


# Under-reporting of COVID-19 in the Northern Health Authority region of British Columbia

Matthew R. P. PARKER<sup>1\*</sup> , Yangming LI<sup>2</sup>, Lloyd T. ELLIOTT<sup>1</sup>, Junling MA<sup>2</sup>, and Laura L. E. COWEN<sup>2</sup>

<sup>1</sup>Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

<sup>2</sup>Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada

*Key words and phrases:* COVID-19; disease analytics; MCMC;  $N$ -mixtures; under-reporting.

*MSC 2020:* Primary 62P10.

**Abstract:** Asymptomatic and pauci-symptomatic presentations of COVID-19 along with restrictive testing protocols result in undetected COVID-19 cases. Estimating undetected cases is crucial to understanding the true severity of the outbreak. We introduce a new hierarchical disease dynamics model based on the  $N$ -mixtures hidden population framework. The new models make use of three sets of disease count data per region: reported cases, recoveries and deaths. Treating the first two as under-counted through binomial thinning, we model the true population state at each time point by partitioning the diseased population into the active, recovered and died categories. Both domestic spread and imported cases are considered. These models are applied to estimate the level of under-reporting of COVID-19 in the Northern Health Authority region of British Columbia, Canada, during 30 weeks of the provincial recovery plan. Parameter covariates are easily implemented and used to improve model estimates. We compare two distinct methods of model-fitting for this case study: (1) maximum likelihood estimation, and (2) Bayesian Markov chain Monte Carlo. The two methods agreed exactly in their estimates of under-reporting rate. When accounting for changes in weekly testing volumes, we found under-reporting rates varying from 60.2% to 84.2%. *The Canadian Journal of Statistics* 49: 1018–1038; 2021 © 2021 Statistical Society of Canada

**Résumé:** Le recours à des protocoles de tests restrictifs et l'existence de formes asymptomatiques et paucisymptomatiques de la COVID-19 contribuent à la non détection de cas COVID-19. Pour comprendre la véritable gravité de l'épidémie, il est primordial d'estimer correctement le nombre de cas non détectés. À cette fin, les auteurs de ce travail proposent un nouveau modèle hiérarchique des dynamiques de la maladie basé sur l'approche de  $N$ -mélanges de population cachée. Ces modèles utilisent trois types de données régionales, à savoir, les nombres de cas déclarés, guéris et décédés. En faisant appel à l'amincissement binomial (binomial thinning) et en traitant les nombres de cas déclarés et guéris comme étant sous-évalués, les auteurs proposent une modélisation de l'état réel de l'épidémie basée sur une partition de la population malade en trois catégories : cas actifs, cas guéris et cas décédés. Cette partition tient compte des cas de propagation intérieure et des cas importés. Les auteurs ont utilisé les données recueillies durant les trente semaines du plan de rétablissement provincial de la région de l'Autorité sanitaire du Nord de la Colombie-Britannique, Canada pour illustrer leur approche et estimer le niveau de sous-déclaration COVID-19 associé. Des covariables peuvent être facilement incorporées au modèle proposé et améliorer la qualité des estimations. Deux méthodes d'ajustement sont retenues: (1) l'estimation par maximum de vraisemblance, et (2) la méthode de Monte Carlo par chaînes de Markov. Les estimations du taux de sous-déclaration obtenues par ces deux méthodes concordent exactement et varient entre 60,2% et 84,2% après ajustement des variations des volumes de tests hebdomadaires. *La revue canadienne de statistique* 49: 1018–1038; 2021 © 2021 Société statistique du Canada

---

Additional Supporting Information may be found in the online version of this article at the publisher's website.

\* Corresponding author: [mrparker909@gmail.com](mailto:mrparker909@gmail.com)

## 1. INTRODUCTION

The ongoing COVID-19 pandemic has already led to around 147,000 confirmed cases and 1,700 deaths [Correction added on 11 November 2021, after first online publication on 1 November 2021: “17,000 deaths” was changed to “1,700 deaths”] in British Columbia (BC), and around 3.9 million deaths worldwide as of the end of June 2021. Disease analytics allow us to estimate the extent of unreported cases. Case counts are available from many government sources. In British Columbia, Canada, the BC Centre for Disease Control releases periodic surveillance reports, which often include COVID-19 case counts for each of five Health Authority regions. However, cases of COVID-19 can go unreported because of controllable factors such as refusal to test or low volumes of virus testing, as well as uncontrollable factors such as asymptomatic or pauci-symptomatic cases, incorrect self-diagnosis or failure to disclose. This poses problems in disease control. For example, it leads to under-reported case counts and thus an underestimate of the severity of the pandemic. Undetected cases also drive community transmission, reducing the effectiveness of contact tracing, quarantine and isolation. In this article, we aim to estimate the true size of an epidemic, given the observed case counts and outcomes.

There is substantial evidence for the existence of unreported cases of COVID-19. For example, Buitrago-Garcia et al. (2020) performed a systematic review and found a 95% confidence interval of 17%–25% for the proportion of truly asymptomatic cases. Several seroprevalence studies have also been conducted to estimate the proportion of unreported cases (Song et al., 2020; Bendavid et al., 2021; Saeed et al., 2021). In particular, Skowronski et al. (2020) found that in May 2020 the number of undetected cases in British Columbia was between 2.25 and 20.5 times greater (a 95% confidence interval) than the reported number of cases. Low ascertainment rates could potentially have been improved using more effective testing strategies (Lawless & Yan, 2021). However, the number of testings required to have only a small percentage of unreported cases would be prohibitive.

Several inter-related models have been developed in recent years to address the problem of estimating abundance in the presence of under-reporting using only case-count data. For example, the INAR (integer autoregressive) hidden population model of Fernández-Fontelo et al. (2016) was used to estimate weekly cases of human papillomavirus in Girona. More recently, Moriña et al. (2021) used a Bayesian hierarchical model to estimate the under-reporting rates of COVID-19 in Spain and found that their results matched seroprevalence data (Spanish Ministry of Health, 2020). Fernández-Fontelo et al. (2020) developed a hidden INAR(1) model designed to analyze the COVID-19 pandemic using only COVID-19 case counts as observed data to inform model-fitting. Their model was used to estimate the under-reporting rates of COVID-19 in several small regions of Spain, and they used susceptible-infectious-removed (SIR) modelling to account for population dynamics. The underlying process was  $X_t = \alpha \circ X_{t-1} + W_t(a_t)$ , where  $X_t$  is the actual number of new cases at time  $t$ ,  $\alpha$  is the binomial thinning probability,  $a_t$  is the new active cases modelled using SIR and  $W_t$  is a Poisson process.

The  $N$ -mixture model is a hierarchical model that can be seen as a series of related models ordered by their conditional probability structure (see Kéry & Royle, 2015, Section 2.3). Often, these models are fitted to data using maximum likelihood estimation (MLE). The  $N$ -mixture models can also be viewed from the Bayesian modelling perspective as hierarchical Bayesian models. Kéry & Royle (2015) used an  $N$ -mixture Bayesian approach to analyze Swiss Great Tits data. They found that the posterior means from a Bayesian analysis (using vague priors) numerically agreed well with the ML estimates.

Hidden INAR models can be viewed as a special case of  $N$ -mixture models in which there is only one site and, with the addition of a mixture distribution for detection, allowing for occasional perfect detection.  $N$ -mixture models have been used extensively since their inception

(Royle, 2004) and have been extended to allow for open population dynamics (Dail & Madsen, 2011). The application of the open  $N$ -mixture modelling framework to disease analytics has been discussed by DiRenzo et al. (2019).

Alternative methods such as capture–recapture (see, e.g., Xu et al., 2014; van Dam-Bates, Fyfe & Cowen, 2016) can produce more precise estimates of detection probability but require more extensive data-gathering (including unique identifiers for tracking, such as personal health numbers), which is not possible during the early stages of a public health crisis such as the COVID-19 pandemic.

Unlike classical MLE, Bayesian methodology can remove the task of integrating the latent abundance parameters ( $N_t$ , for  $t \in \{1, 2, \dots, T\}$ ) from the model likelihood and avoid computational complexity associated with maximizing the likelihood function. Comparisons between Bayesian and ML  $N$ -mixture estimates were made by Toribio, Gray & Liang (2012); however, these comparisons were done only for the closed population models. In this article, we compare MLE and Bayesian Markov chain Monte Carlo (MCMC) model-fitting approaches for an open population model.

We propose a novel model to estimate the levels of under-reporting in regions affected by COVID-19. The model is built on the open population  $N$ -mixtures framework (Royle, 2004; Dail & Madsen, 2011) as well as the hidden INAR framework (Fernández-Fontelo et al., 2016), with the population dynamics modified to allow for domestic spreading of the virus as well as importation of new cases from other regions. We also incorporate a multinomial component in the models to account for active cases, deaths and recoveries. These models are ideally suited to estimate the detection rates when limited data are available.

We applied the new model to data from the Northern Health Authority region of BC and improved the model by incorporating parameter covariates. We found in our applications that the MCMC and the MLE gave comparable results for estimating the probability of detection.

