Incorporating Mobility Data into COVID-19 Forecasting

Dirk Douwes-Schultz and Shuo (Mila) Sun
Supervisors: Dr. Alexandra M. Schmidt and Dr. Erica E. M. Moodie

Department of Epidemiology, Biostatistics and Occupational Health, McGill University

Abstract

Forecasting daily COVID-19 cases is critical for short-term planning of hospital and other public resources. One potentially important piece of information for forecasting COVID-19 cases is cellphone data, which measures the amount of time individuals spend at home. We extend recently developed endemic-epidemic time series models to include a distributed lag model for the effect of cellphone mobility data on the reproductive number, i.e. the number of secondary infections produced per infectious individual in the population, of COVID-19. This analysis is motivated by the hypothesis that less time spent at home will result in more secondary infections and that this effect will be lagged due delays in diagnosis and reporting, and that the use of mobility data will therefore improve forecasting. Endemic-epidemic models are auto-regressive models where the current mean case count is modeled as a weighted average of past case counts multiplied by the reproductive number, plus an endemic component. We introduce a shifted negative binomial weighting scheme for the past counts which is more flexible than previously proposed weighting schemes. Another novel contribution to the recent endemic-epidemic literature is that we preform inference from a Bayesian point of view, which allows all of the uncertainty around unknown parameters to be incorporated into the forecasts. Incorporating a distributed lag model into the specification of the reproductive number allows us to use cellphone data measuring the amount of time spent at home in the current week to help forecast COVID-19 cases in the next week. Bayesian inference and prediction is illustrated in two U.S. counties: King and New York. In both King and New York counties the incorporation of cellphone data through a distributed lag model led to a significantly better fit for the endemic-epidemic model.

Key words : Bayesian prediction; Distributed lag model; Endemic-epidemic model; Reproductive number

1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes the coronavirus disease 2019 (COVID-19), is thought to spread mainly through close contact from person to person by the respiratory route (CDC, 2020a). In March 2020, U.S. states and territories began widely implementing various executive "stay-at-home" orders to mitigate the risk of virus transmission. Understanding the impact of the extent to which people complied with those orders (i.e., population movement was reduced) on the risk of confirmed COVID-19 cases and predicting the

evolution of the pandemic is important for short-term planning and for policy-making to slow the spread of COVID-19.

In this article, we extend the endemic-epidemic (EE) framework to quantify the influence of population movement on the contagiousness of COVID-19 and predicting the dynamics of total number of infected persons in King county and New York county. The EE framework is a class of statistical time series models introduced by Held et al. (2005), where the mean risk is decomposed into endemic and epidemic components. There is a growing literature extending the EE framework (see Paul et al. (2008); Held and Paul (2012); Meyer et al. (2014); Meyer and Held (2017)). We utilize the most recent development from Bracher and Held (2020), which extended the EE framework to include multiple lags of the case counts. As we are modeling daily data, dependence in the case counts likely extends across several days due to the complex dynamics of the disease and delays in testing and reporting (in part due to a high proportion of asymptotic cases).

Endemic-epidemic models are discrete time stochastic susceptible infected recovered (SIR) models (Bauer and Wakefield, 2018). Broadly, in these models, the number of new infections is given by the number of secondary infections produced by recently infected individuals, plus an endemic component which accounts for other contributions to incidence such as imported infections and reporting procedures. The number of secondary infections from recently infected individuals is given by the sum of recently infected individuals, weighted by infectiousness, multiplied by the reproductive number, which is the average number of secondary infections produced by a single infectious individual (Cori et al., 2013).

We focus on estimating the reproductive number for COVID-19 and forecasting new daily cases (i.e. numbers of newly infected people) in the following week, by extending the EE framework through a Bayesian inference procedure, in particular, we use Markov chain Monte Carlo methods to obtain samples from the resultant posterior distribution. Previous applications of the Bracher and Held (2020) EE model have used frequentist inference and prediction, which does not incorporate unknown parameter uncertainty into the forecasting. Additionally, previous weighting schemes for the past case counts have not been very flexible, either only allowing for a decay in the weights or leading to weights very concentrated around their mean. Therefore, we introduce shifted negative binomial weights for the past case counts. In order to incorporate mobility data into the EE forecasts, we model the reproductive number as a function of lagged values of the proportion of individuals staying at home, from cellphone data, using distributed lag models (DLM) (Gasparrini et al., 2010). This is motivated by the hypothesis that less time spent at home will result in more secondary infections and that this effect will be lagged due delays in diagnosis and reporting. The priors for all model parameters are non-informative and we investigate the sensitivity of the results to the priors. In order to examine how the stay-at-home rate can improve the fit, and hence the predictive performance, of EE models, models with different lags are compared, including a first-order EE model and a null model which does not include the stay-at-home rate as a covariate.

