Development of a disease analytic model for estimating the hidden population using the stratified-Petersen estimator

by

Siying Ma
B.Sc., East China Normal University, 2021

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Development of a disease analytic model for estimating the hidden population using the stratified-Petersen estimator

by

Siying Ma
B.Sc., East China Normal University, 2021

Supervisory Committee

Dr. Laura Cowen, Supervisor
(Department of Mathematics and Statistics)

Dr. Junling Ma, Committee member
(Department of Mathematics and Statistics)

ABSTRACT

The COVID-19 pandemic brought the need for novel disease analytic models capable of estimating the true number of infections, including those that evaded detection. Statistical methods, such as the stratified-Petersen estimator, provide effective ways in wildlife population modelling to estimate hard-to-reach population size. We developed a novel disease analytic model to estimate the levels of underreported COVID-19 cases and the true population size based on the idea of developing a Bayesian version of the stratified-Petersen estimator under a state-space formulation using individual-level capture-recapture data. We obtained the capture events from individuals' electronic health records and treated the occurrence of positive SARS-CoV-2 diagnostic test results and 2020 COVID-19-related hospitalizations as the tagging and recapture processes. Applying this model to the data from the Northern Health Authority region in British Columbia, Canada in 2020 by using a Bayesian Markov chain Monte Carlo (MCMC) approach, we found that the estimate of the size of the COVID-19 population ($\hat{N} = 2,967$) is 1.58 (95% CI: (1.53, 1.63)) times greater than the observed cases ($n_{obs} = 1,880$), which is a comparable result to those reported in other studies.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to express my gratitude to:

my supervisor, Laura, for her constant help in both life and study throughout my master life.

my supervisory committee and collaborators, Junling, Lloyd and Kenny, for sharing their expertise and providing professional feedback to my thesis work.

the Visual and Automated Disease Analytics (VADA) program for funding me on this project.

Also, I would like to thank:

my family, especially my mother, for their endless love, support and encouragement at all times.

my partner, Robert, for being my solid backing and source of motivation.

Chapter 1

Introduction

The COVID-19 pandemic brought the need for novel disease analytic models capable of estimating the true number of infections, including ones that evade detection. This is needed to understand the severity of COVID-19.

The COVID-19 pandemic led to 4,952,770 cases and 59,271 deaths by the end of April 30, 2024, in Canada (Government of Canada, 2024). However, reported cases of COVID-19 are generally recognized as a portion of the true number of cases, and underreporting cases of infectious diseases are common in public health studies (Gibbons et al., 2014). For COVID-19, the numbers above may be underestimated due to controllable factors such as limited testing capacity, testing accessibility and availability, and no mandate to report some classes of tests (such as self-administered rapid antigen tests) during later periods of the pandemic. They may also be underestimated due to uncontrollable factors such as asymptomatic or pauci-symptomatic cases or false-negative test results (Parker, M.R.P. et al., 2021; Dougherty et al., 2021; Gandhi et al., 2020). Skowronski et al. (2020) conducted a seroprevalence study in British Columbia in May 2020 and estimated that the true number of cases was 8 times higher than the reported cases in the lower mainland. Understanding the prevalence of COVID-19 and the proportion of underreported cases can provide crucial insights for public health agencies when developing strategies for control and interventions.

The estimation of population size for hard-to-reach populations is a common problem with wildlife population modelling. The Lincoln-Petersen estimator (Seber, 1986; Royle and Dorazio, 2008) is a two-sample capture-recapture method commonly used in ecology to estimate an animal population size, N, when it is impractical to count every individual. On the first visit, defined as the tagging process, n_1 individuals are captured, marked uniquely, and released. On the second visit, defined as the recap-

ture process, n_2 individuals are observed with m of them tagged on the first visit. By the Lincoln-Petersen estimator, the estimate of population abundance, $\hat{N}$, is equal to $n_1 n_2/m$. This method of capture-recapture has been extensively applied to the study of human populations (e.g. Robles et al., 1988; Xu et al., 2014). Schwarz and Taylor (1998) developed the stratified-Petersen estimator from the Lincoln-Petersen estimator to address heterogeneity in catchability as well as avoid assumption violation (equal catchability and complete mixing in the tagging and recovery process). The main idea of this estimator is to use row-level individual data, divide the population into multiple strata based on location or time, and apply the Petersen estimator within each stratum. We will introduce the stratified-Petersen estimator in Chapter 1.1.

State-space formulations have been wildly incorporated into ecological models for estimating population size (Jonsen et al., 2003, 2005; Gimenez et al., 2007). Royle (2008) developed a state-space parameterization of the Cormack-Jolly-Seber (CJS) model with an open population and provided a generic and flexible way to model the transitions between different states of individuals over multiple sampling occasions. King (2012) reviewed Bayesian state-space formulations for capture-recapture models. Bayesian model-fitting algorithms are one of the most common and efficient ways for inferring the unknown states and parameters of state–space models for animal populations (Newman et al., 2009; Link and Barker, 2010; Kéry and Schaub, 2012). Dao (2023) developed a Bayesian state-space model, inspired by the stratified-Petersen estimator, but transformed it into a disease dynamic model to provide a robust estimate of the hidden-population of COVID-19. An introduction to state-space capture-recapture models is provided in Chapter 1.2.

Several other models have also been developed for estimating the population size of COVID-19 using only case-count data. At the beginning of the COVID-19 pandemic, Parker, M.R.P. et al. (2021) developed a single site, hidden Markov model based loosely on N-mixture models (Royle, 2004; Dail and Madsen, 2011) to estimate COVID-19 dynamics using daily count data published by the British Columbia Government. Parker, M.R.P. et al. (2024) further expanded this model to a multi-site framework, which increases the precision compared to the single-site model.

Differing from N-mixture models, the stratified-Petersen estimator requires individual-level data. Individual-level health data is confidential (using a unique identifier for each individual in the database) and much more difficult to access, requiring researchers to receive ethical approval and submit data acquisition requests, while dis-

ease count data is much easier to obtain, often publicly available from websites or government publications. The establishment of individual-level databases for human diseases usually takes a much longer time than establishing a public count database, so it is very hard to have access to individual-level datasets during the early stages of a pandemic. Population Data BC stewards an individual-level database and have been collecting medical data for patients with COVID-19 provided by the Province of British Columbia. Although individual-level data has shortcomings compared to other data types, models with individual-level data usually can produce more precise estimates. Therefore, we expect to obtain a more accurate estimate of population size. The comparison of model accuracy is essential for public health agencies to decide which type of data should be collected for further study.

We developed a novel disease analytic model to estimate the levels of underreported COVID-19 cases and the true population size. The estimator is based on Dao's idea (Dao, 2023) to develop a Bayesian version of the stratified-Petersen estimator under a state-space formulation using individual-level capture-recapture data. The original stratified-Petersen estimator has never been applied to human studies to our knowledge. To transform the estimator into a disease analytic model, we redefine the tagging and recapturing processes based on the COVID-19 diagnosis process (i.e. SARS-CoV-2 PCR tests and hospitalization for treatments) and incorporate Geometric components to address the COVID-19 disease dynamics (i.e. time to being symptomatic, time to be diagnosed, etc.). The model is created for closed populations only and is ideally suited to the outbreak period of infectious diseases with travel restrictions. We applied this model to the data from the Northern Health Authority region in British Columbia, Canada in 2020 using a Bayesian Markov chain Monte Carlo (MCMC) approach. The year 2020 was marked as the beginning stage of the pandemic when the number of cases were highly under-reported. Moreover, many developed models during this stage could give wide estimates due to uncertainty in the aggregated dataset. This brings a need for more refined models using individual-level data. Our analysis revealed that our model generated results that are comparable to those reported in other studies.