## 2. METHODS

### 2.1. Model Development

The classical  $N$ -mixture model developed by Royle (2004) allows estimation of the population abundance  $N$  (which is a latent variable in the model) using under-counted observations  $n$ , which are conditional on  $N$  through a detection thinning process. The abundance  $N_{it}$  at site  $i$ , time  $t$ , is modelled as  $N_{it} \sim \text{Poisson}(\lambda)$ . A detection thinning process is used to generate observed counts  $n_{it}$ , which is modelled as  $n_{it} \sim \text{Binomial}(N_{it}, p)$ . Here,  $\lambda$  is the initial mean site abundance, and  $p$  is the detection probability at time  $t$ . The model extensions of Dail & Madsen (2011) allow for population dynamics, removing the closed population assumption of the original  $N$ -mixture model. The standard dynamics assumption is  $N_{it+1} = S_{it} + G_{it}$ , where  $S_{it}$  models population survival, and  $G_{it}$  models new population gains (immigration) between  $t$  and  $t + 1$ .

The form of the  $N$ -mixture model affords many distributional choices, making it flexible for studying different sorts of populations. To develop a disease analytic version of the  $N$ -mixture model, several modifications are necessary. We will consider only the single-site case, and thus we drop the site subscript  $i$ ; however, we note that under a conditionally independent sites assumption, it is easy to extend these models to multiple sites. We let  $T$  denote the number of sampling occasions:  $t \in \{1, 2, \dots, T\}$ . We make several additions and modifications to the open population  $N$ -mixture models. We specify the model in Equation (1), with the variables defined below, and refer to this specification as we detail the model throughout the remainder of this section.

$$\begin{aligned}
\text{Initial Abundance:} & N_1 \sim \text{Poisson}(\lambda) \\
\text{State Process:} & \{A_t, D_t, R_t\} \sim \text{Mult}(N_t; p_a, p_d, p_r) \\
\text{Observed Active Cases:} & a_t = n_t + a_{t-1} - r_{t-1} - D_{t-1}, a_0 = r_0 = D_0 = 0 \\
\text{Domestic Spread:} & S_t \sim \text{Poisson}(\omega N_{t-1}), \text{ for } t > 1 \\
\text{Imported Cases:} & G_t \sim \text{Poisson}(\gamma), \text{ for } t > 1 \\
\text{Abundance Updates:} & N_t = A_{t-1} + S_t + G_t, \text{ for } t > 1 \\
\text{Observation Process:} & n_t \sim \text{Binomial}(N_t - a_{t-1} + r_{t-1} + D_{t-1}, p) \\
& \{a_t - D_t - r_t, D_t, r_t\} \sim \text{Mult}(a_t; p_a, p_d, p_r)
\end{aligned} \tag{1}$$

Similar to standard  $N$ -mixtures models, we make the assumption that the initial (start of study) active cases is the unknown random variable  $N_1$ . We use a Poisson distribution with mean  $\lambda$  to model this initial population size (Eq. 1: Initial Abundance). This can be thought of as the random state of the dynamic system at the start of data collection.

To model population changes with time, we consider the three possible future states of any individual from time  $t$  to time  $t + 1$ . An individual who is currently infected at time  $t$  can remain an active case, recover, or die by time  $t + 1$ . Thus we partition the total infected individuals  $N_t$  at time  $t$  into the three categories relative to  $t + 1$ : cases who will remain active  $A_t$ ; cases who will recover  $R_t$ ; and cases who will die  $D_t$ . The partitioning is done using a multinomial distribution, with probability of mortality  $p_d$ , probability of recovery  $p_r$ , and probability of remaining an active case of  $p_a = 1 - p_d - p_r$  (Eq. 1: State Process). Using this multinomial model assumes that (i) there are exactly three categories, (ii) all individuals are independent and have the same probabilities for each category, and (iii) the probabilities are constant over time. Assumption (i) seems reasonable (and if another category were to be considered, it could be added to the model without difficulty). Assumption (ii) is a large simplification, since many factors will influence an individual's probability of recovery and death (such as age, genetics, access to health care, etc.). This simplification is necessary due to a lack of individual-level information, and could be alleviated somewhat using, for example, supplementary demographic data such as age along with stratified modelling techniques. Finally, assumption (iii) may be true over short periods; however, the probabilities of recovery and death may change over the course of a pandemic because of such factors as improved understanding of the disease and increased access to treatment. Fortunately, assumption (iii) can be relaxed by including time-varying parameter covariates for  $p_r$  and  $p_d$ . We note that because of the relationship  $p_a = 1 - p_d - p_r$ , including a covariate for  $p_d$  and not for  $p_r$  would imply that any change in  $p_d$  would be reflected entirely in  $p_a$ . So in normal use cases, it would make sense to include the same covariate for  $p_r$  as for  $p_d$ , to allow for some of the deaths to shift into the recovered category rather than remain in the active category. The parameter  $p_d$  could change over time owing to factors such as local hospitals reaching their capacity. We note that for our Northern Health Authority case study, we do not expect assumption (iii) to be a significant factor.

We have established a partitioning of the active cases into categories allowing for deaths and recoveries. We now consider three pathways for producing active cases over time. First, we have from our partitioning the set,  $A_t$ , of cases that will remain active from time  $t$  to time  $t + 1$ . Second, we consider simple importation of cases  $G_t$  (Eq. 1: Imported Cases). We consider importation of cases to be caused by the immigration of external active cases from other regions. We model  $G_t$  as a Poisson random variable, with mean value  $\gamma$ . Thus we have defined  $\gamma$  as the average number of new imported active cases occurring between time  $t$  and time  $t + 1$ . Third, we consider domestic spread via community contact  $S_t$  (Eq. 1: Domestic Spread). Since the rate of domestic spread should be proportional to the current number of active cases (hence exponential growth in active cases is possible), we use the number of active cases at  $t - 1$  to moderate the mean new

infections due to domestic spread. We model  $S_t$  using a Poisson distribution with expected value  $\omega N_{t-1}$ . Thus, we have defined  $\omega$  as the average number of new infections per active case during one sampling period. To determine the number of active cases at time  $t$  (for  $t > 1$ ), we sum the three sources of active cases:  $N_t = A_{t-1} + S_t + G_t$  (Eq. 1: Abundance Updates).

The observation process in Equation (1): Observation Process has two components: a binomial thinning, and a multinomial partitioning. The observed data for this model are the sets of newly observed active cases  $\{n_t\}$  for  $t \in \{1, 2, \dots, T\}$ , newly recovered from previously observed active cases  $\{r_t\}$  for  $t \in \{1, 2, \dots, T-1\}$  and newly deceased cases  $\{D_t\}$  for  $t \in \{1, 2, \dots, T-1\}$ . The term “newly observed” means observed since the previous sampling occasion, so that in our case study  $n_{t+1}$  represents the total cases reported during week  $t+1$  that have been observed after the cases reported during week  $t$ . The newly observed active cases  $n_t$  are assumed to be under-counted. We use binomial thinning to model the under-counting; however, unlike with  $N$ -mixtures, we need to subtract the currently active previously observed cases  $a_{t-1} - r_{t-1} - D_{t-1}$  from the total  $N_t$  prior to thinning. This is because all active cases that have been observed are tracked through time until they either recover or die; they remain active, and cannot be “re-observed” while still active. We calculate the observed active cases at time  $t > 0$  as  $a_t = n_t + a_{t-1} - r_{t-1} - D_{t-1}$ , with  $a_0 = 0$ ;  $a_t$  can be understood as “new active cases at time  $t$ ” plus “previous active cases at time  $t-1$ ” minus “previous active cases which are no longer active at time  $t$ .” The quantities  $a_t$  and  $A_t$  are subtly different;  $A_t$  are the total cases remaining active from time  $t$  to time  $t+1$ , while  $a_t$  are the observed cases which are currently known to be active at time  $t$ . The observed active cases  $a_t$  are partitioned using a multinomial similar to the one used to partition  $N_t$ . The primary difference is that all three categories are fully observed:  $r_t$  are the currently active observed cases who will recover between  $t$  and  $t+1$ ;  $D_t$  are the observed deaths between  $t$  and  $t+1$ ; and  $a_t - r_t - D_t$  are the observed active cases which remain active from  $t$  to  $t+1$ . We use the same three probabilities as in the first multinomial. Thus,  $r_t$  is the observed subset of  $R_t$ ,  $D_t$  is fully observed, and  $a_t - r_t - D_t$  is the observed subset of  $A_t$ . This adds the extra assumption that the observed active cases are equally likely to recover as are the unobserved active cases. This is a critical assumption, as it allows identifiability of the unobserved quantities  $A_t$  and  $R_t$  through the observed data. Relaxing this assumption would require the addition of a known relation between the probability of recovery for observed individuals and that for unobserved individuals. A plate diagram specifying our model is shown in Figure 1, illustrating the data/model interactions. The model has six estimable parameters:  $\lambda$ ,  $\gamma$ ,  $\omega$ ,  $p$ ,  $p_d$  and  $p_r$ .

