AN ABSTRACT OF THE DISSERTATION OF

Ben Brintz for the degree of Doctor of Philosophy in Statistics presented on March 1, 2018.

Title: A Normal Approximation to N-Mixture Models with Applications in Large Abundance Estimation and Disease Surveillance

Abstract approved: _____

Lisa Madsen Claudio Fuentes

N-mixture models provide a structure for making inference about a local population size while accounting for imperfect detection. Using a binomial likelihood, they assume prior distributions on the size parameters and then integrate those parameters out of the full likelihood. For large population sizes, the established frequentist methods have exhibited computational intractability, and the Bayesian methods have exhibited poor convergence and mixing of chains. Additionally, estimability of parameters in these types of models has been criticized in the literature. Although originally used for determining abundance of rare wildlife, we explore using these models for under-diagnosed or under-ascertained infectious diseases which have large prevalence.

We derive an asymptotic approximation of the N-mixture model that does not suffer from computational efficiency and uses information theory to provide a method for diagnosing estimability issues. Additionally, we extend this model to account for spatial

dependency. Simulation studies show improved performance over the established methods in numerous settings, and we successfully apply the asymptotic approximation to model ten years of Oregon Health Authority chlamydia data.

A Normal Approximation to N-mixture Models with Applications in Large Abundance Estimation and Disease Surveillance

by

Ben Brintz

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented March 1, 2018
Commencement June 2018

Doctor of Philosophy dissertation of Ben Brintz presented on March 1, 2018.

APPROVED:

Co-Major Professor, representing Statistics

Co-Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Ben Brintz, Author

ACKNOWLEDGEMENTS

I would like to take this opportunity to extend my gratitude to Professors Lisa Madsen and Claudio Fuentes for advising me throughout my Ph.D and to Professors Charlotte Wickham, Yuan Jiang, and Jeffrey Bethel for their contributions as part of my committee. I would also like to give a special appreciation to Dr. Lauren Brooks, my love, for her utmost confidence in me and for always reminding me to celebrate my achievements.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES (Continued)

A Normal Approximation to N-mixture Models with Applications in Large Abundance Estimation and Disease Surveillance

1 Introduction

Modelling and estimating discrete abundances when individuals are imperfectly detected has various applications including surveillance for both wildlife and epidemiology. In wildlife biology, we may be interested in how many hard-to-spot animals live within an area in order to avoid altering their habitat. In epidemiology, we conduct disease surveillance to determine how many resources we should put towards a disease that is under-detected due to under-ascertainment and hidden populations. We also consider that these individuals are moving in and out of the regions we surveil. Understanding the true abundance of these populations can change how we understand and behave towards the imperfectly detected individuals. Our goal is to utilize publicly available data to obtain better estimates of disease prevalence by using models that consider the possibility of imperfect detection.

In order to estimate imperfectly detected abundances, we sought to utilize the ideas which stem from research that addresses the binomial distribution's unknown size N and probability p estimation problem. Blumenthal and Dahiya (1981) give a comprehensive look on estimating the parameter N of the binomial distribution on the basis of r independent observations and present a modified maximum likelihood estimate for unknown probability p that avoids a negative estimate that can occur if the method of moments

estimator is used. They propose that given the likelihood

$$L(N,p) = \left\{ \prod_{i=1}^{k} \binom{N}{s_i} \right\} p^{\Sigma_i^k s_i} (1-p)^{kN - \Sigma_i^k s_i} \tag{1.1}$$

where N represents the total size, s_i is the i^{th} observation, that there be a prior density on p proportional to

$$p^a (1-p)^b \tag{1.2}$$

the kernel of a beta distribution with integers a and b. They multiply (1.1) and (1.2) and then maximize over the joint likelihood. Carroll and Lombard (1985) address the instability of estimates by discounting data for which p would apparently be near zero using the same process of multiplying (1.1) and (1.2). They then eliminate the nuisance parameter p by integrating the joint likelihood over p and maximize the marginal likelihood for N. Their research explores the stability of the result based on chosen values of a and b. DasGupta and Rubin (2005) also comment on the fundamental difficulties of the problem due to the instability of N under slight perturbations of one or two sample values. Yet they call the Carroll and Lombard (1985) estimate "the best estimate available to date" despite proving the nonexistence of unbiased estimates for N or p, citing the severe underestimation of N. Subsequently, Royle (2004) established what we call "N-mixture models" by popularizing the idea of putting a Poisson prior on N. This assumes there is a natural Poisson process guiding the true abundance that we are interested in rather than having to make the difficult choice of choosing an a and b for the beta density on p. Royle (2004) applies this model to estimating abundance of the North American robin by using data gathered by repeated counts of robins spotted by a single observer across 50 sites and

over 10 days. The model provides a framework for including covariates which influence p rather than treating it as a nuisance parameter. For instance, we may know that the population of humans living near a site may affect the detection of birds. In this case, we could model $\text{logit}(p) = \beta_0 + \beta_1 \cdot population_i$. However, the model that Royle (2004) introduced required a *closed population*, i.e., he assumes there is no movement into or out of the sites within the time period observed. The likelihood to maximize is then simply

$$L(p, \theta | \{n_{it}\}) = \prod_{i=1}^{R} \left\{ \sum_{N_i = max_t n_{it}}^{\infty} \left[\prod_{t=1}^{S} \text{Bin}(n_{it}; N_i, p) \right] f(N_i; \theta) \right\}. \tag{1.3}$$

where θ is the parameter of $f(N_i; \theta)$, $\text{Bin}(n_{it}; N_i, p)$ is the binomial likelihood. In the case of the robins example, R represents the number of sites, S the number of visits, N_i is the total number of robins in site i, n_{it} is the observed number of robins in site i at time t and p is the detection probability. In practice, the sum to infinity must be approximated with some large number k. k is sufficiently large once it reaches a magnitude that increasing it will no longer change the likelihood by more than a chosen threshold. In the case of many animal studies and in the application of disease surveillance over years, we would not expect the closed-population assumption to be satisfied. Dail and Madsen (2011) modified the model of Royle (2004) by allowing dynamic movement of the population between sites of interest. The model, called the Generalized N-mixture Model, induces a temporal correlation structure between the visits of a site using survival and recruitment parameters while assuming independence between sites. However, a literature review revealed there were parameter estimability issues with this model when there is only weak dynamics of each site's population abundance between time periods (**?**). In this case, weak dynamics refers to a lack of autocorrelation between the visits of a single site. The closed population

model assumes you have the same abundance at multiple times so the correlation between visits at each site is always one and the parameters are estimable. We explore this idea further in Chapter 2. We have also quantified this estimability using an asymptotic information method proposed in Chapter 3 in which we assume a multivariate normal marginal likelihood for the observed abundances. We derive the asymptotic information for its parameters and obtain the asymptotic correlation matrix for maximum likelihood estimates of the parameters λ and p, showing a relatively low correlation between the two parameters. In contrast, an increased magnitude of correlation between parameter estimates is observed in models that assume an open metapopulation such as Dail and Madsen (2011) where

$$L(p,\theta|\{n_{it}\}) = \prod_{i=1}^{R}\left\{\sum_{N_{i1}=n_{i1}}^{\infty}\cdots\sum_{N_{iS}=n_{iS}}^{\infty}\left[\prod_{t=1}^{S}\text{Bin}(n_{it};N_{it},p)\right]f(\{N_{i1},N_{i2},\cdots,N_{iS}\};\theta)\right\}. \tag{1.4}$$

S_{it} and G_{it} denote the "survivors" (*i.e.* the number of the N_{it-1} population units that remain at time t) and the "gains" (*i.e.* the number of immigrants at time t), at site i and time t. We have that $N_{it} = S_{it} + G_{it}$, and we can construct the prior $f(\{N_{i1},N_{i2},\cdots,N_{iS}\};\theta\}$ using a hierarchical structure. Typically, $\theta = (\lambda,\omega,\gamma)$ so that

$$N_{i1} \sim \text{Poisson}(\lambda)$$

$$S_{it}|N_{it-1} \sim \text{Bin}(N_{it-1},\omega)$$

$$G_{it}|N_{it-1} \sim \text{Poisson}(\gamma\cdot N_{it-1})$$

for $i = 1,..,R$ and $t = 2,...,S$ where λ is the initial abundance parameter and ω and γ are

interpreted as survival and recruitment dynamic parameters respectively. In addition to the estimation problems with weak dynamics cited in the literature, the integrating out the random variables $\{N_{ij}\}$ leads to computational intractability with the larger abundances which are common in reported infectious diseases. The multiple summations to infinity must be approximated by summing to many large values of k, which can become large with observed abundances in the order of hundreds.

Despite the importance of knowing the true abundance of an imperfectly detected population in its various applications, accomplishing this estimation is a challenge that is not fully understood. In this dissertation I explore and contribute novel ideas to this problem in three parts. In Chapter 2, I describe many of the approaches we explored in order to address abundance estimation with Bayesian methods. In Chapter 3, I propose an asymptotic approximation to the N-mixture model and a method for diagnosing the performance of estimators using asymptotic information theory. In Chapter 4, I propose a spatial extension to the asymptotic N-mixture model which models spatial correlation among neighboring sites. Lastly, in Chapter 5, I present our conclusions on the contributions of the proposed asymptotic models as well as suggest some issues that can be addressed in further research.

2 Bayesian Estimation of Imperfectly Detected Abundances

In order to combat the issues of estimation when there is weak dynamics, or a lack of autocorrelation between the visits of a single site, we aimed to develop a model that could be as close to the closed population model of Royle (2004) as possible while still allowing for a temporal correlation structure. Our intuition for why the simplest multi-visit closed model works is that, given multiple visits, with λ as the initial abundance parameter for all sites, $E[n_i] = V(n_i) = \lambda \cdot p$ and $\text{cov}(n_{i,t}, n_{i,u}) = \lambda \cdot p^2$ so the parameters are estimable, most intuitively, if you consider using sample mean, variance and covariance estimates as method of moment estimators for the combined parameters. Once estimates for these combined parameters are obtained, they can be used to get estimates for the individual parameters. With more observed data from added sites, we obtain more information about λ and p. If we assume a different initial abundance λ_i for each site i, we can rely only on within site data to estimate λ_i while still gaining more information for p with the addition of sites. However, in the open model, each observation does not provide the same information about the parameters because of the structure of the expected mean and covariance which differs from the closed model due to the dynamics parameters. We intuited that if we could model an autocorrelation close to one between the visits of each site while keeping the same mean and variance structure of Royle (2004), we would have just slightly less information about the parameters than in the closed model. This is similar to the idea of effective sample size, the notion that we have a reduced sample size if our

sample's observations are correlated, except we are using the logic that the correlation between parameters suggests we have more than one observation per abundance parameter.

Our idea was to create a likelihood using a Gaussian copula so that we could maintain the discreteness of the abundance parameters $\{N_{it}\}$, while modeling correlation between them. We decided to use a Bayesian method to fit the model because of the likelihood's complexity as well as the possibility of extensions with even more complexity. Additionally, this allowed us to estimate the $\{N_{it}\}$ directly, rather than integrate them out. Furthermore, the Bayesian method would allow us to estimate much larger abundances which lined up with our goals to apply this model to diseases. In this chapter, we present the original conceived model and the adaptations that followed as we frequently encountered issues of extreme estimates. We end with a discussion on addressing the inherent issues with the model and how they lead us to the successful contributions made to N-mixture models in the later chapters.

2.1 Model

In order to address identifiability issues resulting from weak dynamics and find a model more suitable to disease count data while incorporating a temporal correlation component, we explored a new model which we fit using a Metropolis-Hastings in Gibbs Algorithm. Given that the model has abundance parameters N_{it} where $i = 1, \ldots, R$ sites and $t = 1, \ldots, S$ visits plus the parameters that model dependence between visits, we incorporated the Random Sweep Gibbs Sampler presented in **?**, an update scheme which guarantees the reversibility of the chain and intuitively prevents the update order of correlated parameters

from affecting the posterior estimates. The Random Sweep algorithm is:

For each iteration, generate random permutation σ where $\sigma_1, \cdots, \sigma_\delta$ indicates random order of the original δ parameters. Then:

1. $Y_{\sigma_1} \sim g(Y_{\sigma_1}|Y_{\sigma_{-1}})$
2. $Y_{\sigma_2} \sim g(Y_{\sigma_2}|Y_{\sigma_{-2}})$

$\vdots$

δ. $Y_{\sigma_\delta} \sim g(Y_{\sigma_\delta}|Y_{\sigma_{-\delta}})$

where $\sigma_{-\nu}$ indicates all parameters in the permutation but parameter ν, Y_{σ_w} represents the proposed value of parameter σ_w, and g represents the full conditional distribution.

Because the full conditional for each parameter, except detection p and initial abundance λ, do not represent a known posterior distribution, we use the Metropolis-Hastings (M-H) accept-reject algorithm to draw from the parameters' posterior distributions proposed by the random sweep algorithm. Specifically:

1. Propose a move to y from density $q(y|x)$ where x is the current value of the chain Y
2. Calculate $r = \frac{t(y)q(x|y)}{t(x)q(y|x)}$ where $t()$ is the target distribution
3. Accept the new proposal with probability min(1,r), $Y^{t+1} = y$; otherwise $Y^{t+1} = Y^t$

In order for the chains to converge to the posterior distribution quickly, the Metropolis-Hastings proposals need to reflect the structure of the data. In these particular applications, there can be a within-site correlation of the abundance parameters. As such, for the

multivariate parameter $\{N_{i1},...,N_{iS}\}$ for $i = 1,\ldots,R$, we use a Gaussian copula proposal. The copula allows us to induce correlation onto a vector of discrete random variables. The correlation matrix Σ for each multivariate Gaussian proposal is specified as the sample within-site Spearman rank correlation of the 1000 most recent iterations in the chain. The marginal distribution for each individual abundance parameter in a site is a random discrete uniform with lower bound $(N_{ij}^{t-1} - \text{move})$ and upper bound $(N_{ij}^{t-1} + \text{move})$ where move is chosen such that the variance of the discrete uniform proposal distribution is equal to the sample variance of that parameter's chain.

