Statistical methods for species richness estimation using count data from multiple sampling units

by

Angus Gordon Argyle
B.Sc., University of Victoria, 2000

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

DOCTOR OF PHILOSOPHY

in the Department of Mathematics and Statistics

Statistical methods for species richness estimation using count data from multiple sampling units

by

Angus Gordon Argyle
B.Sc., University of Victoria, 2000

Supervisory Committee

Dr. Farouk Nathoo, Supervisor
(Department of Mathematics and Statistics)

Dr. William Reed, Departmental Member
(Department of Mathematics and Statistics)

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

Dr. Min Tsao, Departmental Member
(Department of Mathematics and Statistics)

Dr. Pauline van den Driessche, Departmental Member
(Department of Mathematics and Statistics)

Dr. Bradley Anholt, Outside Member
(Department of Biology)

Dr. Fangliang He, External Examiner
(Department of Renewable Resources, University of Alberta)

Supervisory Committee

Dr. Farouk Nathoo, Supervisor
(Department of Mathematics and Statistics)

Dr. William Reed, Departmental Member
(Department of Mathematics and Statistics)

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

Dr. Min Tsao, Departmental Member
(Department of Mathematics and Statistics)

Dr. Pauline van den Driessche, Departmental Member
(Department of Mathematics and Statistics)

Dr. Bradley Anholt, Outside Member
(Department of Biology)

Dr. Fangliang He, External Examiner
(Department of Renewable Resources, University of Alberta)

ABSTRACT

The planet is experiencing a dramatic loss of species. The majority of species are unknown to science, and it is usually infeasible to conduct a census of a region to acquire a complete inventory of all life forms. Therefore, it is important to estimate and conduct statistical inference on the total number of species in a region based on samples obtained from field observations. Such estimates may suggest the number of species new to science and at potential risk of extinction.

In this thesis, we develop novel methodology to conduct statistical inference, based on abundance-based data collected from multiple sampling locations, on the number

of species within a taxonomic group residing in a region. The primary contribution of this work is the formulation of novel statistical methodology for analysis in this setting, where abundances of species are recorded at multiple sampling units across a region. This particular area has received relatively little attention in the literature.

In the first chapter, the problem of estimating the number of species is formulated in a broad context, one that occurs in several seemingly unrelated fields of study. Estimators are commonly developed from statistical sampling models. Depending on the organisms or objects under study, different sampling techniques are used, and consequently, a variety of statistical models have been developed for this problem. A review of existing estimation methods, categorized by the associated sampling model, is presented in the second chapter.

The third chapter develops a new negative binomial mixture model. The negative binomial model is employed to account for the common tendency of individuals of a particular species to occur in clusters. An exponential mixing distribution permits inference on the number of species that exist in the region, but were in fact absent from the sampling units. Adopting a classical approach for statistical inference, we develop the maximum likelihood estimator, and a corresponding profile-log-likelihood interval estimate of species richness. In addition, a Gaussian-based confidence interval based on large-sample theory is presented.

The fourth chapter further extends the hierarchical model developed in Chapter 3 into a Bayesian framework. The motivation for the Bayesian paradigm is explained, and a hierarchical model based on random effects and discrete latent variables is

presented. Computing the posterior distribution in this case is not straight-forward. A data augmentation technique that indirectly places priors on species richness is employed to compute the model using a Metropolis-Hastings algorithm.

The fifth chapter examines the performance of our new methodology. Simulation studies are used to examine the mean-squared error of our proposed estimators. Comparisons to several commonly-used non-parametric estimators are made. Several conclusions emerge, and settings where our approaches can yield superior performance are clarified.

In the sixth chapter, we present a case study. The methodology is applied to a real data set of oribatid mites (a taxonomic order of micro-arthropods) collected from multiple sites in a tropical rainforest in Panama. We adjust our statistical sampling models to account for the varying masses of material sampled from the sites. The resulting estimates of species richness for the oribatid mites are useful, and contribute to a wider investigation, currently underway, examining the species richness of all arthropods in the rainforest.

Our approaches are the only existing methods that can make full use of the abundance-based data from multiple sampling units located in a single region. The seventh and final chapter concludes the thesis with a discussion of key considerations related to implementation and modeling assumptions, and describes potential avenues for further investigation.

KEY WORDS: *species richness; negative binomial mixture; finite mixture model; data*

augmentation; hierarchical model

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

- Dr. Farouk Nathoo and Dr. William Reed, for their mentorship and tremendous patience.
- The University of Victoria and NSERC, for the fellowships and scholarships.
- Dr. Neville Winchester and the Centre for Tropical Forest Science, for providing their data sets for use in my research.
- My sisters, my friends and my colleagues in the Department of Mathematics and Statistics, for their encouragement and humor.

I would like to especially thank my parents for their incredibly kind and generous support.

Chapter 1

Introduction

1.1 Species Richness

The diversity of living organisms on the planet is tremendous. Nearly two million species have been identified and named by scientists (Mace *et al*., 2005). Most estimates of the number of eukaryotic[1] species on Earth range from 5 million to 30 million (May, 1992). Shrinking the focus to a specific ecosystem, it may still be a daunting task to describe the diversity of life forms in the biological community. An investigator may further restrict attention to the species belonging to a specific taxonomic group. A *taxonomic group* is an assemblage of species that are distinguishable, exist in the same geographical region, and may have similar physical sizes and biological characteristics. An important characteristic of the taxonomic group is the total num-

[1] A eukaryote is an organism whose cell(s) have a membrane called the cell nucleus, which contains the genetic material of the organism.

ber of species in the group, or its *species richness*. For example, Picard, Karembé and Birnbaum (2004) studied the number of woody plant species in an arid savanna, and Lindo and Winchester (2006) studied the number of oribatid mite species living in soil mats in the canopy of old growth western red cedar trees. The taxonomic group may be quite diverse with a large number of species unknown to science. Estimates of species richness are a common measure that ecologists use to compare biological communities (Chiarucci *et al.*, 2003).

Conservation of biodiversity is often framed in terms of species diversity as it is easy for the public to understand this level of diversity (Kiester, 2001). Conservation actions may be framed or structured around the number of species that may be protected. In areas under threat of habitat destruction, knowledge of species richness can aid in estimating the number of species that will be extirpated (May, Lawton & Stork, 1995).

A biological community can cover a very large area relative to the sizes of the organisms under study. For example, Condit (1998) and others studied species of woody plants with a stem diameter of at least 1 cm at chest height in a tropical forest that covers most of Barro Colorado Island (15 km^2 area) in Panama. Due to constraints on resources and time, ecologists are limited to studying areas that make up a small fraction of an ecosystem. On Barro Colorado Island, it took one year for a team of twelve to fourteen people to complete the census of approximately 200,000 woody plants in a 0.5-km^2 study plot (Condit, 1998).

In many taxonomic groups of plants and animals, there are a few common species

that will occur with high frequency in field studies, and the vast majority of species will occur with very low frequencies (Hairston, 1959; Hughes, 1986). For such a taxonomic group, many of the uncommon species will not be observed in a field study. In these circumstances, the data gathered from a study can be used to estimate the total number of species in the taxonomic group.

Estimates of the size of a taxonomic group are particularly helpful for poorly known taxa. The estimates give an indication of the unexplored wealth of biological information stored in the ecosystem. Estimates of the expected number of new species that will be detected with additional sampling effort are helpful in planning future studies (Mao, Colwell & Chang, 2005).

For the purpose of this dissertation, the term *species richness* refers to the total number of species (of the taxonomic group of interest) in the region at the time of sampling. We will assume that the boundaries of the region are well-defined.

The positive association between the size of an area and the number of species observed has been examined since the early 1900's (see, e.g., Arrhnenius, 1921; Gleason, 1922). Statistical theory for the estimation of species richness has early origins (see, e.g., Fisher, Corbet & Williams, 1943; Goodman, 1949). In Section 1.2, we describe the problem in general terms. In Section 1.3, we mention specific instances of this estimation problem arising in other fields. In Section 1.4, we clarify the need for new methodological developments, and Section 1.5 outlines the developments in this dissertation.

1.2 The General Scenario

Consider a large population of objects, and suppose each object belongs to exactly one of S_Ω classes, where S_Ω is an unknown positive integer. A sample of the objects is taken from the population, and it is assumed that the classes can be distinguished from each other. The objective is to infer upon S_Ω, based on the sample of objects. In the context of estimating species richness, the population consists of all specimens of all species (within the taxonomic group) in the biological community, and each species represents a class, so that S_Ω is the number of species.

The fraction of the population sampled is typically small. Moreover, the number of objects in each class (individuals of a species) is unknown, making the estimation of S_Ω challenging (Bunge & Fitzpatrick, 1993).

1.3 Specific Instances of the Scenario

The problem of estimating the number of classes in a population appears in various fields of study including genetics, animal population studies, and astronomy. This scenario also arises in the verification of signatures on a petition, the investigation of an author's vocabulary size, and the study of the size of ancient coin-based economies.

The study of the genetic diversity of a single species may involve investigating the *alleles*, alternative forms of a gene, at a specific position on a chromosome. For example, Huang and Weir (2001) investigated the number of distinct alleles of a gene present in nine populations of a honey bee subspecies. In each population, the genetic

material from a sample of individuals was analysed. The frequencies of occurrence of each allele in the samples were tabulated. Estimates of the total number of different allele types in each population gave an indication of the genetic diversity in each honey bee population.

Agencies involved in the conservation and management of an animal species often frame their goals in terms of the corresponding population size (see, e.g., Williams *et al.*, 2002). These agencies often rely on estimates of animal population sizes to help assess the status of a species. Mark-recapture is a sampling method for estimating the population size of a mobile animal species. Individual animals are observed in a series of capture periods. In each capture period, individuals that are captured receive a tag for future identification; the identities of all detected individuals (previously-marked and first-time captures) are recorded, and the animals released. A capture history exists for each individual that was caught at least once in the series of capture periods. Some individuals will not be caught during the capture periods. In this scenario, each individual animal is considered a unique class. The goal is to estimate the total number of individuals (i.e., classes) in the population, given a record of the captured individuals, and their frequencies of capture.

The problem of estimating the number of classes in a population also appears in astronomy and literary studies. In astronomy, Harwit and Hildebrand (1986) estimated the number of types of observational phenomena that exist in the universe, given the types of phenomena that have been already discovered. In a study of William Shakespeare's vocabulary, Efron and Thisted (1976) estimated the size of

Shakespeare's vocabulary using all of his known works of literature. Subsequently, Thisted and Efron (1987) tested whether an unsigned nine-stanza poem discovered in 1985 could be attributed to Shakespeare. This was accomplished by comparing the frequency of word usage in the poem, with the usage of words in the collected works of Shakespeare. Particular attention was paid to the usage of *new* words, words that did not appear in Shakespeare's known body of work.

Stam (1987) used the archaeological discovery of a cache of ancient coins to estimate the size of a past civilization's economy. Although the coins may have been made with two dies (one die for each side of the coin), attention is restricted to the *obverse side*[2] of the coins. Examining the obverse sides of a cache of coins, the frequency of occurrence of each observed die is recorded. The total number of dies used to mint the coins is estimated, based on the number of observed dies and their frequencies of appearance. The estimated average output of a single die can be obtained from historical records and modern experiments if the method of making the coins can be determined. Multiplying the estimated total number of dies and the estimated average output per die, the product is an estimate of total coinage in the civilization's economy.

As another example, Whiteside and Eakin (2004) considered the problem of verifying the legitimacy of all signatures on a large petition. In some cases, a percentage or sample of the signatures may be validated. Some valid signatures may appear

[2]The obverse side of a coin is the side showing a bust of royalty, a national emblem or year of minting.

multiple times on the petition, but are only counted once. The authors demonstrated how different techniques for accounting for multiple appearances of a valid signature can have a significant effect on the estimated number of valid signatures on a petition.

In all such estimation problems, the sampling methods used to collect data in a given instance will influence the development and/or choice of an estimator. Indeed, even within the realm of estimating the number of species, there are many sampling methods. Chapter 2 discusses the estimation methods that are commonly used for a variety of sampling methods.

1.4 The Need for New Developments

For taxonomic groups of relatively immobile organisms, the field observations may be acquired using area-based sampling. For example, one or more small congruent plots of land or substrate may be selected, and all detected individuals are identified to their species type and recorded (see, e.g., Lindo & Winchester, 2006). The number of individuals of a species in a spatial-based sampling unit is called the *sample abundance* of the species. When sample abundances are recorded for all species observed in one or more sampling units, we will describe the resulting data as *abundance-based* data. The occurrence of individuals of a particular species in a given sampling unit may not be independent; that is, individuals may show a positive or negative spatial association with other members of the same species (Taylor, Woiwood & Perry, 1978; Condit *et al.*, 2000). The spatial associations affect the frequencies of occurrence

within the sampling units (He & Legendre, 2002).

To date, estimators of species richness have not been designed and implemented to make full use of abundance-based data acquired from multiple spatial-based sampling units within a single region of interest (see, e.g., Kéry & Royle, 2008; Bunge & Barger, 2009).

1.5 Outline of Dissertation

In this dissertation, we develop two model-based approaches for abundance-based data from multiple sampling units. In the first model, a classical likelihood-based approach is used for statistical inference on species richness based on a mixed negative binomial model, with an exponential mixing distribution. In the second model, we adopt a hierarchical Bayesian approach for inference with a finite mixture model and discrete latent variables.