## 2.2. Maximum Likelihood Approach

The original approaches to inference for  $N$ -mixture models used MLE (Royle, 2004; Dail & Madsen, 2011), in which the latent variables  $N_t$  are removed from the likelihood via integration over states (summation of the likelihood over  $N_t \in \{n_t, n_t + 1, n_t + 2, \dots, K\}$ , for some suitable upper bound  $K$ ). The computation times involved in this method become problematic for very large  $K$ , since the summations dominate the computing time, which has complexity  $\mathcal{O}(K^3)$ . There are several advantages to using MLE: it is not dependent on correctly specified prior distributions (this can be a disadvantage when reliable prior information exists), and it has excellent large-sample properties, such as consistency, asymptotic normality, and asymptotic efficiency (Bain & Engelhardt, 1992, p. 316). MLE also has the benefit of being deterministic (provided that the optimization method is deterministic, as is the case with the BFGS optimization algorithm; Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

We build the likelihood function from the model described in Section 2.1. In classical  $N$ -mixtures models, integration is done over the possible states of the latent variables  $N_t$ . In our model, we have a second set of latent variables  $R_t$ , necessitating a second integration over states (summation over  $R_t \in \{r_t, r_t + 1, r_t + 2, \dots, N_t - D_t\}$ ). This leads to the likelihood function shown in Equation (2). We note that the likelihood function is written as  $\mathcal{L}$  for brevity,

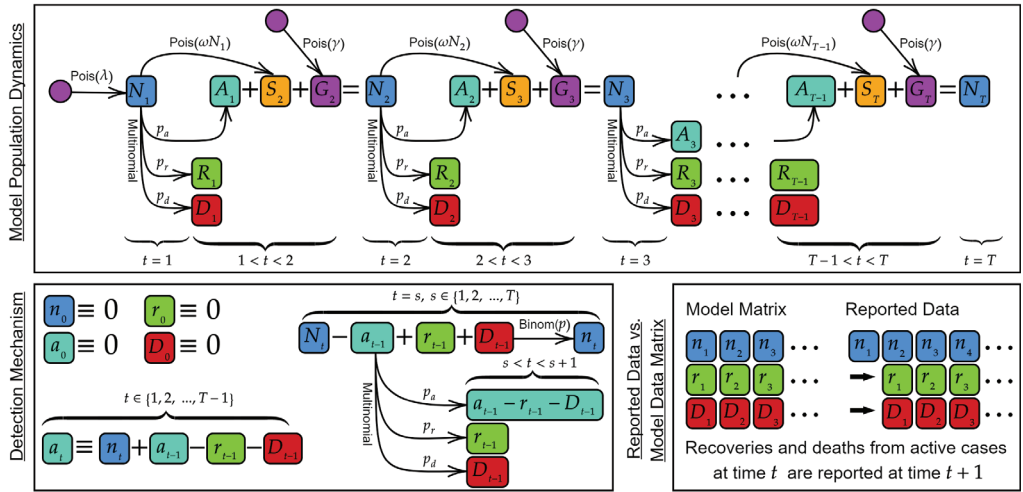


FIGURE 1: The data generating process assumed for our COVID-19 model. Top: the population dynamics are illustrated as a plate diagram (Koller & Friedman, 2009). The  $t = s$  labels indicate the variables defined at time  $t = s$ , while the  $a < t < b$  labels indicate processes which occur between observation time  $a$  and observation time  $b$ . Bottom left: the model detection mechanism, along with the recursive definition for calculated quantity  $a_t$ . Bottom right: the shift in time for recoveries and deaths between the reported data used in practice, and the model data matrix definition is illustrated. Recoveries and deaths associated with active cases at time  $t$  are reported at the next reporting period,  $t + 1$ , rather than at time  $t$  (since they occur between reporting periods).

and should be understood as shorthand for  $\mathcal{L}(\lambda, \gamma, \omega, p, p_d, p_r | \{n_t\}, \{D_t\}, \{r_t\}, K)$ . This model can also be extended to allow for under-counted deaths, in which case a third integration over states would be necessary:  $D_t \in \{d_t, d_t + 1, \dots, N_t - r_t\}$ , where  $d_t$  would be the observed deaths and  $D_t$  would be the latent variable for new deaths; in this case, ‘‘Observed active cases’’ from Equation (1) would have  $D_t$  replaced with  $d_t$ .

$$\begin{aligned}
 \mathcal{L} = & \sum_{N_1=n_1}^K \dots \sum_{N_T=n_T}^K \left\{ \text{Pois}(N_1; \lambda) \cdot \left( \prod_{t=1}^T \text{Binom}(n_t; N_t - a_{t-1} + r_{t-1} + D_{t-1}, p) \right) \right. \\
 & \cdot \left( \prod_{t=2}^T P_{N_{t-1}, N_t} \right) \cdot \left( \prod_{t=1}^{T-1} \text{Mult}(a_t - D_t - r_t, D_t, r_t; a_t, p_a, p_d, p_r) \right. \\
 & \left. \left. \cdot \sum_{R_t=r_t}^{N_t - D_t} \text{Mult}(A_t, D_t, R_t; N_t, p_a, p_d, p_r) \right) \right\}. \tag{2}
 \end{aligned}$$

$$P_{a,b} = \sum_{c=0}^{m=\min\{a,b\}} \text{Pois}(c; \omega a) \cdot \text{Pois}(b - c; \gamma).$$

Here,  $P_{a,b}$  is the transition probability for transitioning from an active number of cases  $a$  to an active number of cases  $b$ . We implemented this model using the R software (R Core Team, 2020). The likelihood function as specified in Equation (2) was programmed in R (archived code is available at <http://dx.doi.org/10.5281/zenodo.5502191>), and optimization was done using the

BFGS optimization algorithm employed by the R function *optim*. The value for  $K$ , the upper bound on summations, was chosen by optimizing the model likelihood with increasing values of  $K$  until the estimated parameter convergence was observed at  $K = 200$ . Computation time for these models, run on a 4.0 GHz AMD Ryzen 9 3900X with 24 logical processors, varies from several hours to several days depending on the model complexity (number and type of parameter covariates). The likelihood function may have large flat regions, especially when no additional covariates are used to inform model fitting. This is important for the estimability of the correlated parameters  $\gamma$  and  $\omega$ , which both inform population growth.

### 2.3. Bayesian Approach

The Bayesian approach to parameter estimation of our COVID-19 model is different from the standard MLE approach beyond the inclusion of priors. Specifically, summations over latent variables are not required for parameter updates; instead, the values of the latent variables are re-sampled. As well, the transition probability calculation  $P_{a,b}$  from Equation (2) is replaced with the distributions for  $S_t$  and  $G_t$ . These changes necessitate the use of stochastic methods such as MCMC for parameter estimation.

If we let  $\pi_\tau$  be the prior distribution for parameter  $\tau$ , then the full joint distribution over the data and parameters is given in Equation (3), where the abundance updates are still given by  $N_t = A_{t-1} + S_t + G_t$  for  $t \in \{2, \dots, T\}$ . Details regarding the dependence between the variables in these models are described in Figure 1.

The Bayesian models were implemented using JAGS (Plummer, 2003) through the RJags package (Plummer, 2019) using the R software (R Core Team, 2020). Computation for these models took several hours when fitting the base model. The full joint distribution  $f$  is given by the following:

$$\begin{aligned}
 f(\lambda, \gamma, \omega, p, p_r, p_d, n_t, r_t, D_t) &= \tilde{\mathcal{L}} \cdot \pi_\lambda \cdot \pi_\gamma \cdot \pi_\omega \cdot \pi_p \cdot \pi_{p_d} \cdot \pi_{p_r} \\
 \tilde{\mathcal{L}} &= \text{Pois}(N_1; \lambda) \cdot \left( \prod_{t=1}^T \text{Binom}(n_t; N_t - a_{t-1} + r_{t-1} + D_{t-1}, p) \right) \\
 &\quad \cdot \text{Mult}(A_t, D_t, R_t; N_t, p_a, p_d, p_r) \cdot \left( \prod_{t=1}^{T-1} \text{Mult}(a_t - D_t - r_t, D_t, r_t; a_t, p_a, p_d, p_r) \right) \\
 &\quad \cdot \left( \prod_{t=2}^T \text{Pois}(G_t; \gamma) \cdot \text{Pois}(S_t; \omega N_{t-1}) \right). \tag{3}
 \end{aligned}$$

#### 2.3.1. Posterior predictive checking

Posterior predictive checking is useful for finding discrepancies between the observed data and the data that can be described by the fitted model. It is a popular Bayesian model-checking approach used by ecologists (see, e.g., Kéry & Schaub, 2011; Gelman et al., 2013). In this paradigm, model fitness is checked by simulating data generated under a fitted model and comparing the simulated data with the observed data. If the observed data is substantially different from the simulated data, then the model must be unable to effectively describe the observed data. Posterior predictive checks can be conducted by examining histograms of the simulated data and comparing them with the observed data, or by computing the posterior predictive  $P$ -values (Gelman et al., 2013). Since our model assumes fully observed  $D_t$ , we generated the simulated observations of  $D_t$  using Equation (1): State Process, and used a binomial rather than a multinomial to generate  $r_t$ ;  $r_t \sim \text{Binomial}(a_t - D_t; p_r)$ . This was necessary to avoid dual specification of  $D_t$ . As well, we

require  $a_t$  to be non-negative, and we accomplish this by rejecting simulated data for which  $a_t$  is negative.

To run the Bayesian models, the adaptation period was 1.4 million iterations. We discarded 1.2 million initial samples as the burn-in, and ran 200,000 additional iterations to obtain the posterior estimates. In order to avoid issues with parameter auto-correlation, we used thinning to keep 1 out of every 200 iterations.

### 3. CASE STUDY

#### 3.1. Data Modalities

The Northern Health Authority region of British Columbia, Canada was chosen for this preliminary study because of the relatively small number of cases, allowing for faster model-fitting. Figure 2 shows the three sets of observed data:  $\{n_t\}$ ,  $\{r_t\}$  and  $\{D_t\}$ , for  $t \in \{1, 2, \dots, 30\}$ .