However, due to the non-symmetry of the copula proposal, the probability density function needs to be computed as part of the M-H algorithm (because $q(x|y) \neq q(y|x)$)which can also be computationally intensive. While this probability density function (PDF) can be calculated exactly using 2^S independent calculations of a multivariate normal cumulative distribution function (CDF) in **?**, this becomes intractable as more visits occur per site. In contrast, the optimized method of Genz and Bretz (2002) outlined in **?**, is a method which only requires S integrals calculated simultaneously and can be accomplished efficiently using a multivariate normal CDF package provided in both R and Python (**?**). Therefore, a Bayesian Metropolis-Hastings in Gibbs algorithm can reasonably be utilized to simulate draws from the posterior distributions of the parameters of interest. The model is shown in Table 2.1.

Table 2.1: New Model

R = number of sites

S = number of visits

$N_{11}, N_{12}, ..., N_{RS}$ represent the R · S abundance parameters where $N_{11}, N_{12}, \ldots, N_{1S}$ are the abundance parameters for site 1.

$p = Pr\{\text{case detected}\}$ - probability of detection

$\lambda_i = E\{N_{i1}\}$ for $i = 1, ..., R$ - expected initial abundance for site i

ρ = correlation between visits

Model: (for $i = 1, ..., R$ and $j = 1, ..., S$)

- $n_{ij} \sim \text{binom}(N_{ij}, p)$
- $p \sim \text{unif}(0, 1)$
- $N_{ij} \sim \text{MVtrpois}_{\Sigma_\rho}(\lambda_i)$
- $\lambda_i \sim \text{gamma}(.0001, 1000)$
- $\Sigma_\rho \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{S-2} & \rho^{S-1} \\ \rho & 1 & \ddots & \ddots & & \rho^{S-2} \\ \rho^2 & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & \rho^2 \\ \rho^{S-2} & & \ddots & \ddots & \ddots & \rho \\ \rho^{S-1} & \rho^{S-2} & \cdots & \rho^2 & \rho & 1 \end{bmatrix}$
- $\rho \sim \text{unif}(0, 1)$

Note: MVtrpois_Σ represents a multivariate distribution with truncated Poisson marginals and correlation induced by a Gaussian copula with dependence matrix Σ.

$$f(N_{it}|\lambda_i, \rho) = \prod_{i=1}^{R} \sum_{j_1=1}^{2} \cdots \sum_{j_S=1}^{2} (-1)^{j_1+\cdots+j_S} \Phi_{\Sigma_\rho}[\Phi^{-1}\{u_{i1j_1}\}, \cdots, \Phi^{-1}\{u_{iSj_S}\}]$$

where $u_{it1} = F_{it}(N_{it})$ and $u_{it2} = F_{it}(N_{it} - 1)$ and F_{it} is the Poisson CDF with rate parameter λ_i truncated by the corresponding n_{it}. Σ_ρ has an AR(1) structure.

An extension of the correlation structure naturally allows for spatial similarities using one extra parameter for adjacent sites. If we consider a model with only two visits ($t = 1,2$) and two independent sites ($i = 1,2$) the copula correlation matrix of $(N_{11}, N_{12}, N_{21}, N_{22})$ would be:

$$\Sigma = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}$$

where ρ is the Spearman correlation between N_{i1} and N_{i2}. However, if we wanted to allow for spatial covariance between the sites, the correlation matrix changes to:

$$\Sigma = \begin{bmatrix} 1 & \rho & \delta & 0 \\ \rho & 1 & 0 & \delta \\ \delta & 0 & 1 & \rho \\ 0 & \delta & \rho & 1 \end{bmatrix}$$

where δ represents the Spearman correlation between abundances N_{1t} and N_{2t} and where we assume independence between N_{it} and $N_{i't'}, t \neq t'$. This structure can be expanded so as to include S time points and R sites allowing for spatio-temporal correlation in the likelihood without the use of latent random variables. Modeling the spatial dependence will increase the dimension of the parameter space.

We found that estimation of the parameters was possible if the correlation parameter ρ was fixed and known. As a preliminary test to compare the effectiveness of the Bayesian Dail/Madsen model versus our new model (NM) from Table 2.1, we conducted

a simulation (R=50 and S=10) with dependence within sites and imperfect detection. For simplicity, we set all parameters constant except $\{N_{it}\}$ and p. Figure 2.1 shows the full traceplots for parameter p resulting from 10,000 iterations of the Gibbs Sampler with a M-H algorithm.

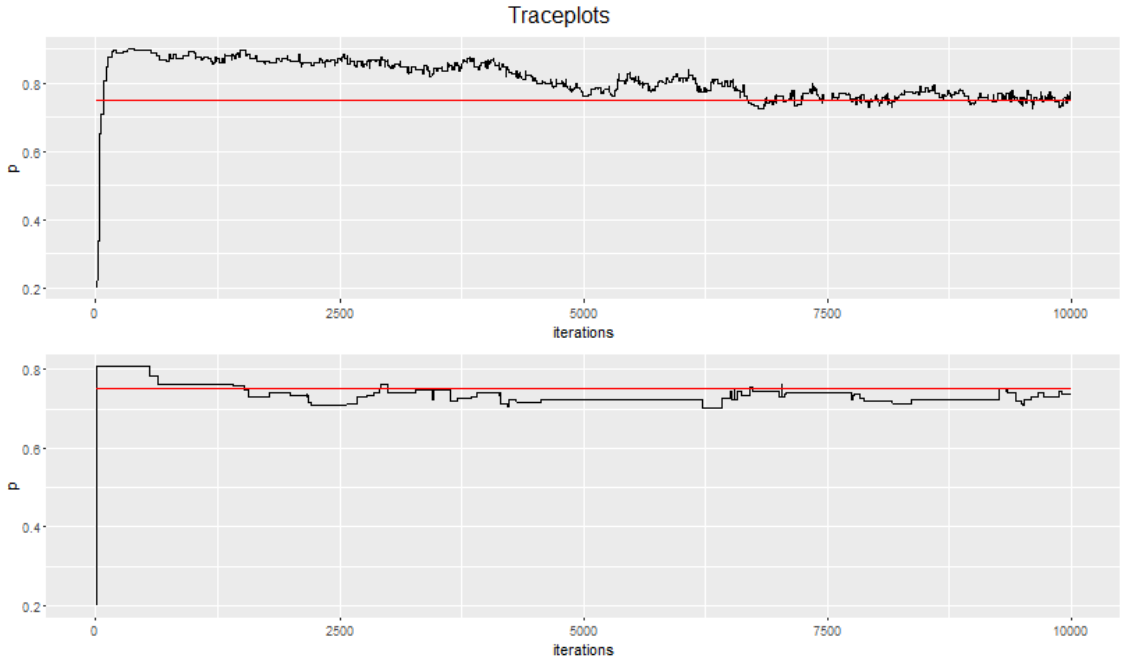

Figure 2.1: Original traceplots for the detection probability p from simulated results of Dail/Madsen (top) and NM (bottom) respectively with the horizontal line as true detection parameter

Both chains for the detection parameter jump above the true value and it does appear that the Dail/Madsen model accepts proposals at a much higher rate. However, the NM plot suggests that the chain converges to the true parameter much more quickly (within 1000 iterations) while the Dail/Madsen model appears to take between 6000 and 7000 iterations.

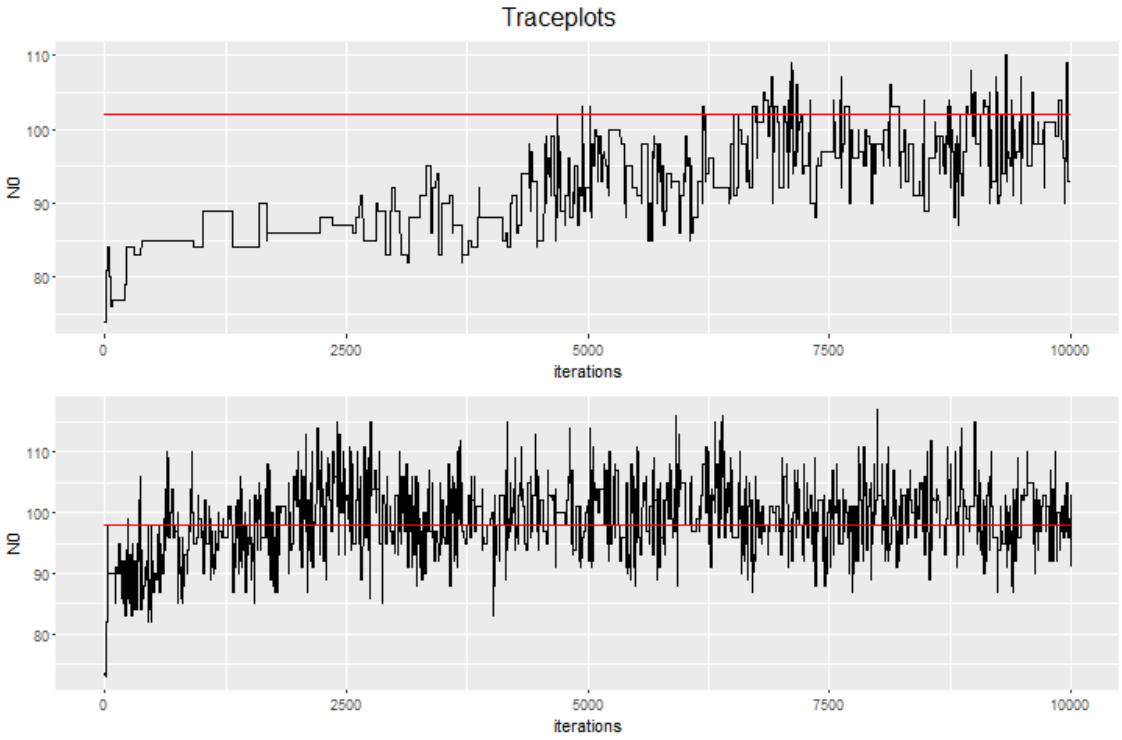

Figure 2.2: Original traceplots from simulated results of Dail/Madsen and NM respectively with dotted lines as the true simulated abundance

Figure 2.2 shows traceplots for one of the abundance parameters, $N0$, for both the DM and NM. It suggests that the NM efficiently attains convergence to the desired posterior abundance in the first site. The other sites look very similar to this one. The true abundance value is 98 and the traceplot achieves stationarity around that true value. In contrast, the Dail/Madsen model does not appear quite as efficient and has not achieved an unbiased posterior median after 10,000 iterations. Later simulations showed that DM model struggles with mixing and convergence of chains to a stationary distribution.

However, using this AR1 structure proved not to work very well when the correlation parameter ρ was estimated instead of fixed. We observed that the estimates for λ_i and

p would go off to opposite extremes during the MCMC due to a lack of a single global maximum in the joint likelihood for the data. **?** suggests this behavior is a consequence of nonidentifiability issues which we have also called estimability issues. We will show an example of the chain for p going to the value one later in this Chapter.

2.2 Copula Prior Implementation

Implementing a copula prior likelihood for high-dimensional discrete data is computationally intensive. Similar to the copula proposal, we originally implemented the Gaussian copula prior using the methods presented in Section 4 of **?** and its continuous extension by **?**. This method requires the calculation of 2^n probabilities where n is the number of observations so it is computationally limited. Ultimately, we implemented the copula prior using the method found in **?** by Genz which only requires n integrations to obtain the copula CDF. While this method does have computational efficiency limitations of its own, it increased the speed greatly over the original implementation.

2.3 Exchangeable Structure

In order to make this model close to that of Royle (2004), we decided to maximize the amount of correlation between the $\{N_{it}\}$ by using an exchangeable structure, a correlation structure which assumes equal correlation between all time points. We posited that we could make the correlation very close to 1 and then slowly lower it to figure out when the model broke down. However, in most cases we observed that the chains for the detection parameter p and $\{N_{it}\}$ were unable to converge to the truth and went off to the extreme

values allowed by the priors. An example of this behavior is shown in Figure 2.3 where we can see that the initial value chosen for p (0.81) was close to the truth (0.80) but after 5000 iterations the chain cascades to an extreme value with a mean close to one.

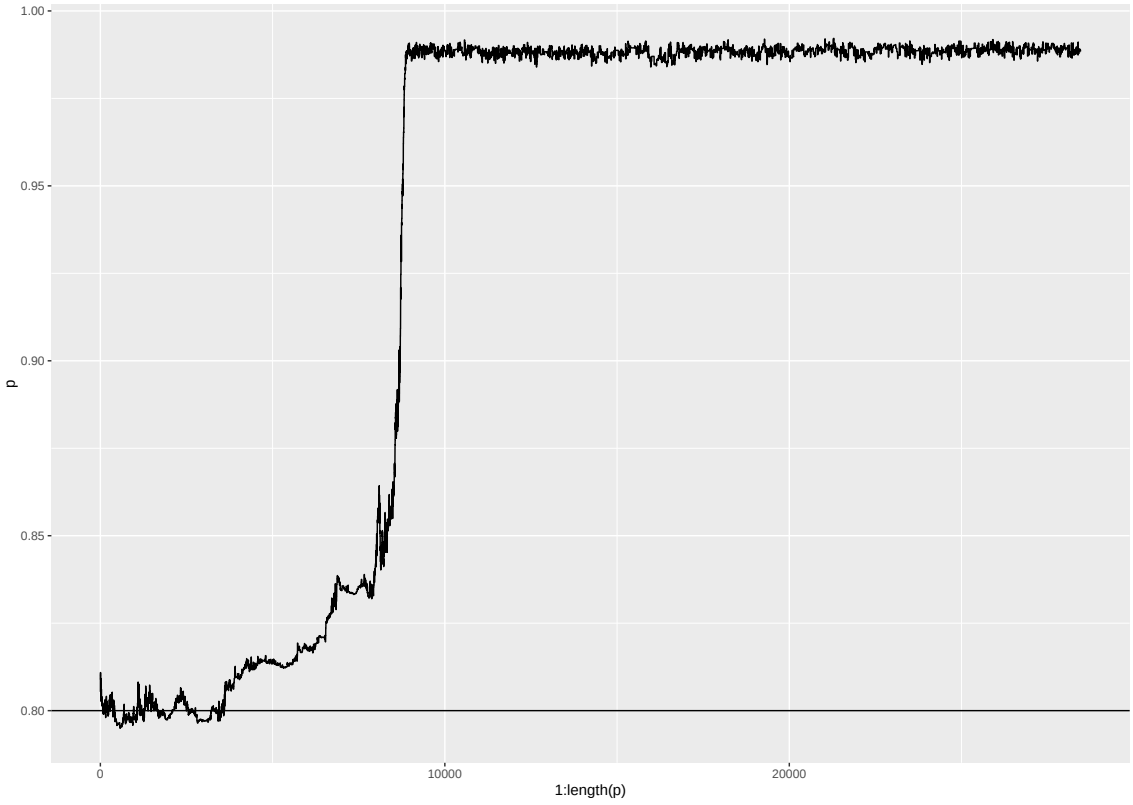

Figure 2.3: Gibbs sampling chain for the detection parameter p when the true value is 0.80

While using the exchangeable structure we tried various adaptive samplers, methods of proposal, and priors on the parameters in order to avoid the sampler from going to extremes but we were not successful in addressing this issue with the Gaussian copula based model.