1.1 Stratified-Petersen estimator

The stratified-Petersen estimator extends the two-sample Lincoln-Petersen estimator to multiple strata. This extension is needed to solve the problem of assumption

violation that will result in severe bias. Two of the key assumptions of the Lincoln-Petersen estimator are often violated in real-world experiments. The first assumption is that the Lincoln-Petersen estimator requires all individuals in the population to have the same probability of capture within samples (the tagging and recovery process). The second assumption is complete mixing of tagged and untagged individuals. The stratified-Petersen estimator developed by Schwarz and Taylor (1998) allows heterogeneity in catchability and mixing when estimating the closed population abundance N (no immigration, emigration, births, or deaths other than due to failure to survive within the study period). In a stratified-Petersen experiment, the population and samples are stratified into s non-overlapping strata at the time of tagging and into t non-overlapping strata at the time of recovery. Then, observed statistics collected from the experiment can be arranged as in Table 1.1, where n_i^c is the number of individuals captured and marked in the tagging stratum i; n_j^r is the number of individuals that are captured in the recovery stratum j regardless of whether they are marked in the tagging process or not; m_{ij} is the total number of individuals marked in the tagging stratum i and recaptured in recovery stratum j, and u_j is the total number of untagged individuals captured in recovery stratum j. Therefore, we have a total of $s + st + t$ statistics.

Table 1.1: Statistics collected from the stratified-Petersen experiment

Tagging	Individuals	Recovery stratum				Not
stratum	tagged	1	2	...	t	recovered
1	n_1^c	m_{11}	m_{12}	...	m_{it}	$n_1^c - m_{1\cdot}$
2	n_2^c	m_{21}	m_{22}	...	m_{2t}	$n_2^c - m_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	n_s^c	m_{s1}	m_{s2}	...	m_{st}	$n_s^c - m_{s\cdot}$
Total of untagged individuals		u_1	u_2	...	u_t	

The goal of this experiment is to estimate the total population size N. The population can be divided as follows: N_i^c (N_j^r) is the total number of individuals from the population present in the tagging stratum i (recovery stratum j), and N_{ij} is the number of individuals from the population moving from tagging stratum i to recovery stratum j. In the stratified-Petersen estimator, we assume that no part of the population will enter a recovery stratum without belonging to one of the tagging strata. In other words, if the population is closed (no individuals die during the experiment),

the target population size N will be equal to the number of the individuals from the population present in all tagging strata ($N_{\cdot}^{c}$) or all recovery strata ($N_{\cdot}^{r}$), i.e.,

$$N = N_{\cdot}^{c} = \sum_{i=1}^{s} \sum_{j=1}^{t} N_{ij} = N_{\cdot}^{r}. \tag{1.1}$$

To estimate the population size, we are interested in p_i^c (p_j^r), the probability of an individual presenting in a certain tagging stratum i (a certain recovery stratum j), as well as θ_{ij}, the probability that an individual presents in the tagging stratum i will survive and move to the recovery stratum j. With these parameters of interest, we can obtain the expected values of statistics by

$$E(n_i^c) = N_i^c p_i^c, \tag{1.2}$$

$$E(m_{ij}) = N_i^c p_i^c \theta_{ij} p_j^r, \tag{1.3}$$

$$E(u_j) = \sum_{i=1}^{s} (1 - p_i^c) N_i^c \theta_{ij} p_j^r. \tag{1.4}$$

Schwarz further developed the R package *SPAS* (Schwarz, 2023) to fit the stratified-Petersen estimator (Darroch, 1961; Plante et al., 1998; Schwarz and Taylor, 1998). A $(s+1) \times (t+1)$ matrix of raw data is required to fit the model, rather than row-level data. The $s \times t$ upper left matrix is the number of individuals marked in row stratum i and recovered in column stratum j. Row $s+1$ contains the total number of unmarked individuals recovered in column stratum j. Column $t+1$ contains the number of individuals marked in each row stratum but not recovered in any column stratum. The estimate of each parameter can be calculated by the function `SPAS.fit.model()`.

1.2 State-space formulation

The capture-recapture model can be specified using a state-space model formulation which separates the state process from the observation processes of individuals observed within the study (King, 2012). Under this formulation, the model is divided into two components: (1) a system process that models the underlying state in each stratum and (2) an observation process that is conditional on the state process to explain the imperfect detection of the system process. After applying a state-space framework to capture-recapture data, we can incorporate the observation error in the

data as a result of a non-perfect recovery process.

Capture-recapture data is usually organized into encounter (or capture) histories of each individual observed in the study. Let $\boldsymbol{h}$ denote the $n \times k$ encounter matrix, in which

$$h_{ij} = \begin{cases} 0 & \text{if individual } i \text{ is unobserved at stratum } j, \\ 1 & \text{if individual } i \text{ is observed at stratum } j, \end{cases} \tag{1.5}$$

for $i = 1, \ldots, n$ and stratum $j = 1, \ldots, k$. Let $\boldsymbol{h}_i$ denote the i^{th} row of matrix $\boldsymbol{h}$, and it represents the encounter history of a certain individual i. For example, when k is equal to 7, we suppose that $\boldsymbol{h}_1 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$ and $\boldsymbol{h}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$. This means that individual 1 is first observed and tagged in the first stratum and recaptured in the fourth stratum and individual 2 is first observed and tagged in the first stratum, but then never captured again. Further, under the state-space framework capture-recapture data $\boldsymbol{h}$ can be addressed as the combination of two processes: a survival process which describes the surviving of the individual between each capture, and an observation process which describes whether or not the individual is captured at a certain time. To define the survival process, let $\boldsymbol{x}$ to be the $n \times k$ matrix where

$$x_{ij} = \begin{cases} 0 & \text{if individual } i \text{ is not alive or has not entered the study yet at stratum } j, \\ 1 & \text{if individual } i \text{ is alive at stratum } j, \end{cases} \tag{1.6}$$

for $i = 1, \ldots, n$ and stratum $j = 1, \ldots, k$. Let $f(i)$ be the first time individual i is observed during the observation period and ϕ_{ij} as the probability individual i is still alive and in stratum $j + 1$. The underlying system can be measured as

$$x_{ij} | x_{ij-1} \sim \text{Bernoulli}(x_{ij-1}\phi_{i,j-1}) \tag{1.7}$$

for stratum $j = f(i) + 1, \ldots, k$.

In the observation process, we denote $\boldsymbol{y}$ as the $n \times k$ matrix corresponding to recapture process such that

$$y_{ij} = \begin{cases} 0 & \text{if individual } i \text{ is not recaptured at stratum } j, \\ 1 & \text{if individual } i \text{ is recaptured at stratum } j, \end{cases} \tag{1.8}$$

for $i = 1, \ldots, n$ and stratum $j = 1, \ldots, k$. Given x_{it}, the observation process, y_{it} can

be obtained by the conditional probability of (notated as |)

$$y_{it}|x_{it} \sim \text{Bernoulli}(p_t^r x_{it}) \quad t = 2, \ldots, k, \tag{1.9}$$

where p_t^r is the probability of recapture in the stratum t (King, 2012). Thus, we successfully separate the observed encounter histories $\boldsymbol{h}$ into the data components $\boldsymbol{x}$, $\boldsymbol{y}$.

Under the state-space framework, the state matrix $\boldsymbol{x}$ shows the underlying state process in each stratum, and we use the observation matrix $\boldsymbol{y}$ to detect the error in the estimation of underlying system. After getting a robust estimate, we then can transfer our target, the population size N, into the number of individuals whose row in the state matrix $\boldsymbol{x}$ does not sum to 0.

Chapter 2

Methods

2.1 Method development

Several modifications are required to transition the stratified-Petersen estimator to a disease analytic model. Instead of collecting data from physical capture events, we achieve encounter histories from individuals' electronic health records. We note that from the traditional stratified-Petersen estimator, the study cohorts are defined on the basis of their first observation (i.e., their first availability) so that the first capture event is measured by the system process in the the state-space formulation described in Chapter 1.2. However, individuals infected with COVID-19 can go for diagnosis anytime after infection. It results in inaccuracy for obtaining the time of individuals' first availability as well as the measurement of the following states and recapture process based on the time of their first diagnosis (capture). Therefore, we apply a temporal stratification and simplify the system process for only measuring their true underlying infection status. We identify the study cohort by the time individuals first show up in the state process and measure both capture and recapture events in the observation process. To do this, we must use another matrix in the observation process corresponding to the capture events. We define the state and observation processes based on COVID-19 progression published on the government of Canada website (Government of Canada, 2023).