The article is structured as follows. Section 2 describes the data sources. Section 3 proposes the model framework. Section 4 presents the model specification and the results, we conclude with a discussion in Section 5.

2 Data description

Two large counties are selected for analysis: King, WA and New York, NY. The USAFacts (https://usafacts.org) COVID-19 time-series data are gathered from the Centers for Disease Control

and Prevention (CDC), state- and local-level public health agencies. County-level data is confirmed by referencing state and local agencies directly. Cases, deaths, and per capita adjustments reflect cumulative totals since January 22, 2020. Confirmed daily cases based on date of report from January to September are collected from USAFacts. The daily percentage of people staying at home data are obtained from the Bureau of Transportation Statistics (BTS), which are estimated by the Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland. The daily travel estimates are from a mobile device data panel from merged multiple data sources that address geographic and temporal sample variation. A multi-level weighting method that employs both device and trip-level weights is applied before travel statistics are computed. County population data are also from USAFacts.

The outcome variable is the daily new confirmed COVID-19 cases. The covariates are the lags of daily cases count and daily stay-at-home rate in each of the two counties. The emergence date of COVID-19 is set as the time of first nonzero cases for each county.

3 Methodology

3.1 The endemic-epidemic model

Let y_t be the COVID-19 case count at day t for $t = p+1, ..., T$ where p is the number of lags fixed, *a priori*. Let $\boldsymbol{y}_{(t-1):(t-p)} = (y_{t-1}, ..., y_{t-p})^T$ be lagged values of the case counts. We model y_t as follows,

$$y_t | \boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}, r \sim \text{NegBin}\left(u_t\left(\boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}\right), r\right),$$

where $\boldsymbol{\theta}$ is a vector of model parameters that we will introduce shortly and r is an overdispersion parameter. We will drop the arguments for u_t from this point to save space. We use the negative binomial 1 parameterization of the negative binomial distribution (Greene, 2008), i.e.

$$p(y_t | \boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}, r) = \binom{y_t + u_t r - 1}{y_t} \left(\frac{r}{r+1}\right)^{u_t r} \left(\frac{1}{r+1}\right)^{y_t} \quad \text{for } y_t = 0, 1, 2, ... \tag{1}$$

In this case, the mean and variance are $E(y_t | \boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}, r) = u_t$ and $V(y_t | \boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}, r) = u_t(1 + r^{-1})$, respectively. In the negative binomial 1 parameterization, the variance increases linearly with the mean instead of quadratically as in the standard negative binomial parameterization, which can lead to unrealistic uncertainty at high counts. We use an endemic-epidemic specification for u_t (Bracher and Held, 2020),

$$u_t = \nu_t + \phi_t \sum_{d=1}^{p} \lfloor \omega_d \rfloor y_{t-d}, \tag{2}$$

where the weights $\lfloor \omega_d \rfloor = \omega_d / \sum_{c=1}^{p} \omega_c$ are normalized and restricted to be positive. The weights ω_d represent the serial interval distribution of the disease, i.e. the distribution of the time between successive cases. Bracher and Held (2020) considered geometric and shifted Poisson weights. However geometric weights only permit a decay with the lags and shifted Poisson weights are not very flexible as they are too concentrated around their mean. Therefore, we propose the use of shifted

negative binomial weights,

$$\omega_d = \frac{\Gamma(d-1+q)}{(d-1)!\Gamma(q)}(1-\kappa)^q\kappa^{d-1}, \quad d = 1,...,p, \tag{3}$$

where $0 < \kappa < 1$ and $q > 0$ are parameters to be estimated. Shifted negative binomial weights allow the serial interval distribution to have a mode at lags greater than 1 and are more flexible than shifted Poisson weights.