2.3.1 Adaptive Metropholis-Hastings and Proposals

We considered that one reason the sampler was going off to extremes was the rate of acceptance of the Metropolis-Hastings proposal. **?** suggest that the selection of the proposal distribution as well as its variance is crucial to the statistical properties of the MCMC and that a poor choice of proposal distribution can result in poor performance of the Monte Carlo estimators, such as estimates with large variances. If the acceptance rate of a chain is too large, we expect that many proposals not in the true stationary distribution are being accepted, while if the acceptance rate of a chain is too small, we expect that many proposals in the true stationary distribution are not being accepted. Therefore, a moderate acceptance rate is desired in order to achieve convergence to the stationary distribution. Due to the connection between the proposal variance and acceptance rate, i.e. a higher variance results in a smaller acceptance rate and a smaller variance results in a higher acceptance rate, we tried increasing or reducing the variance on the proposal based on the variance of the last 1000 iterations of each parameter in the chain. We also explored an algorithm similar to that presented in Section 3 of **?**, an adaptive Metropolis-within-Gibbs which essentially adds or subtracts to the variance of the proposal based on the acceptance rate of the previous 50 iterations where the goal rate is about 44% of the proposals made for each parameter.

In addition to using copulas to define the likelihood of our N-mixture model, we used copulas to make within-site proposals for $\{N_{ij}\}$ in order to achieve convergence to a stationary distribution more quickly due to the proposals having the properties which we expect would maximize the likelihood. Discrete proposals tried were discrete uniform with both constant variance proposals and with proposals dependent on the variance of

the chain. We additionally tested proposals which followed the prior distributions used for $\{N_{ij}\}$ which we will discuss more in length in the next subsection. We used normal proposals for the λ and ρ parameters with both fixed variance and variance dependent on their chains.

2.3.2 Uninformative and Informative Priors

In order to maximize the utility of the Bayesian hierarchical structure and address issues of extreme estimates, we explored various forms of uninformative and informative priors for all parameters. As we presented in Section 2.1, we originally used a uniform(0,1) uninformative prior on p. We also explored using Jeffreys prior, a non-informative prior that is invariant to transformation, for p in the binomial distribution, a beta($\frac{1}{2}, \frac{1}{2}$) shown in Figure 2.4.

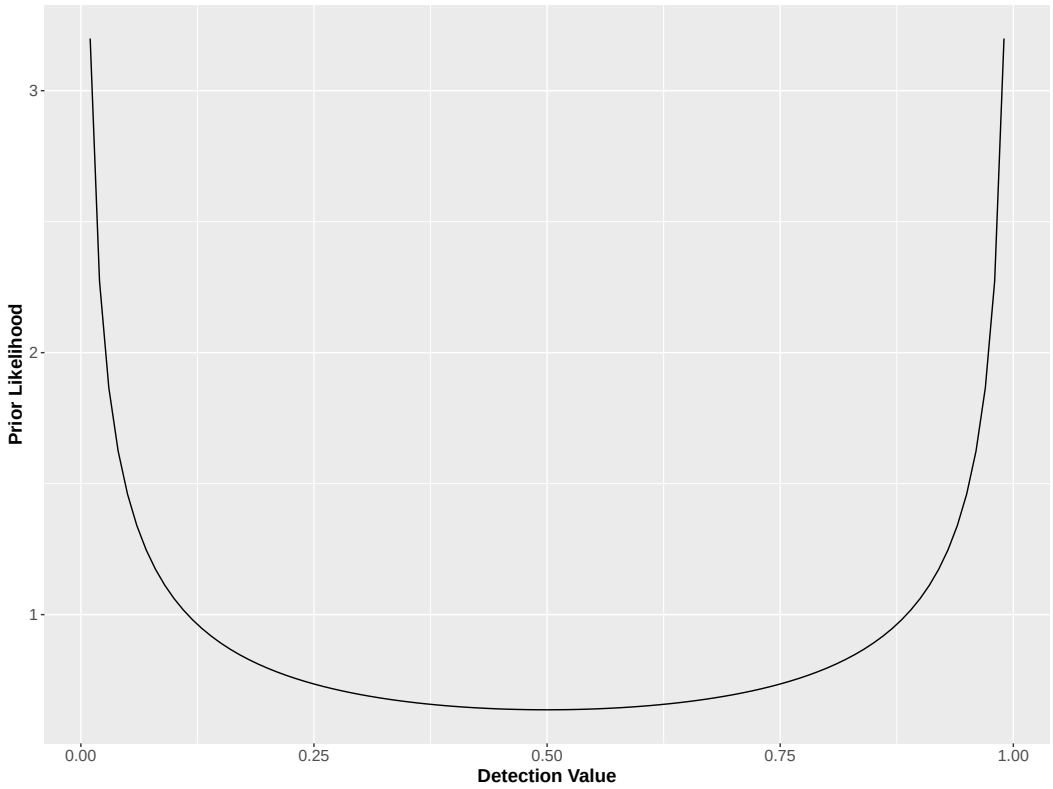

Figure 2.4: Density for Jeffrey's Prior for p from a binomial(N,p) likelihood

Given that our main issue was that p was cascading off to extremes, this prior did not make intuitive sense due to its high prior likelihood at the extremes. As with the uniform prior, the Jeffreys prior resulted in extreme estimates. We also tried informative priors for p to try to avoid going off to extremes such as beta(2,2) and other beta distributions with less variance or skewed away from the extreme of one or zero for p.

As such, we began looking at more informative priors for λ, p and $\{N_{it}\}$. One method

for informing priors for λ proposed by **?** assumes the following:

$$N_{i,t} \sim Poisson(\mu_i) \tag{2.1}$$

$$n_{i,t} \sim Poisson(\lambda_i) \tag{2.2}$$

$$\lambda_i = \mu_i \cdot p \tag{2.3}$$

Using this structure, we can use the data for each site to inform the prior on the λ_i rather than μ_i and then use the estimate of p to inform us on the rate μ.

Similar to **?** above, we wanted to use the observed data to inform our prior on correlation parameter ρ, much like the empirical Bayes method which uses the observed data to inform the prior presented by **?** . We determined that the observed correlation between n_{it} and $n_{it'}$ where $t \neq t'$ within each site would be $p \cdot \rho$. As a result, assuming equal correlation of counts between visits at each site, we could obtain confidence interval of $p \cdot \rho$ by calculating the sample $\text{cor}(n_{i1}, n_{i2}, \cdots, n_{iS})$ and determining the mean and variance of the resulting upper triangle. Instead of a prior on ρ, we put a uniform prior on $p \cdot \rho$ with the minimum and maximum at the minimum and maximum of the confidence interval, respectively. The correlation matrix for the Gaussian copula was then conditional on the $p \cdot \rho$ and p proposal parameters such that the matrix used would be exchangeable with parameter $\frac{p \cdot \rho}{p}$

We also tried to use priors to avoid the tendency for the detection probability to go to extremes. Though we called this an issue of identiability, we were still exploring the idea that it was just an issue of information in the likelihood when N and p are both unknown. If you consider a fixed N and only one visit, the information with respect to p is $Np(1-p)$.

This is maximized when p is at its extremes. This also means that the variance is the lowest at those extremes resulting in the likelihood having the largest peaks when p is at those extremes.

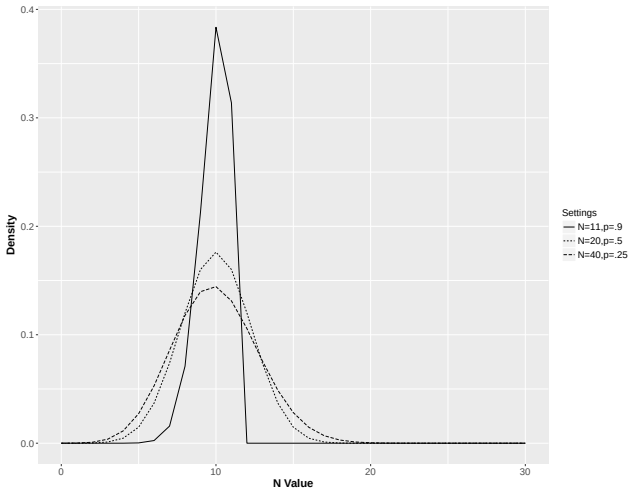

Figure 2.5: A comparison of the binomial density at various N and p settings with a similar expected value.

As can be seen in Figure 2.5, the binomial density is greatest when p and N are both at their extremes. As such, we posited that in a Metropolis-Hastings proposal style algorithm for unknown N and p would encourage chains that trend toward the extreme. Next, we considered using priors to fix this issue. By multiplying the binomial likelihood by $\sqrt{Np(1-p)}$, the likelihood at the expected value should be equalized as seen below in Figure 2.6.

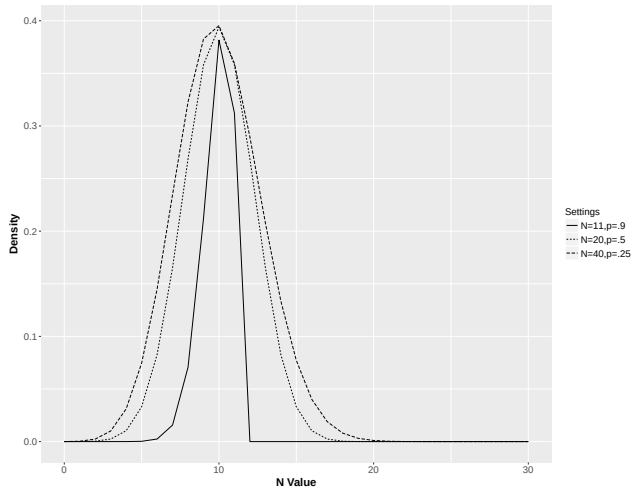

Figure 2.6: A comparison of the binomial density at various N and p settings with a similar expected value and with standardizing prior.

We considered that the MCMC algorithm could make better utilization of the correlation structure by standardizing the variance in the likelihood. The multivariate version of this is to multiply the likelihood by $\prod_{i=1}^{R}\prod_{j=1}^{S}\sqrt{N_{i,t}\cdot p(1-p)}$ which bears a resemblance to Jeffrey's prior for p when N is fixed.

In addition to priors on the other parameters, we tried using informative priors on the $\{N_{it}\}$ within the copula-induced correlation structure. For each marginal N_{it}, we put location shifted (to the observed $\{n_{it}\}$) gamma priors that were skewed positive. Given the negative correlation that we know exists between estimates of $\{N_{it}\}$ and p from the binomial distribution, we tried to account for that correlation in the Gaussian copula prior. By incorporating a $S+1$ dimensional correlation structure for each site we explored setting the $S+1^{st}$ row and column of the correlation matrix to a negative value and also tried encouraging the opposite relationship by setting a positive correlation in the prior likelihood. The effect seemed to delay the chain from going to the extreme but it inevitably did not

work.

2.4 Discussion

Although our explorations started with the successful components of Royle (2004)'s closed population model, the addition of the copula-induced correlation structure without the dynamics structure of Dail and Madsen (2011) lead to its failure. Inducing a less than perfect positive correlation between N_{it}'s using the copula structure did not have the intended effect of increasing the effective sample size and instead resulted in parameter nonidentifiability issues between λ, p and ρ. Without really specific priors (point masses), the parameter chains could never provide reasonable results given simulations settings with a high detection rate (e.g. $p = 0.80$) and correlation close to one (e.g. $\rho = 0.95$).

In Chapter 1 and in the beginning of this Chapter, we commented on the estimability of the parameters in Royle (2004) due to the expected mean and covariance in the model. Using that same reasoning, we magnify the fault in this model's parameterization. Disregarding the copula's imperfect modelling of correlation and assuming a constant initial abundance λ, we find that $E[n_i] = V(n_i) = \lambda p$ and $\text{cov}(n_{i,t}, n_{i,u}) = \lambda p^2 \rho$. Given the form of these properties, we can intuitively see why the parameters are not estimable. For instance, we could reasonably expect to estimate λp using sample mean and variances or $p\rho$ by dividing sample covariances by sample means, but we can not estimate these parameters individually. In this particular problem, the individual parameters are of utmost importance for obtaining overall abundance estimates.