We partition all the infected individuals into symptomatic and asymptomatic cases (Byambasuren et al., 2020). Symptomatic individuals may start experiencing symptoms from 1 to 14 days after they are exposed to the virus and will be recommended to take a lab test and isolate at home immediately. Similarly, local health authori-

ties also suggest potential asymptomatic individuals for taking lab tests after being in contact with COVID-19-infected people to control domestic spread. For symptomatic individuals, if their symptoms are getting worse, such as trouble breathing or persistent pain in the chest, these individuals will be admitted to local hospitals for treatment as soon as possible. We define the individuals with positive COVID-19 lab results or hospitalized for COVID-19 treatment as captured in the observation (tagging or recapture) process. Both of these outcomes are conditional on the state process as whether or not individuals start to show COVID-19 symptoms or potentially identify as asymptomatic individuals. Since it is known that not all individuals infected with COVID-19 require in-patient treatment, we also introduce a separate state process for recapture as to whether or not these symptomatic individuals will develop severe COVID-19 symptoms. We specify the steps of state-space stratified-Petersen experiment in Figure (2.1), with each process defined above.

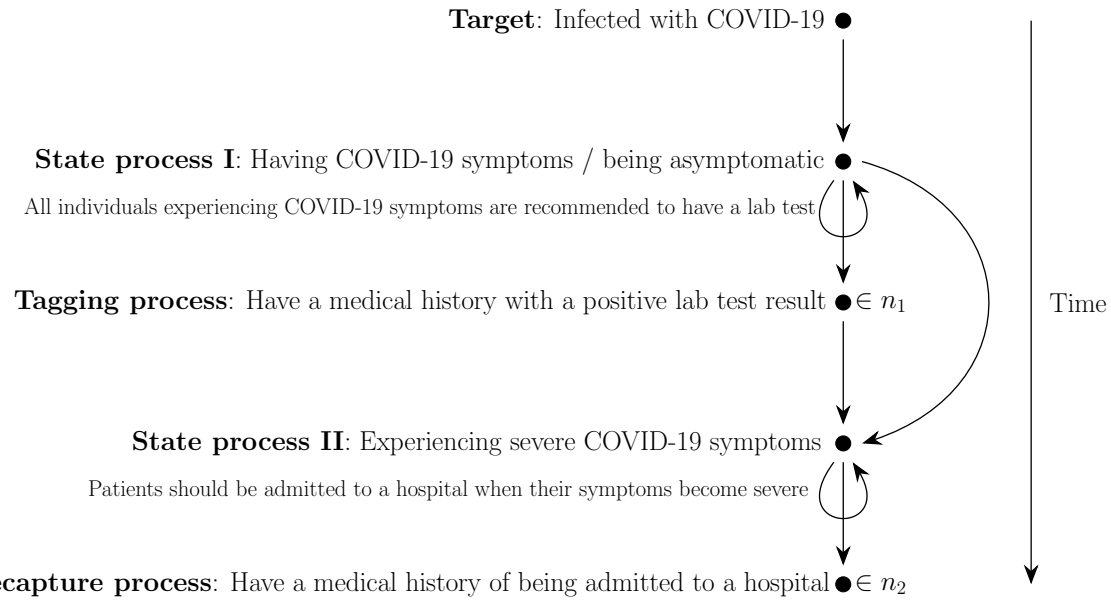

Figure 2.1: Capture-recapture process for the stratified-Peterson experiment of COVID-19 under a state-space formulation.

Data augmentation is an effective approach to estimate population size under the state-space model formulation. Kéry and Schaub (2012) transferred the question of estimating the unknown true population size N from a closed-population model to estimating a detection probability ψ in a zero-inflation model with a fixed upper bound for N, denoted by M. Let n ($n \in [1, N]$) be the number of individuals observed within the study and note that $N - n$ individuals could be captured at least once but

end up being unobserved. In the data augmentation approach, we set the observed data with a fixed dimension of $M \times k$, where k denotes the number of capture events and allows for $M - n$ individuals not captured during the study period; $M - N$ of which are pseudo-individuals.

In this study, the estimator is built on one cohort. In other words, all the study population will show up in the state process in the first stratum. However, under the assumption of independence in each sampling event, it is easy to extend the model to multiple cohorts by expanding the latent variable z_i defined in the equation (2.1) to a multi-variable z_{it} ($t \in [1, T]$) where T is the number of cohorts in the experiment. We expect within a certain period dominated by the same variant (such as the Delta-associated outbreak from July 2021 to September 2021) that the distribution of COVID-19 disease progression would be similar. This allows us to interpret the estimate of parameters with significant real-world values. Therefore, we line up the data within a period by setting the symptom onset date as the time for the first capture event and measure the population size by the latent variable

$$z_i \sim \text{Bernoulli}(\psi). \tag{2.1}$$

The latent variable z_i indicates whether individual i belongs to the study population N, i.e., enters the study in the first stratum and follows the Bernoulli distribution with probability ψ. With the probability ψ, we can calculate the expected value of N by

$$E(N) = M \cdot \psi. \tag{2.2}$$

We can obtain the estimate of the population size by

$$\hat{N} = \sum_{i=1}^{M} z_i, \tag{2.3}$$

which is also equal to $M \cdot \hat{\psi}$.

For the recapture process, we let z_i^b be a latent variable for second state process indicating whether individual i will experience severe symptoms with conditional probability ω,

$$z_i^b | z_i \sim \text{Bernoulli}(z_i \omega). \tag{2.4}$$

According to the disease progression published on the Government of Canada website (Government of Canada, 2023), individuals can only become hospitalized due to the

experience of severe symptoms, which is the condition for individuals to become available for recapture. We denote by s_i the time of individual i from being COVID-19 symptomatic to showing severe symptoms. This time follows a geometric distribution with probability of developing the severe symptoms θ

$$s_i \sim \text{Geometric}(\theta). \tag{2.5}$$

Combined with the variable z_i^b and s_i, we can generate the state matrix for the recapture process, which we denote by $\boldsymbol{b}$:

$$b_{ij}|zi, z_i^b, s_i = z_i \cdot z_i^b \cdot \Big[\underbrace{0 \quad \cdots \quad 0}_{s_i \text{ times}} \quad \underbrace{1 \quad \cdots \quad 1}_{k-s_i \text{ times}} \Big]. \tag{2.6}$$

Since the observation process is limited to at most k capture strata, we restrict the value of s_i to be the minimum of k and the s_i drawn by the geometric distribution above. The observation process in this study can be displayed in the form of the encounter histories of individuals' electronic health records. We generate two observation matrices, $\boldsymbol{h}^{\text{lab}}$ and $\boldsymbol{h}^{\text{hosp}}$, corresponding to capture and recapture processes. We denote by $\boldsymbol{h}^{\text{lab}}$ the encounter histories of individuals' first positive lab result, where

$$h_{ij}^{\text{lab}} = \begin{cases} 1 & \text{if individual } i\text{'s first positive PCR sample collected at stratum } j, \\ 0 & \text{otherwise,} \end{cases}$$

and $\boldsymbol{h}^{\text{hosp}}$ denotes the encounter histories of the individuals' first hospital admission related to COVID-19 treatment,

$$h_{ij}^{hosp} \begin{cases} 1 & \text{if individual } i \text{ is hospitalized for COVID-19 treatment at stratum } j, \\ 0 & \text{otherwise.} \end{cases}$$