In Chapter 2, we review several common sampling models and the associated estimators of species richness. We then clarify the motivation for our developments as more appropriate with abundance-based data from multiple sampling units.

Chapter 3 begins with a description of initial attempts at solving this estimation problem. These initial attempts were designed for abundance-based data from multiple adjacent sampling units that cover a rectangular area. The difficulties of the resulting estimator are discussed, followed by the formulation of a new model for abundance-based data from multiple spatial sampling units that are assumed ran-

domly situated in the region. A hierarchical model based on the negative binomial distribution is used to describe the sample abundance data within a sampling unit. Using this model, we develop the maximum likelihood estimator of species richness S_Ω. We also describe two methods for constructing confidence intervals for S_Ω.

In Chapter 4, we develop an alternative hierarchical Bayes model for the abundance data and discuss a data-augmentation technique for computing the posterior distribution of S_Ω using MCMC. The resulting posterior distribution yields both point and interval estimators of species richness.

In Chapter 5, we evaluate both the maximum likelihood estimator and our Bayesian estimators in two simulation studies. The estimators are compared with commonly-used methods, and the bias and precision are assessed.

In Chapter 6, we present a case study examining oribatid mites sampled from the forest floor and the canopy of the tropical rainforest in the San Lorenzo Protected Area of Panama. Here interest lies in estimating the number of species of oribatid mites on the forest floor and in the canopy of the rainforest. We fit our sampling models to the canopy and ground data sets, and compare our species richness estimates with estimates obtained from other methods.

Potential directions for future research are discussed in Chapter 7.

Chapter 2

Literature Review and Motivation for New Species Richness Estimators

2.1 Introduction

The literature on species richness estimation dates back to the first half of the 20^{th} century (e.g., Gleason, 1922; Preston, 1948). In this chapter, an overview of the common sampling models and the resulting estimators are described in Section 2.2. The key challenges are summarized in Section 2.3. In Section 2.4, we locate our methodology within the general context of species richness estimation.

2.2 Sampling Methods and Estimators

In most studies, the types of organisms under study will influence the method of sampling.

For taxonomic groups of plants or relatively slow-moving animals, it may be feasible to complete an inventory of all individuals in one or more small areas (e.g., Plotkin *et al.*, 2000). On the other hand, for taxonomic groups of mobile animal species, animals may be observed or captured at point locations, or their presence may be detected as a field observer travels along a transect line. For example, in the annual North American Breeding Bird Survey (USGS Patuxent Wildlife Research Center, 2010), observers travel along designated routes on secondary roads; at 0.5-mile intervals, they record all bird species seen or heard in a 3-minute duration. In a study of the diversity of moths at the Rothamsted Experimental Station in the United Kingdom, moths were captured in a light trap (Fisher *et al.*, 1943) situated in one location for four years.

Bunge and Fitzpatrick (1993) and Chao (2005) reviewed the literature on species richness estimation. These reviews organize the estimators according to the types of sampling models that the estimators are derived from. In the following sections, the sampling models are briefly described and commonly-used estimators resulting from each sampling model are mentioned.

2.2.1 Random Sample of Individuals

Suppose a random sample is taken from the population of individuals that reside in a region of interest. Each sampled individual is identified within its species type, and the sample size is fixed in advance. If individuals are randomly selected from the population *without* replacement, then the sample data can be described with a multivariate hypergeometric model (Bunge & Fitzpatrick, 1993).

When sampling *without* replacement from a finite population of known size, an unbiased estimator exists if the sample size is larger than the *regional abundance*[1] of any single species (Goodman, 1949). Unfortunately, the variance of Goodman's estimator is so large that it renders the estimator unusable except for situations where at least 10% of the population is sampled. If the sample size is less than the regional abundance of the most abundant species, then no unbiased estimator exists (Goodman, 1949). Solow (1996) developed a maximum likelihood estimator of the species richness under the assumption that the species have identical numbers of individuals in the population. Unfortunately, if the population is large compared to the sample, then the precision of Solow's estimator is poor.

If the random sample of individuals is drawn from a population that is several orders of magnitude larger than the size of the sample, then the effect of sampling without replacement is negligible and models based on sampling *with* replacement provide an adequate approximation.

[1]The regional abundance of a species is the total number of individual organisms of that species in the region of interest.

We let S_Ω denote the unknown species richness of the taxonomic group in the region of interest, and n_i represent the number of individuals of species i that occur in the sample, for $i = 1, 2, \ldots, S_\Omega$. The size of the sample is $n = \sum_{i=1}^{S_\Omega} n_i$, where $n_i = 0$ implies the absence of species i from the sample; and therefore, its existence within the population remains unknown. When sampling *with* replacement, the *sample abundances* $n_1, n_2, \ldots, n_{S_\Omega}$ have a multinomial distribution where the corresponding cell probabilities $p_1, p_2, \ldots, p_{S_\Omega}$ are equal to the *relative abundances*[2] of the species in the population. If the species have identical relative abundances, a minimum variance unbiased estimator exists provided that the sample size n is at least as large as S_Ω (Darroch, 1958). Unfortunately, in ecological applications, it is seldom realistic to assume equal relative abundances for the species (Bunge & Fitzpatrick, 1993).

In parametric approaches, one can treat the cell probabilities $p_1, p_2, \ldots, p_{S_\Omega}$ as random variables from a probability distribution G, such as the inverse Gaussian distribution (Sichel, 1986). Using the observable sample abundances, maximum likelihood estimation of S_Ω and the parameters of the probability distribution G, which can be viewed as a mixing distribution, can be computed numerically (Sanathanan, 1977).

In Bayesian models, a prior distribution on S_Ω is specified (e.g., Solow, 1994). A prior distribution on the relative abundances of the species can also be specified after conditioning on the value of the species richness (e.g., Lewins & Joanes, 1984). The

[2]The relative abundance of a species is the proportion of individuals in the population that belong to that species.

prior distribution on the relative abundances may depend on hyperparameters, and these hyperparameters can have prior distributions in a hierarchical Bayesian model (Lewins & Joanes, 1984). Point and interval estimates of species richness can be calculated from the posterior distribution of S_Ω (Solow, 1994). It may be possible to derive the posterior distribution of S_Ω analytically when the Bayesian model only uses the number of species observed in the sample and ignores the information arising from the sample abundances (Lewins & Joanes, 1984; Solow, 1994). When the sample abundances are incorporated into the likelihood function, it is typically not possible to derive an analytic expression for the posterior distribution, and Markov chain Monte Carlo methods can be used to sample from the joint posterior of S_Ω and the other model parameters (Zhang & Stern, 2005).

Estimates of S_Ω are sensitive to the choice of the mixing distribution G in the parametric likelihood-based setting, and the choice of prior distributions in the parametric Bayesian setting. For example, in the likelihood-based approaches, two different probability distributions on $p_1, p_2, \ldots, p_{S_\Omega}$ may fit the sample abundances equally well, but lead to significantly different estimates of species richness (Chao, 2005). Link (2003) gave analytical arguments and examples to illustrate this non-identifiability problem in the analogous context of estimating the size of an animal population in a capture-recapture scenario.

Non-parametric estimators of S_Ω have been developed for use with a random sample of individuals, sampled with replacement. Non-parametric estimators avoid assumptions on the mixing distribution. The non-parametric estimators are based on

sample statistics approximating the moment properties of the distribution of relative abundances (e.g., Chao & Lee, 1992).

The number S_{obs} of species observed in the random sample can itself be used as an estimate of the region's species richness (Kéry & Royle, 2008). However, this assumes that all species in the taxonomic group have been observed in the sample, and S_{obs} will be a negatively biased estimate when this assumption is false. Burnham and Overton (1979) used jackknife techniques to develop a series of non-parametric estimators, originally developed for estimating the size of an animal population. The sample abundances of the species can be summarized as frequencies. The frequency f_j denotes the number of species that contribute exactly j individuals to the sample, for $j = 1, 2, \ldots, J$, where $J = max(n_i)$ denotes the maximum of the sample abundances. So, $S_{obs} = \sum_{j=1}^{J} f_j$ and $n = \sum_{j=1}^{J} j \cdot f_j$. Adapted for a sample of individuals, Burnham and Overton's first-order and second-order jackknife estimators are

$$\hat{S}_{Jack1} = S_{obs} + \frac{(n-1)f_1}{n}$$

and

$$\hat{S}_{Jack2} = S_{obs} + \frac{(2n-3)f_1}{n} - \frac{(n-2)^2 f_2}{n(n-1)}\,,$$

respectively.

Chao (1984) developed a non-parametric estimator,

$$\hat{S} = S_{obs} + \frac{f_1^2}{2f_2} \tag{2.1}$$

derived by considering a lower bound for S_Ω. Chao and Lee (1992) developed an alternative non-parametric estimator that uses all of the observed frequencies, f_1, f_2,

$f_3, \ldots, f_J$, assuming that the distribution of relative abundances is fully characterized by their mean and coefficient of variation.

Non-parametric estimators are useful when no prior information on the relative abundances is available. However, it is not atypical to have populations with a large number of species and with very small relative abundances. In this case, as discussed by Engen (1978), the bias of a non-parametric estimator can be unreasonably large, and in fact, unbounded if the only information available is the observed frequencies $f_1, f_2, f_3, ..., f_J$.

2.2.2 Stochastic Abundance Models

In the preceding section, the sample size was assumed fixed. In other scenarios, the sampling may occur for a pre-specified amount of time, area, or volume of space, however, the sample size is a random variable. For example, a sample may consist of all moths caught in a light trap over the course of four consecutive years (Fisher *et al.*, 1943). Each species contributes a random number of individuals to the sample according to a stochastic model. The information in the sample can thus be summarized by the sample size n, the number S_{obs} of species observed, and the frequencies $f_1, f_2, \ldots, f_J$.

The Poisson process is commonly used in stochastic abundance models (e.g., Fisher *et al.*, 1943; O'Hara, 2005). Indexing the species from 1 to S_Ω, individuals of the i^{th} species enter the sample according to a homogeneous Poisson process with rate λ_i, for $i = 1, 2, \ldots, S_\Omega$.

When species have equal rates of entering the sample (i.e. $\lambda_1 = \lambda_2 = \ldots = \lambda_{S_\Omega}$), under the Poisson sampling model, the maximum likelihood estimator can be derived analytically. Darroch and Ratcliff (1980) developed an estimator, $S_{obs}/(1 - f_1/n)$, and demonstrated that this estimator has high efficiency relative to the MLE derived assuming equal rates.

More realistically, the species have different rates of entering the sample, and the Poisson rates $\lambda_1, \lambda_2, \ldots, \lambda_{S_\Omega}$ are treated as independent random variables from a mixing distribution G with probability density function $g(\lambda; \boldsymbol{\theta})$ indexed by $\boldsymbol{\theta}$ (e.g., Fisher *et al.*, 1943; Kempton & Taylor, 1974). The gamma distribution (Fisher *et al.*, 1943), lognormal distribution (Bulmer, 1974), inverse-Gaussian distribution (Ord & Whitmore, 1986) and generalized inverse-Gaussian distribution (Sichel, 1997) have been suggested as the mixing distribution for the Poisson sampling model. Having imposed a mixing distribution G on the Poisson rates, the marginal distribution of the proportion of species appearing r times in the sample is

$$P_G(r; \boldsymbol{\theta}) = \int_0^\infty \frac{e^{-\lambda}\lambda^r}{r!} g(\lambda; \boldsymbol{\theta})\, d\lambda \,,$$

for $r \geq 0$. The likelihood function for the unknown S_Ω and parameter vector $\boldsymbol{\theta}$ is

$$L(S_\Omega, \boldsymbol{\theta}) = \frac{S_\Omega!}{(S_\Omega - S_{obs})! \prod_{r\geq 1}(f_r!)} \left[P_G(0; \boldsymbol{\theta})\right]^{S_\Omega - S_{obs}} \prod_{r\geq 1} \left[P_G(r; \boldsymbol{\theta})\right]^{f_r} \,. \tag{2.2}$$

The maximum likelihood estimates of S_Ω and $\boldsymbol{\theta}$ are calculated numerically (e.g., Bulmer, 1974; Ord & Whitmore, 1986; O'Hara, 2005), as the MLE of $\boldsymbol{\theta}$ cannot, in general, be obtained analytically. The MLE of S_Ω under the likelihood in Equation

2.2 can be derived as

$$\hat{S}_\Omega = \left\lfloor \frac{S_{obs}}{1 - P_G(0; \hat{\boldsymbol{\theta}})} \right\rfloor ,$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ (Sanathanan, 1977) and the floor function $\lfloor x \rfloor$ maps a real number x to the largest integer not greater than x.

An alternative to the Poisson sampling model is the negative binomial sampling model (O'Hara, 2005). The negative binomial sampling model permits wider variability in the sample abundances of a species relative to the expected value, and this sort of overdispersion is typical. The number of individuals of the i^{th} species that enter the sample can be treated as a negative binomial random variable with mean rate λ_i and dispersion parameter k. O'Hara (2005) used lognormal and gamma probability distributions for the mixing distribution of the rates $\lambda_1, \lambda_2, \ldots, \lambda_{S_\Omega}$ on the species.