Using the method which we propose and explore in Chapter 3, we can confirm this in-

tuition, showing that the three parameters of interest $[\lambda, p, \rho]$ are either perfectly positively or negatively correlated. In order to estimate these parameters in the Bayesian context, we would need very precise informative priors on at least one of these parameters in order to remove its confoundedness. For instance, if we put a strong prior on the ρ parameter, we could see similar success to the closed population model or similar to the example chain we provided in Figure 2.1. As a result, we will continue our exploration for models suitable for disease prevalence by starting with the more dynamic structure of the generalized N-mixture model of Dail and Madsen (2011).

3 An Asymptotic Approximation to the N-Mixture Model for the Estimation of Disease Prevalence

Submitted to Biometrics

3.1 Introduction

The estimation of abundance is an important component of both ecological and epidemiological applications. The generalized N-mixture model of Dail and Madsen (2011) allows estimation of abundance of an open metapopulation while accounting for imperfect detection. Given a set of observed counts $\{n_{it}\} = \{n_{11}, \ldots, n_{RS}\}$, the likelihood is defined as

$$L(\{N_{it}\}, p|\{n_{it}\}) = \prod_{i=1}^{R}\prod_{t=1}^{S} \text{Bin}(n_{it}; N_{it}, p), \tag{3.1}$$

where, $\text{Bin}(\cdot; M, \theta)$ denotes the density function of a Binomial distribution with parameters M and θ. Hence, the likelihood $L(\{N_{it}\}, p|\{n_{it}\})$ in (3.1) is simply the product of binomial densities where, for $i = 1, \ldots, R$ and $t = 1, \ldots, S$, the parameter N_{it} corresponds to the abundance at site i in the sampling period t, and p represents the detection probability. Since the abundance parameters $\{N_{it}\}$ are unknown, they are typically taken as random quantities, and therefore, given a prior density $f(\{N_{i1}, N_{i2}, \cdots, N_{iS}\}; \theta)$, the abundance

parameters can be integrated out from the joint likelihood as

$$L(p,\theta|\{n_{it}\}) = \prod_{i=1}^{R}\left\{\sum_{N_{i1}=n_{i1}}^{\infty}\cdots\sum_{N_{iS}=n_{iS}}^{\infty}\left[\prod_{t=1}^{S}\text{Bin}(n_{it};N_{it},p)\right]f(\{N_{i1},N_{i2},\ldots,N_{iS}\};\theta)\right\}. \tag{3.2}$$

Then, if S_{it} and G_{it} denote the "survivors" (*i.e.* the number of the N_{it-1} population units that remain at time t) and the "gains" (*i.e.* the number of immigrants at time t), at site i and time t we have that $N_{it} = S_{it} + G_{it}$, and we can construct the prior $f(\{N_{i1},N_{i2},\ldots,N_{iS}\};\theta\}$ using a hierarchical structure. Typically,

$$N_{i1} \sim \text{Poisson}(\lambda) \tag{3.3}$$

$$S_{it}|N_{it-1} \sim \text{Bin}(N_{it-1},\omega) \tag{3.4}$$

$$G_{it}|N_{it-1} \sim \text{Poisson}(\gamma N_{it-1}) \tag{3.5}$$

for $i = 1,\ldots,R$ and $t = 2,\ldots,S$ where λ is the abundance parameter and ω and γ are interpreted as survival and recruitment dynamic parameters respectively. In a more general setting the conditional recruitment parameter could be taken as any function of the previous time $t-1$ in site i, *e.g.* $G_{it}|N_{it-1} \sim \text{Poisson}(g(N_{it-1}))$. Obtaining a closed form expression for the likelihood in (3.2) is intractable, and in practice, it is approximated by an expression of the form

$$L(p,\theta|\{n_{it}\}) \approx \prod_{i=1}^{R}\left\{\sum_{N_{i1}=n_{i1}}^{K}\cdots\sum_{N_{iS}=n_{iS}}^{K}\left[\prod_{t=1}^{S}\text{Bin}(n_{it};N_{it},p)\right]f(\{N_{i1},N_{i2},\ldots,N_{iS}\};\theta)\right\}, \tag{3.6}$$

where K is chosen to be sufficiently large, subjected to computational constraints.

Since their introduction in Royle (2004), these models have been largely discussed in

the literature and a number of extensions have been proposed. For instance, **?** illustrate issues of parameter confounding in the single-visit model and suggest that any information about the p, ω, and γ parameters must come from a dynamic structure or auto-correlation between visits. **?** expands upon the model to allow for density dependency and environmental stochasticity in both the survival and recruitment parameters of the original model. On the other hand, **?** extends the class of models by proposing alternate dynamics using exponential or Ricker-logistic processes, random effects, and accounting for zero-inflation. **?** explores the effect that the choice of K has on the results of the generalized N-mixture model when the number of visits or the detection rate is low, and proposes a model which chooses K automatically, suggesting the use of the sample covariance as a diagnostic for parameter estimability. More recently, **?** presents the N-mixture model likelihood in a closed form version by expressing the infinite sum embedded in the likelihood in terms of a hypergeometric function in order to resolve the *choice-of-K* problem.

Notice that the N-mixture models have also a natural connection to disease surveillance in epidemiology, as infectious diseases have a dynamic structure where diseases can both "survive" in their reservoirs, while also "recruiting" from other reservoirs for that disease. In addition, even though the prevalence of a disease can greatly exceed the abundance of rare wildlife, they are both imperfectly detected. In the case of disease, underdetection can occur due to latency of symptoms of the disease and due to hidden populations which are unwilling or unable to report contraction of the disease. In the field of disease surveillance, sampling methods have been developed to address imperfect detection by making hidden populations, such as those with HIV or other stigmatized sexually transmitted infections (STI), more accessible through respondent-driven sampling as in **?** and **?**. These methods

have shown to improve estimation of aggregate characteristics, but they require successive sampling and costs of subject recruitment that is not already built into the existing disease reporting infrastructure.

In this paper we discuss the implementation of a multivariate normal asymptotic approximation of the generalized N-mixture model that does not suffer from computational inefficiency with larger abundances and does not require setting a predetermined value for K as in (3.6). Furthermore the multivariate normal asymptotic approximation model enables the researcher to diagnose the estimability of the parameters that occur, usually due to weak dependence between visits in the observed data, using the asymptotic information matrix. This structure also allows for a seamless implementation of a re-parametrized model. We compare this asymptotic approximation to the Dail-Madsen model via simulations and evaluate its performance using the American Robin data from Dail and Madsen (2011). Finally we demonstrate the usefulness of the N-mixture models in disease surveillance by looking at the prevalence of chlamydia in the state of Oregon from 2007-2016. In the application, we focus only on the temporal component of the problem and assume no spatial dependence.

3.2 Chlamydia Data in Oregon

The Oregon Health Authority website maintains yearly records with the number of cases per county of several reportable diseases, including chlamydia. Chlamydia is a common sexually transmitted infection that (in most cases) do not show any symptoms and therefore is imperfectly detected. When symptoms are present, they often do not appear

Table 3.1: Oregon County populations and observed counts of chlamydia 2007-2016, ordered by population

	County	Population	2007	2008	⋯	2015	2016
1	Multnomah County	799766	2924	3205	⋯	4664	5144
2	Washington County	582779	1011	1183	⋯	2025	2267
3	Clackamas County	408062	741	795	⋯	1168	1250
⋮	⋮	⋮	⋮	⋮		⋮	⋮
34	Gilliam County	1854	2	0	⋯	4	3
35	Sherman County	1710	2	1	⋯	3	6
36	Wheeler County	1344	0	0	⋯	3	0

for several weeks suggesting a potential spread of the disease over time. Although chlamydia can be cured when diagnosed and treated, the untreated disease can lead to permanent damage to the reproductive system in women and long-term pelvic pain in both men and women. Additionally, repeat infections are common so a follow-up screening is recommended 3 months after treatment. The spreading and curing of chlamydia suggests that a model with both survival of the disease and recruitment is appropriate.

From the **?** website (link available in the references), we obtained the yearly cases of chlamydia from 2007 to 2016 for each of its 36 counties as depicted in Table 4.1. The observed abundances for some of the bigger counties are in the thousands while those of some of the smaller counties are in the single digits. As a result, current models that can handle only small abundances such as the Dail-Madsen (DM) N-mixture model are inappropriate for these data.

Figure 4.1 suggests that there may be spatial dependence for chlamydia given that counties close to each other have a similar reported case value. However, in this paper we will ignore the possible spatial component of the problem, noting that some of the similarity between adjacent counties may simply be due to similar population size.

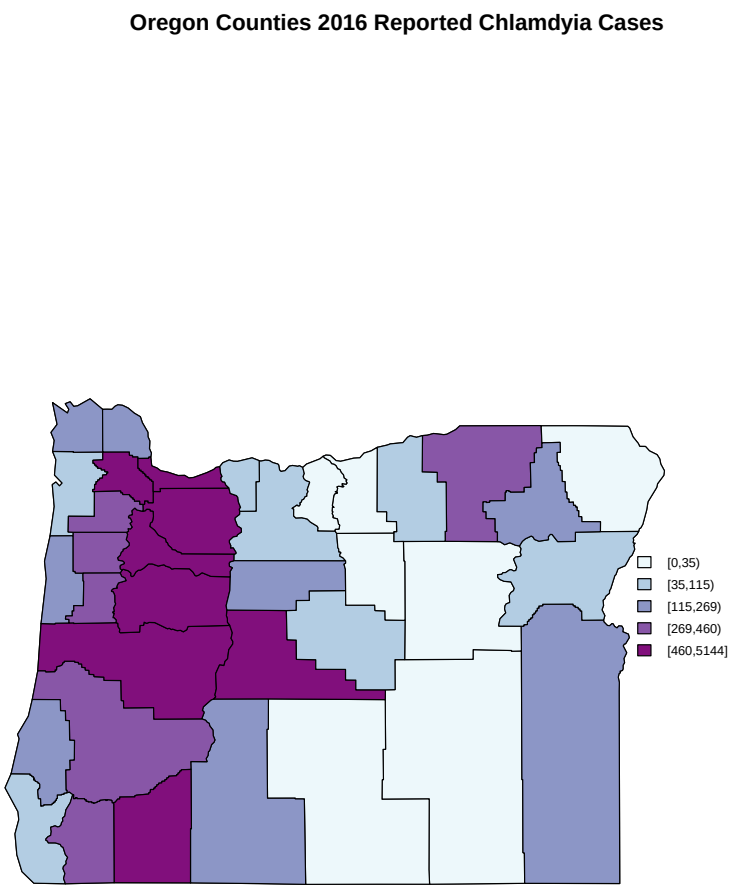

Figure 3.1: A heat map of the 2016 Oregon counties reported chlamydia cases.

3.3 Asymptotic Approximation

Dail and Madsen (2011) use optimization to derive estimates for the detection, rate and population dynamics parameters in their model. The implementation avoids directly computing abundance by integrating the abundance parameters out using summations (2) and then provides various methods for using the estimated parameters to back-calculate abundance at time t. Although others have implemented the DM N-mixture model using alternative methods such as Bayesian MCMC in JAGS (**?** and **?**) which directly calculates the abundance parameters, we will do so by making the assumption that the observed abundance counts are large and therefore follow an approximate Gaussian distribution.

Each of the sites' observed visits are asymptotically normally distributed. We assume that each site's observed vector is asymptotically distributed as multivariate normal with mean and covariance determined by the original rate and dynamics structure of the DM model, so that

$$L(\lambda, p, \omega, \gamma | \{n_{it}\}) \approx \prod_{i=1}^{R} MVN(\mu, \Sigma) \tag{3.7}$$

where μ is the S-dimensional mean vector and Σ is the $S \times S$-dimensional covariance matrix for each time i. In order to calculate the mean parameter μ_t for each time t, we first start with $\mu_1 = E[n_1] = E[E[n_1|N_1]] = \lambda p$, by the law of total expectation. With similar reasoning and assuming a constant survival rate ω and constant recruitment rate γ, i.e. $g(N_{it-1}) = \gamma$, the t-th visit's mean in the i-th site is calculated as

$$\mu_{it} = p(\mu_{it-1}\omega + \gamma). \tag{3.8}$$

Each element of the diagonal of the covariance matrix Σ is equivalent to μ_{it} for each site

i due to the mean-variance relationship of the Poisson distribution and the constant rate parameters. Using the law of total covariance, the off-diagonals are calculated as

$$\text{cov}(E(n_{it}), E(n_{it^*})) = p^2 \,\text{cov}(E(N_{it}|N_{it-1}), E(N_{it^*}|N_{it^*-1})) = p \cdot \mu_{min(i,v)} \cdot \omega^{|i-v|} \quad (3.9)$$

assuming the survivals and gains are independent. Assuming constant $\lambda, p, \omega, \gamma$ across all sites, this results in an identical S-dimensional multivariate normal structure for each site. A similar model can be defined for a negative binomial initial abundance rather than Poisson, or for other extensions of the original model. Additionally, fixed effect covariates can be included on the parameters by plugging a regression equation into the likelihood instead of the one parameter. For example, the inverse logit of $\eta_0 + \eta_1(sensitivity)$ could be used instead of p if we're dealing estimating the prevalence of an underdiagnosed disease, and we have some general knowledge on the sensitivity of that disease.

In this setting, the magnitude of the abundance does not affect the computational difficulty of estimating the model parameters, and we can explore whether abundance affects the relative error of parameter estimates and identifiability. Additionally, using the asymptotic model, we can explore how the estimates are affected by low detection or lack of correlation among visits and the true parameter values' relationship with estimability of the parameters in this model.