We note that lab tests can only occur on or after the symptom onset date (equivalently, the first stratum), and hospitalization can only occur after experiencing severe symptoms. In other words, hospitalization can only take place after the first testing stratum, i.e., the second hospitalization stratum, to leave sufficient time for individuals to develop symptoms. Therefore, for the initial time, i.e., stratum $j = 1$, the encounter history of the lab test is only conditional on the latent variable z_i. We

describe the initial capture process as follows:

$$h_{i1}^{\text{lab}}|z_i \sim \text{Bernoulli}(z_i p_a), \tag{2.7}$$

where p_a is the probability of an individual infected with COVID-19 being tested and having a positive lab test result. Based on the assumption that the encounter history of hospital admission can only occur from the second stratum, we set the initial recapture process as follows:

$$h_{i1}^{\text{hosp}}|z_i, z_i^b = 0. \tag{2.8}$$

For stratum j ($j \geq 2$), the encounter history of lab test and hospital admission is described as follows:

$$h_{ij}^{\text{lab}}|z_i, h_{ij-1}^{\text{lab}}, h_{ij}^{\text{hosp}} \sim \text{Bernoulli}\left(z_i \prod_{l=1}^{j-1}(1-h_{il}^{\text{lab}}) \prod_{l=1}^{j}(1-h_{il}^{\text{hosp}})(1-p_a)^{j-1}p_a\right), \tag{2.9}$$

$$h_{ij}^{\text{hosp}}|z_i, z_i^b, b_{ij}, s_i, h_{ij-1}^{\text{hosp}} \sim \text{Bernoulli}\left(z_i \cdot z_i^b \cdot b_{ij} \cdot \prod_{l=1}^{j-1}(1-h_{il}^{\text{hosp}})(1-p_b)^{j-s_i-1}p_b\right), \tag{2.10}$$

where p_b is the probability of an individual infected with COVID-19 being hospitalized for treatment of COVID-19. The probability of detecting in the tagging stratum j given that this individual has been COVID-19 symptomatic is $(1-p_a)^{j-1}p_a$ for $j \geq 2$. Similarly, the probability of detecting in the recapture stratum j given that this individual has shown severe COVID-19 symptoms at stratum s_i is $(1-p_b)^{j-s_i-1}p_b$ for $j \geq s_i + 1$. In other words, h_{ij}^{lab} and h_{ij}^{hosp} can also be structured following geometric distributions. We set the condition that the observation captured at stratum j relies on the observation at stratum $j-1$ since we only focus on the first time of individuals being tested and being hospitalized and do not measure the distribution of the following encounter histories. We introduce an exception for patients who take the lab test after being admitted to the hospital. We assume that these patients take the tests directly in the hospitals rather than external labs, so we do not count these lab test encounter histories as parts of the tagging process, which results in h_{ij}^{lab} conditional on h_{ij}^{hosp}.

2.2 Bayesian approach

The Bayesian Markov chain Monte Carlo (MCMC) approach is useful for parameter estimation. We denote by π_α the prior distribution for the parameter α and obtain the full posterior joint distribution from Equation (2.11).

$$f(\psi, \omega, \theta, p_a, p_b | \boldsymbol{h}^{\text{lab}}, \boldsymbol{h}^{\text{hosp}}) = \tilde{\mathcal{L}} \cdot \pi_\psi \cdot \pi_\omega \cdot \pi_\theta \cdot \pi_{p_a} \cdot \pi_{p_b}, \tag{2.11}$$

where $\tilde{\mathcal{L}}$ is the likelihood function over the data and is given in Equation (2.12),

$$\begin{aligned}\tilde{\mathcal{L}} = \prod_{i=i}^{N} \Big\{ & \text{Bernoulli}(\psi) \cdot \text{Bernoulli}(\omega) \cdot \text{Geometric}(\theta) \cdot \text{Bernoulli}(p_a) \\ & \cdot \prod_{j=2}^{k} \text{Bernoulli}\Big(\prod_{l=1}^{j-1} (1 - h_{il}^{\text{lab}}) \prod_{l=1}^{j} (1 - h_{il}^{\text{hosp}})(1-p_a)^{j-1} p_a) \Big) \\ & \cdot \prod_{j=2}^{k} \text{Bernoulli}\Big(z_i \cdot z_i^b \cdot b_{ij} \cdot \prod_{l=1}^{j-1} (1 - h_{il}^{\text{hosp}})(1-p_b)^{j-s-1} p_b \Big) \Big\}.\end{aligned} \tag{2.12}$$

2.2.1 Prior distribution

The Bayesian MCMC algorithm requires us to specify prior distributions for each parameter in the model. The summary of the prior distributions for each parameter is shown in Equation (2.13).

$$\begin{aligned} \pi_\psi &= \text{Uniform}(0,1) & \pi_{p_a} &= \text{Beta}(1,1) \\ \pi_\omega &= \text{Uniform}(0,1) & \pi_{p_b} &= \text{Beta}(1,1) \\ \pi_\theta &= \text{Beta}(1,1) \end{aligned} \tag{2.13}$$

In this study, we assume that we do not know any prior information about the parameters, so we apply the non-informative uniform prior for all the parameters (Beta(1,1) is equivalent to Uniform(0,1)). We note that for the parameters of θ, p_a and p_b, they are the parameters of geometric distributions from Equation (2.5), (2.9) and (2.10). Therefore, if we obtain any prior information in further studies, conjugate beta prior could be a good choice which has support from 0 to 1 and are very flexible as prior distributions (Banner et al., 2020).

2.2.2 Partially observed data

In this study, the observation matrices provide information about individuals' presence for testing or hospitalization, while the true underlying infectious states are not fully observed. Similar to observation history, we use 0s and 1s to represent whether they belong to the study population or the group of severe patients. In the Bayesian MCMC approach, we derive the true infectious state and severe symptom state from the observation matrices and treat them as the partially observed data: (1) If individual i is not observed in both testing and hospitalization, i.e., the matrices $\boldsymbol{h}^{\text{lab}}$ and $\boldsymbol{h}^{\text{hosp}}$ are both 0s within the observation period, we replace the value of the latent variable z_i from 0 to NA and input the initial value equal to 0. (2) If individual i is not observed in the hospitalization regardless of his lab test results, i.e., the matrix $\boldsymbol{h}^{\text{hosp}}$ is 0s within the observation period, we substitute the value of the second state process z_i^b from 0 to NA and set the initial value equal to 0.

2.3 Simplified Model

The model stated in Chapter 2.1 requires data from both encounter histories and state information. We note that identification issues for the parameters will occur if we fail to provide the necessary data. In our model, observation matrices $\boldsymbol{h}^{\text{lab}}$ and $\boldsymbol{h}^{\text{hosp}}$ as well as state variable $\boldsymbol{z}$ are essential to obtain the target population size N, while the state matrix $\boldsymbol{b}$ (i.e., information about severe symptoms $\boldsymbol{s}$) is not available for some databases.

If the information corresponding to severe COVID-19 symptoms is missing, the Bayesian model will estimate the values of parameters θ and p_b from only the distribution of observation matrix $\boldsymbol{h}^{\text{hosp}}$, which results in obtaining implausible estimates of parameters θ and p_b. Therefore, a simplified model structure is required to fit the data without symptom severity information. We will further discuss an example of this identification issue in Chapter 3.2.2. From Equation (2.5) and (2.10), we note that both variables of s_i and h_{ij}^{hosp} follow the geometric distribution. We remove the measurement of severe symptoms and only focus on the distribution of time from symptom onset to hospitalization. Thus, we redefine with parameter p_b as the probability of being captured by hospitalization given that individual i has already shown COVID-19 symptoms. The value of the newly defined p_b will be close to the product of θ and p_b from the original model. Under the simplified model, since the measure-

ment of s_i from Equation (2.5) is removed, we also simplify the second state process for recapture by only measuring whether the individual will experience severe symptoms with $\boldsymbol{z}^b$ and remove the original state matrix for recapture process $\boldsymbol{b}$. Therefore, we describe the observation process $\boldsymbol{h}^{\text{hosp}}$ for the stratum $j \geq 2$ in Equation (2.14).

$$h_{ij}^{\text{hosp}} | z_i, z_i^b, h_{ij-1}^{\text{hosp}} \sim \text{Bernoulli}\left(z_i \cdot z_i^b \cdot \prod_{l=1}^{j-1} (1 - h_{il}^{\text{hosp}})(1 - p_b)^{j-1} p_b \right) \tag{2.14}$$

2.4 Compute Solution

In the Northern Health Authority case study, we gathered the data provided by the BC Centre for Disease Control (BCCDC) and the Ministry of Health via Population Data BC. Access to data provided by the Data Stewards is subject to approval but can be requested for research projects through the Data Stewards or their designated service providers. The following data sets were used in this study: consolidation - registry, COVID-19 Testing, and hospital separations (DAD). You can find further information regarding these data sets by visiting the PopData project webpage at: `https://my.popdata.bc.ca/project_listings/21-016/collection_approval_dates`. All inferences, opinions, and conclusions drawn in this publication are those of the author(s), and do not reflect the opinions or policies of the Data Steward(s).

