0.024
Parks % change	0.010
Transit % change	0.001
Gov. response index	-4.956
Containment index	4.456
Economic support index	0.633
Stringency index	-0.027
Week	0.155
Population density	-0.001
% over 65	-17.024
sigma_u	0.821
simga_e	0.783

Table 2: Posterior summaries of the parameters

Parameter	Posterior Mean	95% Credible Interval
beta1	-0.035	(-0.041 -0.028)
beta2	0.006	(0.005 0.007)
beta3	0.024	(0.02 0.029)
beta4	-4.95	(-4.971 -4.927)
delta	-12.741	(-15.066 -10.614)
gamma	0.00018	(-2e-05 0.00038)

beta1: mobility changes for retail and recreation

beta2: mobility changes for parks

beta3: mobility changes for transit stations

beta4: government response index

delta: percentage of elderly population

gamma: population density

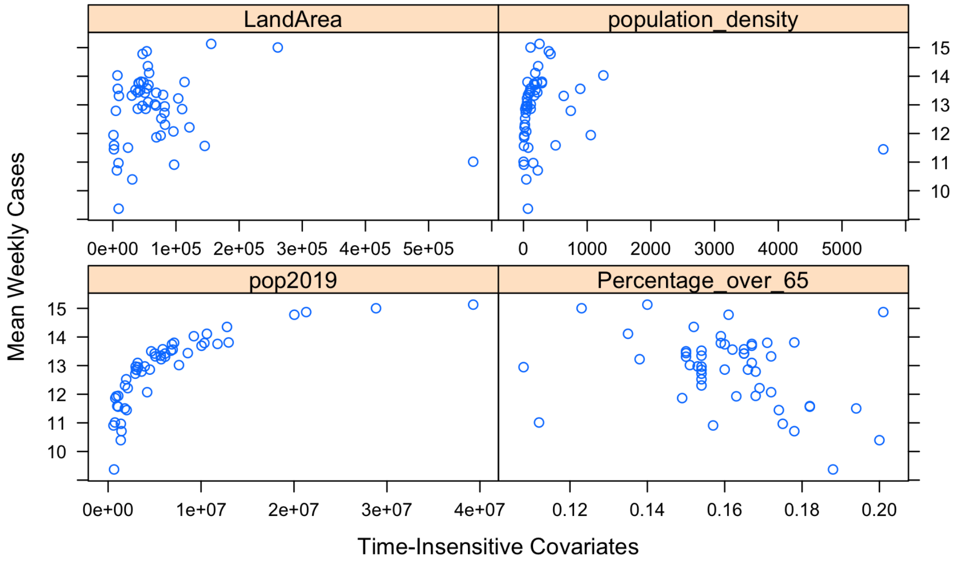