3.3.1 Estimability

? and **?** explore the assumptions of the DM model and the choice of K, and suggest that the model may have parameter estimability issues due to unidentifiability when detection

is low (p close to 0) or when counts from the same site show no or weak dynamical patterns (low ω and/or high γ). Intuitively, we would expect to have more information about the sites' abundance if there is a strong correlation between the visits, the strongest case being the Royle (2004) closed population model. Using our multivariate normal model, we can explore more exactly the cause of these issues beyond describing them as "weak dynamics." Let I denote the Fisher Information matrix. The m,n^{th} entry of I in the multivariate normal model is

$$I_{m,n} = \frac{\partial \mu^T}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n} + \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_m} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_n} \right), \tag{3.10}$$

where μ_{ij} and $\Sigma_{ij} = \sigma_{ij}$ are defined in (4.3), (3.8) and (3.9) and $[\theta_1\ \theta_2\ \theta_3\ \theta_4] = [\lambda\ p\ \omega\ \gamma]$ (**?**). Using the inverse of the information matrix, calculated analytically, we can obtain the asymptotic variance-covariance matrix to determine under what settings the MLE's of the parameters of interest become correlated highly enough to impede estimation. We consider that a perfectly negative or positive correlation between two parameters would suggest that they are completely confounded with one another.

Because a near-singular information matrix corresponds to high correlation among parameters estimates, we can detect settings in which we cannot estimate parameters well by using the determinant of the information matrix as a single-number summary of estimability. This is not the same as expected orthogonality of **?** (having zeros in the off-diagonals of the information matrix) but, intuitively, an information matrix with a lot of off-diagonals close to zero will have a large determinant, while those with off-diagonals close to the values of the diagonal will have a small determinant. This summary allows us to better understand the scenarios in which the model may perform poorly.

We describe two hypothetical scenarios in detail. The first sets $S = 10, \lambda = 10, p = 0.8, \omega = 0.5,$ and $\gamma = 2$. The asymptotic correlation matrix shows a strong correlation of -0.919 between expected initial abundance, $\hat{\lambda}$, and detection rate, $\hat{p}$, and of -0.659 between $\hat{p}$ and survival rate, $\hat{\omega}$. The matrix has a mildly positive correlation between $\hat{\lambda}$ and recruitment rate, $\hat{\gamma}$, and a mildly negative correlation between $\hat{\omega}$ and $\hat{\gamma}$. The determinant of I is 11.25. In the second scenario we kept all settings the same except for an increase in ω from 0.5 to 0.9. In this scenario, the correlation between $\hat{\lambda}$ and $\hat{p}$ is -0.609 and the correlation between $\hat{p}$ and $\hat{\omega}$ is -0.577. Note the increase in ω flips the correlation between $\hat{\lambda}$ and $\hat{\gamma}$ from a positive correlation to a mildly negative correlation and strengthens the negative correlation between $\hat{\omega}$ and $\hat{\gamma}$ to -0.933. Our intuition is that as ω increases, it becomes more difficult to tease apart ω and γ's effect on abundance. Here, the determinant of I is 306.43. Given the decrease in correlation between $\hat{\lambda}$ and $\hat{p}$, those estimates will have a smaller asymptotic variance, and therefore an improved precision in their respective confidence intervals.

One way to address issues of estimability is to consider various re-parametrizations that might increase the determinant of I, for instance, where the four parameters are $\xi(\theta) = [\lambda\, p\; p\; \omega\; \gamma]$. Using information theory from **?**, we determined that

$$\frac{|I^*(\xi)|}{|I(\theta)|} = |J|^2 \tag{3.11}$$

where $J = \left|\left|\frac{\partial \theta_j}{\partial \xi_i}\right|\right|$. Given this formula, we can see that the re-parametrization above increases the determinant by a factor of $\frac{1}{p^2}$. This suggests that this parametrization may help with estimability when p is small.

3.4 Simulation Study

In order to show the advantages of the asymptotic approximation model we will compare its performance to the DM model at small abundances. Since Dail and Madsen (2011)'s implementation cannot handle large abundances, we implement the DM model in JAGS and compare to the proposed asymptotic model. We first show that the asymptotic model performs similarly to the DM model at various small abundance settings. Simulation settings and results for $R = 36$ sites and $T = 10$ visits are summarized in Table 3.2. We used the `pcountOpen` function in *R* (**?**) from the unmarked package (**?**) with $K = 100$ to run the DM model. For the asymptotic model, we use the Python optimizer function `minimum` in the Scipy optimize module (**?**) with bounded optimization method L-BFGS-B ($\lambda \in (0,\infty), p \in (0,1), \omega \in (0,1), \gamma \in (0,\infty)$). We simulated data at various mean abundances ($\lambda = 5, 10$) as well as a range of levels of detection and dynamics in order to compared the performances of the DM and asymptotic models. Performance was evaluated by taking the average of the absolute error (AE) and absolute relative error (ARE) between the true total abundance at time i, $N_{i\cdot}$, and the estimated total abundance at time i, $\hat{N}_{i\cdot}$:

$$AE = \frac{1}{R}\sum_{i=1}^{R} |N_{i\cdot} - \hat{N}_{i\cdot}| \tag{3.12}$$

$$ARE = \frac{1}{R}\sum_{i=1}^{R} \frac{|N_{i\cdot} - \hat{N}_{i\cdot}|}{N_{i\cdot}}. \tag{3.13}$$

We additionally show the average time in seconds it takes for the optimizer to complete at each setting on a server with a 2.50 GHz processor and 256GB RAM.

The results in Table 3.2 suggest that at $\lambda = 5$ and $\lambda = 10$, the performances are very

Table 3.2: Each row is based on 100 simulated data sets using the given λ, p, ω and γ values and R=36,S=10. AE=Mean Absolute Error, ARE=Mean Absolute Relative Error

Model	λ	p	ω	γ	AE	ARE	Time(sec.)
DM	5	0.25	0.50	1	55.92	0.59	15.84
Asy.					61.60	0.66	3.65
DM	5	0.25	0.90	0	23.44	0.20	21.45
Asy.					32.20	0.27	3.54
DM	5	0.50	0.50	1	28.50	0.31	15.89
Asy.					28.70	0.31	3.87
DM	5	0.50	0.90	0	8.13	0.07	23.49
Asy.					9.90	0.08	3.56
DM	5	0.80	0.50	1	13.88	0.15	19.92
Asy.					14.44	0.16	3.67
DM	5	0.80	0.90	0	3.58	0.03	22.64
Asy.					4.58	0.04	3.55
DM	10	0.25	0.50	2	121.09	0.66	17.54
Asy					244.53	1.32	3.88
DM	10	0.25	0.90	0	53.47	0.22	29.33
Asy					62.14	0.25	3.31
DM	10	0.50	0.50	2	43.93	0.24	19.23
Asy					45.04	0.24	3.75
DM	10	0.50	0.90	0	18.34	0.08	22.08
Asy					20.40	0.09	3.56
DM	10	0.80	0.50	2	22.66	0.12	18.54
Asy					22.99	0.13	3.72
DM	10	0.80	0.90	0	5.93	0.03	22.98
Asy					7.80	0.03	3.95

similar when $p = 0.5$ or $p = 0.8$ but the DM model outperforms the asymptotic approximation model on the settings where $p = 0.25$. The asymptotic model provides estimates with much more computational efficiency. Considering the asymptotic approximation model assumes a large abundance, and choosing a large enough K for the DM model is an issue with larger abundances, these results are expected. They do suggest a similar performance for both models at this abundance, and we know through simulations (not shown) that the DM model becomes computationally intractible at higher abundances.

We compare the asymptotic approximation model to the DM model implemented in JAGS since the Bayesian MCMC does not require K to be chosen and can therefore handle higher abundances. In addition to comparing the AE and ARE between the asymptotic approximation model and the JAGS model, we also compare the confidence coverage of $\hat{N}_{\cdot 10}$, the estimated total abundance of all sites at the last time period. For the asymptotic approximation model, we have assumed an asymptotic multivariate normally distributed mean and variance-covariance matrix. Therefore, given their fully defined distribution, we determine a 95% interval by generating 10,000 samples of $[\hat{\lambda}, \hat{p}, \hat{\omega}, \hat{\gamma}]$ and derive estimates for 10,000 $\hat{N}_{\cdot 10}$ using

$$\hat{N}_{\cdot 1} = R\hat{\lambda} \tag{3.14}$$

$$\hat{N}_{\cdot t} = \hat{\omega}\hat{N}_{\cdot t-1} + R\hat{\gamma} \tag{3.15}$$

from Dail and Madsen (2011) and taking their 2.5^{th} and 97.5^{th} percentiles. In JAGS, we can simply parametrize the sum of the the last time periods abundances and obtain the 95% posterior credible interval for this parameter.

In Table 4.4, we compare results of the asymptotic approximation model to a JAGS

Table 3.3: Each row is based on 100 simulated data sets using the given λ, p, ω and γ values. AE=Mean Absolute Error, ARE=Mean Absolute Relative Error, $\widehat{N_{\cdot 10}}$ cov is the proportion of 95% CIs that included the true value of $N_{\cdot 10}$, $E[\widehat{N_{\cdot 10}}]$ is the approximate expected value of $\widehat{N_{\cdot 10}}$, and CI Width is the width of the $\widehat{N_{\cdot 10}}$ CI, PSRF or potential scale reduction factor is the Gelman-Rubin statistic, Time(sec.) is the average computation time in seconds, R=36,S=10

Model	λ	p	ω	γ	AE	ARE	$\widehat{N_{\cdot 10}}$ Cov.	$E[\widehat{N_{\cdot 10}}]$	CI Width	PSRF	Time(sec.)
Asy.	50	0.25	0.50	10	353.5	0.38	0.87	722	1960.6		5.6
JAGS					1408.6	1.52	0.91	722	3819.8	1.95	1912.9
Asy.	50	0.25	0.90	0	231.0	0.20	0.95	697	826.7		5.5
JAGS					759.1	0.64	0.71	697	1330.0	1.67	1401.3
Asy.	50	0.50	0.50	10	188.8	0.20	0.94	722	757.4		5.5
JAGS					419.8	0.45	0.93	722	1412.2	1.25	1747.6
Asy.	50	0.50	0.90	0	92.8	0.08	0.92	697	278.5		5.6
JAGS					121.3	0.10	0.93	697	346.2	1.07	1283.3
Asy.	50	0.80	0.50	10	91.4	0.10	0.91	722	322.6		5.5
JAGS					106.9	0.11	0.94	722	402.6	1.04	1530.4
Asy.	50	0.80	0.90	0	32.9	0.03	0.99	697	116.8		5.5
JAGS					31.3	0.03	0.93	697	88.3	1.01	1226.2
Asy.	200	0.25	0.50	40	1898.8	0.51	0.90	2888	9829.4		5.5
JAGS					3103.4	0.83	0.82	2888	7253.3	2.10	2069.3
Asy.	200	0.25	0.90	0	1094.1	0.23	0.89	2789	2974.6		5.5
JAGS					971.0	0.21	0.89	2789	2363.0	1.94	1463.6
Asy.	200	0.50	0.50	40	770.3	0.21	0.96	2888	2983.5		5.5
JAGS					1379.1	0.37	0.88	2888	3812.0	1.39	1902.4
Asy.	200	0.50	0.90	0	434.0	0.09	0.91	2789	1085.4		5.6
JAGS					514.5	0.11	0.91	2789	1208.1	1.25	1411.4
Asy.	200	0.80	0.50	40	416.7	0.11	0.85	2888	1253.4		5.6
JAGS					496.3	0.13	0.89	2888	1589.5	1.11	1713.9
Asy.	200	0.80	0.90	0	100.7	0.02	0.94	2789	351.7		5.2
JAGS					109.3	0.02	0.92	2789	328.2	1.02	1478.6

model when simulating from larger abundances. We used 100,000 iterations with a burn-in of 50,000 and a thinning of 20 in JAGS. At these settings the asymptotic approximation model does as well or outperforms the JAGS model in terms of the absolute error. Additionally, the 95% CI coverage for $N_{\cdot 10}$ is more consistently close to its nominal level (excepting cases $\lambda = 50, p = 0.25, \omega = 0.50, \gamma = 10$ and $\lambda = 200, p = 0.80, \omega = 0.50, \gamma = 40$) with similar or smaller CI widths. Finally, the asymptotic approximation model takes about 5 seconds to produce results while the JAGS model takes much longer (usually around 300 times longer). We note that the Gelman-Ruben statistic (PSRF) column indicates that more iterations are required in most cases but this will add even further to the computational time. The results also show that the JAGS mean process time is quite variable based on the settings and that process time is positively related to the AE. Simulations not shown in this paper showed that the $\hat{N}_{\cdot 10}$ did not have a $\hat{R} < 1.1$ for certain settings even after 1,000,000 iterations.

3.5 Case Studies

We conducted two case studies to demonstrate the usefulness of the asymptotic approximation model. The first study models American Robin data to compare results with the DM model at small abundances and the second models Oregon chlamydia data to show results from larger abundances.

Table 3.4: Results of both the Dail-Madsen N-mixture Model fit (K=100) and the Asymptotic N-mixture Model fit and their 95 % CIs on the American Robin data

Model	$\hat{\lambda}$	$\hat{p}$	$\hat{\omega}$	$\hat{\gamma}$
Dail-Madsen	2.31 (1.17 , 4.53)	0.52 (0.22, 0.81)	0.70 (0.34, 0.92)	0.48 (0.21, 1.14)
Asymptotic	2.51 (1.33, 3.70)	0.49 (0.26, 0.72)	0.70 (0.47, 0.93)	0.50 (0.15, 0.86)

3.5.1 North American BBS American Robin Data

In order to compare the asymptotic model with the DM model with real data, we fit the asymptotic model to the same American Robin Data from North American BBS as in Dail and Madsen (2011). We decided to only fit the model with the lowest AIC, the Poisson prior for the initial abundance distribution, and no covariates for the detection, survival, or recruitment parameters. The Dail-Madsen model was fit using the `pcountOpen` function in R, while the asymptotic model was fit using `Scipy` minimize function in Python.

Table 4 shows very similar results for both models. However, it is worth noting the time each model took to fit. With K=100, the Dail-Madsen model took 9.05 seconds, while the asymptotic N-mixture model only took 0.04 seconds. Dail and Madsen (2011) used K=40, and at this setting, the time required to fit the models was only 1.24. We used a slightly larger number to demonstrate the impact of the abundance on the computing time of this model. Increasing K to 300 requires 3.75 minutes. The Dail-Madsen intervals are not symmetric because of the log and logit transformations of the parameters. Despite small observed counts, the asymptotic model achieves very similar results to the DM model suggesting that the normal approximation is adequate.