We gathered publicly available data from the BC CDC Surveillance Reports (BC Centre for Disease Control, 2020). For the 30 weeks starting 26 March 2020 and ending 15 October 2020, we used weekly counts aggregated on Thursdays. The 30-week time period was chosen because the data definitions in the public reports were relatively stable over this period, making data comparable between weeks. We note that the reporting period shifted from Thursdays to Fridays after the 15 October 2020 reporting date. The start dates for each phase of the provincial recovery plan are shown in Table 1. The end of each phase of the provincial recovery plan caused changes/reductions in the provincial protective measures (such as reopening of local businesses in Phase 2), which led to inhomogeneity in the dynamics parameters. To account for this inhomogeneity, we fitted the parameters dependent on time-varying covariates using link functions. A summary of the public health measures taken for each phase is shown in Table 1. Additional data used as covariates include the number of new COVID-19 tests administered per

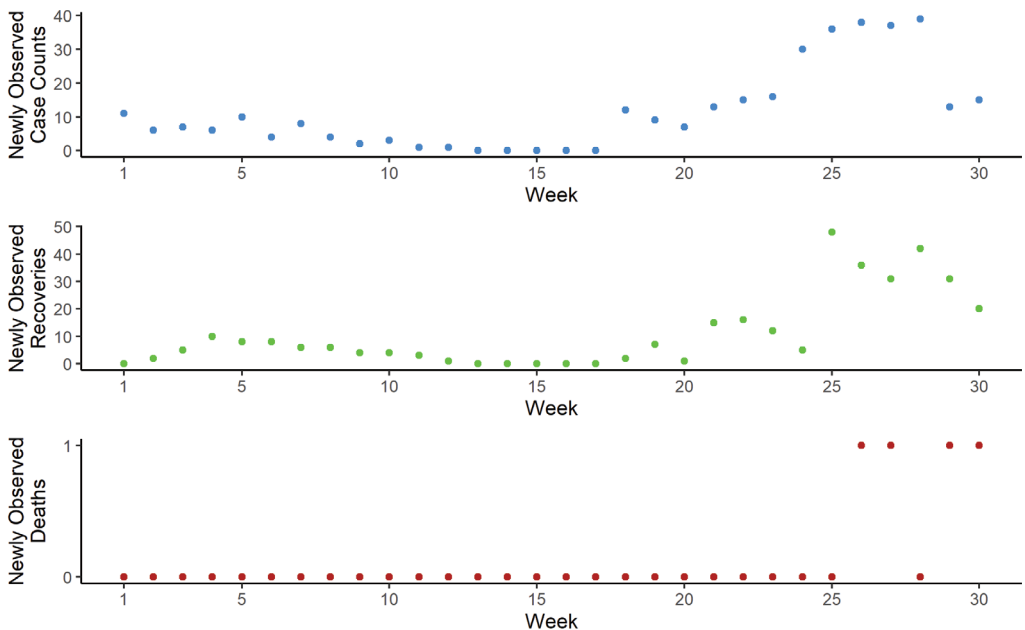


FIGURE 2: Plots of the three datasets: case counts  $\{n_t\}$  (top), recoveries  $\{r_t\}$  (middle) and deaths  $\{D_t\}$  (bottom). Data are for the 30 weeks from 26 March 2020 to 15 October 2020 for the Northern Health Authority region of British Columbia, Canada.



TABLE 1: Summary of the BC recovery plan phases and associated measures.

Recovery plan phase	Summary of measures
Phase 1: “Implementation” Start date: 26 March 2020	<ul style="list-style-type: none"> <li>• Public health emergency declared</li> <li>• Mandated physical distancing</li> <li>• Banned gatherings of more than 50 people</li> <li>• Closed dine-in service at bars and restaurants, and non-essential personal services</li> <li>• Closed all BC provincial parks</li> <li>• US/Canada border closure</li> </ul>
Phase 2: “Initial relaxation” Start date: 21 May 2020	<ul style="list-style-type: none"> <li>• Part time return to school for K-12 students</li> <li>• Businesses and sectors reopened with extra safety precautions and physical distancing measures</li> <li>• Employers required to develop COVID-19 safety plans</li> </ul>
Phase 3a: “Further relaxation” Start date: 25 June 2020	<ul style="list-style-type: none"> <li>• Allowing non-essential travel in BC</li> <li>• Reopening of hotels and movie theatres</li> </ul>
Phase 3b: “Start of school year” Start date: 17 September 2020	<ul style="list-style-type: none"> <li>• Same as 3a, plus start of school year</li> </ul>

*Note:* See Government of British Columbia (2020) and the BC Centre for Disease Control (2020) for complete lists of enacted measures and protocols.

week (Province of British Columbia, 2020)—which is a strong indicator of the detection rate  $p$ —and Google regional mobility data (Google LLC, 2020)—which is a potential indicator of the domestic spread ( $\omega$ ). The phase boundaries were used to demarcate categorical covariates for each phase.

Publicly available data are often intensely aggregated, poorly reported (Barone, 2020), or susceptible to technical issues such as data loss (Fetzer & Graeber, 2020). In the case of the Surveillance Report data, there are several shortcomings to consider. Reporting dates may be delayed from the date of detection, or from the date of infection. Reporting delays are also not necessarily the same between case counts, recovery counts and death counts. The number of observed cases is dependent on the testing methodology (number of tests administered, effectiveness of tests, etc.), and this testing methodology can change over time. The data also suffer from ad hoc reporting times, so that counts are not always available for each day, and some days contain lump sum data dumps. Owing to these particular data issues, we chose to use weekly aggregated counts rather than daily counts. These data limitations are concerning, and more reliable data could improve the accuracy of our results.

## 3.2. Results

### 3.2.1. Base model comparison MLE versus MCMC

To compare the MLE and MCMC methods, we fitted base models with no covariates. Both the MLE and the MCMC base models were run on Compute Canada’s WestGrid. The fitted model parameters are shown in Table 2. For both methods,  $p_d$  was estimated to be essentially zero, as the Northern Health Authority region of BC had no deaths until Phase 3b. Several parameter estimates are dissimilar between the MLE and MCMC methods. The estimates for  $\gamma$  and  $\omega$  are significantly different, with no overlap in their uncertainty intervals. This is likely due to the

TABLE 2: Parameter estimates for the base model (with no parameter covariates) fitted to the Northern Health Authority COVID-19 data.

Parameter	MLE	MCMC
$\lambda$	52.58 (38.91, 71.05)	29.36 (16.44, 47.71)
$\gamma$	$4 \times 10^{-9}$ (0, $\infty$ )	1.60 (0.14, 4.00)
$\omega$	1.01 (0.96, 1.08)	0.62 (0.54, 0.70)
<b><math>p</math></b>	<b>0.30 (0.21, 0.41)</b>	<b>0.30 (0.22, 0.41)</b>
$p_d$	0.0024 (0.0012, 0.0048)	0.0037 (0.0012, 0.0080)
$p_r$	0.48 (0.46, 0.50)	0.62 (0.58, 0.66)

Note: Included for comparison are classic  $N$ -mixture maximum likelihood estimates (MLE) and Bayesian model estimates (MCMC), with 95% confidence intervals shown in parentheses for MLE parameter estimates, and 95% credible intervals shown for MCMC estimates. Parameters are initial mean abundance parameter  $\lambda$ , mean imported cases parameter  $\gamma$ , mean domestic spread parameter  $\omega$ , probability of detection  $p$ , probability of mortality  $p_d$ , and probability of recovery  $p_r$ . We note that  $p$  (in bold) is our primary parameter of interest for estimating under-detection.

similarity between models with small population growth (when both parameters  $\gamma$  and  $\omega$  are small, this may cause a non-identifiability issue for these parameters). However, the Bayesian estimate of the detection probability,  $\hat{p} = 0.30$ , is identical to that of the classic  $N$ -mixture estimate,  $\hat{p} = 0.30$ . Parameter errors are estimated using the estimated Hessian matrix for the MLE method, and using the posterior credible intervals for the MCMC method.

The Bayesian approach requires specification of prior distributions for each parameter in the model. The prior distributions we used are summarized in Equation (4). For the parameter  $\lambda$ , we use the gamma distribution with mean 15. For the parameter  $\gamma$ , we chose to use the uniform distribution with lower bound equal to 0 and upper bound equal to 30. For the detection probability parameter  $p$  and the probability of death  $p_d$ , we used the Uniform(0,1) distribution to place equal probability on all possible values. Since  $p_r$  and  $p_d$  are dependent, we use  $p_d$  as the upper bound for the uniform distribution of  $p_r$ . We chose a weakly informative prior for  $\omega$ , the uniform distribution with lower bound 0 and upper bound 5.

For the Bayesian approach we used the mean as the test quantity for posterior checking (see Figure 3). The red lines indicate the observed means, which are close to the middle of the simulated distributions, showing a good fit. The posterior predictive  $P$ -values for observed cases, recovered cases, and deaths are 0.37, 0.42 and 0.42, respectively. Each of the  $P$ -values is close to 0.5, which also shows that the model has a good fit.