The parametric maximum likelihood approach can be viewed as an empirical Bayesian approach if one considers the mixing distribution G to be a prior distribution whose parameters $\boldsymbol{\theta}$ are estimated (Chao, 2005). Rodrigues, Milan, and Leite (2001) described empirical Bayesian and hierarchical Bayesian estimators under the Poisson sampling model; the Poisson rates had a gamma prior distribution and prior probability distributions were imposed on the parameters of the gamma distribution. Rodrigues *et al.* considered two different prior distributions on S_Ω: a prior proportional to $(S_\Omega)^{-1}$, and a Poisson prior on S_Ω. For the hierarchical Bayesian approach, Rodrigues *et al.* estimated the posterior distribution of S_Ω using Markov chain Monte Carlo simulations.

Maximum likelihood estimators and Bayesian estimators in the stochastic abun-

dance model setting face the same challenges as the estimators based on a random sample of individuals. Two different candidates for the mixing distribution or the prior distribution on the rates of species contributing individuals to the sample may adequately fit the sample data but yield significantly different estimates of species richness. Furthermore, a good fit to the data does not necessarily imply that the estimate of S_Ω is adequate (Chao, 2005).

Under a Poisson sampling model, if one conditions on the size of the sample, the sample abundances have a multinomial distribution (Bunge & Fitzpatrick, 1993). Consequently, the non-parametric estimators from the preceding section can be used. However, in Section 2.2.1, the sample size n was fixed, so the variance of a non-parametric estimator will be different under the two sampling models (Chao, 2005).

Norris and Pollock (1998) developed a non-parametric maximum likelihood estimator of species richness using a Poisson sampling model. Instead of assuming a parametric form for the mixing distribution, they used a discrete distribution with a finite number of support points for the Poisson rates. They employed an EM algorithm to estimate the number of support points, and the values and weights of the support points. Mao and Lindsay (2004) took finite mixtures of the Poisson sampling model a step further by looking at samples from multiple regions; one sample was taken from each region and species were permitted to have different Poisson rates of entering the samples in the different regions.

2.2.3 Multiple Bernoulli Samples

In Sections 2.2.1 and 2.2.2, the number of individuals of each species observed in a sample was recorded. For some sampling methods in ecology, it is difficult to count the number of individuals of a species. Only the presence or successful detection of the species is recorded. For example, on a road-side route in the annual North American Breeding Bird Survey (USGS Patuxent Wildlife Research Center, 2010), the detection of a bird species can be quickly noted, but accurately determining the number of individuals of a species requires more effort.

This section considers multiple observation or sampling sessions where only the detection of species are recorded in each session. The detection of a species in a sampling session can be considered a Bernoulli trial, and the probability of detection may change with the sampling session and with the species. Each sampling session can be modelled as consisting of a collection of Bernoulli trials, one for each of the S_Ω species. These multiple Bernoulli sampling models are appropriate for taxonomic groups of mobile species where it may be difficult to acquire accurate counts of observed individuals of each species during the sampling sessions. Records of detection can also be obtained from some sampling methods (e.g., quadrat sampling – inspecting several congruent plots) of surveying plants and terrestrial and aquatic organisms that are slow-moving relative to the pace of sampling (e.g., Ugland, Gray, and Ellingsen, 2003).

The class of estimators based on multiple Bernoulli samples originates from re-

search on the estimation of the size of animal populations (e.g., Burnham & Overton, 1979). In mark-capture-recapture studies, animals of one species are caught, marked for identification and released in a series of capture periods. The records of capture for all the animals captured at least once are used to estimate the total number of animals in the population. If the series of capture periods takes place over a short interval of time, then the population may be assumed to be *closed*[3] in order to simplify the analysis.

Under the assumption of a closed population, the simplest multiple Bernoulli model, M_0, treats all the animals as having identical probabilities of capture and this common probability of capture for the individuals is used for each sampling session. More sophisticated closed-population models (e.g., Otis *et al.*, 1978) allow capture probabilities to vary with the sampling sessions, to vary by behavioural response to capture, and/or to vary by individual animal. For example, in the model M_h, the capture probabilities are permitted to vary among the individuals, but an individual has the same capture probability in each sampling session.

Returning now to the issue of estimating the total number of species based on the records of detection from multiple sampling sessions, a species is detected or captured in a sampling session when at least one specimen of the species is detected in the sampling session. The taxonomic group of species in the region is treated as a closed population for the duration of the sampling sessions. In other words, it is

[3]For the duration of the capture periods, the population does not change; there is no migration and no births or deaths.

assumed that the species composition in the region does not change (i.e., no losses or additions of species) during the period of sampling.

The probability of the occurrence and successful detection of a species during a sampling session can be expressed as the probability of occurrence in the area or volume of space under going sampling multiplied by the conditional probability of being detected in the sampling session given that the species is present. In the following discussion, the term "probability of detection" refers to the probability of the occurrence *and* successful detection of a species during a sampling session. The probability of detection of a species depends on several factors including the regional abundance of the species, its spatial distribution of individuals across the region, its biological characteristics (e.g., size, colouring, vocalizations, etc.), and the ability of the field observers (Boulinier *et al.*, 1998).

With the simplest capture-recapture model M_0, all species in all sampling sessions are treated as having the same probability of being detected. Under model M_0, the maximum likelihood estimator of S_Ω does not have a closed-form expression, and is computed numerically.

In some species assemblages such as tropical ants and other insect groups, some species will have large regional abundances but many other species will be scarce (Mao & Colwell, 2005). As a result, the abundant species tend to be detected in nearly all sampling sessions; whereas, many rare species will only be detected in a handful of sampling sessions, if at all. Under these circumstances, when relative abundances are unequal, using estimators derived from M_0 will result in negatively biased estimators

(Boulinier *et al.*, 1998). Thus, estimators based on a more elaborate model (e.g., M_h) are appropriate.

In the context of species richness estimation, the M_h model permits species to have different probabilities of being detected. The probability of detection of a species under the M_h model remains the same for each sampling session. We suppose that there are τ sampling sessions, and that a species may be detected in 0, 1, 2, ..., $\tau - 1$, or τ sampling sessions. The number of sampling sessions in which a species is captured, under appropriate assumptions, is a binomial random variable.

Under M_h, Yip (1991) proposed a class of parametric estimators based on a beta distribution to model the probabilities of detection of the species, and Martingale theory to derive an estimating equation for species richness.

Dupuis and Joachim (2006) explored a Bayesian approach to estimating species richness using the detection history of species in a random sample of quadrats. Dorazio *et al.* (2006) and Royle, Dorazio, and Link (2007) also developed a Bayesian approach; their sampling protocol requires multiple sites, each to be visited a fixed number of times, and the detection of species recorded on each visit. The resulting replication arising from repeated visits allow for the modelling of the probability of occurrence of a species in a site and the probability of its detection given occurrence. Rather than directly impose a prior distribution on the species richness S_Ω, Dupuis and Joachim (2006), Dorazio *et al.* (2006) and Royle *et al.* (2007) view the number of species in the region as being a subset of a larger collection of species, where the number of species in this larger collection is known.

Burnham and Overton (1979) developed a family of non-parametric jackknife estimators for the M_h model. Based on the original estimator being the total number of species observed, they developed the k^{th} order jackknife estimators, for $k = 1, 2, \ldots, \tau$ where τ is the total number of capture periods or sampling sessions. Non-parametric estimators from other sampling models, like Chao's estimator presented in Equation 2.1, have been adapted for detection history data (Chao, 1987).

2.2.4 Species-Area Curves and Species-Accumulation Curves

A positive relationship between the area of a site and the number of species in the site is common (Arrhenius, 1921; Gleason, 1922). These relationships have been studied for different groups of organisms and subsequently have been modelled with mathematical equations. One purpose of modelling the relationship is to estimate the total number of species in the region where the surveyed sites are situated (e.g., Palmer, 1990; Grassle and Maciolek, 1992).

This section introduces this approach, also called species-area curves and species-accumulation curves. The curves model the relationship between the number of species observed and the corresponding sampling effort. The sample abundances of the species do not need to be recorded; only the detection of the species in the sites or sampling units is used.

Species-Area Curves

When sites of different areas are surveyed, the numbers of species found in the sites can be plotted against the areas of the sites. A mathematical equation fit to the resulting species-area graph is called a species-area curve. A species-area curve is a monotonically increasing function of the area of sites. The power equation $S(A) = c_1 A^{c_2}$ was the first mathematical description of a species-area curve (Arrhenius, 1921); $S(A)$ represents the mean or expected value of the number of species found in random sites with area A, and c_1 and c_2 are positive constants. Gleason (1922) proposed using an exponential equation $S(A) = c_3 + c_4 \cdot ln(A)$ for the species-area curve, where c_3 and c_4 are constants. Arrhenius and Gleason both suggested empirical support for the power and exponential species-area curves, respectively.

The species-area curve should have a sigmoidal shape when the region has a fixed number of species whose individuals are independently and randomly distributed across the region (He & Legendre, 1996). He and Legendre (1996) related the exponential, power, and logistic equations in a general model for the relationship between the number of species and the area. The fit of the power, exponential, logistic, and other curves depends on the interval of values for A and the scales of variation in the environment (He & Legendre, 1996; Gurevitch *et al.*, 2002).

After fitting a curve to the species-area data, the horizontal asymptote of the curve is the estimate of the region's species richness. For curves that do not have a horizontal asymptote (e.g., the power and exponential equations), the estimate of the

region's species richness is the value of $S(A)$ corresponding to the region.

The species-area graph is sensitive to the way in which the data are gathered. The species-area graph will tend to increase faster if separate sites are used rather than neighbouring or nested sites as the species composition will generally be more similar in neighbouring and nested sites, compared with sites that are separated from each other (Palmer, 1990).

Different models of the relative abundance and spatial patterns of the species can result in the same species-area curve. For example, He and Legendre (1996) and Harte *et al.* (1999) discussed two different models that both result in a power equation for the species-area curve.

Species-Accumulation Curves

Depending on the sampling method, the sampling effort may be measured by the number of specimens captured, the number of quadrats that have been surveyed, the amount of area that has been examined, the volume of soil/sediment/water that has been screened, the number of hours of observation, and so on. In addition, the identification of new species is recorded when they are first observed. The accumulation of species can be considered alongside the amount of sampling effort. The resulting curves, called species-accumulation curves, can be fit to the graph of the accumulated number of species versus sampling effort.

Like the species-area curves, the species-accumulation curves can be used to estimate the number of species in a region. Species-accumulation curves can be applied

to many sampling schemes (Chao, 2005), including a random sample of individuals or the detection history of species from multiple sampling sessions.

At least nine species-accumulation curves have been used in the past (Flather, 1996) including the power, exponential and logistic equations mentioned in the previous discussion on species-area curves. Two other common species-accumulation curves are the Michaelis-Menten equation and the negative exponential equation (Colwell & Coddington, 1994). Let $S(t)$ represent the expected number of species accumulated after t units of sampling effort. The Michaelis-Menten equation is

$$S(t) = \frac{S_\Omega \cdot t}{c_5 + t} \, ,$$

where c_5 is a positive constant. The negative exponential equation can be expressed as

$$S(t) = S_\Omega \cdot \left(1 - e^{-c_6 t}\right) \, ,$$

where c_6 is a positive constant.

When a mathematical curve is fit to the species-accumulation graph (e.g., using least squares), the estimate of the region's species richness is obtained as the horizontal asymptote of the curve. For curves without horizontal asymptotes, if T units of sampling effort would acquire the entire population, then the species-accumulation curve is evaluated at $t = T$ to estimate the total number of species in the region.

The order in which sample units are accumulated does affect the species-accumulation graph (Colwell & Coddington, 1994). To avoid the order having an effect, species-accumulation graphs can be generated for all or a large number of random orderings

of the sampling units. The mean number of species observed can then be computed at each value t of sampling effort.

Estimating species richness from species-area curves and species-accumulation curves suffer from several related problems (Chao, 2005). Although a curve may fit the data well, the estimate of species richness in the region is beyond the range of the available sample data, and hence, an extrapolation. If the amount of sampling is insufficient to reveal an asymptote in the graph, then different curves may adequately fit the sample data, but produce significantly different estimates (Chao, 2005). The expected shape of a species-accumulation graph depends upon the relative abundances of the species (Colwell & Coddington, 1994) and the spatial patterns of the species (He & Legendre, 2002) in situations where the sampling effort has a spatial component. In addition, the variance of an extrapolated estimate of species richness obtained from a species-accumulation curve cannot be justified without making assumptions about the organisms under study, such as their relative abundances and spatial patterns (Chao, 2005).

Bunge and Fitzpatrick (1993) suggested that rather than curve-fitting, more efficient estimators would result from explicitly using the parametric statistical model that these curves are derived from. This also allows for formal statistical inference.

2.3 Challenges

Ideally, a description of the species diversity of a taxonomic group would include the species richness, a list of all species in the group, the relative abundances of the species and their spatial distributions across the region of interest. However, due to constraints on time and resources, field surveys usually cover less than 1% of the region of interest (Chiarucci *et al.*, 2003). The relative abundances of the species in the region are unknown, aside from the information gathered during sampling. In addition, as the true species richness is seldom known, the actual bias of any estimate of species richness, in real applications, is difficult to assess (O'Hara, 2005).