Table 3.5: Output of python Optimizer. MLE is the optimizer's estimate and Asy. Var. is the diagonal of the inverse of the asymptotic information matrix calculated using the MLE

parameter	MLE	Asy. Var.
β	0.392E-02	0.16E-06
p	0.668	0.46E-02
γ	1.072	0.58E-05

3.5.2 Oregon Health Authority Chlamydia Data

In order to account for the large range of observed abundances in the chlamydia data, we used actual population for each of the 36 counties as a covariate for the initial abundance parameter λ. Additionally, we no longer consider a constant recruitment rate by conditioning the recruitment parameter $G_{it}|N_{it-1}$ on the previous time period's abundance as we would expect the number of new cases of an infectious disease to depend on the previous time period's prevalence. As a result, we assume

$$\lambda_i = \beta \cdot population_i \tag{3.16}$$

and the recruitment parameter is defined as in (5). The parameter point estimates and asymptotic variances we obtained are in Table 5.

The table estimates the abundance $\hat{\lambda}_i = 0.00392 \cdot population_i$ with with a very small asymptotic variance. Additionally, $\hat{\gamma} = 1.0738$, also with small variance, suggests an increase in abundance over time. We estimate $\hat{p} = 0.668$ with 95% CI (0.535,0.801). This suggests that there is evidence for less than perfect detection of Chlamydia cases in Oregon counties between 2007 and 2016. Practical reasoning for this imperfect detection may be latency of the infection or possibly hard-to-reach or hidden populations including drug

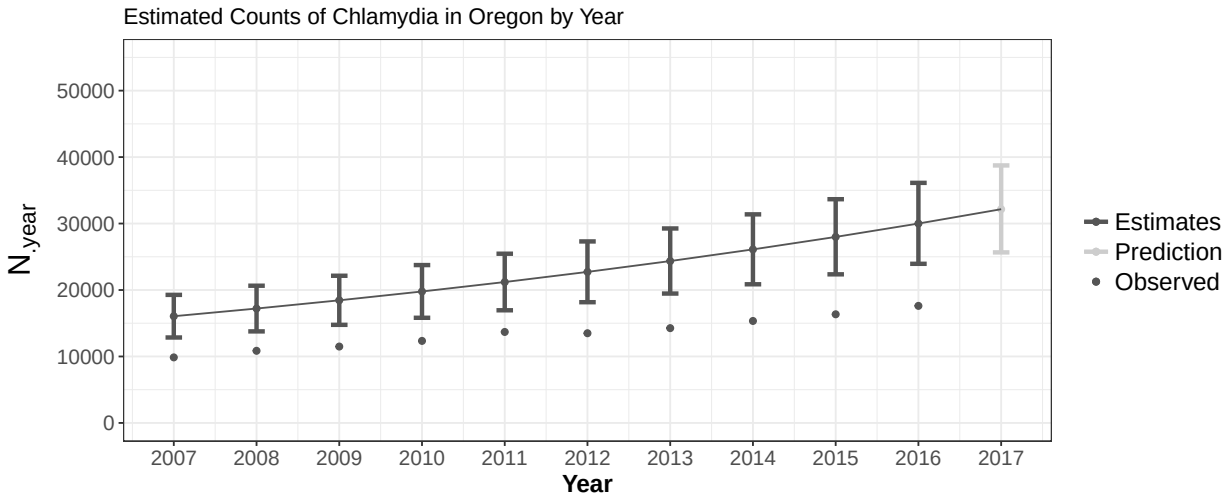

Figure 3.2: Both observed total counts $n_{\cdot t}$ and estimated total counts $\widehat{N_{\cdot t}}$ of chlamydia in Oregon and CIs for 2007-2016. Predicted counts and CI for 2017.

users, sex workers, and others having the infection. Given the evidence for underdetection of Chlamydia, we estimated the total prevalence of Chlamydia in Oregon between the years of 2007 to 2016.

Figure 4.2 shows a positive trend in both observed and estimated total prevalence in Oregon. It suggests that the total prevalence is actually greater than what was reported in each of those years. The estimates and CIs assume a constant level of underdetection and constant recruitment each year.

3.6 Discussion

There are two main advantages to making the asymptotic approximation to the N-mixture model. The first, related to implementation, is that one doesn't need to choose a K

value for the upper bound of integration to approximate the infinite sums in (2). Secondly, by estimating parameters based on a multivariate normal distribution, we can diagnose identifiability issues which limit estimation in the model using the asymptotic information matrix as described in Section 3.

As a result of these improvements, the asymptotic model is not restricted to moderately small n_{it}, and computationally efficiently gives estimates. This lack of restriction makes the asymptotic approximation a useful tool for estimating the prevalence of disease. **?** proposed a method using sample covariance to determine when an infinite number of abundances ($\hat{\lambda}$) may be made from their Multivariate Poisson model. Our method adds to this diagnosis tool using information and mutual information for the individual parameters in a model. Derivation of the correlation and covariance of the maximum likelihood estimates allows us to determine unidentifiability due to a high or perfect correlation among MLEs. It also assigns an appropriate standard error to the MLEs.

From the simulation in Section 4 we can see that for low abundances (λ=10), the results of the Dail-Madsen model and the asymptotic approximation model are similar. This suggests that the asymptotic approximation may be adequate at quite a low abundance. Further simulations demonstrate that the asymptotic model performs well with larger abundances with less bias and better coverage than the JAGS model. We also observe the flexibility of the asymptotic model from our cases studies where it achieves similar results to the Dail-Madsen model using the robin data but can also be applied to chlamydia data which has abundances in the thousands. The results of the case studies on disease prevalence data can help inform policy decisions regarding the disease. Implementing the model in JAGS as in **?** serves as an alternative for efficiently estimating larger abundances but lacks

the same parameter identifiability diagnosis tools and, as demonstrated in our simulation, results in more bias in the simple model considered. In addition, using JAGS takes more computing time and often resulted in Markov chain convergence difficulties.

Future uses of the asymptotic model include testing new parametrizations used in the N-Mixture model type. The normal assumption allows for easy exploration of re-parametrizations that could be useful under different circumstances. Another model enhancement would assume a spatial covariance structure. In this model, the asymptotic variance-covariance matrix would no longer have a block-diagonal structure.

4 A Spatial Extension to the Asymptotic Approximation of the N-mixture Model

4.1 Introduction

Estimating the true prevalence of under- reported or under- ascertained diseases is difficult to accomplish without knowledge of local demographics or extensive studies (**?**). Additionally, understanding the dynamics of a disease is an important step in understanding changes in prevalence over time (**?**). With certain diseases, such as infectious diseases, we would expect there to be a spatial dependency on the prevalence of diseases for locations near each other. We aim to provide a spatially explicit statistical method for surveilling imperfectly detected diseases using the reportable disease infrastructure already established through the Center for Disease Control (CDC) and the Oregon State Health Authority.

In the previous chapter, we established the use of the asymptotic N-mixture model as an appropriate model to make disease prevalence estimates for high abundance imperfectly detected or underreported diseases while disregarding any dependence of the disease in space. Other extensions of the N-mixture model have used factors such as weather, habitat, density dependence, to account for differences or similarities in space (**?**). In **?**, we see the first spatially explicit N-mixture model which models survival, reproduction, emigration, and abundance while representing movement among adjacent habitat patches. They model the spatial relationship using an immigration random variable, suggesting that the number of immigrants that site i will receive is the proportion of those emigrating from

adjacent site j, weighted by the number of j's adjacent sites. While this structure make sense for its application of wildlife, that is not necessarily the case in the spread of disease among humans. In the context of infectious disease, we cannot assume that a discrete proportion of adjacent sites' cases leaving one site will move to their neighbors, especially with vastly different populations of neighboring sites. Therefore, we propose a spatially explicit model that includes covariance information between measured sites by using adjacent sites' prevalence, rather than their abundances. In this Chapter we will present the chlamydia data in Section 2 and the spatial asymptotic N-mixture model in Section 3. In Section 4, we compare estimability between the non-spatial and spatial asymptotic models using the information method established in Chapter 3. In Section 5 we use simulated data with spatial covariance to test the new model's performance and also compare its performance to a model without spatial covariance information. Lastly, in Section 6 we include the results of the new spatial model on the Chlamydia data from Chapter 3.

4.2 The Chlamydia Data

We use the same Oregon Health Authority website chlamydia data in our analysis as in Chapter 3. Chlamydia is a sexually transmitted infection that is imperfectly detected because most people who have it do not have symptoms. Because chlamydia can be cured if detected a model that accounts for the reduction or survival of the disease over time as well as the recruitment through sexual transmission is appropriate. It also suggests that people in close proximity may have an effect on each other or that one county's rate will affect another's rate.

Table 4.1: Oregon County populations and observed counts of chlamydia 2007-2016, ordered by population

	County	Population	2007	2008	⋯	2015	2016
1	Multnomah County	799766	2924	3205	⋯	4664	5144
2	Washington County	582779	1011	1183	⋯	2025	2267
3	Clackamas County	408062	741	795	⋯	1168	1250
⋮	⋮	⋮	⋮	⋮		⋮	⋮
34	Gilliam County	1854	2	0	⋯	4	3
35	Sherman County	1710	2	1	⋯	3	6
36	Wheeler County	1344	0	0	⋯	3	0

From the **?** website, we obtained the yearly cases of chlamydia from 2007 to 2016 for each of its 36 counties as seen in Table 4.1. The observed abundances for some of the bigger counties are in the thousands while those of some of the smaller counties are in the single digits.

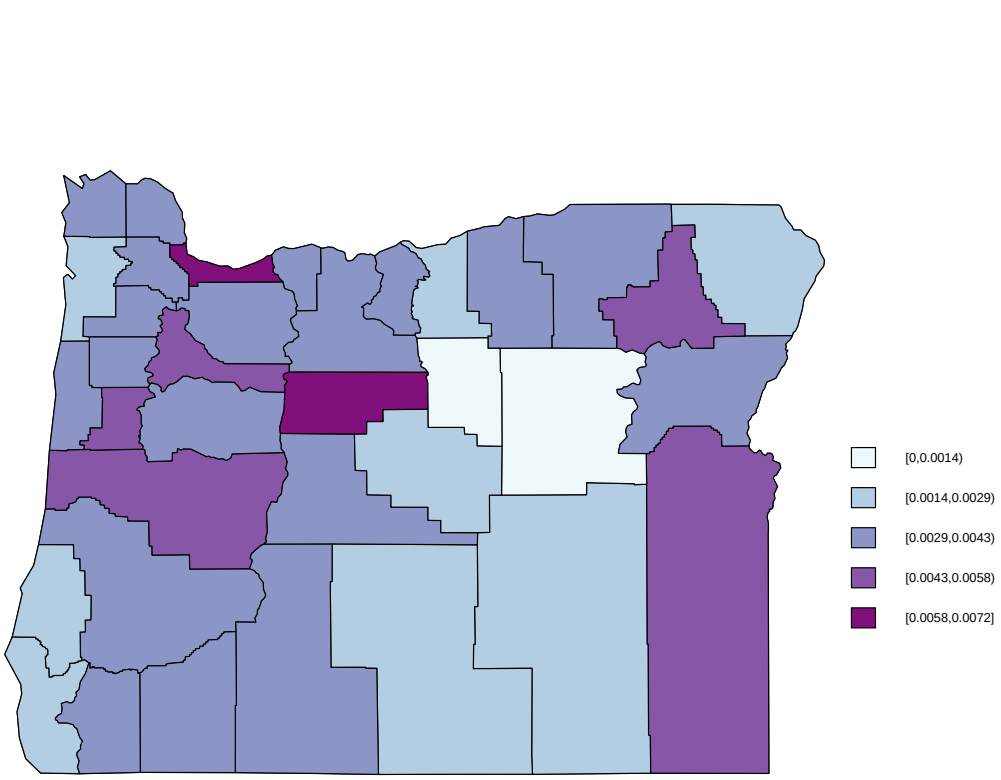

Figure 4.1: A heat map of the 2016 Oregon counties reported chlamydia prevalance

Figure 4.1 suggests that there may be spatial dependence for chlamydia given that counties close to each other have a similar prevalence. We will utilize that prevalence in each county in our spatial component of the model.

4.3 Model

Similar to the Asymptotic N-Mixture Model, the spatial model makes use of the multivariate normal structure to estimate parameters. In the Asymptotic N-mixture Model we had a block diagonal covariance matrix where the blocks represented dependence in time within a site but the diagonal structure reflected independence between sites. In the spatial model, we no longer have a block diagonal matrix, allowing independence between sites.