We implemented the models in R (R Core Team, 2022) using the NIMBLE package (de Valpine et al., 2017) on the Population Data BC server. In the model simulation study, we generated simulation data and applied the models using the Digital Research Alliance of Canada platform(`alliancecan.ca`). Code for the Northern Health Authority case study is available at `https://github.com/siying-ma/COVID-DiseaseModel-SP.git`.

Chapter 3

Case Study

3.1 Data modality

We obtained COVID-19 data for the Northern Health Authority region, British Columbia, Canada from February 6, 2020 (the earliest date of COVID-19-related records in the data sources) to December 31, 2020 (Government of Canada, 2024). During this period, the total number of observed cases in the Northern Health Authority region was relatively small, which allows for fast parameter estimation (Parker, M.R.P. et al., 2021). We used a unique patient identifier to link the testing and hospitalization data from two data sources. Since the BCCDC datasets do not contain the information for severe symptoms, we applied the simplified model defined in Chapter 2.3 for the following analysis.

There were 1,880 individuals with medical records indicating either an external positive lab test result or a COVID-19-related hospitalization. We identified patients with in-patient treatments through COVID-19-related diagnostic codes (U071 and U072) defined by the Canadian Institute for Health Information (2023). Among the 1,880 individuals, 1,384 of them (73.62%) contain full information, while the remaining 496 ones (26.38%) did not contain information about symptom onset dates. For the 1,384 individuals with symptom onset dates, we aggregated their capture histories, $\boldsymbol{h}^{\text{lab}}$ and $\boldsymbol{h}^{\text{hosp}}$, by including their earliest positive lab test and earliest hospital admission after symptomatic onset date respectively. Consequently, there were 1,301 individuals (94.00%) exclusively sampled by lab testing, no patient (0.00%) exclusively sampled through hospitalization and 83 patients (6.00%) with both a lab test and hospitalization record. While 16 of the 83 had a lab test followed by hospital

admission, the remaining 67 patients were tested after hospital admission.

In the stratified-Petersen experiment, the two sampling sites of tagging and recapture processes must be spatially distinct to reduce potential bias resulting from population movements. Since the testing data from Population Data BC did not specify testing sites, we assumed that individuals who did COVID-19 tests after hospitalization were likely to be tested as part of the hospital admission process and did not have external lab tests. Therefore, we considered these patients as the ones only captured from hospitalization and excluded their lab test records from the capture history, $\boldsymbol{h}^{\text{lab}}$. Similarly, for the 496 individuals without recorded symptom onset dates, we observed 420 individuals exclusively from lab testing, 46 patients exclusively from hospitalization and 30 patients from both lab testing and hospitalization. Within the 30 patients captured by both lab tests and hospitalization, 6 of them took the lab tests after admission to hospitals and were further considered as the ones only captured from hospitalization. Since we cannot distinguish these individuals from asymptomatic individuals, we considered all of the 496 individuals as ones with missing symptom onset dates. Consequently, we input a sequence of NAs in the rows of observation matrices, $\boldsymbol{h}^{\text{lab}}$ and $\boldsymbol{h}^{\text{hosp}}$, for these 496 individuals. In the Bayesian MCMC algorithm, missing data are considered random variables whose posterior distributions can be obtained by the same prior distributions on the parameters (Ma and Chen, 2018). The latent variables, $\boldsymbol{z}$ and $\boldsymbol{z}^b$, were thus generated based on individuals' electronic health records by the rules described in Chapter 2.2.2.

We aligned the data for the 1,384 individuals without missing data by setting their symptom onset date as the first capture event. In this study, we set the cutoff point for our observation period to be 21 days after the case symptom onset dates. Figure 3.1 shows the distribution of time for study individuals from being symptomatic to taking the first lab test as well as to getting the first hospital admission. From Figure 3.1, the test volume remains high within the first 6 days from showing symptoms, starts to decline sharply from the seventh day, and stays close to zero from the fifteenth day. However, the number of observed hospitalizations is much smaller than the number of tests. The distribution of time from symptom onset to first hospitalization is relatively flat. In general, the number of hospitalized cases gradually increased during the first 8 days and then started to decrease. We note that the number of hospitalized cases is less than or equal to 3 cases per day after the fifteenth day. Therefore, we assume that a 21-day observation period is sufficient for capturing the pattern of COVID-19 information for this study.

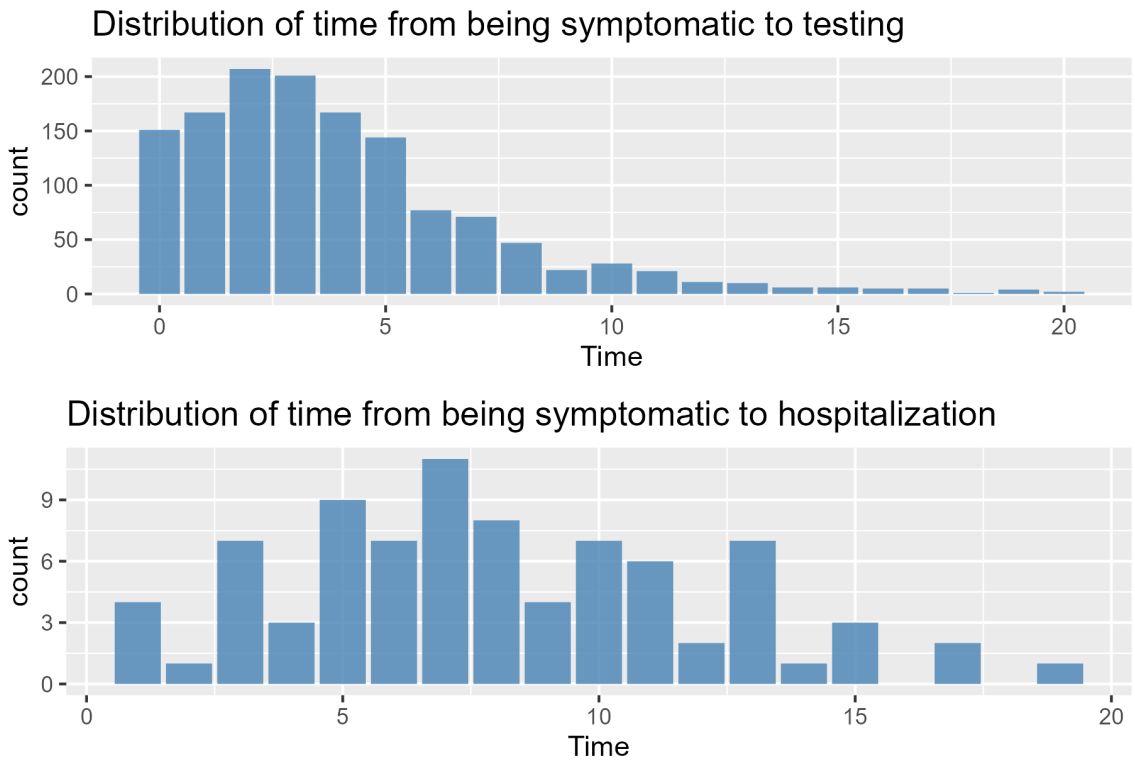

Figure 3.1: Distribution of time (days) from being symptomatic to test/hospitalization from the observed data $\boldsymbol{h}^{\text{lab}}$ and $\boldsymbol{h}^{\text{hosp}}$ from 2020-02-06 to 2020-12-31

Infectious disease data often has underreporting or delayed reporting (Stoner et al., 2023), especially during the early stage of a pandemic. Also, during the early stage of a pandemic, data collection is imperfect. Either a severe delay or data collection error may result in obtaining impossible disease progression records. One of the most common such situations is an unreasonable gap between symptom onset date and testing date. For example, an individual's electronic health record may show that they took a lab test on September 1, 2020, but their reported symptom onset date was March 10, 2020. We dealt with these records as poorly reported data, excluded them from capture histories and set the observation matrix as missing data.

3.2 Results

3.2.1 Fitted model results

We estimated parameters using a Bayesian MCMC approach. To perform inference with the Bayesian models, we fit the model to the data with 1 chain and 50,000 iterations of the adaptation period. We set 10,000 initial samples as the burn-in and ran 40,000 additional iterations to obtain the posterior estimates.