The parametric approaches to species richness estimation require simultaneous inference on the relative abundances (or equivalently, the regional abundances) of the species. However, with a small sample, inference is difficult, particularly in the lower tail of the distribution of relative abundances. Species that have the smallest relative abundances tend to appear infrequently, if at all, in the sample. Different choices for the distribution of relative abundances can lead to significantly different estimates of species richness (Chao, 2005).

The estimators in Section 2.2.1 of this chapter assume that the data are from a random sample of individuals. In ecology, sampling usually occurs at one or more spatial locations in the region (e.g., light traps for capturing moths – Kempton & Taylor, 1974; observation points at regular intervals on routes in the North American Breeding Bird Survey – USGS Patuxent Wildlife Research Center, 2010). As

a result, individuals that are closer to the sampling locations are more likely to be included in the sample. Consequently, the spatial distribution of each species should be considered, because individuals of a species tend not to be completely randomly distributed (He & Legendre, 2002) and the aggregation of *con-specific*[4] individuals does affect species richness estimators (Chao *et al.*, 2009).

If sampling is conducted at multiple locations in a region, locations that are in close proximity will have observations that exhibit *spatial autocorrelation*[5]. Dormann *et al.* (2007) discussed methods to account for the spatial autocorrelation. The application of these methods has so far been limited to the species *observed* in the field study. Given the group of species that were observed in the initial sampling effort, the number of these observed species to occur in other parts of the region is predicted using covariates such as vegetation indices, topography, and satellite imagery (Luoto *et al.*, 2004; St-Louis *et al.*, 2009).

Due to the difficulties with making inference about the relative abundances and spatial distributions of the species, one may settle for lower and upper bounds on the species richness in a region. The number, S_{obs}, of species observed in the sample is a strict lower bound on the species richness. Chao's (1984) non-parametric estimator was designed to be a lower bound. Determining an upper bound on species richness is more difficult. Without making an assumption about the parametric form of the distribution of relative abundances, one cannot exclude the possibility of a very large

[4] *con-specific* means belonging to the same species

[5] Spatial autocorrelation is an increasing lack of independence between observations as the distance between the observations decreases.

number of species with very small relative abundances in the population (Harris, 1959). Mao and Lindsay (2007) showed that non-parametric approaches may suffer from confidence intervals with unbounded upper limits, in some cases.

2.4 Rationale for New Estimators

When abundance-based data is acquired from multiple sampling units of identical shape and size such as plots of land, one or more of the three following methods have been used to work with the data:

1. An estimate of species richness is computed based solely on the sample abundances within a single sampling unit. This is repeated for each sampling unit. These separate estimates do not take advantage of the data from the other sampling units, and therefore the estimates are imprecise (Kéry & Royle, 2008).

2. The abundance-based data from all sampling units are pooled together and treated as one large sample. For each species, its sample abundances are summed over the sampling units. Estimators designed for a single sample of individuals can then be applied. However, combining all the observations together loses the information specific to each sampling unit in the process (Kéry & Royle, 2008).

3. The data is reduced to detection histories for the species (e.g., Ugland *et al.*, 2003; Lindo & Winchester, 2006). For each species observed during sampling, it

is detected in a sampling unit if its sample abundance is greater than zero for the sampling unit. Estimation methods based on multiple Bernoulli models and/or species-accumulation curves are applied to the detection histories. Reducing the abundance-based data to detection histories makes use of some, but not all of the information from the sampling units.

These existing approaches do not make full use of abundance-based data acquired from multiple sampling units randomly distributed within one region.

Chao *et al.* (2000) and Chao, Shen and Hwang (2006) developed estimators for the number of species shared in common between two communities. Pan, Chao, and Foissner (2009) developed an estimator of a lower bound on the number of species shared in common among a set of multiple communities. These methods assume that one sample of individuals is obtained from each community, and individuals are assumed to be randomly sampled with replacement. The abundance-based data in a sample was treated as having a multinomial distribution. These "multiple community" estimators could be applied to multiple samples taken from a single community. Under the corresponding multinomial sampling models, if a species has the same multinomial cell probability associated with each sample, then the samples could be pooled together to form one large sample. Estimators based on a single random sample of individuals (e.g., estimators from Section 2.2.1) could then be used instead of the estimators proposed by Chao *et al.* (2000, 2006) and Pan *et al.* (2009). Unfortunately, the resulting multi-sample estimators offer no advantage over estimators based on a single sample, when all samples are drawn from only one community.

Mao and Lindsay (2004) proposed a non-parametric maximum likelihood estimator of the total number of species that exist in a collection of communities. One sample is gathered from each community. A species contributes individuals to a sample according to a Poisson process, and the rate of the Poisson process is permitted to vary for this species across the different communities. One might consider applying Mao and Lindsay's estimator to multiple samples taken from one community. In that case, a given species has the same Poisson rate associated with each sample. The samples can be combined together to form a single sample, and the single sample Poisson models (e.g., Section 2.2.2) would be applicable. Under the Poisson and multinomial sampling models, using multiple samples from the same community offers no advantage over a single sample of comparable size.

We have discussed how current species richness estimation methods would need to combine or simplify the abundance-based data from the multiple sites, and therefore, the existing methods lose some or all of the site-specific information. Given the sparse sampling that is typical in most studies, such information loss is undesirable.

Estimating species richness using *multiple samples* from a *single community* is an open area of research (Bunge & Barger, 2009). In the next chapter, we develop a stochastic abundance model for such data and discuss likelihood-based inference.

Chapter 3

A Negative Binomial-Exponential Mixture Model

In this chapter, a new parametric maximum likelihood estimator of species richness is developed. The estimator relies upon abundance-based data gathered from the examination of spatial-based sampling units randomly located within the region of interest.

A previous attempt to estimate species richness using a grid of contiguous sampling units is outlined in Section 3.1. The difficulties encountered with this initial attempt prompted a change in the sampling method and the rigorous development of a statistical sampling model. In Section 3.2, we explain the rationale for the negative binomial distribution and we also discuss probability distributions for approximating the distribution of regional abundances. Subsequent sections present the hierarchical negative binomial model and develop likelihood-based inference for this model.

3.1 Initial attempt: Observations from a Grid of Quadrats in One Site

Picard, Karembé and Birnbaum (2004) developed a species richness estimator that explicitly incorporated the spatial patterns of the species observed in a study site. Their method required the spatial coordinates of each specimen in the site. We attempted to modify their estimator to apply to data with a coarser resolution of spatial information (i.e., abundance-based data from quadrats that form a grid, covering the site). A major difficulty with this modified estimator was the development of formal inference on the sampling distribution of the estimator.

3.1.1 PKB's method for species richness estimation

Picard, Karembé and Birnbaum (PKB) were interested in estimating the number of woody plant species in a tree savanna in Mali. The field observations consisted of a census of a 0.5-hectare site. All plant specimens with a base girth greater than 10 cm were assumed to be detected and correctly identified in the census.

PKB's estimation method was based on three steps. In the first step, PKB analysed the *spatial point pattern*[1] of each species observed in a 0.5-hectare site using Ripley's K-function (Ripley, 1981). For a species that exhibited completely random spatial patterns, its spatial distribution was modelled by a homogeneous Poisson point

[1]The spatial point pattern of a species is the set of coordinates denoting the locations of all observed individuals of the species in the study site.

process. A *Matérn point process*[2] was used to model the point pattern of each species that exhibited spatial aggregation. For species not observed in the site, PKB assumed they had random spatial patterns modelled by homogeneous Poisson point processes.

PKB assumed that a model of the spatial distribution of a species extends to similar habitat in the tree savanna beyond the study site. In addition, it was assumed that the species are mutually independent, so that the parameters of a point process model can be estimated for a species without accounting for dependence on other species.

The second step involved inference on the *spatial densities*[3] of the woody plant species. PKB ranked the *sample spatial densities*[4] of the species found in the study site from largest to smallest. They fit a geometric series to the ranked sample spatial densities. For the species absent from the study site, PKB extrapolated the geometric series for the spatial densities of the species absent from the site, assuming they would have the smallest spatial densities of all species.

[2]A Matérn point process (Stoyan & Stoyan, 1994) is a model for the construction of one type of aggregated point pattern. A set of discs of radius R are distributed randomly in a two-dimensional plane according to a homogeneous Poisson process with spatial density ω. Points are independently and uniformly distributed inside the discs. The number of points associated with each disc is an independent Poisson random variable with mean λ. The collection of points created from the Matérn point process is an instance of a point pattern.

[3]The spatial density of a species refers to its average number of individuals per unit area. The spatial density of a species is equal to its regional abundance divided by the area of the region.

[4]The sample spatial density of a species is the total number of individuals of a species in the study site divided by the area of the site.

Given a point process model and an estimated spatial density of a species, the third step estimated the probability of the species occurring in a disc-shaped area of any size. If a new site labelled A is the same size and shape as the original 0.5-hectare study site, the expected number, $E(S_A)$, of species in A should be close to the number, S_{obs}, of species observed in the original study site. Using the point process models and estimated spatial densities for all species, PKB determined the number $\tilde{S}_\Omega$ of species in the region needed for $E(S_A)$ to equal S_{obs}.

3.1.2 Modifying PKB's Estimation Method

The steps of PKB's estimation method were modified to accommodate a change in the format of the data. Rather than require the spatial coordinates of all individuals observed in a study site, the rectangular-shaped site is partitioned into a grid of uniformly-sized contiguous square quadrats, such that the quadrats form rows and columns parallel with the sides of the site. In each quadrat, the sample abundances of the species are recorded. This change in data format is applicable to scenarios where it is not possible to specify the precise spatial coordinates of individuals.

In the first step of the modified estimation method, a separate spatial analysis is performed for each observed species. The spatial analysis uses an agglomerative method (Greig-Smith, 1952) on the sample abundances in the grid of quadrats and a randomization test (Cressie, 1993) to detect significant spatial aggregation. Like PKB, we use homogeneous Poisson point processes and Matérn point processes to model random and aggregated spatial distributions, respectively. We also assumed

species absent from the site had completely random spatial distributions.

In the second step, we use the sample spatial densities of the species as estimates of their true spatial densities. For inference on the spatial densities of the species absent from the site, we pooled the sample abundances from the quadrats together. Then, treating the spatial densities of the species as arising from a lognormal distribution, we fit a lognormal-mixed Poisson sampling model (see Kempton & Taylor, 1974) to the single sample of abundance data in the site. In the third step, we estimate the number of species in the region in a similar manner as PKB.

We investigated the use of a first-order jackknife resampling method (Cox & Hinkley, 1974) to reduce the bias of our estimator and to construct confidence intervals for the species richness. We tested our modified PKB estimation method in two scenarios. In the first scenario, fifty replicate data sets were constructed. In each data set, the species had completely random spatial patterns; the spatial densities of the species were randomly generated from a lognormal distribution. The modified PKB method consistently overestimated the species richness. Interestingly, the jackknife resampling method actually increased the bias and the variance of the species richness estimates and produced confidence intervals whose coverage probabilities did not come close to meeting the nominal coverage level. A single-sample maximum likelihood estimator based on a lognormal-mixed Poisson sampling model outperformed the modified PKB method with little bias and comparable variance.

In the second scenario, fifty replicate data sets were constructed. In each data set, each species had an aggregated spatial pattern generated from a Matérn point

process; the spatial densities of the species were randomly generated from a lognormal distribution. In this scenario, the modified PKB method had smaller bias and MSE than the MLE from the lognormal-mixed Poisson sampling model; however, the jackknife resampling still did not reduce the bias of the modified PKB estimates.

Investigation into the poor performance of the jackknife resampling revealed that the assumption of normality of the jackknife pseudo-values (Cox & Hinkley, 1974) was not satisfied.

A general observation in ecology is that as the distance between locations shrinks, the similarity in observations between the locations tends to increase (Dormann *et al.*, 2007). The presence of this spatial autocorrelation implies that observations in neighbouring quadrats should not be treated as independent. The lack of independence makes inference on the species richness difficult when using a grid of contiguous quadrats.

The multiple steps in the estimation method and the likely presence of spatial autocorrelation make it difficult to find a satisfactory method of estimating the variance of the modified PKB estimator, and producing reliable confidence intervals for species richness. Our approach also requires the estimation of a large number of parameters in the process of estimating the species richness. The method fits separate point process models to the observed species. Therefore, the total number of parameters used in the estimation of S_Ω is proportional to the number of species observed.

The difficulties with this initial attempt motivated the development of a species richness estimator that is described in the remainder of this chapter. Like $\tilde{S}_\Omega$, the new

estimator uses abundance-based data acquired from the inspection of spatial-based sampling units. However, the sampling units are assumed randomly located in the region (i.e., not adjacent to each other) to minimize spatial autocorrelation.

3.2 Candidates for the Statistical Sampling Model

The new estimator of species richness is developed for the study of multiple randomly-placed spatial-based sampling units within the region of interest. The sampling units are assumed to have identical shape and size (e.g., a two-dimensional square quadrat, or a fixed volume of substrate). The sampling units are also assumed to be sufficiently far apart as to not overlap. As well, the spatial autocorrelation between sampling units is assumed to be negligible such that observations between sampling units can be treated as independent. Observers conduct a study of each sampling unit for organisms from the taxonomic group of interest. Each specimen is identified to its species type. The number of individuals (i.e., sample abundance) of each species observed in a sampling unit is recorded.