Given that the main issue with N-mixture models is parameter identifiability, we wanted to incorporate spatial dependence in the model that didn't include extra parameters that might be confounded with parameters from the non-spatial components of the model. In order to accomplish this, we use a conditional auto-regressive (CAR) model based on adjacency much like **?**. We decided that using an equal weighted CAR model was not appropriate due to the possibility of adjacent counties having very different population sizes. With equal weighting, the prevalence in really big or small populations could disproportionately affect their neighbors. Therefore, we adjusted the CAR model to use a weighted prevalence model. Much like in Chapter 3 where $i = 1, \ldots, R$=number of sites and $t = 1, \ldots, S$=number of visits, we will consider

$$n_{it} \sim \text{Binomial}(N_{it}, p) \tag{4.1}$$

$$N_{i1} \sim \text{Poisson}(\lambda_i) \tag{4.2}$$

$$\lambda_i = \beta \cdot pop_i \tag{4.3}$$

$$S_{it}|N_{it-1} \sim \text{Bin}(N_{it-1}, \omega) \tag{4.4}$$

where p represents the constant detection rate and S_{it} represents those who still have the

disease from time $t-1$ to t in county i with surival rate ω, but the recruitment process will now incorporate the previous time period of all neighbors so that

$$G_{it}|\tilde{N}_{it-1} \sim \text{Poisson}(\gamma \cdot \tilde{N}_{it-1}) \tag{4.5}$$

where

$$\tilde{N}_{it} = pop_i \cdot \left(\Sigma_{j \in \partial_i} \frac{\frac{N_{jt}}{pop_j}}{k_i} \right) \tag{4.6}$$

where k_i is county $i's$ population and ∂_i is the set of its neighbors. ∂_i is determined using a $R \times R$ neighbors matrix $\mathbf{A}$ such that $A_{mm} = 1$ and $A_{mn} = A_{nm} = 1$ if sites m and n are neighbors. Other elements are 0. The recruitment of site i at time t is a multiple γ of the average of the prevalence (inside the parentheses) from the previous time period of the site and all of its adjacent sites times the population of the site i. With this formulation, we assume an increased prevalence in one site will have a positive effect on the assumed prevalence of its adjacent sites. Additionally, since the neighbors only affect each other's prevalance, a large county will not have an unduly great of effect on a small neighbor's recruitment in terms of added abundance. There is no structure that induces dependence on the first time period. We assume that the population size will account for the spatial similarities between counties or that there is no dependence at time zero.

In order to implement this model, we must derive the asymptotic mean and variance-covariance matrix of this structure. We assume the joint likelihood of all the data is asymptotically distributed as multivariate normal with mean and covariance determined by the

original rate and dynamics structure of the defined model above, so that

$$L(\lambda, p, \omega, \gamma | \{n_{it}\}) \approx MVN(\mu, \Sigma)$$

where μ is the vector of means with elements $\mu_{i,t}$. The $R \cdot S \times R \cdot S$ dimensional variance-covariance matrix Σ can be broken up into S $R \times R$ matrices where each $R \times R$ matrix represents the spatial covariance between sites at time period t. In order to best explain the asymptotic representation of the matrix, we will index a single element as $\sigma_{i,j}^{t,u}$ where t and u index the R time periods and i and j index the sites within that time period, e.g., $\sigma_{1,1}^{1,2} = \Sigma_{1,R+1}$, the first row and $R+1^{st}$ column of Σ. In order to calculate the multivariate mean parameter μ_{it} for each site i at time t, we first start with $\mu_{i1} = E[n_{i1}] = E[E[n_{i1}|N_{i1}]] = \lambda_i \cdot p$ by the law of total expectation, where λ_i is defined as in (4.3). With similar reasoning and assuming a constant survival rate ω and constant recruitment rate γ, the t^{th} visit's mean in the i^{th} site is calculated as

$$\mu_{it} = p \cdot (\mu_{it-1} \cdot \omega + \gamma \cdot \tilde{\mu}_{it-1}) \tag{4.7}$$

where $\tilde{\mu}_{it-1}$ is defined similarly to $\tilde{N}_{it-1}$ in Equation 4.6.

The diagonal elements of the $R \times R$ covariance matrix $\Sigma_{i,i}^{1,1}$ at the first time period are all equal to λ_i and the off-diagonal elements are all 0. The covariance between observed counts at site k, time t, and site l, time u, $k \neq l$ and $t \neq u$ is

$$\Sigma_{k,l}^{t,u} = p^2 \cdot (\omega \cdot \sigma_{k,l}^{t,u-1} + \gamma \cdot \tilde{\sigma}_{k,l}^{t,u-1}). \tag{4.8}$$

The main diagonal is

$$\sigma_{i,i}^{t,t} = \mu_{t-1} \cdot \omega \cdot (1-\omega) + \tilde{\mu}_{t-1} \cdot \gamma + \sigma_{i,i}^{t-1,t-1} \cdot \omega^2 + 2 \cdot \omega \cdot \gamma \cdot \tilde{\sigma}_{k,l}^{t-1,t-1} + \gamma^2 \cdot \left(\frac{pop_i}{k_i}\right)^2 \cdot \left(\Sigma_{j \in \partial_i} \frac{\sigma_{j,j}^{t-1,t-1}}{pop_j^2} + 2 \cdot \Sigma_{j<k \in \partial_i} \frac{\sigma_{j,k}^{t-1,t-1}}{pop_j \cdot pop_k}\right). \quad (4.9)$$

The resulting mean vector and covariance matrix are of the form

$$\mu = \begin{bmatrix} \mu_{1,1} \\ \mu_{2,1} \\ \vdots \\ \mu_{R,1} \\ \mu_{1,2} \\ \vdots \\ \mu_{R,2} \\ \vdots \\ \mu_{R,S} \end{bmatrix} \Sigma = \begin{bmatrix} \sigma_{1,1}^{1,1} & \cdots & \sigma_{1,R}^{1,1} & \sigma_{1,1}^{1,2} & \cdots & \sigma_{1,R}^{1,2} & \cdots & \sigma_{1,1}^{1,S} & \cdots & \sigma_{1,R}^{1,S} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \\ \sigma_{R,1}^{1,1} & \cdots & \sigma_{R,R}^{1,1} & \sigma_{R,1}^{1,2} & \cdots & \sigma_{R,R}^{1,2} & \cdots & \sigma_{R,1}^{1,S} & \cdots & \sigma_{R,R}^{1,S} \\ \sigma_{1,1}^{2,1} & \cdots & \sigma_{1,R}^{2,1} & \sigma_{1,1}^{2,2} & \cdots & \sigma_{1,R}^{2,2} & & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \sigma_{R,1}^{2,1} & \cdots & \sigma_{R,R}^{2,1} & \sigma_{R,1}^{2,2} & \cdots & \sigma_{R,R}^{2,2} & & \vdots & & \vdots \\ \vdots & & \vdots & & & & \ddots & \vdots & & \vdots \\ \sigma_{1,1}^{S,1} & \cdots & \sigma_{1,R}^{S,1} & \cdots & \cdots & \cdots & \cdots & \sigma_{1,1}^{S,S} & \cdots & \sigma_{1,R}^{S,S} \\ \vdots & & \vdots & & & & & & & \\ \sigma_{R,1}^{S,1} & \cdots & \sigma_{R,R}^{S,1} & \cdots & \cdots & \cdots & \cdots & \sigma_{R,1}^{S,S} & \cdots & \sigma_{R,R}^{S,S} \end{bmatrix}$$

In order to make prevalance estimates we use a similar process as in Chapter 3 where we first estimate the first time period abundance using the initial abundance coefficient β and the known populations and then step through the other time periods using the ω and γ

parameters such that

$$\widehat{N_{\cdot 1}} = \beta \cdot \Sigma_{i=1}^{R} pop_i \tag{4.10}$$

$$\widehat{N_{\cdot t}} = \widehat{N}_{it-1} \cdot \omega + \gamma \cdot \tilde{\widehat{N}}_{it-1}. \tag{4.11}$$

For the asymptotic approximation model, we have assumed an asymptotic multivariate normally distributed mean and variance-covariance matrix. Therefore, given the asymptotic distribution of the MLEs, we determine a 95% interval by generating 10,000 samples of $[\hat{\beta}, \hat{p}, \hat{\omega}, \hat{\gamma}]$ and derive estimates for 10,000 $\widehat{N_{\cdot 10}}$ using (4.10) and (4.11).

4.3.1 Other Models

Before settling on the weighted prevalence based on neighbors structure for γ, we explored other methods for modelling spatial dependence. The first idea was to do a conditional auto-regressive type model to account for spatial covariance using neighbor averages. This model is defined as

$$N_{i1} \sim \text{MVN}(\lambda_i, \Sigma) \tag{4.12}$$

$$G_{it} | N_{it-1} \sim \text{Poisson}(\gamma \cdot \overline{N}_{it-1}) \tag{4.13}$$

where $\overline{N}_{it-1} = \Sigma_{j \in \partial_i} \frac{N_{jt-1}}{k_i}$ where k_i is the number of neighbors of site i. This method revealed its inadequacies both when I was simulating data and trying to fit the model to observed data. In the simulations small counties with big neighbors were increasing prevalence too quickly and big counties with mostly small neighbors were shrinking. None

of the simulated data looked like real observed data and fitting the model to real data was resulting in extreme parameter fits. The weighted model based on prevalence defined in Equation (4.5) does not let the neighbors affect each other disproportionately and fits more naturally in the field of epidemiology where prevalences are usually measured as rates rather than counts to account for population size.

The other idea we considered was to induce correlation between neighbor sites using the γ variable, making it a random variable, distributed as multivariate log-normal. This would create either a positive or negative correlation between sites due to the similiarity or dissimilarity of their recruitment. The implementation of this model was extremely computationally intensive because it required precise estimates of large normal moments. We could only implement this model at smaller sample sizes within a reasonable amount of time and the implementation did not perform well.

4.4 Estimability

Using our multivariate normal model, we can explore and compare estimability of the spatial asymptotic model with the non-spatial asymptotic model because both utilize the multivariate normal distribution and use the same parameters. We note the different interpretation in the γ parameter between both models since γ serves as the parameter which determines the strength of the spatial relationship among neighbors in the spatial model. Let I denote the Fisher Information matrix. The m, n^{th} entry of I in the multivariate normal model is

$$I_{x,y} = \frac{\partial \mu^T}{\partial \theta_x} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_y} + \frac{1}{2} \mathrm{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_x} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_y} \right), \tag{4.14}$$

where μ_{it} and Σ_{mn} are defined above and $[\theta_1 \ \theta_2 \ \theta_3 \ \theta_4] = [\lambda \ p \ \omega \ \gamma]$ (**?**). Using the inverse of the information matrix, calculated analytically, we can obtain the asymptotic variance-covariance matrix to determine under what settings the MLE's of the parameters of interest become correlated highly enough to impede estimation. We consider that a perfectly negative or positive correlation between two parameters' MLE's would suggest that they are completely confounded with one another.

One improvement in this model over the non-spatial model is that the ω and γ parameters are no longer completely confounded for estimates of ω close to zero. As an example, we plug in the same results for the chlamydia data from Chapter 3, i.e. $\lambda = 0.00392, p = 0.668, \omega = 0, \gamma = 1.072$, and we find that $\mathrm{cor}(\omega, \gamma) = -0.9992$. While this is highly negative correlated, they are no longer completely confounded with each other and their asymptotic variance can be estimated using the model.

4.5 Simulations

We conducted two simulation studies to determine the performance of the spatially explicit model as well as its robustness compared to the non-spatial model. The first study simulates data from the model in Section 4.3. We used the neighbors matrix as well as the county populations from the chlamydia data and then simulated data using two levels of β (0.005, 0.05), three levels of p (0.4, 0.7, 0.9), two levels of ω (0.5, 0.8), and two levels of γ (0.3, 0.6). We conducted 1000 simulations for each setting and then determined

the average relative error ($ARE = \frac{1}{R}\sum_{i=1}^{R} \frac{|N_{i\cdot} - \hat{N}_{i\cdot}|}{N_{i\cdot}}$) of the prevalence estimates (equations (4.10) and (4.11)) as well as the confidence interval coverage on the last time period. The combinations of the settings chosen reflect the observed slow increase in prevalence over time of the infectious diseases on the Oregon Health Authority website.

Table 4.2: Each row is based on 1000 simulated data sets using the given β, p, ω and γ values. MnARE=Mean Absolute Relative Error, MdARE=Median Absolute Relative Error,$\widehat{N_{\cdot 10}}$ Cov. is the proportion of 90% CIs that included the true value of $N_{\cdot 10}$ and CI Width is the width of the $\widehat{N_{\cdot 10}}$ CI, R=36,S=10

β	p	ω	γ	MnARE	MdARE	$\widehat{N_{\cdot 10}}$ Cov.	CI Width
0.005	0.40	0.50	0.60	0.23	0.17	0.91	41941.67
0.005	0.40	0.80	0.30	0.17	0.13	0.87	30621.17
0.005	0.70	0.50	0.60	0.13	0.10	0.89	22136.88
0.005	0.70	0.80	0.30	0.08	0.06	0.89	14081.31
0.005	0.90	0.50	0.60	0.08	0.06	0.86	13628.08
0.005	0.90	0.80	0.30	0.05	0.03	0.88	8053.37
0.05	0.40	0.50	0.60	0.25	0.17	0.89	433104.41
0.05	0.40	0.80	0.30	0.17	0.13	0.90	313207.81
0.05	0.70	0.50	0.60	0.11	0.09	0.90	214159.92
0.05	0.70	0.80	0.30	0.07	0.06	0.88	132778.59
0.05	0.90	0.50	0.60	0.07	0.05	0.88	127797.52
0.05	0.90	0.80	0.30	0.04	0.03	0.88	74848.27

Table 4.2 shows that we see less error in the abundance estimates for a higher true detection level p. It also shows that an increased ω and a smaller γ improves performance

by reducing ARE. This suggests that a stronger spatial dependence reduces the ability to estimate the parameters in the model because a larger γ reflects a stronger spatial dependence between neighboring counties. The smaller median ARE than mean ARE in all cases reflects the right skewedness in the ARE estimates and suggests that some simulations performed much worse than the average. Additionally, we note that the coverage is less than nominal in a majority of the settings. In our exploration of this phenomenon, we conjectured that the spatial component was reducing our effective sample size below what we needed for the asymptotic information matrix. However, doubling the sample size (R=72) did not help the coverage levels (simulations not shown). We tried two other methods for deriving confidence intervals for the prevalance at the last time period in order to determine if this particular method was the issue. In the first we simulated 10,000 samples of $\hat{p}$ and divided $\Sigma_{i=1}^{R} n_{i,10}$ by each sample $\hat{p}$ to create a bootstrap distribution for the sum of prevalances at the last time period. The second used a second order Taylor expansion to get an estimate for $\frac{1}{\hat{p}}$ and $SE(\frac{1}{\hat{p}})$ and derived the confidence interval using $\frac{\Sigma_{i=1}^{R} n_{i,10}}{\hat{p}} \pm Z_{0.95} \cdot SE(\frac{\Sigma_{i=1}^{R} n_{i,10}}{\hat{p}}\hat{p})$. Neither alternative method performed closer to the nominal level than the first method. In our exploration of this issue, we grouped the results by whether $\hat{p} > p$. This grouping showed that if $\hat{p} > p$ the coverage was generally well below 90% and if $\hat{p} < p$ the coverage was generally above 90%.