In (2), v_t is the endemic component, which represents infection not due to local sources (e.g. imported infections) and can also account for difference in reporting across time. Following Bracher and Held (2020), we model v_t as a log linear function of fixed effects, $\boldsymbol{\zeta}$, and temporal covariates, $\boldsymbol{z}_t$, i.e. $\log v_t = \boldsymbol{z}_t^T\boldsymbol{\zeta}$ (see (7)). The advantage of the EE formulation is that it can be derived as a discrete SIR model (Bauer and Wakefield, 2018) and, therefore, provides a means of incorporating the stay-at-home rates into the forecast in an epidemiologically sound manner. This is done through modeling the reproduction number of the disease, ϕ_t in (2), as a lagged function of the stay-at-home rate.

3.2 Local reproductive number

The parameter $\phi_t > 0$, which is a multiplier on the weighted previous daily cases, represents the instant reproductive number of the disease at time t, or the average number of susceptible individuals that will be infected by an infectious individual if conditions remain the same as at time t (Cori et al., 2013). We assume that the reproductive number ϕ_t is affected by lagged versions of the stay-at-home rate. The rationale here is that population mobility can increase human interaction, which raises the risk of COVID-19 transmission. Additionally, the effect of changes in population mobility on reported case counts should be lagged, as there will be a delay between infection, diagnosis and reporting. By modelling ϕ_t as a function of the stay-at-home rate in a lagged manner, the model implicitly assumes an interaction effect between the lags and previous cases on current case count. In other words, we assume the population mobility is an effect modifier, which would change the magnitude of the forecast power of previous daily cases on current time case count.

We propose to use the distributed lag model to characterize ϕ_t. This method allows the effect of stay-at-home rate to be distributed over a specific period of time, with certain coefficients describing each lag's contribution. However, due to high correlation between those lags in adjacent days, simply using a linear combination of those lags would result in collinearity which leads the precision of estimation to be quite poor (Zanobetti et al., 2002; Gasparrini et al., 2010). We extend the EE model from Bracher and Held (2020) by adding some constraints to the distributed lags, such as imposing a smooth curve using polynomial functions or splines, to gain more precision.

Let s_t denote the stay-at-home rate at day t. We model the reproduction number ϕ_t using a distributed lag model with fixed L lags, in matrix notation we have

$$\log(\phi_t) = \alpha + \boldsymbol{s}_{t\cdot}^\top \boldsymbol{C}\boldsymbol{\eta}, \tag{4}$$

where he fixed $L \geq 1$, $\boldsymbol{s}_{t\cdot} = [s_{t-l}, s_{t-(l+1)}, ..., s_{t-(l+L-1)}]^\top$ for a fixed integer $l \in \{0,1,2,...\}$. Furthermore, $\boldsymbol{\eta}$ is the vector of unknown parameters with length v_l, $\boldsymbol{C}$ is a $L \times v_l$ matrix of basis variables derived from the application of a specific function to the lag vector $\boldsymbol{s}_{t\cdot}$. For example, if $\boldsymbol{C} = \boldsymbol{1}$ (i.e., a vector of ones), (4) is the moving average model; if $\boldsymbol{C}$ is a $L \times L$ identity matrix, it becomes a simple

linear model. $\boldsymbol{C}$ could also be defined as a polynomial or splines function to describe a non-linear curve along lags. The effects of each lag $\boldsymbol{\beta}$ is represented as $\boldsymbol{\beta} = \boldsymbol{C\eta}$. It can be seen as adding constraint to the shape of the distributed lag effects $\boldsymbol{\beta}$ (Gasparrini et al., 2010). Note that, from (4), $\phi_t > 0$ and therefore u_t in (2) is always positive.

3.3 Inference procedure

Let $\boldsymbol{\theta}$ represent the vector of all model parameters specifying u_t (i.e. $\boldsymbol{\theta} = \left(\kappa, q, \boldsymbol{\zeta}, \boldsymbol{\beta}\right)^T$) and let $\boldsymbol{y} = (y_{p+1}, ..., y_T)^T$ be the vector of case counts starting after the first lag p. The likelihood function is given by,

$$p(\boldsymbol{y}|\boldsymbol{\theta}, r) = \prod_{t=p+1}^{T} p(y_t|\boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}, r), \tag{5}$$

where $p(y_t|\boldsymbol{y}_{(t-1):(t-p)}, \boldsymbol{\theta}, r)$ is given by (1). As there are no latent variables that need to be marginalized from the likelihood, maximum likelihood estimation would be straightforward. However, we take a Bayesian approach as all uncertainty in the estimation of the parameters is accounted for in the forecasts. In contrast, a frequentist approach relies on plug-in forecasts which underestimate parameter uncertainty (Bracher and Held, 2020).