Figure 2: Average Weekly Cases vs Time-Insensitive Covariates

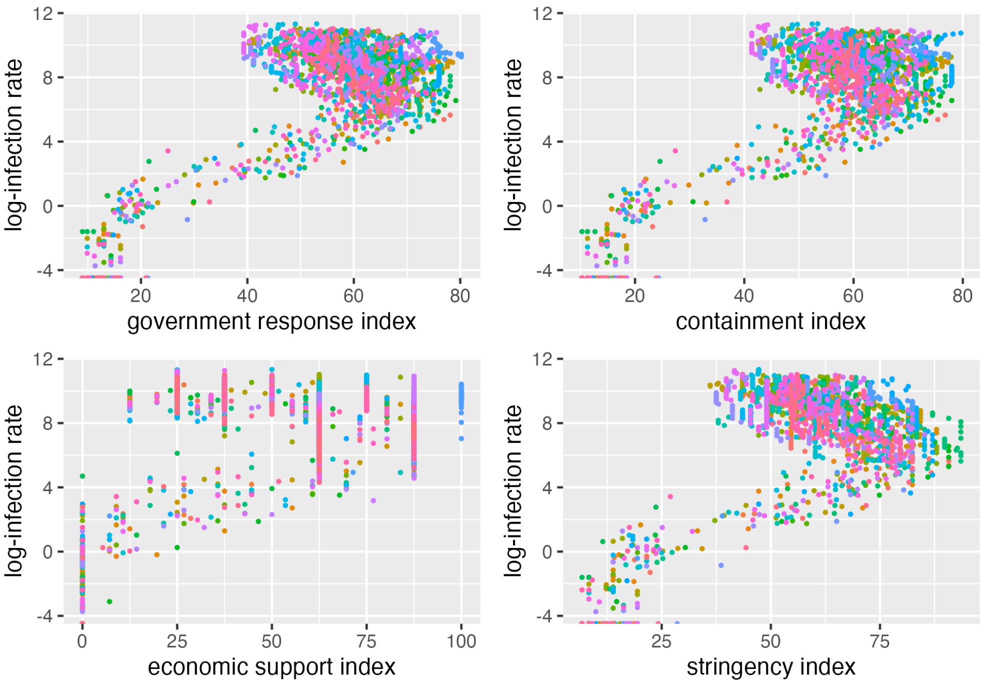

Figure 3: Log-Infection Rate vs Government Response Indexes

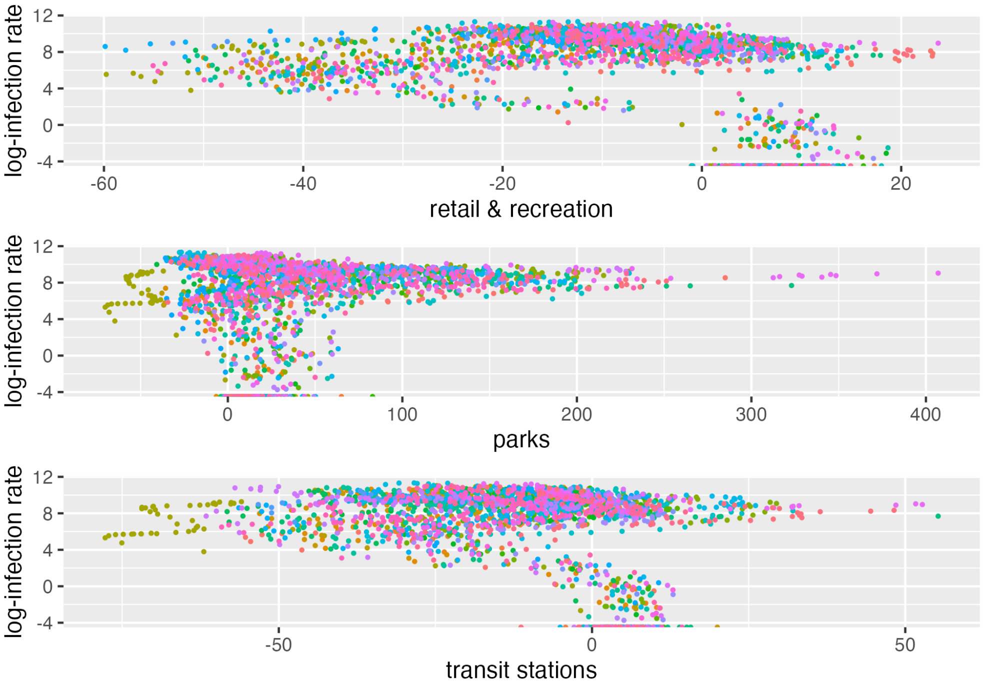

Figure 4: Log-Infection Rate vs Mobility Percentage Changes

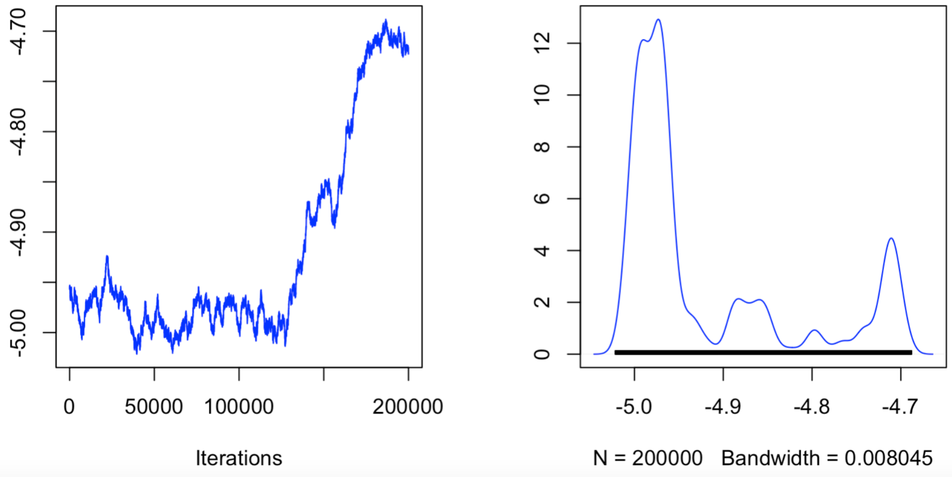

Figure 5: Trace plot and density plot for government response