Our development of a statistical model for the sample abundances of species in the sampling units has two stages. In the first stage, a number of probability distributions are considered for modelling the number of individuals of a species to occur in a sampling unit. The second stage explores models for the distribution of the regional abundances and this is a mixing distribution.

3.2.1 Modelling the Aggregation of Con-specific Individuals

The sampling units are randomly sampled without replacement from the region. However, the region is assumed to be large enough that the observations between sampling units are considered to be roughly independent. In addition, the spatial distribution of a species is assumed to be consistent across the region; that is, a species exhibits the same spatial pattern everywhere in the region.

If a species has a completely random spatial pattern where individuals are independently and randomly placed in the region, the number of its individuals that occur in a sampling unit will have a binomial distribution. The binomial distribution could be approximated with a Poisson distribution when the sampling unit is a small fraction of the entire region (Plotkin & Muller-Landau, 2002).

A problem with the use of a Poisson distribution arises from the fact that individuals of a species tend to occur in clusters or patches in nature. This tendency toward spatial aggregation has been observed for many types of organisms including crustaceans, insects, fish, birds, orchids (Taylor, Woiwood & Perry, 1978) and woody plants (Condit *et al.*, 2000). As a result, the sample abundances of a species obtained from a random sample of sampling units will tend to have a sample variance that is larger than the sample mean, making the Poisson assumption untenable.

Discrete probability distributions with variances larger than their means can be constructed from mixtures of Poisson distributions. For example, if the mean of the Poisson distribution is itself treated as a random variable with a gamma distribu-

tion, then the resulting gamma-mixed Poisson model, more commonly known as the negative binomial distribution, allows for overdispersion. The lognormal distribution (Cassie, 1962) and generalized inverse Gaussian distribution (Sichel, 1975) can also be used as continuous mixing distributions on the Poisson mean. However, the resulting lognormal-Poisson mixture and generalized inverse Gaussian-Poisson mixture have more complicated probability mass functions than that of the negative binomial distribution, which has a closed-form expression.

Several discrete probability distributions have been used to model the sample abundance of a species when its individuals exhibit spatial aggregation. The Thomas (Thomas, 1949), Neyman Type A (Neyman, 1939), Pólya-Aeppli (Pólya, 1930) and negative binomial distributions arise when modelling individuals as occurring in clusters such that the clusters are distributed across the region according to a homogeneous Poisson process. In the Thomas and Neyman Type A distributions, the number of individuals in a cluster has a Poisson distribution. In the Pólya-Aeppli distribution, the number of individuals in a cluster has a geometric distribution. The negative binomial distribution arises when the number of individuals in a cluster has a logarithmic distribution (Elliott, 1977). Several other processes result in a negative binomial distribution (Johnson, Kotz & Kemp, 1993), suggesting that the negative binomial distribution provides a general model of sample abundances for a variety of aggregated spatial distributions (Gaston *et al.*, 2006).

Through simulations, Pielou (1957) found the Neyman Type A and Thomas distributions to be inadequate for modelling the sample abundances of a species unless

the individuals formed very tight clusters which would lie entirely inside or outside a sampling unit. Martin and Katti (1965) found the Neyman Type A and negative binomial distributions to be widely applicable to 35 data sets from a variety of organisms that exhibited spatial aggregation. The Pólya-Aeppli distribution has a rather complicated probability mass function involving the confluent hypergeometric function. Furthermore, the Pólya-Aeppli distribution originates from a specialized set of circumstances and is not as widely applicable as the negative binomial distribution (Elliott, 1977).

Taylor's Power Law covers a wider range of spatial distributions for the individuals of a species than the negative binomial model (Elliott, 1977). In Taylor's Power Law (Taylor, 1961), the variance σ^2 of the numbers of individuals of a species appearing in the sampling units is a power function of the mean μ: $\sigma^2 = a\mu^b$ where a and b are real numbers with $a > 0$. This common empirical relationship corresponds to the Tweedie family of probability distributions (Dunn & Smyth, 2005) for values of b outside the interval $(0, 1)$. Unfortunately, closed-form expressions of the probability functions for Tweedie distributions only exist for $b = 0$, 1, 2, and 3, resulting in the Gaussian distribution, Poisson distribution (when $a = 1$), gamma distribution and the inverse Gaussian distribution, respectively (Dunn & Smyth, 2005). Furthermore, the Gaussian, gamma and inverse Gaussian distributions are continuous probability distributions which excludes them from consideration for modelling the integer-valued sample abundances.

The spatial pattern of a species can have multiple scales and the pattern may

not be consistent across the region under study. Changing the size or shape of the spatial-based sampling units can change the fit of a model (Greig-Smith, 1983). There is no single probability distribution that will be appropriate in all situations for all species. Nevertheless, if we assume the spatial pattern of a species is approximately consistent across the region, then modelling the sample abundances of a species in uniformly-shaped sampling units is still useful for inferential purposes (Krebs, 1999).

We choose the negative binomial distribution to model the sample abundances of a species in the sampling units because of its wide approximation to the spatial distributions of species (Martin & Katti, 1965; Elliott, 1977; Gaston *et al.*, 2006). The negative binomial distribution also has a relatively simple probability mass function compared with the other probability distributions under consideration, which will reduce the time and complexity required to complete numerical computations such as the maximization of the log-likelihood function in Section 3.5.

3.2.2 Distribution of Regional Abundances of the Species

A species with a large regional abundance (number of individuals of the species in the region of interest) will have a higher average number of individuals in the sampling units than a species with a small regional abundance. Thus, the mean number of individuals a species contributes to a sampling unit will depend on the regional abundance of the species. As the second stage in developing a statistical sampling model for the observations in the sampling units, this section explores probability distributions for modelling the distribution of regional abundances of the species. These

probability distributions represent mixing distributions on the mean of the negative binomial model.

The regional abundances of the species are discrete positive integers. However, for the sake of computations, it is easier to use a continuous probability distribution to approximate the distribution of regional abundances (Kempton & Taylor, 1974). Kempton and Taylor argued that a continuous probability distribution is an adequate approximation because "the range of abundances is normally large, and the probability of any species having a given abundance correspondingly small."

Candidate probability distributions for the regional abundances have been proposed based on empirical field studies, ecological theories and statistical sampling theory.

The lognormal distribution fits a large number of data sets (Diserud & Engen, 2000); however, it has been shown to fit poorly in some applications (Diserud & Engen, 2000). Some ecological theories, such as the sequential broken-stick niche-partitioning model (Bulmer, 1974) using the central limit theorem (May, 1975), support the lognormal distribution. The lognormal distribution also arises from some stochastic abundance models of population dynamics (e.g., Engen and Lande, 1996a).

The gamma distribution was first used by Fisher (Fisher *et al.*, 1943) and fits reasonably well with several data sets (Kempton & Wedderburn, 1978). It, too, can result in a poor fit for some data sets (Diserud & Engen, 2000). Engen and Lande (1996b) developed a stochastic abundance model for the regional abundances which results in a gamma distribution.

The inverse Gaussian distribution (Ord & Whitmore, 1986), generalized inverse Gaussian distribution (Sichel, 1997) and other positively skewed probability distributions have also been considered such as the Pareto distribution and finite mixtures of exponential distributions (Bunge & Barger, 2008). Hubbell (2001) and others (see McGill, Maurer, and Weiser, 2006) have proposed zero-sum multinomial distributions derived under *neutral theories*[5] of biodiversity.

There is no consensus in empirical studies and ecological theories on the choice of probability distribution to model the regional abundances of species (Williamson & Gaston, 2005; McGill *et al.*, 2006; Bunge & Barger, 2008). Different probability distributions for the regional abundances may adequately fit a given sample of observations, but they can lead to significantly different estimates of species richness (Norris & Pollock, 1998; Chao, 2005).

We used the lognormal distribution in earlier work with our modified version of PKB's estimation method. However, we encountered numerical difficulties when the sample contained species with sample abundances exceeding 160, although Bulmer's (1974) approximation was helpful when using a Poisson sampling model. Bunge and Barger (2008, 2009) have noted that the lognormal, generalized inverse Gaussian, and Pareto distributions present computational difficulties that require numerical approximations when used with a Poisson sampling model.

We attempted to use the gamma distribution to model the distribution of re-

[5]Neutral theories propose that all individuals of all species in a taxonomic group have identical per capita rates of birth, death, dispersal, and speciation.

gional abundances. However, testing on some data sets revealed numerical difficulties when using the gamma distribution. We explain the numerical difficulties in Section 3.4.1 after the abundance-based sampling model is introduced in Section 3.4. As a simplification, we implemented an exponential distribution (i.e., fixing the gamma distribution's shape parameter equal to one) as a model of the distribution of regional abundances. No numerical problems were encountered when using the exponential distribution in initial tests, and we use the exponential distribution as a model of the distribution of regional abundances.

3.3 Assumptions

The assumptions involve the data collection, the spatial patterns of the species, and the interaction between species.

A1 *Uniform sampling effort*

The sampling units are assumed to have identical shape and dimensions. The total number of individuals expected to occur in the sampling units will be approximately equal.

A2 *Sampling units contain independent sets of observations*

The sampling units are assumed to have been randomly located in a region that is very large relative to the size of the sampling units. It is also assumed that the boundaries of the sampling units do not intersect each other. With sufficient distance between sampling units, spatial autocorrelation between sampling units

will be negligible. Therefore, the sampling units will be treated as independent.

A3 *Closed Population*

Over a short period of time, each sampling unit is thoroughly examined for all individuals that are in the desired taxonomic group. The total time to conduct the study is assumed to be sufficiently short such that, during the study, no new species enter the region nor does any existing species disappear from the region.

A4 *100% Probability of Detection and Correct Identification*

In the following developments, it will be assumed that all individuals that occur in a sampling unit are detected and correctly identified to their species types. In ecological studies, the probability of detecting an individual may depend on a number of factors such as the physical characteristics of the species, the season or time of year, and the skill of the observer (e.g., North American Breeding Bird Survey, USGS Patuxent Wildlife Research Center, 2010). Without the assumption of 100% probability of detection, multiple visits to each sampling unit would be required in order to model both the occurrence and the detection of species (Dorazio & Royle, 2005). The 100% probability of detection of all individuals in a sampling unit may be close to being met for organisms that are immobile and easily visible to observers, such as the woody plant specimens with a base girth of 10 cm in the 0.5-hectare site in Mali (Picard *et al.*, 2004). The assumption may also be reasonable when the entire contents of the sampling

units are extracted (e.g., samples of soil or marine substrate) and taken to a laboratory for careful sorting and inventory. If a detected organism belongs to a species that is new to science, then it is assumed the organism is correctly distinguished from the existing list of known species and any future individuals of the new species are correctly identified.

A5 *Species are mutually independent*

We assume that the spatial distribution of one species is independent of all other species. This assumption is not realistic as two organisms cannot occupy the same physical space. Furthermore, in nature, some species tend to occur together and other species do not – these associations may be the result of direct interactions between species, or an indirect result as the species respond to environmental conditions (Dormann *et al.*, 2007). Because there are usually tens or hundreds of species in the taxonomic group, it would be challenging to model these inter-species interactions. To keep the work analytically tractable, estimators of species richness usually assume the species are mutually independent (Dorea & Mingoti, 2006), and this is a simplifying assumption.

A6 *Spatial distributions of species are stationary*

For each species, the spatial distribution of its individuals is assumed to follow a consistent pattern across the region, and each species has the entire region as its potential geographic range. These assumptions are common in work with species-area curves (Picard *et al.*, 2004). They allow the comparison of the

variation and mean of the con-specific sample abundances among the sampling units to indicate the tendency of the individuals to cluster.

3.4 Statistical Sampling Model for Abundance-based Data from Multiple Sampling Units

In a large region with total area Ω containing S_Ω species, Q spatial-based sampling units of identical shape and dimensions have been randomly selected. Over a short period of time, each sampling unit is thoroughly examined for all individuals that are in the desired taxonomic group. The number of individuals of the i^{th} species occurring in the q^{th} sampling unit will be represented by $n_{i,q}$, for $i = 1, 2, \ldots, S_\Omega$, and $q = 1, 2, \ldots, Q$.

Based on the discussion in Section 3.2.1, the number of individuals of a species in a sampling unit will be modelled as negative binomial random variable. The sampling units are assumed to be sufficiently far apart that spatial autocorrelation is negligible and the observations from the Q sampling units are treated as mutually independent. Assuming the spatial distribution of the i^{th} species is consistent across the region, its sample abundances $\boldsymbol{n_i} = \langle n_{i,1}, n_{i,2}, \ldots, n_{i,Q} \rangle \overset{iid}{\sim}$ Negative Binomial(p_i, k_i) where $0 < p_i < 1$ and the dispersion parameter $k_i > 0$. We then have

$$P_{NB}\left(n_{i,q}\,;\, k_i, p_i\right) = \frac{\Gamma(k_i + n_{i,q})}{\Gamma(k_i) \cdot n_{i,q}!} p_i^{n_{i,q}} \left(1 - p_i\right)^{k_i},$$

for $i = 1, 2, \ldots, S_\Omega$ and $q = 1, 2, \ldots, Q$, where $\Gamma(\cdot)$ denotes the gamma function.