Prior Distributions:

$$\begin{aligned}
 \pi_\lambda &= \text{Gamma}(\text{shape} = 15, \text{rate} = 1) & \pi_p &= \text{Uniform}(0, 1) \\
 \pi_\gamma &= \text{Uniform}(0, 30) & \pi_{p_d} &= \text{Uniform}(0, 1) \\
 \pi_\omega &= \text{Uniform}(0, 5) & \pi_{p_r} &= \text{Uniform}(0, p_d)
 \end{aligned} \tag{4}$$

### 3.2.2. Bayesian sensitivity analysis

We performed a sensitivity analysis to determine the impact of the  $\lambda$  and  $\gamma$  prior distributions on parameter estimates. Since the parameter  $\lambda$  informs the initial population size, it is crucial that the prior not be overly informative; therefore, we performed a sensitivity analysis for  $\pi_\lambda$ . We also performed a sensitivity analysis for  $\pi_\gamma$ , since the discrepancy between the MLE and

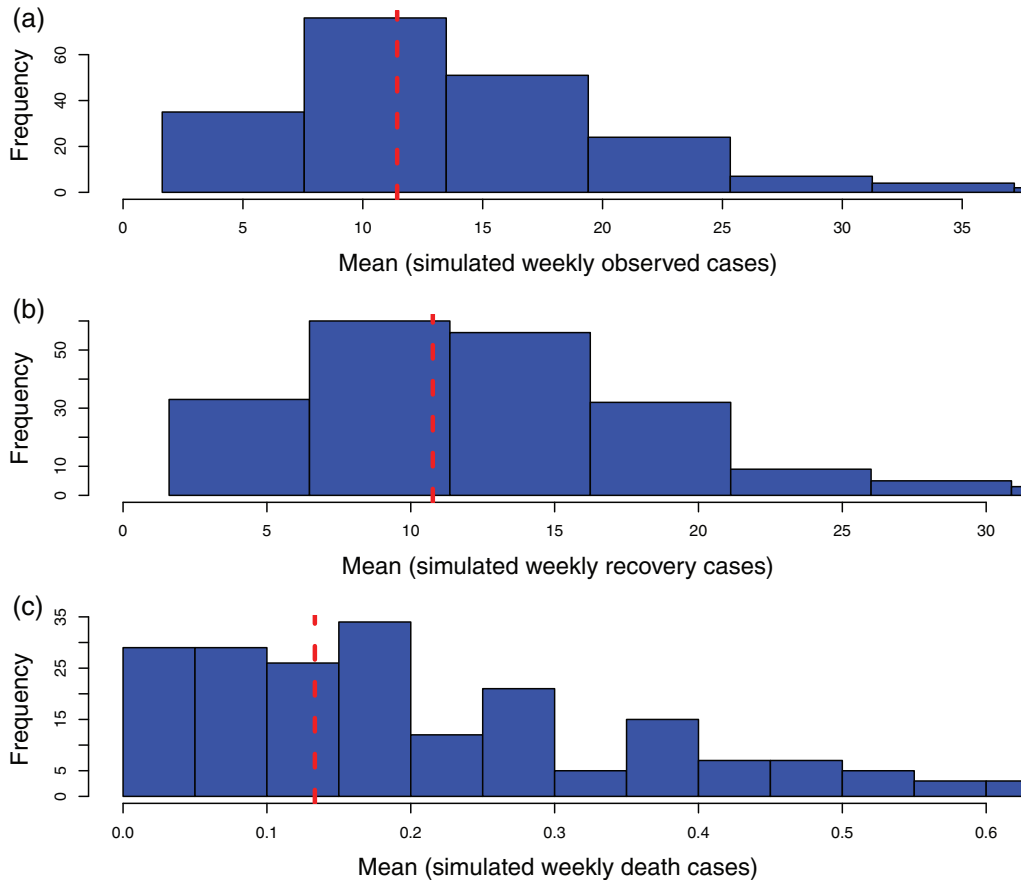


FIGURE 3: Posterior model checking for COVID-19 cases in the Northern Health Authority region showing (a) observed cases, (b) observed recoveries and (c) observed deaths. The posterior means are depicted as vertical red dashed lines.

MCMC estimates for  $\gamma$  led to the concern that the prior may be overly informative. Both prior distributions were found to have low impact on the parameter estimates.

For the parameter  $\lambda$ , we set the mean of the gamma distribution to be 5, 10 and 20 in the analysis. We set the variance of the prior distribution to 200 in order to have large variation in  $\lambda$ . The medians and the relative 95% credible intervals for the parameter estimates are shown for each  $\lambda$  prior distribution in Table 3.

For the parameter  $\gamma$ , we used a uniform distribution with minimum 0 and the maximum varying from 5 to 30. The medians and the relative 95% credible intervals for the parameter estimates are shown for each  $\gamma$  prior distribution in Table 4.

### 3.3. Bayesian Simulation Study

In order to ascertain the ability of the model to provide adequate parameter estimates, we performed a simulation study by setting ground truth values for the model parameters. We set the parameter values close to the parameter estimates for the Northern Health Authority region Bayesian model, and we set the number of sampling occasions to be the same, i.e., 30.

TABLE 3: Sensitivity analysis for the  $\lambda$  prior  $\pi_\lambda$  in the base model (no parameter covariates) fitted to the Northern Health Authority COVID-19 data.

Lambda prior	Mean 5 variance 200		Mean 10 variance 200		Mean 20 variance 200	
$\hat{\lambda}$	29.91	(16.81, 49.24)	29.62	(16.65, 48.27)	29.09	(16.93, 46.25)
$\hat{\gamma}$	1.66	(0.13, 4.01)	1.65	(0.15, 4.01)	1.59	(0.11, 3.89)
$\hat{\omega}$	0.62	(0.54, 0.70)	0.62	(0.54, 0.70)	0.62	(0.54, 0.71)
$\hat{p}$	0.31	(0.22, 0.41)	0.30	(0.21, 0.40)	0.31	(0.23, 0.41)
$\hat{p}_d$	0.0049	(0.0036, 0.0057)	0.0037	(0.0014, 0.0056)	0.0046	(0.0018, 0.0076)
$\hat{p}_r$	0.62	(0.58, 0.66)	0.62	(0.58, 0.66)	0.62	(0.58, 0.66)

Note: Parameters are initial mean abundance parameter  $\lambda$ , mean imported cases parameter  $\gamma$ , mean domestic spread parameter  $\omega$ , probability of detection  $p$ , probability of mortality  $p_d$ , and probability of recovery  $p_r$ . The  $\lambda$  parameter was given a gamma prior distribution. The posterior medians and 95% credible intervals for the parameter estimates are shown.

TABLE 4: Sensitivity analysis for the  $\gamma$  prior  $\pi_\gamma$  in the base model (with no parameter covariates) fitted to the Northern Health Authority COVID-19 data.

Gamma prior	Uniform(0, 5)		Uniform(0, 10)		Uniform(0, 30)	
$\hat{\lambda}$	29.86	(16.69, 48.09)	29.26	(16.37, 48.44)	30.04	(1.68, 48.53)
$\hat{\gamma}$	1.64	(0.13, 4.14)	1.64	(0.13, 3.99)	1.58	(0.15, 3.99)
$\hat{\omega}$	0.62	(0.54, 0.70)	0.62	(0.54, 0.71)	0.63	(0.55, 0.71)
$\hat{p}$	0.30	(0.21, 0.40)	0.31	(0.22, 0.41)	0.30	(0.21, 0.76)
$\hat{p}_d$	0.0024	(0.0009, 0.0044)	0.0040	(0.0024, 0.0056)	0.0026	(0.0007, 0.0044)
$\hat{p}_r$	0.62	(0.58, 0.66)	0.62	(0.58, 0.66)	0.62	(0.58, 0.66)

Note: Parameters are initial mean abundance parameter  $\lambda$ , mean imported cases parameter  $\gamma$ , mean domestic spread parameter  $\omega$ , probability of detection  $p$ , probability of mortality  $p_d$ , and probability of recovery  $p_r$ . The  $\gamma$  parameter was given a uniform prior distribution. Posterior medians and 95% credible intervals for the parameter estimates are shown.

We generated random populations and observations based on the ground truth parameter values and the model structure shown in Figure 1. The randomly generated observation data included observed cases, recoveries and deaths. We used the generated observations to fit the base model using JAGS (Plummer, 2003) via the RJags package (Plummer, 2019) and the prior distributions in Equation (4). We repeated the process 150 times, calculating the mean of the 150 posterior medians (Table 5). The coverage probability for the 150 credible intervals for the detection probability is close to 0.95, indicating good model performance. For the other parameters, the coverage probabilities are smaller; however, considering the small number of replicates, the coverage is reasonably close to 0.95. The simulation study was conducted on Compute Canada’s WestGrid.

### 3.3.1. Best fitted model results

We incorporated parameter covariates into our model, and we considered several nested models, starting with a base model with no covariates. MLE was used to fit each of the nested models.

TABLE 5: Results of the Bayesian model simulation study with 150 independent replicates to compare parameter estimates with true values.

Parameter	Ground truth	Parameter estimates	Coverage probability
$\lambda$	10	9.07	0.93
$\gamma$	2	4.03	0.89
$\omega$	0.5	0.44	0.90
$p$	0.5	0.48	0.97
$p_d$	0.005	0.0053	0.85
$p_r$	0.5	0.50	0.91

Note: Parameters are initial mean abundance parameter  $\lambda$ , mean imported cases parameter  $\gamma$ , mean domestic spread parameter  $\omega$ , probability of detection  $p$ , probability of mortality  $p_d$ , and probability of recovery  $p_r$ . Parameter estimates are simulation means of the 150 posterior medians, with 95% coverage probability.

TABLE 6: Model covariates, log-likelihood ( $\ell$ ), number of parameters (Q), Akaike information criterion (AIC),  $\Delta$ AIC, and small sample AIC (AICc).