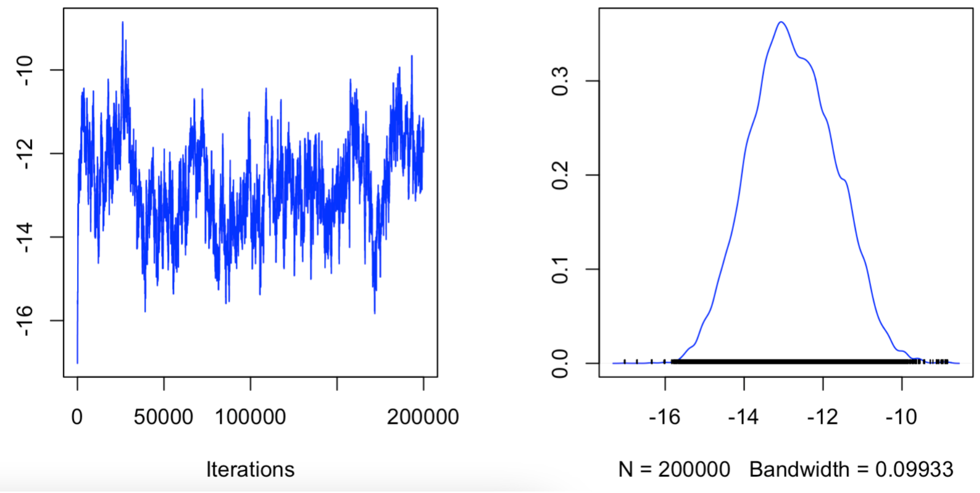

Figure 6: Trace plot and density plot for elderly population percentage

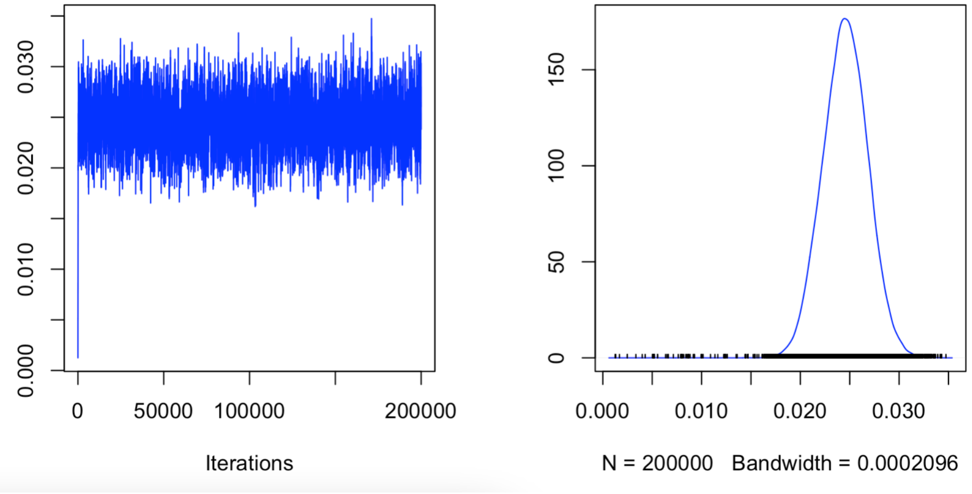

Figure 7: Trace plot and density plot for mobility change

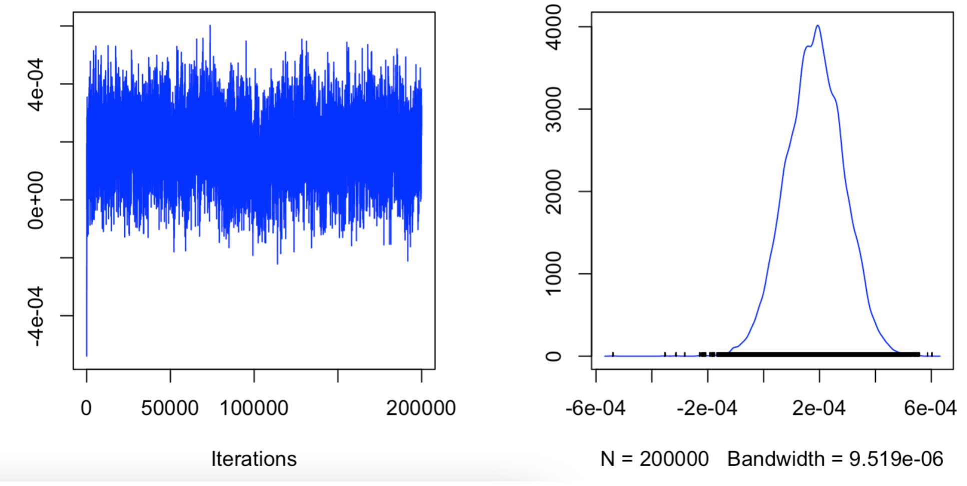

Figure 8: Trace plot and density plot for population density

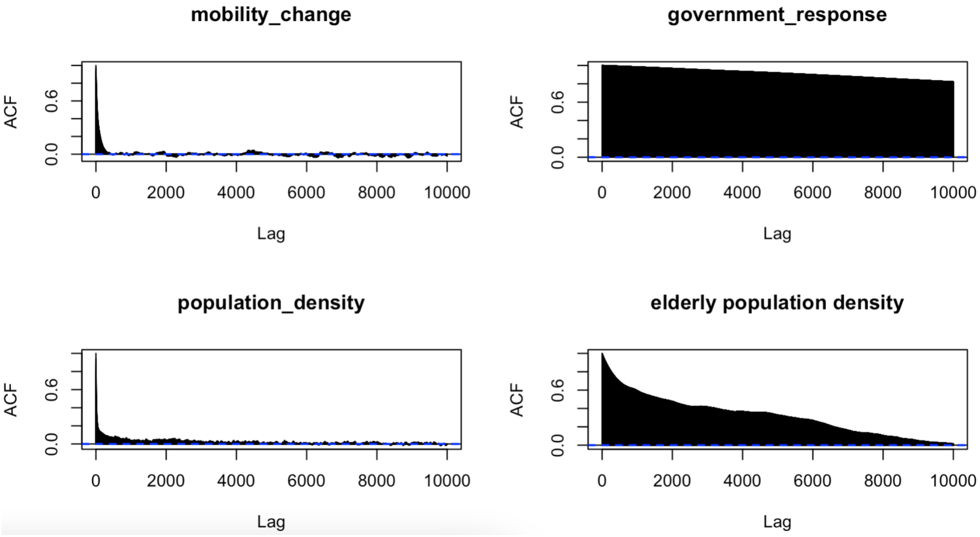

Figure 9: ACF plot with maximum lag at 10000