Let v denote the area of a sampling unit divided by the area, Ω, of the region. Given the regional abundance, N_i, of the i^{th} species, the expected number of its individuals in a sampling unit is $\mu_i = v \cdot N_i$. The negative binomial model for the i^{th} species can be reparameterized in terms of its mean μ_i and the dispersion parameter k_i with $\mu_i = \frac{p_i k_i}{1-p_i}$. The parameter k_i characterizes the overdispersion and can be interpreted as the strength of the spatial aggregation of a species at the scale of the size of the sampling units. Values of k_i close to zero correspond to strong spatial aggregation. As $k_i \to \infty$, the negative binomial reduces to the Poisson distribution, and represents a species having a completely random spatial pattern (He & Legendre, 2002). Given μ_i and k_i, we have

$$P_{NB}\left(n_{i,q}\,;\,k_i,\mu_i\right) = \frac{\Gamma(k_i+n_{i,q})}{\Gamma(k_i)\cdot n_{i,q}!}\left(\frac{\mu_i}{\mu_i+k_i}\right)^{n_{i,q}}\left(\frac{k_i}{\mu_i+k_i}\right)^{k_i}, \tag{3.1}$$

for $n_{i,q} = 0, 1, 2, \ldots$.

Conditioning on μ_i and k_i, we have

$$P\left(\boldsymbol{n_i} \mid k_i,\mu_i\right) = \prod_{q=1}^{Q} P_{NB}\left(n_{i,q}\,;\,k_i,\mu_i\right)\ . \tag{3.2}$$

Therefore, the conditional probability of the i^{th} species occurring zero times in each of Q sampling units (i.e., $\boldsymbol{n_i} = \mathbf{0}$) is

$$\begin{aligned} P\left(\mathbf{0} \mid k_i,\mu_i\right) &= \prod_{q=1}^{Q} P_{NB}\left(0 \mid k_i,\mu_i\right) \\ &= \prod_{q=1}^{Q}\left\{\frac{\Gamma(k_i+0)}{\Gamma(k_i)\cdot 0!}\left(\frac{\mu_i}{\mu_i+k_i}\right)^{0}\left(\frac{k_i}{\mu_i+k_i}\right)^{k_i}\right\} \\ &= \left(\frac{k_i}{\mu_i+k_i}\right)^{Qk_i}, \end{aligned} \tag{3.3}$$

for $i = 1, 2, \ldots, S_\Omega$.

Given the data, k_i and μ_i could be estimated by maximizing $P(\boldsymbol{n_i} \mid k_i, \mu_i)$ in Equation 3.2. A problem with this approach arises from the fact that there will be species in the region that did not occur in any of the sampling units. In this case, a mean value $\mu_i = 0$ maximizes the probability of a species occurring zero times in each sampling unit. That is, to maximize the joint probability in Equation 3.3, a species has a regional abundance N_i equal to zero, which means the species does not exist in the region. This would incorrectly suggest that there are no species in the region other than the ones observed in the sampling units, so that S_{obs} is the estimate of S_Ω.

Rather than fit a separate model for each species, the parameters (k_1, μ_1), (k_2, μ_2), $\ldots$, $(k_{S_\Omega}, \mu_{S_\Omega})$ are treated as a random sample from a continuous bivariate probability distribution G, having density function g and indexed by $\boldsymbol{\theta}$. Using a mixing distribution to model the variation in the (k_i, μ_i) pairs has a clear advantage over fitting individual negative binomial distributions to each species. The principal advantage is that we now allow inference on the species absent from the sampling units based on the species that are observed in the sampling units. The overall variation in (k_i, μ_i) is characterized by $\boldsymbol{\theta}$.

The mixing distribution G can be interpreted as a model of the relationship between the regional abundance $(N_i = \mu_i / v)$ of a species and its tendency to aggregate (expressed through k_i) at spatial scales comparable to the size of the sampling units (Picard *et al.*, 2004). For example, Condit *et al.* (2000) describe data sets where the species with low abundances exhibit a greater tendency for con-specific aggregation

than the species with large abundances; the mixing distribution G can be chosen to reflect the negative association between the abundances and the tendencies for spatial aggregation. In Section 3.4.1, choices for G are discussed.

For simplicity of presentation, the subscript i is dropped when referring to the negative binomial parameters μ_i and k_i unless necessary. The notation μ and k will denote the negative binomial parameters of a species, where the pair (k, μ) is generated from the mixing distribution G.

Having imposed a mixing distribution G on the parameters (k, μ), the marginal distribution of the proportion of species contributing $\boldsymbol{x}$, a vector of Q con-specific sample abundances, is

$$P_G\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}\right) = \int_{k=0}^{\infty}\int_{\mu=0}^{\infty} P\left(\boldsymbol{x} \mid k, \mu\right) \cdot g(k, \mu; \boldsymbol{\theta})\, d\mu\, dk\,, \tag{3.4}$$

where $P\left(\boldsymbol{x} \mid k, \mu\right)$ is given in Equation 3.2. We note that this marginal distribution averages over the different (k, μ) in the population: $P_G\left(\boldsymbol{x}; \boldsymbol{\theta}\right) = E_g\left[P(\boldsymbol{x} \mid k, \mu)\right]$, so that, marginally, the expected number of species contributing $\boldsymbol{x}$ to the sampling units is $S_\Omega \cdot P_G\left(\boldsymbol{x}; \boldsymbol{\theta}\right)$.

The joint probability mass function expressed in Equation 3.4 is used in the construction of the likelihood function for S_Ω and $\boldsymbol{\theta}$.

3.4.1 Mixing Distributions on μ and k

We had concerns about the numerical integration required to compute the multivariate p.m.f. in Equation 3.4, particularly when trying to maximize the logarithm of the

likelihood function. As a result, we start with a tractable simple distribution, G, for the negative binomial parameters μ and k associated with the species. We apply a univariate probability distribution on μ, and fix k at a common value for all species.

The gamma distribution was initially chosen to model the regional abundances. A gamma distribution on the regional abundances is equivalent also to modeling $\mu_1, \mu_2, \ldots, \mu_{S_\Omega}$ as random variables from some other gamma distribution with shape α and rate β. With the negative binomial parameter k fixed at a common value for all species, the multivariate p.m.f. in Equation 3.4 becomes

$$P_G\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}\right) = \int_{\mu=0}^{\infty} P\left(\boldsymbol{x} \mid k, \mu\right) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu}\, d\mu\,, \tag{3.5}$$

where $\boldsymbol{\theta} = \langle \alpha, \beta, k \rangle$, $\alpha > 0$ and $\beta > 0$.

This negative binomial sampling model with a gamma mixing distribution on μ was tested on a small number of realistic data sets. In some cases, in the course of maximizing the log-likelihood function, the estimates of α were small enough (i.e., < 0.01) to cause numerical problems when evaluating the integral in Equation 3.5. The log-likelihood is maximized also numerically, but this does require the accurate and reliable computation of many probabilities of the form shown in Equation 3.5. For small values of α, the gamma distribution places significant mass in the vicinity of zero, which causes numerical problems. We attempted to approximate the integral in Equation 3.5, particularly for intervals of μ with a lower endpoint at zero. Despite some success approximating the integral when $\boldsymbol{x} = \mathbf{0}$, numerical problems continued to be encountered when $\boldsymbol{x} \neq \mathbf{0}$. As a result, the log-likelihood function could not be

reliably computed.

Instead of using the gamma distribution, we model the means μ_1, μ_2, ..., μ_{S_Ω} as i.i.d. draws from an exponential distribution, $\mu_i \overset{iid}{\sim} \text{Exponential}(\beta)$ with rate $\beta > 0$. Keeping the negative binomial parameter k fixed at a common value for all species, the multivariate p.m.f. is

$$P_G\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}\right) = \int_0^\infty P\left(\boldsymbol{x} \mid k, \mu\right) \cdot \beta e^{-\beta\mu}\, d\mu\,, \tag{3.6}$$

where $\boldsymbol{\theta} = \langle \beta, k \rangle$.

The value of the negative binomial parameter k will vary between species and also depend on the size of the sampling units (Plotkin & Muller-Landau, 2002). For now, the common value of k assumed for all species can be interpreted as a measure of the overall tendency of the species assemblage to exhibit con-specific spatial aggregation at the scale of the sampling units.

3.5 Likelihood Function and MLEs

The steps used in the construction of the likelihood function are applicable for any mixing distributions on (k, μ). Let S_{obs} represent the total number of species that are observed at least once in the sampling units. Without loss of generality, the S_Ω species in the region will be indexed with the positive integers $1, 2, \ldots, S_\Omega$ in a manner such that the species observed in the sampling units are labelled with the integers $i = 1, 2, \ldots, S_{obs}$ and the species absent from the sampling units are labelled with the integers $i = S_{obs} + 1, S_{obs} + 2, \ldots, S_\Omega$, with S_Ω itself being unknown.

Assuming species occur in the sampling units independently of one another, the marginal probability of S_{obs} is

$$P(S_{obs} \mid S_\Omega, \boldsymbol{\theta}) = \binom{S_\Omega}{S_{obs}} \left[1 - P_G\left(\mathbf{0}\,;\,\boldsymbol{\theta}\right)\right]^{S_{obs}} P_G(\mathbf{0}\,;\,\boldsymbol{\theta})^{S_\Omega - S_{obs}}, \tag{3.7}$$

where $P_G(\mathbf{0}\,;\,\boldsymbol{\theta})$ is obtained from Equation 3.4 as $E_g\left[P(\mathbf{0} \mid k, \mu)\right]$, the marginal probability of absence from the sampling units. The general form of $P_G(\mathbf{0}\,;\,\boldsymbol{\theta})$ when (k, μ) has mixing distribution G with parameter vector $\boldsymbol{\theta}$ and density function g is found using Equation 3.4:

$$\begin{aligned} P_G\left(\mathbf{0}\,;\,\boldsymbol{\theta}\right) &= \int_{k=0}^{\infty}\int_{\mu=0}^{\infty} P\left(\mathbf{0} \mid k, \mu\right) \cdot g(k, \mu; \boldsymbol{\theta})\, d\mu\, dk \\ &= \int_{k=0}^{\infty}\int_{\mu=0}^{\infty} \left(\frac{k}{\mu + k}\right)^{Qk} \cdot g(k, \mu; \boldsymbol{\theta})\, d\mu\, dk\,, \end{aligned}$$

where $P\left(\mathbf{0} \mid k, \mu\right)$ has been expanded on the second line of the equation using Equation 3.3. For an exponential distribution on μ, and k fixed at a common value for all species, $P_G(\mathbf{0}\,;\,\boldsymbol{\theta})$ is computed using the multivariate p.m.f. in Equation 3.6:

$$\begin{aligned} P_G\left(\mathbf{0}\,;\,\boldsymbol{\theta}\right) &= \int_{\mu=0}^{\infty} P\left(\mathbf{0} \mid k, \mu\right) \cdot \beta e^{-\beta\mu}\, d\mu \\ &= \int_{\mu=0}^{\infty} \left(\frac{k}{\mu + k}\right)^{Qk} \cdot \beta e^{-\beta\mu}\, d\mu\,. \end{aligned} \tag{3.8}$$

Given that a species is observed, the marginal probability that it has vector of sample abundances $\boldsymbol{x}$ is

$$P\left(\boldsymbol{x} \mid \boldsymbol{x} \neq \mathbf{0}\right) = \frac{P_G\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}\right)}{1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})} \tag{3.9}$$

where $P_G\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}\right)$ is computed with Equation 3.4 for general mixing distributions on (k, μ), or with Equation 3.6 for the exponential mixing distribution on μ and a common value of k. Using the assumption that the species are mutually independent,

the joint probability of $\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}}$, conditioning on the fact that S_{obs} species occurred in the sampling units, is

$$P(\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}} \mid S_{obs}, \boldsymbol{\theta}) = c \cdot \prod_{i=1}^{S_{obs}} \frac{P_G\left(\boldsymbol{n_i}\,;\,\boldsymbol{\theta}\right)}{1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})}\,, \tag{3.10}$$

where c is a multinomial coefficient involving S_{obs} and the number of distinct vectors among $\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}}$.

Multiplying the probabilities in Equations 3.7 and 3.10 together gives the marginal probability of S_{obs} species occurring in the sampling units *and* the observed species having sample abundances $\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}}$,

$$\begin{aligned} &P(\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}}, S_{obs} \mid S_\Omega, \boldsymbol{\theta}) \\ &\quad = \binom{S_\Omega}{S_{obs}} \left[1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})\right]^{S_{obs}} P_G(\mathbf{0}\,;\,\boldsymbol{\theta})^{S_\Omega - S_{obs}} \cdot c \prod_{i=1}^{S_{obs}} \frac{P_G\left(\boldsymbol{n_i}\,;\,\boldsymbol{\theta}\right)}{1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})} \\ &\quad = \binom{S_\Omega}{S_{obs}} P_G(\mathbf{0}\,;\,\boldsymbol{\theta})^{S_\Omega - S_{obs}} \cdot c \prod_{i=1}^{S_{obs}} P_G\left(\boldsymbol{n_i}\,;\,\boldsymbol{\theta}\right)\,. \end{aligned} \tag{3.11}$$

Using Equation 3.11, the likelihood function for S_Ω and $\boldsymbol{\theta}$ is

$$L(S_\Omega, \boldsymbol{\theta}) \propto \binom{S_\Omega}{S_{obs}} P_G(\mathbf{0}\,;\,\boldsymbol{\theta})^{S_\Omega - S_{obs}} \cdot \prod_{i=1}^{S_{obs}} P_G\left(\boldsymbol{n_i}\,;\,\boldsymbol{\theta}\right)\,. \tag{3.12}$$

Unfortunately, the MLE of $\boldsymbol{\theta}$ cannot usually be solved analytically. If the MLE of $\boldsymbol{\theta}$ is known, then the MLE of S_Ω can be found analytically. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, the MLE of S_Ω is then the smallest positive integer for S_Ω such that

$$\frac{L(S_\Omega + 1, \hat{\boldsymbol{\theta}})}{L(S_\Omega, \hat{\boldsymbol{\theta}})} < 1\,. \tag{3.13}$$

The MLE[6] of S_Ω is thus

$$\hat{S}_\Omega = \left\lfloor \frac{S_{obs}}{1 - P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})} \right\rfloor , \tag{3.14}$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. The form of the MLE $\hat{S}_\Omega$ in Equation 3.14 is a standard result for maximum likelihood estimators of species richness (Sanathanan, 1977).