Model covariates	Q	$\ell$	AIC	$\Delta$ AIC	AICc
( <i>pha</i> , <i>vol</i> )	10	-270.45	560.90	0	572.47
( <i>pha</i> )	9	-274.19	566.39	5.49	575.39
( <i>mob</i> , <i>vol</i> )	13	-270.74	567.49	6.59	590.24
( <i>mob</i> , <i>pha</i> , <i>vol</i> )	16	-268.31	568.61	7.71	610.46
( <i>mob</i> )	12	-273.36	570.72	9.82	589.08
( <i>vol</i> )	7	-290.92	595.83	34.93	600.92
(No covariates)	6	-298.07	608.14	47.24	611.79

Note: Parameter covariates are indicated by shorthand: *mob* for Google mobility data as covariates for  $\omega$  (+6 covariates); *vol* for testing volume as covariate for  $p$  (+1 covariate); and *pha* for BC Recovery Plan phase as covariates for  $\omega$  (+3 covariates).

Covariates were incorporated into the model using log transforms (to limit the parameter range from 0 to  $\infty$ ) for parameters  $\lambda$ ,  $\gamma$  and  $\omega$ , and logit transforms (to limit the parameter range from 0 to 1) for parameters  $p$ ,  $p_d$  and  $p_r$ . For example, consider the model with normalized weekly test volumes  $V_t$  as a covariate for the probability of detection  $p$ . This would be included in the model using the logit transform  $\text{logit}(p_t) = \beta_0 + \beta_1 V_t$ . In this way,  $\beta_0$  would be the baseline coefficient for  $p_t$ ,  $\beta_1$  would be the effect on probability of detection due to number of tests administered, and  $p_t$  would be constrained to take values between 0 and 1. We used Akaike’s information criterion (AIC; Akaike, 1974) to compare models (Table 6). We also considered the small-sample version of AIC, the AICc, which produced the same top model as AIC. The nested models were all run on a 4.0 GHz AMD Ryzen 9 3900X with 24 logical processors.

We considered several possible sets of parameter covariates, indicated by the following shorthand: *mob* for Google mobility data as covariates for  $\omega$  (+6 covariates); *vol* for testing volume as covariate for  $p$  (+1 covariate); and *pha* for BC Recovery Plan phase as covariates

for  $\omega$  (+3 covariates). The best fitted model was  $(pha, vol)$ , which included a total of 10 model parameters.

Since the covariates for  $p$  and  $\omega$  are time-varying, the parameters themselves are also time-varying. For this reason, the parameter estimates for  $p_t$  and  $\omega_t$  are summarized graphically for each week  $t$  in Figures 4 and 5. The remaining parameter estimates are summarized in Table 7. The estimated weekly probability of detection was used to estimate  $N_t$  through a Horvitz–Thompson-type estimator,  $\hat{N}_t = n_t/\hat{p}_t$  (Figure 6).

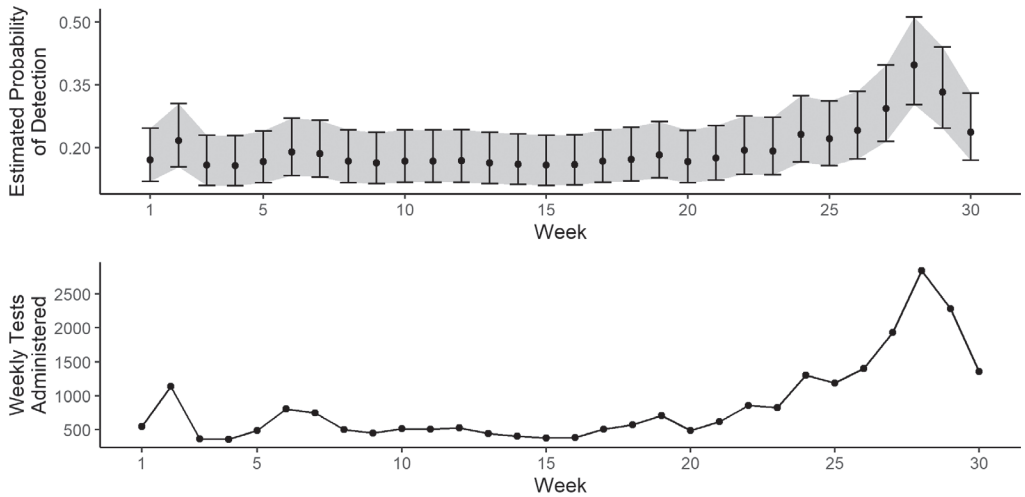


FIGURE 4: Estimated probability of detection  $\hat{p}$  from the best fitted model  $(pha, vol)$ , as chosen by AIC. Top: weekly detection probability, along with weekly 95% confidence intervals. Bottom: covariate weekly testing volume. Trends in weekly testing volume can be seen to match trends in probability of detection, since they are related through  $\text{logit}(p) = \beta_0 + \beta_1 V_t$ , where  $V_t$  is the normalized weekly testing volume, and  $\beta_i$  are the covariate coefficients.

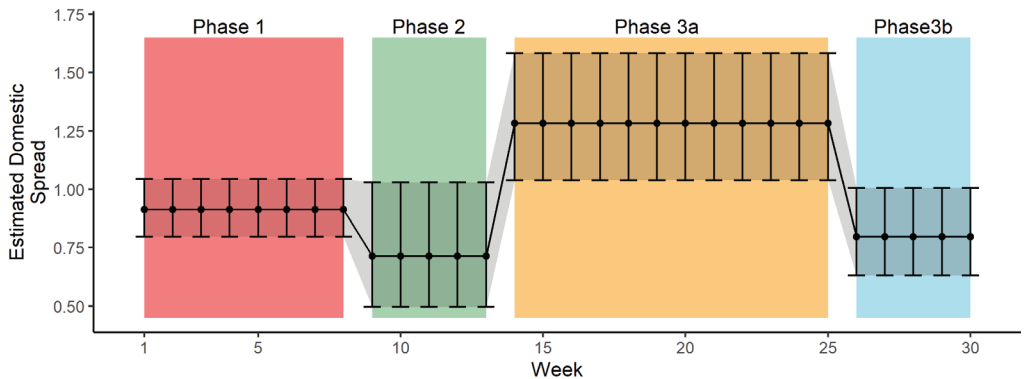


FIGURE 5: Estimated average weekly infections per infected individual  $\hat{\omega}$  from the model  $(pha, vol)$  per week, with 95% confidence intervals. BC Recovery Plan phases are indicated by shaded and labelled regions. Phase covariates are related to  $\omega$  through  $\text{log}(\omega) = \beta_1 I_1 + \beta_2 I_2 + \beta_{3a} I_{3a} + \beta_{3b} I_{3b}$ , where  $I_i$  are indicator variables indicating Phase  $i$ , and  $\beta_i$  are the covariate coefficients.

TABLE 7: Parameter estimates and 95% confidence intervals for the best fitted model (*pha, vol*).

Parameter	Estimate	CI
$\lambda$	64.87	(43.98, 95.67)
$\gamma$	$4 \times 10^{-9}$	(0, $\infty$ )
$p_d$	0.0020	(0.0010, 0.0042)
$p_r$	0.48	(0.46, 0.50)

Note: Parameters are initial mean abundance parameter  $\lambda$ , mean imported cases parameter  $\gamma$ , probability of mortality  $p_d$ , and probability of recovery  $p_r$ . Time-varying parameters  $p$  and  $\omega$  are not included.

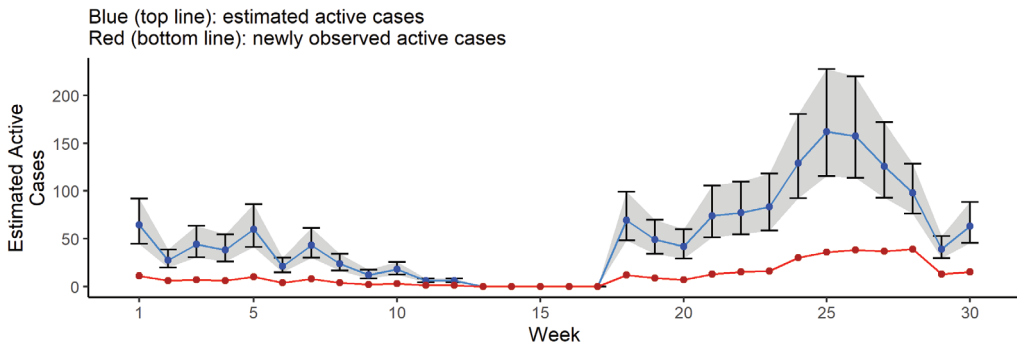


FIGURE 6: Estimated active cases  $\hat{N}_t$  per week from the best fitted model (*mob, vol*), as chosen by AIC. Bottom line (red): newly observed active cases. Top line (blue): estimated active cases with 95% confidence intervals.  $\hat{N}_t$  are calculated from the estimated probability of detection  $\hat{p}_t$  and newly observed active cases by  $\hat{N}_t = n_t / \hat{p}_t$ .

We note that all fitted models had  $\hat{\gamma} \approx 0$ , which is on the boundary of the parameter space. This led to 95% confidence intervals of  $(0, \infty)$  and an inability to estimate uncertainty for  $\gamma$ .

#### 4. DISCUSSION

We developed a novel model for disease analytics, accounting for population dynamics and under-detected cases, by incorporating three datasets that are commonly available publicly during a pandemic. The model also incorporates parameter covariates, which allow for time-varying parameters and parameter change points. We have estimated the level of under-detection of COVID-19 cases in the BC Northern Health Authority region, and found that there is substantial evidence of undetected COVID-19 cases during the first 30 weeks of the pandemic in BC. We have compared two methods of implementing a base model with no covariates, and found that both the MLE and the MCMC approaches are able to describe the population dynamics and estimate levels of under-reporting. We then improved upon the parameter estimates by incorporating several parameter covariates for the MLE implementation. The best fitted model, (*pha, vol*), as chosen by AIC, was found to have the probability of detection dependent on the volume of weekly tests administered, and domestic spread dependent on changes in phase for the BC Recovery Plan.