The posterior distribution of $(\boldsymbol{\theta}, r)^T$, $p(\boldsymbol{\theta}, r|\boldsymbol{y})$, is proportional to $p(\boldsymbol{y}|\boldsymbol{\theta}, r)p(\boldsymbol{\theta}, r)$ where $p(\boldsymbol{\theta}, r)$ is the prior distribution of $(\boldsymbol{\theta}, r)^T$. We used independent uninformative priors for all parameters. More specifically, all unbounded parameters were assigned normal priors with mean 0 and variance 100, q was assigned a gamma prior with shape and scale equal to 0.1, κ was assigned a uniform prior from 0 to 1 as it is a probability parameter, and r was assigned a uniform prior from 0 to 50. We investigated the sensitivity of the model fitting to the priors for r and q, and found the same results held under several different priors.

As the posterior $p(\boldsymbol{\theta}, r|\boldsymbol{y})$ has no closed form, Markov chain Monte Carlo (MCMC) methods were used to obtain samples from the posterior distribution. More specifically, we used the R package JAGS which utilizes a Gibbs sampler with some steps of the slice sampling algorithm (Neal, 2003) to draw from the joint posterior (Plummer, 2003).

3.4 Temporal predictions

From a Bayesian point of view, predictions for future instants in time are obtained through the posterior predictive distribution, which naturally accounts for the uncertainty in the estimation of the parameter vector $\boldsymbol{\theta}$. Here the goal is to obtain K-step ahead forecasts of the number of cases. In this case, the posterior predictive posterior distribution, $p(y_{T+K}|\boldsymbol{y})$, is given by,

$$\begin{aligned} p(y_{T+K}|\boldsymbol{y}) = \int & p(y_{T+K}|y_{T+K-1}, ..., y_{T+K-p}, \boldsymbol{\theta}, r) \times p(y_{T+K-1}|y_{T+K-2}, ..., y_{T+K-p-1}, \boldsymbol{\theta}, r) \\ & \times \cdots p(y_{T+1}|y_T, ..., y_{T+1-p}, \boldsymbol{\theta}, r) p(\boldsymbol{\theta}, r|\boldsymbol{y})\, dy_{T+k-1} \cdots dy_{T+1} d\boldsymbol{\theta} dr. \end{aligned}$$

The above integral can be approximated through Monte Carlo integration, that is,

$$p(y_{T+K}|\boldsymbol{y}) \approx \frac{1}{Q-M} \sum_{m=M+1}^{Q} p(y_{T+K}|y_{T+K-1}^{[m]}, ..., y_{T+K-p}^{[m]}, \boldsymbol{\theta}^{[m]}, r^{[m]}), \tag{6}$$

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

In practice we use a simulation procedure to draw from the posterior predictive distributions. The sampling algorithm is described in Algorithm 1. For $k = 1, ..., K$, Algorithm 1 will obtain realizations from the resultant posterior predictive distribution $p(y_{T+k}|\boldsymbol{y})$. The prediction procedure requires $\boldsymbol{s}_{T+K} = [s_{T+K-l}, s_{T+K-l-1}, ..., s_{T+K-l+L-1}]^\top$, so that we must have $K = l$. In our applications we considered $l = 7$, so that we predict cases in the next week using cellphone data from the current week.

Algorithm 1: Posterior Predictive Simulation

for m *in* $M+1 : Q$ **do**
  for k *in* $1 : K$ **do**
    Draw $y_{T+k}^{[m]}$ from $p(y_{T+k}|y_{T+k-1}^{[m]}, ..., y_{T+k-p}^{[m]}, \boldsymbol{\theta}^{[m]}, r^{[m]})$.
  end
end

4 Data Analysis

4.1 Model Specification and Fitting

In the observed data, there are a small number of days with zero daily cases after the emergence date, which we posit is likely due to under- or delayed reporting. We distributed those zeros by averaging the most recent five days ahead. Two hundred days (i.e. 28.5 weeks) after emergence date are used for model fitting, we forecast the reproductive number and new daily cases one week ahead.