When the MLE of $\boldsymbol{\theta}$ is not known ahead of time, it may be possible to numerically maximize the logarithm of the likelihood function simultaneously for $\hat{S}_\Omega$ and $\hat{\boldsymbol{\theta}}$. Because S_Ω is an integer-valued parameter and the multivariate p.m.f. involves the evaluation of an integral, it is computationally more feasible to find estimates of S_Ω and $\boldsymbol{\theta}$ numerically in two stages using *conditional likelihood* and *profile likelihood*.

3.5.1 Conditional MLEs

In the conditional likelihood method (Sanathanan, 1977), the likelihood function in Equation 3.12 is expressed as the product of two functions,

$$L(S_\Omega, \boldsymbol{\theta}) = L_{bin}(S_\Omega, \boldsymbol{\theta}) \cdot L_{cond}(\boldsymbol{\theta}) \, .$$

The function $L_{bin}(S_\Omega, \boldsymbol{\theta})$ is the likelihood of S_{obs} species occurring in the sampling units:

$$L_{bin}(S_\Omega, \boldsymbol{\theta}) = \binom{S_\Omega}{S_{obs}} \left[1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})\right]^{S_{obs}} P_G(\mathbf{0}\,;\,\boldsymbol{\theta})^{S_\Omega - S_{obs}},$$

[6] If $\frac{S_{obs}}{1-P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})}$ is an integer, then the likelihood function has the same maximum value at $\hat{S}_\Omega = \frac{S_{obs}}{1-P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})} - 1$ and at $\hat{S}_\Omega = \frac{S_{obs}}{1-P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})}$.

which has the form of $P(S_{obs} \mid S_\Omega, \boldsymbol{\theta})$ stated in Equation 3.7. The likelihood $L_{bin}(S_\Omega, \boldsymbol{\theta})$ can be viewed as treating S_{obs} as a binomial random variable with index S_Ω and the marginal (population-averaged) probability of occurrence in the sampling units is $1-P_G(\mathbf{0};\boldsymbol{\theta})$. The function $L_{cond}(\boldsymbol{\theta})$ is the likelihood of the non-zero sample abundance vectors $\{\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}}\}$, conditioning on the number, S_{obs}, of species observed:

$$L_{cond}(\boldsymbol{\theta}) = \prod_{i=1}^{S_{obs}} \frac{P_G\left(\boldsymbol{n_i}\,;\,\boldsymbol{\theta}\right)}{1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})}\,.$$

The conditional likelihood $L_{cond}(\boldsymbol{\theta})$ has the form of the probability $P(\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}} \mid S_{obs}, \boldsymbol{\theta})$ stated in Equation 3.10.

The first step in the conditional likelihood method is to estimate $\boldsymbol{\theta}$. The conditional likelihood function $L_{cond}(\boldsymbol{\theta})$ does not involve the species richness S_Ω. The *conditional MLE* of $\boldsymbol{\theta}$ is computed numerically by choosing the value of $\boldsymbol{\theta}$ that maximizes $ln\, L_{cond}(\boldsymbol{\theta})$. We let $\hat{\boldsymbol{\theta}}_{CMLE}$ denote the conditional MLE of $\boldsymbol{\theta}$. These numerical computations were performed in the R software environment (R Development Core Team, 2011). An initial value of $\boldsymbol{\theta}$ is required in R's optimization routine *optim* in order to numerically maximize $ln\, L_{cond}(\boldsymbol{\theta})$.

For the exponential mixing distribution on μ and a common value of k for all species, $\boldsymbol{\theta} = \langle \beta, k \rangle$. Initial values were computed for k and the exponential distribution's rate parameter β using an approach based on the method of moments. To compute the initial values of β and k, attention was restricted to the species that occurred in the sampling units. Let $\bar{n}_i = \sum_{q=1}^{Q} n_{i,q}/Q$ and $s_i^2 = \sum_{q=1}^{Q} (n_{i,q} - \bar{n}_i)^2/(Q-1)$ represent the sample mean and sample variance, respectively, of the sample abundances

for species i, for $i = 1, 2, \ldots, S_{obs}$. For each species observed, if its sample mean $\bar{n}_i$ is less than its sample variance s_i^2, then a method of moments routine *theta.mm* in R computes an estimate, $\tilde{k}_i$, of the negative binomial dispersion parameter k_i associated with species i. If $\bar{n}_i > s_i^2$, then the routine *theta.mm* will not work and an estimate, $\tilde{k}_i$, of k_i proportional to $\bar{n}_i$ is used. The initial value of k is $k^{(0)} = \frac{1}{S_{obs}} \sum_{i=1}^{S_{obs}} \tilde{k}_i$.

For the species observed in the sampling units, the expected values (μ_1, μ_2, ..., $\mu_{S_{obs}}$) of their sample abundances are estimated by their respective sample means $\bar{n}_1$, $\bar{n}_2$, ..., $\bar{n}_{S_{obs}}$. The average of $\bar{n}_1, \bar{n}_2, \ldots, \bar{n}_{S_{obs}}$ is then used as an estimate of $E(\mu_i)$. The exponential distribution with rate parameter β has a mean of $1/\beta$. Substituting the estimate of $E(\mu_i)$ into the equation $E(\mu_i) = 1/\beta$, we arrive at an initial value of β equal to $\frac{S_{obs}}{\sum_{S_{obs}}^{i=1} \bar{n}_i}$.

In the production of initial values for β and k, only the non-zero sample abundance vectors $\{\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_{S_{obs}}}\}$ were used. These initial estimates are undoubtedly biased as the species absent from the sampling units have not been included in the computations. Nevertheless, these rough estimates have proved adequate for use as the initial values in the numerical optimization routines that maximize $ln\, L_{cond}(\boldsymbol{\theta})$ to produce $\hat{\boldsymbol{\theta}}_{CMLE}$.

After computing $\hat{\boldsymbol{\theta}}_{CMLE}$, the second step in the conditional likelihood method (Sanathanan, 1977) is to estimate S_Ω. $\hat{\boldsymbol{\theta}}_{CMLE}$ is substituted into the binomial likelihood $L_{bin}(S_\Omega, \boldsymbol{\theta})$, and $L_{bin}(S_\Omega, \hat{\boldsymbol{\theta}}_{CMLE})$ is maximized for S_Ω by choosing the smallest positive integer for S_Ω such that the likelihood ratio $\frac{L_{bin}(S_\Omega+1, \hat{\boldsymbol{\theta}}_{CMLE})}{L_{bin}(S_\Omega, \hat{\boldsymbol{\theta}}_{CMLE})}$ is less than 1.

The conditional MLE of S_Ω is thus

$$\hat{S}_{\Omega_{CMLE}} = \left\lfloor \frac{S_{obs}}{1 - P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}}_{CMLE})} \right\rfloor \,. \tag{3.15}$$

The form of $\hat{S}_{\Omega_{CMLE}}$ in Equation 3.15 is identical to the form of the unconditional MLE $\hat{S}_\Omega$ in Equation 3.14 except a different estimate of $\boldsymbol{\theta}$ is used. The term $P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}}_{CMLE})$ in Equation 3.15 can be interpreted as the estimated proportion of species in the region that are absent from the sampling units, based on the conditional MLE of $\boldsymbol{\theta}$.

Maximizing the conditional likelihood for $\boldsymbol{\theta}$ and then maximizing the binomial likelihood for S_Ω is less challenging numerically than maximizing the original unconditional likelihood $L(S_\Omega, \boldsymbol{\theta})$ over all parameters.

The estimated marginal probability of absence using the conditional maximum likelihood approach will be greater than when using the unconditional maximum likelihood approach (see, e.g., Sanathanan, 1977). In the current context with multiple sampling units, this means $P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}}_{CMLE}) > P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})$. Therefore, the conditional MLE $\hat{S}_{\Omega_{CMLE}}$ will be greater than or equal to the unconditional MLE $\hat{S}_\Omega$.

3.5.2 Profile Likelihood and Unconditional MLEs

Given the conditional MLEs of S_Ω and $\boldsymbol{\theta}$, we use a method involving profile likelihoods to find the *unconditional* MLEs $\hat{S}_\Omega$ and $\hat{\boldsymbol{\theta}}$. For any positive integer $S^* \geq S_{obs}$, the profile log-likelihood of S_Ω evaluated at $S_\Omega = S^*$ is

$$l_P(S^*) = sup_{\boldsymbol{\theta}}\, ln\, L(S^*, \boldsymbol{\theta}) \,, \tag{3.16}$$

where $L(S^*, \boldsymbol{\theta})$ is the unconditional likelihood function evaluated at $(S^*, \boldsymbol{\theta})$. The conditional MLE $\hat{\boldsymbol{\theta}}_{CMLE}$ is used as the initial value of $\boldsymbol{\theta}$ in R's optimization routine *optim* for the computation of the profile log-likelihoods. By definition, the profile log-likelihood function has its maximum at the unconditional MLE of species richness. In fact, the profile log-likelihood at $\hat{S}_\Omega$ is equal to the maximum value of the logarithm of the original likelihood function $L(S_\Omega, \boldsymbol{\theta})$. That is, $l_P(\hat{S}_\Omega) = ln\, L(\hat{S}_\Omega, \hat{\boldsymbol{\theta}})$.

From Sanathanan (1977), the MLE of species richness, $\hat{S}_\Omega$, is bounded between the number of species observed in the sampling units and the conditional MLE of species richness, $\hat{S}_{\Omega_{CMLE}}$. Therefore, the profile log-likelihood can be computed for every integer S^* between S_{obs} and $\hat{S}_{\Omega_{CMLE}}$, inclusive, in an exhaustive search for the MLE $\hat{S}_\Omega$ and the corresponding MLE $\hat{\boldsymbol{\theta}}$. In our work, the profile log-likelihood of S_Ω has been observed to be monotonically increasing from S_{obs} to its global maximum at $\hat{S}_\Omega$, and the profile log-likelihood is monotonically decreasing from $\hat{S}_\Omega$ to $\hat{S}_{\Omega_{CMLE}}$. Taking advantage of the monotonicity, a more efficient binary search for $\hat{S}_\Omega$ is conducted on the profile log-likelihood over the set $\left\{S_{obs}, S_{obs}+1, \ldots, \hat{S}_{\Omega_{CMLE}}\right\}$. Maximizing the profile log-likelihood, we acquire the unconditional MLEs $\hat{S}_\Omega$ and $\hat{\boldsymbol{\theta}}$.

In the initial testing of the exponential mixing distribution on μ on realistic data sets, no problems were encountered numerically computing the integral in Equation 3.6. The maximum likelihood estimates of S_Ω and $\boldsymbol{\theta} = \langle \beta, k \rangle$ were also successfully computed numerically for each data set tested.

.

3.6 The Sampling Distribution of $\hat{S}_\Omega$

To conduct inference on S_Ω, we require the sampling distribution of the maximum likelihood estimator, $\hat{S}_\Omega$. In general, finite-sample exact distributional results in the area of species richness estimation are infeasible. Bunge and Barger (2009) stated that "no exact small-sample variance results are known for any version" of the species richness problem. Therefore, we focus on estimating characteristics of the sampling distribution of $\hat{S}_\Omega$ using asymptotic methods.