Choosing the best fitted model using AIC has the benefit of parsimony: we selected the model with the smallest number of model parameters, which explained the most variability

in the observed data. The AIC approach did not select the full model (*mob, pha, vol*) over the best fitted model because of this parsimony. This is not surprising, since a likely effect of the different BC Recovery Plan phases should be to affect changes in population mobility. Thus, some of the information contained in the mobility data is likely accounted for in the phase data.

Estimates of weekly probability of detection for the model (*pha, vol*) can be seen to closely follow the trend in weekly administered tests (Figure 4). We can see clearly that under-reporting was the lowest at the end of the 30-week period (minimum of 60.2% at week 28) and the highest between weeks 3 and 20 (maximum of 84.2% at week 4).

As illustrated in Figure 6, there is a period (weeks 13–17) with zero observed new active cases. This leads to estimates of  $\hat{N}_t = 0$  for that time period. However, it is important to note that it is unlikely that  $N_t$  is actually zero; rather, we had insufficient data for that period. This is a period with testing volume insufficient to detect the smaller number of active cases adequately. The bottom plot of Figure 4 indicates the low testing volume from week 13 to week 16.

An additional goal of this study was to identify any change in  $\omega$  over each of the phases of the BC Recovery Plan. We show in Figure 5 the results of estimating  $\omega$  for the model (*pha, vol*). The four recovery plan phases are labelled in the plot, and the effect of each phase on domestic spread is clear. For Phase 1,  $\hat{\omega}_1 = 0.91$ , 95% CI: (0.80, 1.05); for Phase 2,  $\hat{\omega}_2 = 0.71$ , 95% CI: (0.50, 1.03); for Phase 3a,  $\hat{\omega}_{3a} = 1.28$ , 95% CI: (1.04, 1.58); and for Phase 3b,  $\hat{\omega}_{3b} = 0.80$ , 95% CI: (0.63, 1.00). Thus Phase 3a saw the largest average domestic spread out of the four considered periods.

Regions with many COVID-19 cases may have experienced higher rates of under-detection than the Northern Health Authority region because of a lack of capacity for testing. However, our results are consistent with a serological study (Skowronski et al., 2020) in the lower mainland of BC, which used two sampling snapshots (one in March and one in May 2020). Skowronski et al. (2020) found that for May 2020, the 95% confidence interval for total cases was between 2.25 and 20.5 times greater than the reported number of cases, while we found for May 2020 (weeks 6–10) a 95% confidence interval of 3.69–8.75 times greater than the reported number of cases. While these confidence intervals are consistent with each other, and we believe the under-detection rate could be similar across regions (when testing volume has been accounted for), further studies applying these methods to other regions would be necessary for confirmation.

Owing to the summation over states to remove the latent variables  $N_t$  and  $R_t$  from the likelihood, the likelihood function is computationally demanding, with computation times roughly proportional to  $K^3$ . This is one important reason to consider the Bayesian MCMC approach over the MLE approach. The Bayesian approach is more computationally tractable for large  $N_t$ , as the MCMC algorithm explores the space of possible latent variable states stochastically, without resorting to integrating over states. For comparison, in the Northern Health Authority case study, both the MLE approach and the MCMC approach took close to 6 h of computing time (both were computed using Compute Canada's WestGrid). However, for larger regions, the MLE approach may become intractable because of the dependence of the computing time on the population size.

In choosing the priors for the MCMC approach, it would be beneficial to incorporate results from other comparable studies. In particular, there is little prior knowledge informing the range of the spread rate  $\omega$  and the mean importation  $\gamma$ . Improving the prior distributions would be helpful in reducing the posterior variance and improving the accuracy of estimates (Kéry & Royle, 2015, Section 2.5.3).

Currently, our model cannot account for the differences between symptomatic and asymptomatic cases. The data we used to inform the model involve aggregated case counts of primarily



symptomatic cases. However, our model attempts to estimate the total number of COVID-19 cases (including both symptomatic and asymptomatic cases). This is possible because asymptomatic cases will still contribute to the spread of the virus, increasing the number of symptomatic cases (Ganyani et al., 2020; Johansson et al., 2021). Thus the population growth parameters  $\gamma$  and  $\omega$  are informed by both the symptomatic and the asymptomatic cases (and the probability of detection  $p$  will be deflated by the presence of asymptomatic cases). However, our model does not treat the two categories separately, thus making the implicit model assumption that both categories share identical population dynamics. In future work, this could be addressed by including additional information about asymptomatic cases, which to our knowledge is not widely available from regional public data. In the hospitalization records model of Pullano et al. (2021), researchers looked at under-detection of COVID-19 in France, and their approach included information on numbers of symptomatic and asymptomatic cases observed through testing, allowing them to assess population dynamics for those case categories. Similar data collection could be done for other regions, and the information could be incorporated into our model using techniques similar to those of Pullano et al. (2021). If the proportion of asymptomatic cases was known over time, then that information could be used as a time covariate for relevant parameters in our model.

Our model does not make use of any additional data to distinguish between domestic spread and importation. If we had access to reliable data on the proportion of observed cases that are due to domestic spread versus importation, then we could also incorporate these data as parameter covariates and further improve the estimates of the population dynamics parameters  $\gamma$  and  $\omega$ . This could improve our model fit, as our current estimate for the apparent mean importation is  $\hat{\gamma} \approx 0$ , which is not likely to be the true mean importation. As an example, consider as auxiliary data the number of incoming travellers who have tested positive per week. The auxiliary data would give a lower bound on  $\gamma$  for each time point, leading to better estimates of both  $\gamma$  and  $\omega$ . The auxiliary data would be incorporated in the MLE models by including an indicator variable in the likelihood function and in the MCMC models as additional prior knowledge for  $\gamma$ .

An important limitation of our model is the requirement for enough reporting periods  $T$ . Initially we looked at only the Phase 1 data for the Northern Health Authority region. However, since Phase 1 contained only 8 weeks of data, both the MLE and the MCMC models failed to converge. For the MLE model, increasing  $K$  caused  $\hat{\lambda}$  to increase without bound, while  $\hat{p}$  decreased asymptotically to zero. For the MCMC model, the failure was evident in that the prior distributions chosen for both  $\lambda$  and  $p$  became overly informative.

Another shortcoming of the model, identified by an anonymous reviewer, is that the number of deaths  $D_t$  is assumed fully observed, and in the model specified in Equation (1) we make the simplifying assumption that  $p_d$  is the same between the two multinomial components (the state process and the observation process) of the model. This implies that  $\hat{p}_d = E\left[\frac{D_t}{N_t}\right]$ , and also that  $\hat{p}_d = E\left[\frac{D_t}{a_t}\right]$ . However, this is inconsistent when  $D_t > 0$ , since  $N_t > a_t$ . This inconsistency will have little to no effect on our case study, since  $\hat{p}_d$  is small. However, for larger values of  $p_d$ , this issue could be solved by adding an additional parameter  $\alpha$  to the model, which would allow  $\alpha p_d$  (probability of death for observed cases) to be larger than  $p_d$  (probability of death for total cases). In this case, in Equation (1): Observation Process, we would need to replace  $p_d$  with  $\alpha p_d$ . Several options exist for modelling  $\alpha$ . It could be another free parameter to be estimated by the model fitting process, or it could be taken to be the reciprocal of the probability of detection:  $\alpha = 1/p$ . The latter case seems preferable, as it has the potential to inform estimates of the probability of detection better when  $p_d$  is large; and when  $p$  approaches 1,  $\alpha p_d$  approaches  $p_d$ . This latter case

leads to the following generative process:

$$\begin{aligned}
 \text{Initial Abundance:} & \quad N_1 \sim \text{Poisson}(\lambda) \\
 \text{State Process:} & \quad \{A_t, D_t, R_t\} \sim \text{Mult}(N_t; p_a, p_d, p_r) \\
 \text{Observed Active Cases:} & \quad a_t = n_t + a_{t-1} - r_{t-1} - D_{t-1}, a_0 = r_0 = D_0 = 0 \\
 \text{Domestic Spread:} & \quad S_t \sim \text{Poisson}(\omega N_{t-1}), \text{ for } t > 1 \\
 \text{Imported Cases:} & \quad G_t \sim \text{Poisson}(\gamma), \text{ for } t > 1 \\
 \text{Abundance Updates:} & \quad N_t = A_{t-1} + S_t + G_t, \text{ for } t > 1 \\
 & \quad n_t \sim \text{Binomial}(N_t - a_{t-1} + r_{t-1} + D_{t-1}, p) \\
 \text{Observation Process:} & \quad \{a_t - D_t - r_t, D_t, r_t\} \sim \text{Mult}(a_t; p_a^*, \alpha p_d, p_r)
 \end{aligned} \tag{5}$$

where  $p_a^* = 1 - \alpha p_d - p_r$  is the probability of remaining active in the observed subset of cases. We note that this implies an additional constraint:  $p > p_d$ . However, the constraint is already accounted for by the assumption that deaths are fully observed.