We set the augmented dataset with the size, M, equal to the double size of the observed dataset, $3,760$. We fit the simplified model. For patients with lab test or hospitalization records but without valid symptom onset dates, since we input NAs for their observation data, we set two constraints in the MCMC algorithm so that they must have a corresponding test or hospitalization record in their observation matrices during the observation period in each iteration. We estimated the true population size of COVID-19 cases from 2020-02-06 to 2020-12-31 as 2,967 with a 95% credible interval (CI) of (2,878, 3,059). The posterior parameter estimates are displayed in Table 3.1.

Table 3.1: Parameter estimates of the Bayesian simplified model fitted to the Northern Health Authority COVID-19 data from 2020-02-06 to 2020-12-31.

Parameter	Mean	Median	St.Dev.	95% CI
ψ	0.789	0.789	0.014	(0.761, 0.817)
ω	0.106	0.105	0.011	(0.087, 0.130)
p_a	0.113	0.113	0.005	(0.104, 0.122)
p_b	0.060	0.060	0.010	(0.041,0.080)

We note that the estimate of ψ itself is not informative, but based on Equation (2.2), we can obtain the expected value of the estimated population size $\hat{N}$ through the following equation:

$$E(\hat{N}) = M \cdot \hat{\psi} = 3760 \cdot 0.789 = 2966.64.$$

This value is identical to the posterior mean of the true population size. The estimate of the probability of severe symptom indicates that 10.6% of symptomatic individuals will further experience severe COVID-19 symptoms and be admitted to the hospital for COVID-19 treatment. The estimate of probability of lab testing capture, $\hat{p_a}$, shows that symptomatic individuals or potential asymptomatic individuals have a

higher possibility (56.8% with 95% CI of (53.6%, 59.8%)) of testing within 7 days from first showing symptoms or exposure to COVID-19, and the test volume will decrease gradually after 7 days, which is aligned with the data shown in Figure 3.1. We note that the estimate of probability of hospitalization capture, $\hat{p_b}$, is very low, which further addresses the reason that the trend of time from being symptomatic to hospitalization is relatively flat (Figure 3.1).

3.2.2 Bayesian simulation study

Posterior predictive validation is a method widely used in Bayesian analysis to identify discrepancies between the observed data and results obtained from the fitted model. This method is also used to validate the model efficiency (Gelman et al., 2003). In this study, to evaluate the model's capability of providing accurate parameter estimates, we conducted a Bayesian simulation study in which the true values of model parameters were the estimates from the Northern Health Authority experiment in Chapter 3.2.1. We applied the model defined in Chapter 2 to generate simulated encounter histories and used partially observed simulated datasets to obtain the estimates. To measure the discrepancies, we took an example of simulated dataset selected at random and compared the distribution of simulation data with the actual encounter histories. Significant differences between the observed and simulated data represent that the model may not effectively capture the observed patterns. Second, we ran the Bayesian algorithm with this simulation data and compared the estimates with the true parameter values. Finally, to further validate the model adequacy, we repeated the Bayesian simulation study 100 times and observed coverage rates of the target and parameters. Both the full and simplified models defined in Chapter 2.1 and 2.3 were studied.

Under the Simplified Model

We generated an observed dataset using the simplified model described in Chapter 2.3 and set the values of model parameters equal to the mean of the parameter estimates obtained by the data from the Northern Health Authority case study in Chapter 3.2.1. Thus, the true population size N was set to 2,967, the augmented population size M was set to 3,760, the probability of developing severe symptoms ω was set to 0.1057, the probability of lab test capture p_a was set to 0.1128, and the probability of hospitalization capture p_b was set to 0.0599. To obtain a more accurate simulation, we

expanded the observation period to 28 days, meaning that we had 28 sampling events in the simulation study. As there were some individuals in the case study data who did not have a valid symptom onset date, we randomly selected 30% of individuals from our simulation dataset and treated their symptom onset dates as missing.

To provide an example of our simulated data compared to the case study data, we randomly selected one simulation dataset (seed: 8144392) and compared the result obtained from the Bayesian model with the case study data. We observed 1,936 individuals either from testing or hospitalization in this simulated dataset, while 1,845 of them tested positive for COVID-19 and 164 of them had the records from hospitalization. The comparison of the distribution of observed data $\boldsymbol{h}^{\text{hosp}}$ and $\boldsymbol{h}^{\text{lab}}$ with the true distribution from Population Data BC is shown in Figure 3.2.

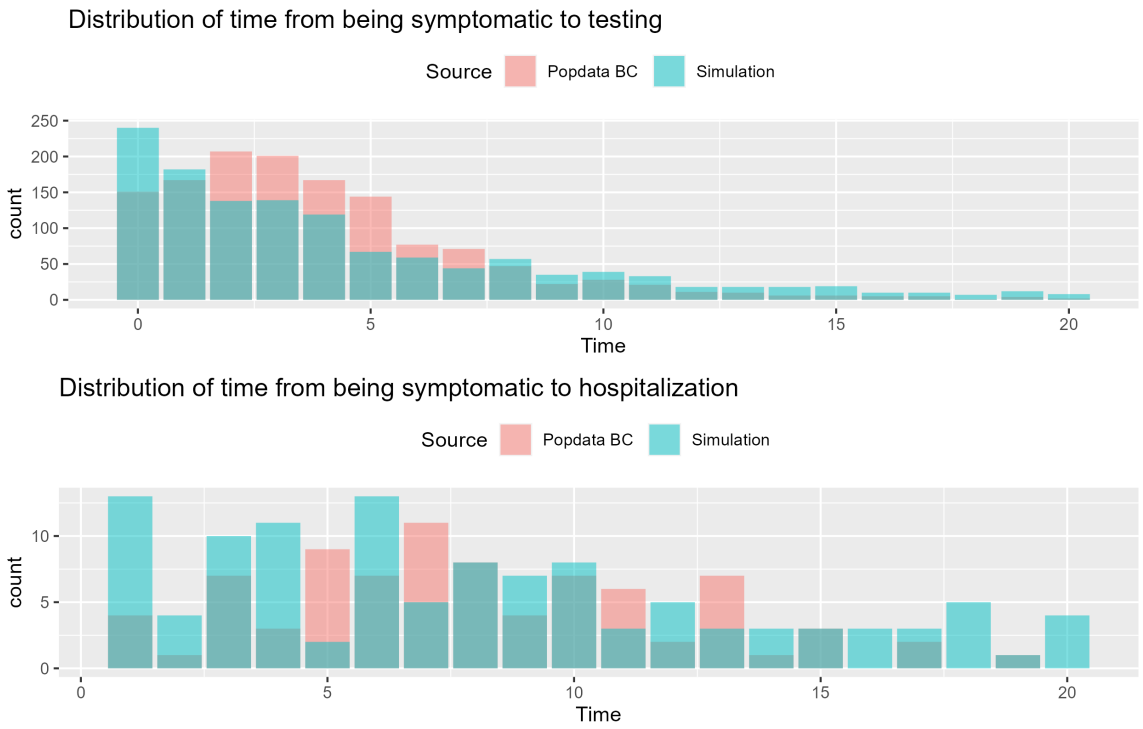

Figure 3.2: Comparison of the distribution of time (days) from being symptomatic to test/hospitalization from the simulated observed data (blue) $\boldsymbol{h}^{\text{lab}}$ (top) and $\boldsymbol{h}^{\text{hosp}}$ (bottom) with that of the Northern Health Authority region case study data (red).

Applying the Bayesian model to this simulated dataset we found that the estimated posterior mean value of population size $\hat{N}$ is equal to 2,979 with 95% CI of (2,896, 3064). The estimated value of N is very close to the true population size and falls into the range of the 95% credible interval. The estimates of the parameters are

shown in Table 3.2

Table 3.2: Parameter estimates of the Bayesian simplified model fitted with the one example simulation dataset.

Parameter	Mean	St.Dev.	95% CI
ψ	0.792	0.013	(0.766, 0.818)
ω	0.102	0.009	(0.085, 0.122)
p_a	0.110	0.004	(0.102, 0.119)
p_b	0.054	0.008	(0.038, 0.071)

We found that the true value of parameters all fell in the 95% credible interval, indicating good model performance and efficiency.