Examining patterns in incidence by day of the week, we notice some days, like Wednesday, often report fewer cases then the surrounding days which could be due to reporting protocols for example, Wednesday reporting may represent Sunday testing. Therefore, we assume v_t is a function of the day of the week,

$$\log v_t = \texttt{wkdayeff}_{\texttt{wkday}_t}, \tag{7}$$

where $\texttt{wkdayeff}_c$ is the fixed effect of weekday c and $\texttt{wkday}_t$ is the day of the week at time t.

Estimation is performed using JAGS software. For fitting the models in JAGS, we ran 100,000 MCMC iterations on 3 chains (each started from different points in the parameter space), each with a burn-in of 10,000 iterations. Convergence was checked using the Gelman Rubin statistic (upper bound <1.05 for all estimated parameters) and the minimum effective sample size (>1000), as well as by examining trace plots (Plummer et al., 2006).

Model (2) represents a general form with many possible specifications. Herein, we fit and compare several models for King and New York counties using the deviance information criteria (DIC) (Spiegelhalter et al., 2002). The model with the lowest DIC is considered to have the best fit and generally a difference of 5 or more in the DIC is considered meaningful. First, we compared $p = 7, 14$ and 21, and found no significant difference in the DIC for either county. Therefore, we used $p = 7$ for both counties as it allows the most data to be incorporated into the fitting. Table 1 gives some results

of the model comparisons, specifically for determining the optimal lags in the DLM of the stay-at-home rate and for determining the optimal weighting scheme for the serial interval distribution. We compared shifted negative binomial and shifted Poisson weights for the serial interval distribution and found the shifted negative binomial weights significantly improved the fit in both counties. We use natural cubic spline for the basis matrix for stay-at-home rate in (4). For each county, we concentrate on four variants of Eq. (4) which differed in $\boldsymbol{s}_{t\cdot} = [s_{t-l}, s_{t-l-1}, \ldots, s_{t-l-L+1}]^\top$: $l = 0, L = 1$; $l = 7, L = 1$; $l = 7, L = 8$; $l = 7, L = 15$; and compare these to a null model excluding stay-at-home rate $\boldsymbol{s}_{t\cdot}$ as a covariate. For both counties we found an optimal starting lag (l in (4)) of 1 week for the stay-at-home rate in the DLM model (4), and an optimal ending lag (L in (4)) of 2 weeks for King county and 3 weeks for New York county. This means the stay-at-home rate between 1 and 2 or 3 weeks in the past is the most relevant for determining current incidence.

Table 1: DIC of the fitted endemic-epidemic models with different stay-at-home lag and weighting settings for each county. The best model with lowest DIC for each county is underlined.

County	Weighting scheme	Lags of stay-at-home rate at time t ($\boldsymbol{s}_{t\cdot}$)				
		Without covariate	s_t	s_{t-7}	$[s_{t-7}, \ldots, s_{t-14}]^\top$	$[s_{t-7}, \ldots, s_{t-21}]^\top$
King	shifted Poisson				1956	
	shifted NB	1946	1949	1948	<u>1938</u>	1944
New York	shifted Poisson					2099
	shifted NB	2103	2090	2090	2091	<u>2085</u>

4.2 Model Results

Table 2 gives the estimates for the best fitting model in each county (underlined model in Table 1), where we exclude spline coefficients as they are not informative (see Figure 2 instead). The overdispersion parameter is twice as large in New York county compared to King county. Cases initially increased very quickly in New York county which led to a large amount of overdispersion in the cases. The weights for past cases are nearly the same in both counties (see Figure 1 as well). The weekday effects are for the most part not significant. As we are considering cases from the previous 7 days, the autoregressive component of the model may be accounting for differences in reporting between weekdays instead, such that the effects reported in the table account only for residual variation due to weekday not already taken into account.

Figure 1 plots the posterior weights ω_d in (3) for the optimal EE model with shifted NB weights for King and New York county, including 95% posterior credible intervals for the estimated weights. Similar monotonic patterns of decaying curves are observed for the two counties. Longer lags show lower weights, suggesting that the farther from the current time, the less the forecasting power of previous daily cases. The 95% posterior credible intervals show a non-monotonic pattern where the first lag has the largest posterior credible interval and the narrowest interval appears around in the middle of the lags. The serial interval distribution of COVID-19 is thought to have a mode at 3-5 days (Nishiura et al., 2020). Our weighting scheme does allow for this but a constant decay in the weights was estimated instead. An explanation is that we are viewing the results of a complex process that not only involves new infections but also variable delays in diagnosis and changing reporting procedures.