For future applications to larger populations, we recommend introducing the  $\alpha$  parameter specified in Equation (5). To illustrate the similarity between the  $\alpha = 1$  model and the  $\alpha = 1/p$  model when the number of deaths is relatively small, we show in Table 8 the parameter estimates from fitting both models using the Bayesian framework. Parameter estimates are nearly identical between the two models.

For future applications of this model to more recent data, in particular to the current situation with the COVID-19 pandemic, we also recommend including total administered vaccinations as a time-varying parameter for  $\omega$ , since a larger proportion of vaccinated individuals in a population reduces the rate of spread of the disease (Haas et al., 2021).

In the  $N$ -mixtures disease analytics work of DiRenzo et al. (2019), data are required for modelling both infected and non-infected individuals, as sampling is assumed to occur at random in a mixed population. In contrast, we only consider infected individuals, as the publicly available COVID-19 counts data do not contain any non-infected individuals. It is possible for us to ignore the uninfected population because of the implicit assumption that the uninfected population is large compared to the infected population (so that we do not account for population

TABLE 8: Bayesian parameter estimates for the base model (with no parameter covariates) fitted to the Northern Health Authority COVID-19 data.

Parameter	Model $\alpha = 1$	Model $\alpha = 1/p$
$\lambda$	29.36 (16.44, 47.71)	30.30 (16.62, 48.51)
$\gamma$	1.60 (0.14, 4.00)	1.71 (0.13, 3.98)
$\omega$	0.62 (0.54, 0.70)	0.62 (0.54, 0.70)
$p$	0.30 (0.22, 0.41)	0.31 (0.21, 0.41)
$p_d$	0.0037 (0.0012, 0.0080)	0.0030 (0.0013, 0.0055)
$p_r$	0.62 (0.58, 0.66)	0.62 (0.58, 0.66)

Note: Included for comparison are estimates with 95% credible intervals for the  $\alpha = 1$  model from Equation (1) and the  $\alpha = 1/p$  model from Equation (5). Parameters are initial mean abundance parameter  $\lambda$ , mean imported cases parameter  $\gamma$ , mean domestic spread parameter  $\omega$ , probability of detection  $p$ , probability of mortality  $p_d$ , and probability of recovery  $p_r$ . Note that for the Northern Health Authority data, the difference in estimates between the two models is negligible.

saturation effects, such as running out of susceptible members of the population to infect). While this implicit assumption is reasonable during early stages of the pandemic, it can be removed by including time-varying covariates for  $\omega$ , which would allow  $\omega$  to decrease to zero as the population becomes saturated.

We have looked at a single Health Authority region in British Columbia, Canada. We will continue to apply these methods to each of the remaining four Health Authority regions so as to compare the results across the province. By extending these methods to multi-site modelling, we intend to evaluate the situation for the province of BC as a whole, treating each Health Authority region as an independent site. This will allow each site to borrow population dynamics information from the other regions.

## ACKNOWLEDGEMENTS

This study was enabled in part by the support provided by WestGrid ([www.westgrid.ca](http://www.westgrid.ca)) and Compute Canada/Calcul Canada ([www.computeCanada.ca](http://www.computeCanada.ca)). We would like to acknowledge the Michael Smith Foundation for Health Research and the Victoria Hospitals Foundation for support through a COVID-19 Research Response grant, as well as a Canadian Statistical Sciences Institute Rapid Response Program—COVID-19 grant to LC that supported this research. This manuscript was greatly improved by comments and feedback from two anonymous reviewers, as well as from the journal editors.

## REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Bain, L. & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*, 2nd ed., Brooks/Cole, Cengage Learning, Belmont, CA.
- Barone, V. (2020). Chile counts coronavirus deaths as ‘recovered’. *New York Times*. Retrieved from <https://nypost.com/2020/04/14/chile-counts-coronavirus-deaths-as-recovered/>. Accessed 15 April 2020.
- BC Centre for Disease Control. (2020). BC COVID-19 data [surveillance reports]. Retrieved from <http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data>. Accessed 20 October 2020.
- Bendavid, E., Mulaney, B., Sood, N., Shah, S., Bromley-Dulfano, R., Lai, C., Weissberg, Z., et al. (2021). COVID-19 antibody seroprevalence in Santa Clara County, California. *International Journal of Epidemiology*, 50, 410–419.
- Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms: 2. The new algorithm. *IMA Journal of Applied Mathematics*, 6, 222–231.
- Buitrago-Garcia, D., Egli-Gany, D., Counotte, M. J., Hossmann, S., Imeri, H., Ipekci, A. M., Salanti, G., & Low, N. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Medicine*, 17, e1003346.
- Dail, D. & Madsen, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, 67, 577–587.
- DiRenzo, G. V., Che-Castaldo, C., Saunders, S. P., Grant, E. H. C., & Zipkin, E. F. (2019). Disease-structured  $N$ -mixture models: A practical guide to model disease dynamics using count data. *Ecology and Evolution*, 9, 899–909.
- Fernández-Fontelo, A., Cabaña, A., Puig, P., & Morriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, 35, 4875–4890.
- Fernández-Fontelo, A., Morriña, D., Cabaña, A., Arratia, A., & Puig, P. (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS One*, 15, e0242956.
- Fetzer, T. & Graeber, T. (2020). Does contact tracing work? Quasi-experimental evidence from an Excel error in England. CAGE working paper no. 521.
- Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13, 317–322.
- Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., & Hens, N. (2020). Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, 25, 2000257.

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL.
- Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24, 23–26.
- Google LLC. (2020). Google COVID-19 community mobility reports. <https://www.google.com/covid19/mobility/>. Accessed 12 April 2021.
- Government of British Columbia. (2020). Phase 1: BC's restart plan. Retrieved from <https://web.archive.org/web/20201124233822/https://www2.gov.bc.ca/gov/content/safety/emergency-preparedness-response-recovery/covid-19-provincial-support/phase-1>. Accessed 26 July 2021.
- Haas, E. J., Angulo, F. J., McLaughlin, J. M., Anis, E., Singer, S. R., Khan, F., Brooks, N., et al. (2021). Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: An observational study using national surveillance data. *The Lancet*, 397, 1819–1829.
- Johansson, M. A., Quandelacy, T. M., Kada, S., Prasad, P. V., Steele, M., Brooks, J. T., Slayton, R. B., Biggerstaff, M., & Butler, J. C. (2021). SARS-CoV-2 transmission from people without COVID-19 symptoms. *JAMA Network Open*, 4, e2035057.
- Kéry, M. & Royle, J. A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press, London, UK.
- Kéry, M. & Schaub, M. (2011). *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Academic Press, New York, NY.
- Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.
- Lawless, J. F. & Yan, P. (2021). On testing for infections during epidemics, with application to Covid-19 in Ontario, Canada. *Infectious Disease Modelling*, 6, 930–941.
- Moriña, D., Fernández-Fontelo, A., Cabaña, A., Arratia, A., Ávalos, G., & Puig, P. (2021). Cumulated burden of COVID-19 in Spain from a Bayesian perspective. *European Journal of Public Health*, 31, 917–920.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vol. 124, Vienna, Austria, 1–10.
- Plummer, M. (2019). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10.
- Province of British Columbia. (2020). B.C. COVID-19 – Laboratory information. Retrieved from <https://governmentofbc.maps.arcgis.com/home/item.html?id=ba047e4a9bd24beb9ca6e94c05eddef9>. Accessed 19 April 2021.
- Pullano, G., Di Domenico, L., Sabbatini, C. E., Valdano, E., Turbelin, C., Debin, M., Guerrisi, C., et al. (2021). Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature*, 590.
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Royle, J. A. (2004).  $N$ -mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60, 108–115.
- Saeed, S., Drews, S. J., Pambrun, C., Yi, Q.-L., Osmond, L., & O'Brien, S. F. (2021). SARS-CoV-2 seroprevalence among blood donors after the first COVID-19 wave in Canada. *Transfusion*, 61, 862–872.
- Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24, 647–656.
- Skowronski, D. M., Sekirov, I., Sabaiduc, S., Zou, M., Morshed, M., Lawrence, D., Smolina, K., et al. (2020). Low SARS-CoV-2 sero-prevalence based on anonymized residual sero-survey before and after first wave measures in British Columbia, Canada, March–May 2020. medRxiv: 2020.07.13.20153148.
- Song, S.-K., Lee, D.-H., Nam, J.-H., Kim, K.-T., Do, J.-S., Kang, D.-W., Kim, S.-G., & Cho, M.-R. (2020). IgG seroprevalence of COVID-19 among individuals without a history of the coronavirus disease infection in Daegu, Korea. *Journal of Korean Medical Science*, 35, e269.
- Spanish Ministry of Health. (2020). Estudio ene COVID-19: Segunda ronda estudio nacional de sero-epidemiología de la infección por SARS-CoV-2 en España. Technical report.

- Toribio, S. G., Gray, B. R., & Liang, S. (2012). An evaluation of the Bayesian approach to fitting the  $N$ -mixture model for use with pseudo-replicated count data. *Journal of Statistical Computation and Simulation*, 82, 1135–1143.
- van Dam-Bates, P., Fyfe, M., & Cowen, L. L. E. (2016). Applying open population capture–recapture models to estimate the abundance of injection drug users in Victoria, Canada. *Journal of Substance Use*, 21, 185–190.
- Xu, Y., Fyfe, M., Walker, L., & Cowen, L. L. E. (2014). Estimating the number of injection drug users in greater Victoria, Canada using capture–recapture methods. *Harm Reduction Journal*, 11, 1–9.
- 

*Received 30 December 2020*

*Accepted 12 September 2021*