The second simulation study explores the same settings but simulates data both with the model described in Section 4.3 as well as from the non-spatial model in Chapter 3. It also fits both models to both types of simulations to see how robust the method is for data simulated from the wrong model (spatial/non-spatial). We simulated 500 data sets for each type of data.

Table 4.3: Each row is based on 500 simulated data sets using the given β, p, ω and γ values. Simulation Type specifies whether the row's data was simulated using spatial dependency or not, Spatial Fit ARE=Mean Absolute Relative Error for the data fit to the spatial model, Nonspatial Fit ARE=Mean Absolute Relative Error for the data fit to the nonspatial model,R=36,S=10

Simulation Type	β	p	ω	γ	Spatial Fit ARE	Nonspatial Fit ARE
Spatial	0.005	0.40	0.50	0.60	0.24	3.17
Nonspatial					0.26	0.14
Spatial	0.005	0.40	0.80	0.30	0.17	1.18
Nonspatial					0.21	0.14
Spatial	0.005	0.70	0.50	0.60	0.13	1.68
Nonspatial					0.05	0.08
Spatial	0.005	0.70	0.80	0.30	0.07	0.22
Nonspatial					0.06	0.07
Spatial	0.005	0.90	0.50	0.60	0.08	0.80
Nonspatial					0.11	0.05
Spatial	0.005	0.90	0.80	0.30	0.06	0.13
Nonspatial					0.07	0.05
Spatial	0.05	0.40	0.50	0.60	0.25	3.55
Nonspatial					0.27	0.14
Spatial	0.05	0.40	0.80	0.30	0.18	1.35
Nonspatial					0.21	0.14
Spatial	0.05	0.70	0.50	0.60	0.11	1.63
Nonspatial					0.06	0.07
Spatial	0.05	0.70	0.80	0.30	0.07	0.22
Nonpatial					0.05	0.05
Spatial	0.05	0.90	0.50	0.60	0.07	0.82
Nonspatial					0.09	0.06
Spatial	0.05	0.90	0.80	0.30	0.05	0.10
Nonspatial					0.06	0.05

As shown in Table 4.3, when the simulation type is spatial, the spatial fit performs much better than the nonspatial fit. When the simulation type is nonspatial, the nonspatial fit most commonly does better than the spatial fit but not by nearly the magnitude as in the previous comparison. In some cases, the spatial fit performance was equal to or even outperformed the nonspatial fit when the data was simulated without spatial dependency. In both scenarios, the level of ω and γ affects the magnitude of the performance difference, i.e., a bigger ω and smaller γ reduces the magnitude of the difference on average. These results suggest that the spatial model is quite robust to whether the data has spatial dependency or not. Still, fitting the data with the appropriate model is preferable.

4.5.1 Computing

We used the high performance computing (HCP) cluster in order to run many simulations at once. Because we simulated 1000 datasets per each of the 12 settings, we needed to simulate and fit 12,000 datasets. Given that the cluster's statistics queue can run 144 jobs at once, we aimed to drastically reduce the amount of time those 12,000 jobs took by submitting them to the queue. However, while conducting the simulations in the cluster for this chapter, we encountered some issues with computational efficiency. The Python minimizing function Scipy.minimize automatically uses multiple threading, a process with which Python can do multiple tasks at once. However, this process does not increase speed of a program if it already uses 100% CPU time. As a result, the number of threads chosen automatically by the minimize function caused the load average of each queue in the cluster to increase well above the total number of available nodes. This caused a major

slowdown as the job queues need to cool down to a small load average in order to take more jobs from the pending queue. We were able to address this issue by limiting the max number of threads to 2 using the mkl.set_num_threads function in Python. We assigned 2 cores per job in the cluster as well and this allowed the cluster to work at a more appropriate load average.

4.6 Chlamydia Results

We fit the chlamydia data to the the spatially explicity model. The parameter point estimates and asymptotic variances are in Table 4.4.

Table 4.4: Output of Python Optimizer, MLE is the optimizer's estimate, Asy. Var. is the diagonal of the inverse of the asymptotic information matrix calculated using the MLE

parameter	MLE	Asy. Var.
β	0.00378	0.6E-07
p	0.73	0.23E-02
ω	0.862	0.155E-02
γ	0.253	0.155E-02

We estimate $\hat{p} = 0.73$ with 95% CI (0.636, 0.825) suggesting a less than perfect rate for detecting chlamydia cases in Oregon. The results also include non-zero estimates for both the survival parameter ω and the spatial recruitment parameter γ in the final model, allowing a potentially more interpretable model than the non-spatial model which estimated $\omega = 0$. Given that we are assuming a less than perfect detection rate, including those with latent symptoms or those in hidden populations, we would not expect all cases of chlamydia to be cured or to leave the county. With the spatial model, we estimate $\hat{\omega} = 0.862$ with 95% CI (0.784, 0.939) suggesting a high level of survival and a lack of

relocation from year to year for those with chlamydia. We estimate $\hat{\gamma} = 0.253$ with 95% CI (0.176, 0.330). The non-zero estimate of γ suggests that there is a spatial effect on recruitment from the adjacent counties.

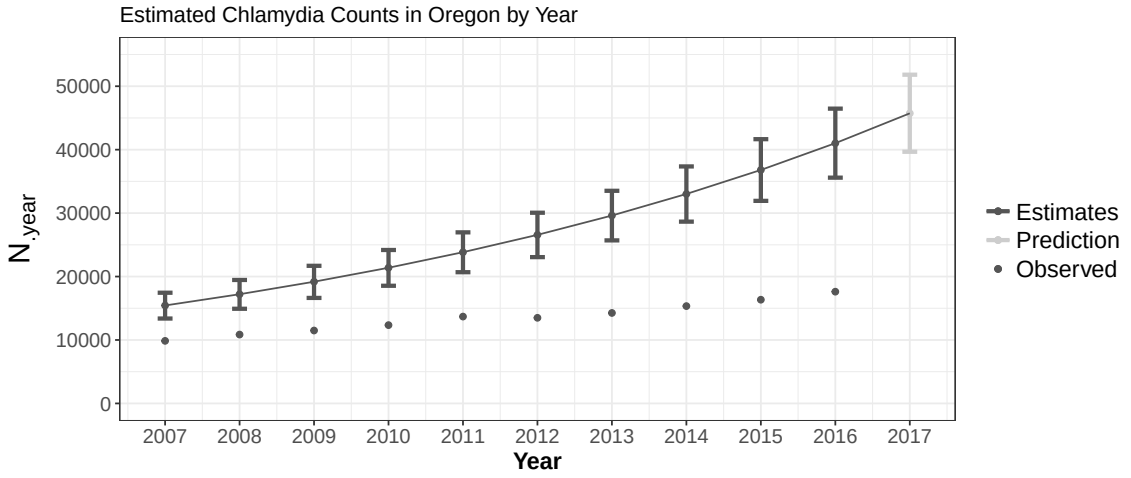

Figure 4.2: Both observed total counts $n_{\cdot t}$ and estimated total counts $\widehat{N_{\cdot t}}$ of Chlamydia in Oregon and CIs for 2007-2016. Predicted total counts and CI for 2017.

We can see a similar increasing trend of chlamydia in Figure 4.2 as in the non-spatial model from Chapter 3. However, the spatial model provides us with much more precise esimates of the total prevalence in each year from 2007-2016 and the prediction in 2017. We considered that the higher level of precison could be explained by undercoverage. The simulation settings closets to our results ($\beta = 0.005, p = 0.70, \omega = 0.80, \gamma = 0.30$ or $\beta = 0.005, p = 0.90, \omega = 0.80, \gamma = 0.30$) had a coverage of .89 and .88 respectively, suggesting that at those settings, the coverage is close to nominal. In 2017, we predict a total chlamydia prevalence of 45,752 with 95% CI (39,667, 51,814).

4.7 Discussion

We developed the spatial extension to the asymptotic N-mixture model as a natural next step for dealing with infectious disease data. The human to human connection suggests that a high prevalence in one location could encourage a high prevalence in neighboring locations. As such, we induced the spatial dependence using a neighbors type model. By weighting the spatial dependence on the prevalence by population rather than total count over a year, we avoid letting the populations of neighboring counties from disproportionately affecting each other in the model. This assumes that the recruitment of one county is based on the prevalence of its neighboring counties and itself at the previous time period as opposed to the nonspatial model in which each county only conditions recruitment on its own previous time period's prevalence.

We have shown through simulation that the spatial asymptotic approximation of the N-mixture model performs well in common scenarios and outperforms the nonspatial asymptotic approximation model with spatially dependent data. Due to the spatial γ structure, the ω and γ parameters can no longer be confounded and the interpretation of these parameters are more separated than in the nonspatial model. The spatial model also shows robustness in cases where the data is not spatially dependent.

In the Oregon Health Authority chlamydia data, we see somewhat similar estimates between the spatial and nonspatial models but smaller variance in both the initial abundance parameter β and the detection parameter p. However, with a detection rate $p = 0.73$, we would not expect the survival of the disease $\omega = 0$ as suggested by the nonspatial model as some of the cases would be hidden or undiagnosed and not be treated. We also might not expect a 86% survival rate as suggested in the spatial model, but the interpretation of

the spatial model fits the story better than the nonspatial model and further contributes to the evidence of imperfect detection of chlamydia in Oregon. Future research can be done to better tailor the spatial and dynamic aspects of the N-mixture model for applications in infectious diseases.

5 Conclusion

We have shown that disease surveillance is a natural application of N-mixture models as the field of disease surveillance has issues of under-ascertainment and under-diagnosis. In order to apply N-mixture models to larger abundances such as in disease surveillance, we developed the asymptotic approximation to the N-mixture model. While this idea is simple, it avoids the necessity of prohibitively slow integration to obtain the likelihood for maximization and does not depend on the convergence of Bayesian MCMC parameter chains. Our simulations in Chapter 3 show that a Bayesian MCMC implementation of the Dail and Madsen (2011) likelihood does not perform as well as the asymptotic approximation to N-mixture models for large abundances.

Beyond enabling use for larger abundances, the multivariate normal approximation enables the user to diagnose the parameters' estimability using information theory. Given the known derivation of the information matrix for re-parametrizations of the multivariate normal distribution, we can easily obtain the information matrix for any parameters in the model and therefore derive the asymptotic covariance and correlation matrices of the MLE's. The asymptotic information and correlation matrices provide information on parameter estimability. For instance, a correlation of 1 or -1 determines nonidentifiability between two parameters and a correlation of 0 determines orthogonality between two parameters. As such, this method can be useful for testing the utility of new model parametrizations of N-mixture models.

Given that the asymptotic method relies only on the asymptotic mean vector and variance-covariance matrix, it can be applied to any N-mixture model extension that has finite expected means, variances, and covariances. The spatial extension to the asymptotic approximation of the N-mixture model, which we explored in Chapter 4 with simulations and the Oregon Health Authority year chlamydia prevalence data set, is one such possible extension. While the derivation of the expected means and covariances can be complicated, once they are derived, implementing the model by filling out the mean vector and covariance matrix is easily accomplished. Additionally, MLEs can be obtained efficiently using an optimizing function despite the complexity of the underlying model. Still, this extension exhibited some undercoverage on the last time periods prevalence estimate when p is overestimated. Despite trying various alternative methods for building confidence intervals, we were unable to attain nominal coverage. This undercoverage is an issue to be explored in future research.

Both the spatial and nonspatial models' analyses provided evidence for an imperfect detection of Chlamydia in Oregon from the yearly reported counts from 2007 to 2016. However, the interpretation of the dynamics estimates are very different. In the nonspatial model, we estimated a survival rate of the disease of zero. It makes sense for this to be very low given the curability of the disease, but perhaps doesn't make sense given the assertion of imperfect detection. If some carriers of chlamydia are not detected, we would expect some to continue carrying the disease and to survive in a given county over a year. In the model with the spatial recruitment parameter, we estimated a survival rate of the disease to be 0.862. While there is a change of interpretation of the recruitment parameter between the nonspatial and spatial models, the change in the estimated survival rate and

its interpretation are hard to explain. Given the curability of the chlamydia, we would not expect 86% of cases to continue carrying the disease, especially given our estimated detection rate of 73%. For future work, we need to further consider the meaning of the survival rate parameter in a model accounting for imperfect detection while in the context of a highly curable disease. In general, we caution against using the estimates of the dynamics parameters to make inference. Instead, we treat them as nuisance parameters so that we don't need to assume a closed population and because the data doesn't contain much information specifically about survival and recruitment of the disease.

Although not necessary in all settings, the asymptotic approach is a contribution to the N-mixture class of models. It helps address the question of parameter estimability in the presence of weak dynamics, an issue which is addressed in some of the literature but not fully defined. Our asymptotic information method specifically points to parameters that are highly correlated and therefore, harder to estimate. It enables the use of these models to improve count estimates in disease surveillance as well as other large abundance applications while remaining cost effective. Additionally, there are many possible epidemiological extensions to the asymptotic approximation to the N-mixture model. For instance, future ideas for research in this area include N-mixture models for diseases with more complex dynamics such as seasonality or a large lag in detection as is seen over time with the Zika virus.

Bibliography

Blumenthal, S. and Dahiya, R. C. (1981). Estimating the binomial parameter n. *Journal of the American Statistical Association*, 76(376):903–909.

Carroll, R. J. and Lombard, F. (1985). A note on n estimators for the binomial distribution. *Journal of the American Statistical Association*, 80(390):423–426.

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

DasGupta, A. and Rubin, H. (2005). Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference*, 130(1):391–404.

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