Table 2: Posterior means and 95% posterior credible intervals (in brackets) of the parameters (excluding spline coefficients, see Figure 2) from the best fitting models in King and New York counties.

Variable	King	New York
r in (1)	0.08 [0.06, 0.10]	0.04 [0.03, 0.04]
κ in (3)	0.84 [0.51, 0.99]	0.81 [0.43, 0.99]
q in (3)	0.90 [0.42, 2.23]	0.92 [0.43, 2.30]
$\texttt{wkdayeff}_c$ in (7)		
$c = 1$	-5.69 [-21.46, 2.42]	-6.41 [-21.90, 2.03]
$c = 2$	-6.86 [-21.75, 1.68]	-5.81 [-21.35, 2.46]
$c = 3$	-6.52 [-21.87, 2.06]	-3.41 [-19.77, 3.27]
$c = 4$	2.65 [-5.05, 3.75]	-1.48 [-18.63, 3.50]
$c = 5$	-6.29 [-21.55, 2.24]	-4.51 [-20.63, 3.17]
$c = 6$	3.39 [2.70, 3.86]	-3.42 [-19.95, 3.35]
$c = 7$	-3.90 [-20.20, 3.02]	-7.03 [-22.15, 1.58]

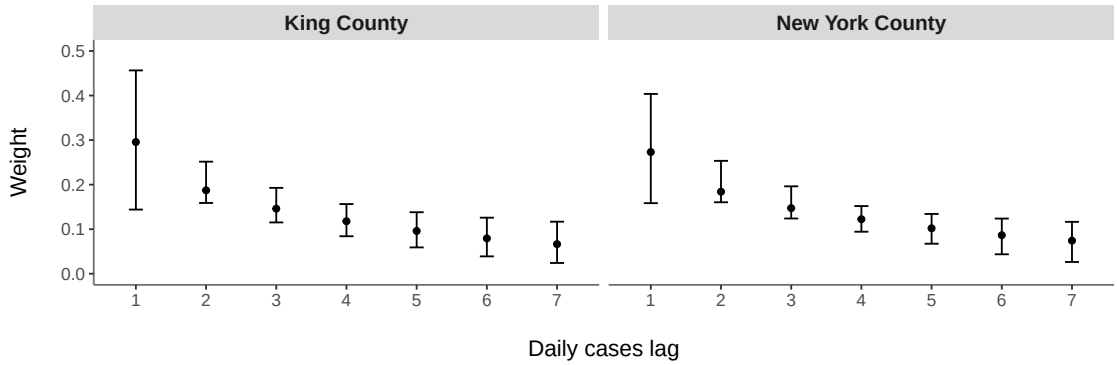

Figure 1: Plot of posterior weights ω_d in (3) with 95% posterior credible intervals from EE models with optimal stay-at-home rate lags.

Figure 2 shows, for each county, the estimated reproductive number ratio for stay-at-home rate at optimal lags (i.e. 7-14 for King county and 7-21 for New York county). The overall lag effects of stay-at-home rate on daily cases are negative, while specific lag effects tend to be different. This supports the assumption that more travel outside the home would increase the risk of COVID-19 infection. The relationship between the stay-at-home rate and the reproductive number, ϕ_t, seems to change along lags, and the shapes of the reproductive number ratio relationship tend to be different between the two counties. Narrower credible intervals are observed with more lags included. According to the CDC, the incubation period for COVID-19 is thought to extend to 14 days, with a median time of 4-5 days from exposure to symptoms onset (CDC, 2020b). Also considering the delays of diagnosis and test reporting, it is reasonable to assume that ϕ_t is affected by lags of stay-at-home rate. Figure 2 confirms these delayed effects: for King county, the 95% credible interval of lag 14 doesn't contain 1, while for New York county, 1 is neither contained in the 95% credible intervals of lag 7 nor of lag 21.