To further evaluate the model performance, we repeated the simulation study 100 times independently and calculated the mean of 100 posterior parameter estimates (Table 3.3). The coverage probability for the 100 credible intervals of the targeted population size N and probability of developing severe symptoms are 0.97 and 0.95 respectively, indicating strong model performance. For the other two parameters corresponding to capture probabilities, the coverage probabilities were smaller. The reason for the smaller coverage probabilities of p_a and p_b is related to the significant amount of missing data (30% of the total observed population).

Table 3.3: Average parameter estimates, coverage probability, bias, and root mean square error of the Bayesian model simulation study under the simplified model with 100 independent replicates.

Parameter	Mean	Coverage Probability	Bias	Root Mean Square Error
N	2969.45	0.97	2.45	43.15
ω	0.110	0.95	0.005	0.010
p_a	0.113	0.85	< 0.001	0.004
p_b	0.056	0.85	-0.004	0.011

Under the full model

As we discussed in Chapter 2.3, the value of p_b in the simplified model should be close to the product of θ and p_b in the full model defined in Chapter 2.1. According to work on COVID-19 disease progression (Government of Canada, 2023), symptoms may

often appear 2-14 days after exposure to the virus, and severe symptoms can occur anytime after showing symptoms. However, timely in-patient medical treatments are strongly recommended by the BCCDC for patients experiencing severe symptoms. Therefore, we assume that the combination of a relatively large value of p_b and a relatively small value of θ is reasonable for the true values of parameters in the full model while the product of θ and p_b is equal to 0.06. In this case, we set the true value of θ to 0.15 and the true value of p_b to 0.4. Similar to our simulation results for the simplified model, we observed 1,955 individuals from the simulation dataset, while 1,750 of them were detected exclusively from testing, 119 of them were detected exclusively from hospitalization and 86 patients were observed with both of the records. From the model results, we obtained the estimated population size $\hat{N}$ as 2,970 with 95% credible interval (2,892, 3,052). The estimates of parameters are shown in Table 3.4.

Table 3.4: Parameter estimates of the Bayesian full model fitted with simulated data.

Parameters	Mean	St.Dev.	95% CI
ψ	0.790	0.013	(0.765, 0.815)
ω	0.102	0.007	(0.088, 0.116)
θ	0.157	0.011	(0.135, 0.179)
p_a	0.117	0.003	(0.111, 0.124)
p_b	0.414	0.023	(0.370, 0.460)

Since the estimated population size (2,970) is close to the true population size (2,967), and true values of model parameters fall in the 95% credible interval, we conclude that the full model has good model performance.

We ran the model with the same dataset again, but this time we removed information about severe symptoms. This was done to validate the need for the simplified model. We obtained the estimated value of population size N equal to 2,969 with 95% credible interval (2,891, 3,049), which indicates that the estimate of our target population size is still accurate even without data related to severe symptoms. The estimates of parameters are shown in Table 3.5.

The estimate of testing capture probability, $\hat{p_a}$, is still significant since the estimate of p_a is only related to the lab test observation, $\boldsymbol{h}^{\text{lab}}$ (Table 3.5). However, we found that the estimates of parameters related to hospitalization are implausible. The estimated value of θ is 0.975 with 95% credible interval (0.912, 0.999), and the estimated value of p_b is 0.089 with 95% credible interval (0.073, 0.106). Both of the

Table 3.5: Parameter estimates of the Bayesian full model fitted with a simulated dataset without information about severe symptoms.

Parameters	Mean	Median	St.Dev.	95% CI
ψ	0.789	0.789	0.013	(0.765, 0.814)
ω	0.114	0.114	0.008	(0.099, 0.130)
θ	0.975	0.982	0.023	(0.912, 0.999)
p_a	0.117	0.117	0.003	(0.111, 0.124)
p_b	0.089	0.089	0.009	(0.073, 0.106)

true values of θ and p_b did not fall in the credible intervals and the estimates were not significant. Therefore, this demonstrates the necessity of applying the simplified model for datasets in which information about severe symptoms are lacking.

We then repeated the simulation study (without missing data) under the full model 100 times independently and calculated the mean of the 100 parameter estimates to further evaluate model performance (Table 3.6). The coverage probability of 100 credible intervals of the target population size N and all parameters are above 0.9. Particularly, the coverage probability for the population size is 0.97, indicating good model performance.

Table 3.6: Average parameter estimates, coverage probability, bias and root mean square error of the Bayesian model simulation study under the full model with 100 independent replicates.

Parameter	Mean	Coverage Probability	Bias	Root Mean Square Error
N	2962.72	0.97	-4.28	38.48
ω	0.106	0.92	< 0.001	0.008
p_a	0.113	0.96	< 0.001	0.003
p_b	0.403	0.96	0.003	0.023
θ	0.153	0.94	0.003	0.012

Chapter 4

Discussion

During the height of the SARS-CoV-2 pandemic, case enumeration was a difficult problem due to under-reporting. Population size estimation is crucial to understand the severity of COVID-19, and to understand future pandemics. We developed a Bayesian state-space capture-recapture model based on the stratified-Petersen estimator and applied this disease dynamics model to estimate the population size of COVID-19 cases in the Northern Health Authority region of British Columbia in 2020. We incorporated two confidential individual-level datasets from the BC Centre for Disease Control and the Ministry of Health via Population Data BC that contain lab test and hospitalization information. Our estimate of the size of the COVID-19 case population ($\hat{N} = 2,967$) is 1.58 (95% CI: (1.53, 1.63)) times greater than the observed cases ($n_{obs} = 1,880$), with an estimated detection rate of 0.63 (95% CI: (0.62,0.65)). This means that only 63.4% of individuals infected with COVID-19 in the Northern Health Authority region had electronic health records indicating a positive lab test results or COVID-related hospitalizations.

Our estimated population size aligns with the findings of other studies. Parker, M.R.P. et al. (2021) obtained similar results using public case count data, and Olobatuyi et al. (In Preparation) achieved extremely similar estimates using a Susceptible-Infectious-Recovered multi-event capture-recapture (SIR-MECR) model using the same datasets and observation period, further validating our model.

Beyond the population size, other parameter estimates provide novel insights into the province's COVID-19 situation. The probability of severe symptoms (ω) is a direct reflection of disease severity and provides an effective way of measuring the need for hospital resources. In the case of the Northern Health Authority, the allocation of hospitalization resources caused by COVID-19 in 2020 is approximately 10.6% of the

true COVID-19 population size. The estimation of parameters θ, p_a and p_b provides a model of disease progress by taking into consideration the aspects of COVID-19 symptoms, disease diagnosis (lab test) and treatment (hospitalization). Note that we assumed θ, p_a, and p_b all follow geometric distributions. Since θ is a latent variable, we cannot obtain any information from the encounter histories. However, data of time from being symptomatic to lab testing or hospitalization are partially observed in our case study, which can provide some insights on the distribution of p_a and p_b. From Figure 3.2, we can find that in general, observed data from Population Data BC has a similar distribution to our simulated data generated under geometric distributions. However, it is possible that more accurate estimates can be produced under other distributions.

Our model requires individual-level datasets. In contrast to publicly available datasets, individual-level datasets are confidential and require both ethics approval and privacy training to access. However, our results show that individual-level data can offer more precise estimates for both the population size and model parameters. Our coverage probabilities for population size in the simulation study are 0.97 under both the simplified and full models even with missing data. The coverage probabilities are higher than the ones obtained from the model using count data (Parker, M.R.P. et al., 2021), which speaks to the advantage of using individual-level data. Moreover, our model produced narrower credible intervals for all parameters, showing that the estimated values are stable.

Another advantage of our Bayesian model compared to traditional stratified-Petersen estimators is that our model has the potential to incorporate covariates. This can by done by using logit links as in Williams et al. (2002). Several factors have been shown to have significant relationships with the disease progress of COVID-19. Romero Starke et al. (2021) showed that the risk of hospitalization increased by 3.4% per age year and obtained high confidence of evidence of the increase in COVID-19 disease severity due to age. Pshenichnaya et al. (2022) found that the chance of hospitalization has a significant increase for patients with certain comorbidities (by 1.496 times with oncology, by 1.502 times with cardiovascular pathology, etc.). We expect to improve our estimates by incorporating covariates such as age and gender in future studies.