Figure 3 shows the estimated and 7-day predicted reproductive number series (i.e., ϕ_t in (4)) for the optimal EE models. The average ϕ_t is around 0.93, ranging from 0.66 to 1.47 for King county, and 1.06 ranging from 0.72 to 1.95 for New York county. Though with different scales,

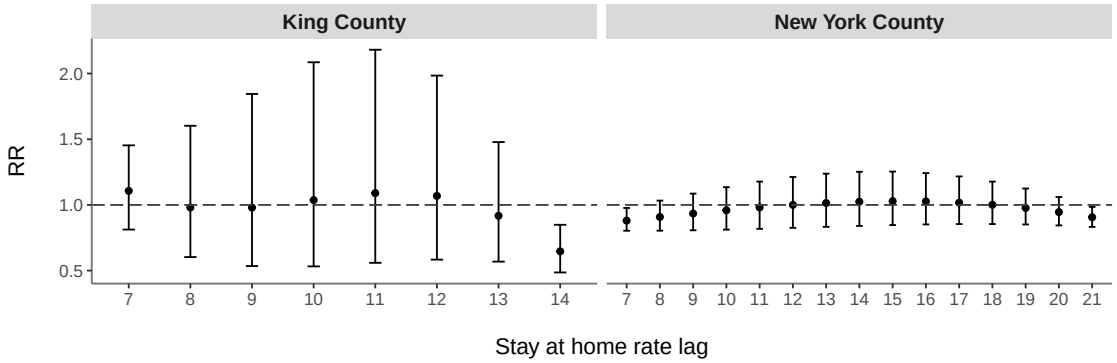

Figure 2: Plot of reproductive number ratio (RR) for every 0.1-unit increase in stay-at-home rate (i.e., $0.1 \times \boldsymbol{\beta} = 0.1 \times \boldsymbol{C\eta}$ in (4)) with 95% posterior credible intervals.

the two series for King and New York display a similar trend: start with a large value in March, then decrease in early April, and finally become relatively stable around 1 until September. The largest ϕ_t appears around mid-March, the time when daily new cases increase steeply at the very beginning of COVID-19 emergence (see Figure 4). The small ϕ_t value in early April is likely due to the statewide community mitigation activities, including closing non-essential business and issuance of orders encouraging (if not mandating) residents to stay at home. Then ϕ_t begins to increase again, it's highly because of COVID-19 fatigue leading to less discipline and less compliance.

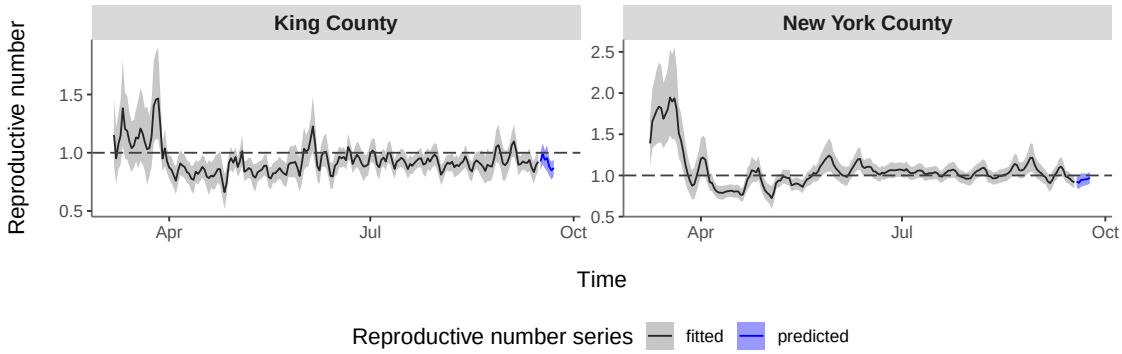

Figure 3: Reproductive number series ϕ_t in (4) with 95% posterior credible intervals from EE models with optimal stay-at-home rate lags.

Figure 4 examines the fitted versus the observed count, and the predicted versus the observed count for the following 7 day period, along with 95% credible intervals using the distributed observed daily cases. The figure suggests that the EE model captures the daily trend well. The good fit is even more pronounced for the one week ahead forecasts.

5 Discussion

We extended endemic-epidemic time series models to include a distributed lag model in the specification of the reproductive number, in order to incorporate lagged effects of cellphone mobility data into the forecasting of daily COVID-19 cases. This method is based on sound epidemiological theory about the effect of mobility on the generation of secondary cases: less time spent at home should