In our model, we collected individuals' daily events in a temporal stratification. As for the stratified-Petersen estimator, it is possible to have overlaps between tagging and recapture strata in a temporal stratification, provided that the transition from

capture to recapture is ensured. For example, in fisheries, recapture events typically do not begin until several weeks after tagging has begun (Schwarz and Taylor, 1998). However, in the case of COVID-19, it is possible for individuals to have both lab tests and hospital admission within the same week. This results in an important limitation of our Bayesian model in that we must apply our model to daily data. If we want to use the traditional stratified-Petersen estimator to analyze the true population size in a year, we must have more than 350 strata. It is difficult to obtain estimates with such small bin widths. For example, because of this model limitation, Dao (2023) used 7-day widths. The best solution for getting the population size in a long-term period is to align the data based on symptom onset dates. This reduces the number of strata. Note that a necessary assumption for such data aggregation is that the distribution of COVID-19 disease progression is similar during the course of the observation period. Another benefit for aggregating the data is to ensure there are enough observations in both the capture and recapture processes in order to provide accurate estimates.

In future studies, we propose to address non-homogeneity of disease progression by splitting the data into 6 subsets based on the BC recovery plan phases in 2020 (BC Centre for Disease Control, 2021). This requires us to extend our model to multiple cohorts. We also intend to analyze the influence of potentially related factors, such as age, by incorporating them as covariates of our Bayesian model.

Bibliography

Banner, K. M., Irvine, K. M., and Rodhouse, T. J. (2020). The use of Bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889.

BC Centre for Disease Control (2021). British columbia (bc) covid-19 situation report - week 1: January 3 – january 9, 2021. `http://www.bccdc.ca/Health-Info-Site/Documents/COVID_sitrep/BC_COVID-19_Situation_Report_Jan_15_2021.pdf`.

Byambasuren, O., Cardona, M., Bell, K., Clark, J., McLaws, M., and Glasziou, P. (2020). Estimating the extent of asymptomatic covid-19 and its potential for community transmission: Systematic review and meta-analysis. *Journal of the Association of Medical Microbiology and Infectious Disease Canada*, 5(4):223–234.

Canadian Institute for Health Information (2023). Covid-19: Locating the icd-10-ca/cci code [job aid]. *Ottawa, ON: CIHI*.

Dail, D. and Madsen, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, 67(2):577–587.

Dao, V. (2023). Estimating the prevalence of the COVID-19 pandemic in the Vancouver Island Health Authority region using the stratified Petersen model. Master's project, Department of Mathematics and Statistics, University of Victoria.

Darroch, J. N. (1961). The two-sample capture-recapture census when tagging and sampling are stratified. *Biometrika*, 48(3/4):241–260.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Dougherty, B., Smith, B., Carson, C., and Ogden, N. (2021). Exploring the percentage of covid-19 cases reported in the community in canada and associated case fatality ratios. *Infectious Disease Modelling*, 6:123–132.

Gandhi, R., Lynch, J., and Del Rio, C. (2020). Mild or moderate covid-19. *New England Journal of Medicine*, 383(18):1757–1766.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Press, Boca Raton, Fl.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., and Kretzschmar, M. E. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*, 14(1):147.

Gimenez, O., Rossi, V., Choquet, R., Dehais, C., Doris, B., Varella, H., Vila, J.-P., and Pradel, R. (2007). State-space modelling of data on marked individuals. *Ecological Modelling*, 206(3):431–438.

Government of Canada (2023). Covid-19: Symptoms, treatment, what to do if you feel sick. `https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.html`. Date accessed: 2023-01-27.

Government of Canada (2024). Covid-19 epidemiology update: Summary. `https://health-infobase.canada.ca/covid-19/`. Date accessed: 2024-05-01.

Jonsen, I. D., Flemming, J. M., and Myers, R. A. (2005). Robust state–space modeling of animal movement data. *Ecology*, 86(11):2874–2880.

Jonsen, I. D., Myers, R. A., and Flemming, J. M. (2003). Meta-analysis of animal movement using state-space models. *Ecology*, 84(11):3055–3063.

King, R. (2012). A review of Bayesian state-space modelling of capture-recapture-recovery data. *Interface Focus*, 2(2):190–204.

Kéry, M. and Schaub, M. (2012). *Bayesian Population Analysis using WinBUGS: A Hierarchical Perspective*. Academic Press.

Link, W. A. and Barker, R. J. (2010). *Bayesian Inference: with ecological applications*. Academic Press, London, UK.

Ma, Z. and Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, 47(3):297–313.

Newman, K. B., Fernández, C., Thomas, L., and Buckland, S. T. (2009). Monte carlo inference for state–space models of wild animal populations. *Biometrics*, 65(2):572–583.

Olobatuyi, K., Ma, J., Brown, P., and Cowen, L. (In Preparation). Multi-event dynamic capture-recapture model for Big Data: Estimating undetected COVID-19 Cases in British Columbia, Canada.

Parker, M.R.P., Cao, J., Cowen, L.L.E., Elliott, L.T., and Ma, J. (2024). Multi-site disease analytics with applications to estimating COVID-19 undetected cases in Canada. *Annals of Applied Statistics*, in print.

Parker, M.R.P., Li, Y., Elliott, L.T., Ma, J., and Cowen, L.L.E. (2021). Underreporting of COVID-19 in the Northern Health Authority region of British Columbia. *Canadian Journal of Statistics*, 49(4):1018–1038.

Plante, N., Rivest, L.-P., and Tremblay, G. (1998). Stratified capture-recapture estimation of the size of a closed population. *Biometrics*, 54(1):47–60.

Pshenichnaya, N., Lizinfeld, I., and Zhuravlev, G. (2022). Factors influencing on hospitalization of covid-19 patients with comorbidity. *International Journal of Infectious Diseases*, 116:S39.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robles, S. C., Marrett, L. D., Aileen, C. E., and Risch, H. A. (1988). An application of capture-recapture methods to the estimation of completeness of cancer registration. *Journal of Clinical Epidemiology*, 41(5):495–501.

Romero Starke, K., Reissig, D., Petereit-Haack, G., Schmauder, S., Nienhaus, A., and Seidler, A. (2021). The isolated effect of age on the risk of covid-19 severe outcomes: a systematic review with meta-analysis. *BMJ Global Health*, 6(12):e006434.

Royle, J. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, J. A. (2008). Modeling individual effects in the Cormack-Jolly-Seber model: A state-space formulation. *Biometrics*, 64(2):364–70.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, San Diego, California.

Schwarz, C. and Taylor, C. (1998). Use of the stratified-Petersen estimator in fisheries management: estimating the number of pink salmon (*Oncorhynchus gorbuscha*) spawners in the Fraser River. *Canadian Journal of Fisheries and Aquatic Sciences*, 55(2):281–296.

Schwarz, C. J. (2023). *SPAS: Stratified-Petersen Analysis System*. R package version 2023.3.31.

Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics*, 42(2):267–292.

Skowronski, D. M., Sekirov, I., Sabaiduc, S., Zou, M., Morshed, M., Lawrence, D., Smolina, K., Ahmed, M. A., Galanis, E., Fraser, M. N., Singal, M., Naus, M., Patrick, D. M., Kaweski, S. E., Mill, C., Reyes, R. C., Kelly, M. T., Levett, P. N., Petric, M., Henry, B., and Krajden, M. (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*.

Stoner, O., Halliday, A., and Economou, T. (2023). Correcting delayed reporting of COVID-19 using the generalized-Dirichlet-multinomial method. *Biometrics*, 79(3):2537–2550.

Williams, B., Nichols, J., and Conroy, M. (2002). *Analysis and Management of Animal Populations*. Elsevier Science.

Xu, Y., Fyfe, M., Walker, L., and Cowen, L.L.E. (2014). Estimating the number of injection drug users in greater Victoria, Canada using capture-recapture methods. *Harm Reduct Journal*, 11:9.