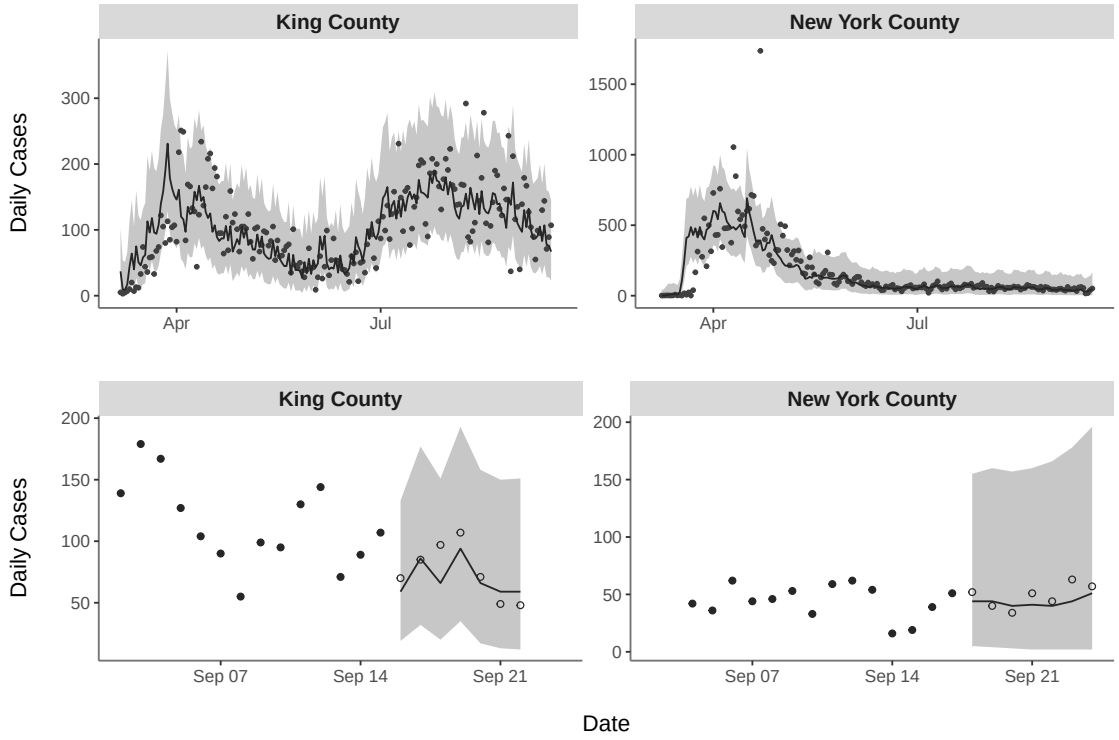

Figure 4: Daily cases series from EE models with optimal stay-at-home rate lags. Top row: Fitted values with 95% posterior credible intervals. Bottom row: Predicted values with 95% posterior credible intervals. Dots denote the distributed observed cases, the solid dots are used for modeling.

increase the number of secondary cases produced by the current infectious population and this effect should be lagged due to delays between infection and diagnosis. The incorporation of the distributed lag model led to significantly better fits in New York and King counties, with stay-at-home rate at lags of 1 to 2-3 weeks to be most predictive of current incidence. We also introduced a novel weighting scheme for the serial interval distribution, shifted negative binomial weights. Shifted negative binomial weights are more flexible than weights proposed in Bracher and Held (2020). Another significant contribution we make is the use of Bayesian inference and prediction, which has the advantage that all unknown parameter uncertainty is incorporated into the forecasts.

Despite their advantages in forecasting, Bayesian models are often avoided due to their perceived long computational times. However, we found our models converged and produced forecasts in 5-10 minutes, so that our proposed method can readily be used for daily online prediction of new COVID-19 cases. Our model can help guide policymakers by quantifying likely spikes in cases due to increased mobility that is picked up by cell phone data.

In our analysis, we observed consistency in the weights across counties. In future work, this similarity could be exploited in a larger hierarchical model to combine data from multiple counties into a single model to further improve prediction.

References

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*, 67(5):1379–1398.

Bracher, J. and Held, L. (2020). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *arXiv:1901.03090 [stat]*.

CDC (2020a). *How COVID-19 Spreads*. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html.

CDC (2020b). *Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19)*. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512.

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, 99(3):585–590.

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

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

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

Meyer, S., Held, L., et al. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612–1639.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–767.

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.

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

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11. Number: 1.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4):583–639.

Zanobetti, A., Schwartz, J., Samoli, E., Gryparis, A., Touloumi, G., Atkinson, R., Le Tertre, A., Bobros, J., Celko, M., Goren, A., et al. (2002). The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement. *Epidemiology*, 13(1):87–93.