3.6.1 Asymptotics

Sanathanan (1977) developed an asymptotic variance formula for the MLE of species richness in a very general setting that can be applied to many sampling scenarios. In the context of our negative binomial-exponential mixture model, Sanathanan's formula for the asymptotic variance of $\hat{S}_\Omega$ is based on

$$\text{Asymptotic Var}\left(\frac{\hat{S}_\Omega - S_\Omega}{\sqrt{S_\Omega}}\right) = \left(a_{00} - \boldsymbol{a}_0^T \boldsymbol{A}^{-1} \boldsymbol{a}_0\right)^{-1} , \tag{3.17}$$

where $a_{00} = \frac{1-P_G(\mathbf{0}\,;\,\boldsymbol{\theta})}{P_G(\mathbf{0}\,;\,\boldsymbol{\theta})}$, $\boldsymbol{a}_0 = \frac{1}{P_G(\mathbf{0}\,;\,\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \{1 - P_G(\mathbf{0};\boldsymbol{\theta})\}$, and $\boldsymbol{A}$ is the Fisher information matrix about $\boldsymbol{\theta}$. The gradient vector $\nabla_{\boldsymbol{\theta}} \{1 - P_G(\mathbf{0};\boldsymbol{\theta})\}$ is

$$\nabla_{\boldsymbol{\theta}} \{1 - P_G(\mathbf{0};\boldsymbol{\theta})\} = \begin{bmatrix} \frac{\partial}{\partial \beta} \{1 - P_G(\mathbf{0};\boldsymbol{\theta})\} \\ \\ \frac{\partial}{\partial k} \{1 - P_G(\mathbf{0};\boldsymbol{\theta})\} \end{bmatrix} = \begin{bmatrix} -\frac{\partial}{\partial \beta} P_G(\mathbf{0};\boldsymbol{\theta}) \\ \\ -\frac{\partial}{\partial k} P_G(\mathbf{0};\boldsymbol{\theta}) \end{bmatrix} ,$$

and the Fisher information matrix is

$$\boldsymbol{A} = \begin{bmatrix} E\left\{\left(\frac{\partial \, ln \, L(S_\Omega,\boldsymbol{\theta})}{\partial \beta}\right)^2\right\} & E\left\{\left(\frac{\partial \, ln \, L(S_\Omega,\boldsymbol{\theta})}{\partial \beta}\right)\left(\frac{\partial \, ln \, L(S_\Omega,\boldsymbol{\theta})}{\partial k}\right)\right\} \\ \\ E\left\{\left(\frac{\partial \, ln \, L(S_\Omega,\boldsymbol{\theta})}{\partial \beta}\right)\left(\frac{\partial \, ln \, L(S_\Omega,\boldsymbol{\theta})}{\partial k}\right)\right\} & E\left\{\left(\frac{\partial \, ln \, L(S_\Omega,\boldsymbol{\theta})}{\partial k}\right)^2\right\} \end{bmatrix} ,$$

where the expectations are over values of $\boldsymbol{x}$ with respect to the probability function $P_G(\boldsymbol{x};\boldsymbol{\theta})$, given S_Ω and $\boldsymbol{\theta}$.

To estimate the asymptotic variance of $\hat{S}_\Omega$, we substitute the MLEs of S_Ω and $\boldsymbol{\theta}$ into the right side of Equation 3.17. Fixing $S_\Omega = \hat{S}_\Omega$, the log-likelihood function $ln\ L(\hat{S}_\Omega, \boldsymbol{\theta})$ is maximized over $\boldsymbol{\theta}$ using the L-BFGS-B method (Byrd *et al.*, 1995), a modification of the BFGS quasi-Newton method with bounds on $\boldsymbol{\theta}$. In addition to $\hat{\boldsymbol{\theta}}$, the optimization yields the observed information matrix $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ for $\boldsymbol{\theta}$ evaluated at the MLEs $(\hat{S}_\Omega, \hat{\boldsymbol{\theta}})$. The estimated asymptotic variance of $\hat{S}_\Omega$ is then

$$\widehat{Var}(\hat{S}_\Omega) = \hat{S}_\Omega \cdot \left(\hat{a}_{00} - \hat{\boldsymbol{a}}_0^T \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{a}}_0\right)^{-1}, \tag{3.18}$$

where $\hat{a}_{00} = \frac{1-P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})}{P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})}$ and $\hat{\boldsymbol{a}}_0 = \frac{1}{P_G(\mathbf{0}\,;\,\hat{\boldsymbol{\theta}})} \nabla_{\boldsymbol{\theta}|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \{1 - P_G(\mathbf{0}\,;\,\boldsymbol{\theta})\}$ with $\nabla_{\boldsymbol{\theta}} \{1 - P_G(\mathbf{0};\boldsymbol{\theta})\}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Following the asymptotic theory from Sanathanan's general setting, the MLEs $(\hat{S}_\Omega, \hat{\boldsymbol{\theta}})$ are asymptotically normally distributed and unbiased, which, in conjunction with $\widehat{Var}(\hat{S}_\Omega)$ from Equation 3.18, allows construction of approximate confidence intervals for S_Ω. Measures of uncertainty based on asymptotic results are only appropriate for large samples. When applied to small samples, the lower limit of an approximate confidence interval for species richness based on asymptotic normality can, in fact, lie below the number of species observed (Bunge & Barger, 2009).

3.6.2 Likelihood Intervals

An alternative approach from constructing approximate confidence intervals for species richness is through profile likelihood intervals. The profile log-likelihood function l_P was introduced in Equation 3.16. This function evaluated at a positive integer S^* is the maximum value of the log-likelihood function, $ln\, L(S_\Omega, \boldsymbol{\theta})$, with $S_\Omega = S^*$ fixed.

A $100p\%$ *profile likelihood interval* for S_Ω comprises all values of S_Ω that satisfy the inequality $l_P(S_\Omega) - l_P(\hat{S}_\Omega) \geq ln\, p$. In other words, the $100p\%$ profile likelihood interval contains values of S_Ω that achieve likelihoods (for some values of $\boldsymbol{\theta}$) of at least $100p\%$ of the maximum value $L(\hat{S}_\Omega, \hat{\boldsymbol{\theta}})$ of the likelihood function.

The $100p\%$ profile likelihood interval has an approximate coverage probability of $P(\chi^2_{(1)} \leq -2\, ln\, p)$ (Kalbfleisch, 1985), meaning that a 14.7% profile likelihood interval for S_Ω will have an approximate coverage probability of 95%. This approximation is derived from the asymptotic distribution of the generalized likelihood ratio test.

Since S_Ω is an integer-valued parameter, the lower endpoint of the profile likelihood interval is found numerically by computing the profile log-likelihoods for integer values of S_Ω between S_{obs} and $\hat{S}_\Omega$, inclusive. When computing the profile log-likelihood, the numerical optimization can be time-consuming, in particular when the number of sampling units is large or when the number, S_{obs}, of species observed is large. However, the profile log-likelihood will typically have a unimodal shape when plotted versus S_Ω, with a maximum at $\hat{S}_\Omega$. To reduce the computational time, a binary search algorithm is employed to find the lower endpoint of the profile likelihood interval. The binary

search requires computing the profile likelihood for on the order of $log_2(\hat{S}_\Omega - S_{obs})$ values of S_Ω. The lower limit of the $100p\%$ profile likelihood interval is the smallest value of S_Ω such that $l_P(S_\Omega) \geq l_P(\hat{S}_\Omega) + ln(p)$.

The upper endpoint of the $100p\%$ profile likelihood interval is the largest value of S_Ω such that $l_P(S_\Omega) \geq l_P(\hat{S}_\Omega) + ln(p)$. To compute the upper endpoint of the profile likelihood interval, bounds on the value of the upper endpoint are first found. The profile log-likelihoods are computed for an increasing geometric sequence of values $S_\Omega^{(1)}, S_\Omega^{(2)}, \ldots$, where $S_\Omega^{(j)}$ and $S_\Omega^{(j+1)}$ differ by $2^{j-1} \cdot 200$, and $S_\Omega^{(1)} = \hat{S}_\Omega$. Let $S_\Omega^{(m)}$ denote the smallest value in the sequence such that $l_P(S_\Omega^{(m)}) < l_P(\hat{S}_\Omega) + ln(p)$. Then, the upper limit of the likelihood interval is bounded between $S_\Omega^{(m-1)}$ and $S_\Omega^{(m)}$. A binary search algorithm then is employed to find the upper limit of the profile likelihood interval between these two bounds.

In Chapter 5, the interval estimates constructed from profile likelihood intervals will be compared with those based on the asymptotic normality of $\hat{S}_\Omega$.

3.7 Inference on $\boldsymbol{\theta}$

In addition to inference on S_Ω, inference on $\boldsymbol{\theta} = \langle \beta, k \rangle$ is also of interest. Inference on the rate β of the exponential mixing distribution translates to inference on the distribution of regional abundances of the species. The dispersion parameter k relates to the overall tendency for con-specific sample abundances to have a variance greater than their expected value.

Having previously computed the observed information matrix $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$, the square roots of the main diagonal elements of $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})$ are the estimated asymptotic standard errors of the MLEs of β and k. Approximate confidence intervals for β and k can be constructed based on the asymptotic normality of their MLEs.

3.8 Summary

The estimator $\hat{S}_{\Omega}$ introduced in this chapter assumes randomly-placed sampling units of uniform size. This sampling method is more feasible in ecological field studies for some taxonomic groups (e.g., benthic species – Ugland *et al.*, 2003; microarthropods – Lindo & Winchester, 2006) than acquiring a truly random sample of individuals. The sample abundances of species are recorded for each sampling unit.

We have developed a mixture model to accommodate the tendency of con-specific individuals to occur in spatial clusters. The sample abundances of a species in the sampling units are treated as conditionally independent negative binomial random variables. An exponential mixing distribution is applied to the means of the negative binomial distributions of the species, and this permits inference on the unobserved species. The negative binomial dispersion parameter k is assumed constant across the species.

Attempts to apply a gamma mixing distribution on μ encountered computational difficulties with numerical integration, as did attempts to use the lognormal distribution in an earlier model. Alternative techniques of numerical integration, perhaps

with numerical approximations, are likely required in order to use the gamma and lognormal distributions as mixing distributions on μ.

Based on the statistical model, the likelihood function is developed assuming that the species are mutually independent. Adopting a profile likelihood approach, the MLEs of S_Ω and $\boldsymbol{\theta} = \langle \beta, k \rangle$ are computed numerically.

Approximate interval estimates for S_Ω are based on either the asymptotic normality of $\hat{S}_\Omega$ or profile likelihood methods. Asymptotic standard errors for the components of $\hat{\boldsymbol{\theta}}$ are estimated from the numerical computation of the observed information matrix.

The estimator $\hat{S}_\Omega$ developed in this chapter is the only estimator of species richness designed for abundance-based data from multiple samples (i.e., randomly-located spatial-based sampling units) taken from a single region. As discussed by Bunge and Barger (2009), this is an important and open area of research for the estimation of species richness when several samples are taken from a single region.

Chapter 4

Bayesian Species Richness Estimation with Abundance-Based Data from Multiple Sampling Units

‘

In this chapter, the negative binomial mixture model is extended to allow for greater flexibility, with inference based on the Bayesian framework. Extending the model framework into the Bayesian setting has several benefits. The Bayesian approach provides an entire discrete posterior distribution for S_Ω, which can be used to obtain integer-valued point estimates such as the posterior median and posterior mode. Large sample approximations such as asymptotic normality are not required.

The asymmetry of the posterior distribution is incorporated when developing interval estimates. In addition, interval estimates based on the posterior distribution will remain within the parameter space; in particular, the lower limit of an interval estimate will never be less than S_{obs}, which was a problem in Section 3.6.1 for the frequentist method. Furthermore, the computational implementation of the Bayesian model can be quickly adjusted to accommodate modifications to the model, and this will be demonstrated in the case study considered in Chapter 6.

For readers unfamiliar with Bayesian statistical analysis, we have included a concise overview of Bayesian analysis and computation with Markov chain Monte Carlo (MCMC) methods in Appendix B. The contents of Appendix B have been reproduced with permission from Section 2.4 of Hong Li's (2008) thesis.

4.1 Bayesian computation and the Number of Parameters

In the methodology of Chapter 3, we assume that the S_Ω species independently contribute individuals to each sampling unit with expected values $\boldsymbol{\mu} = \langle \mu_1, \ldots, \mu_{S_\Omega} \rangle$ and $\mu_i \sim \text{Exponential}(\beta)$, $i = 1, \ldots, S_\Omega$.

Bayesian hierarchical models typically do not permit an analytic analysis of the posterior distribution. Therefore, to make full use of the sample data, numerical approaches using Markov chain Monte Carlo (MCMC) are necessary. However, if a prior distribution is placed explicitly on S_Ω, Bayesian computation is complicated

by the fact that the number of the parameters (in our case, the number of means, $\mu_1, \ldots, \mu_{S_\Omega}$) will change as S_Ω changes in each iteration of the MCMC sampler.

Reversible jump MCMC algorithms can accommodate the fluctuating number of parameters between iterations (King *et al.*, 2010). However, reversible jump methods can be difficult to implement and require considerable tuning. As an alternative, we use a data augmentation technique (Royle *et al.*, 2007) that expands the model in such a way so that the number of parameters remains fixed. This permits Bayesian computation using standard MCMC methods based on random walk Metropolis Hastings and Gibbs sampling. The resulting algorithms can be implemented in the freely available WinBUGS (Lunn *et al.*, 2000) and JAGS (Plummer, 2003) software for Bayesian computation.

4.2 Hierarchical Bayesian model with Data Augmentation

To develop our data-augmented modelling formulation, we begin by conceptualizing a hypothetical 'super-community' encompassing the taxonomic group of under study. The super-community occupies a large geographic area (containing the region of interest) and the number of species in the super-community is $M >> S_\Omega$. The *precise value* of M is not of importance, and we will assign M to be a fixed known number that is larger than any reasonable value for the number of species in *our* region of interest. We use the concept of super-community to elicit prior information on M,

which is essentially a gross upper bound on the true unknown S_Ω. Assigning M a priori avoids computational difficulties arising from variable-dimension parameter spaces (see, e.g., Royle *et al.*, 2007). Values of M are easily elicited in practical applications (see Chapter 6), and robustness to the choice of M can be examined through sensitivity analyses.

For each species in the super-community, we define a binary variable $z_i \in \{0, 1\}$, $i = 1, \ldots, M$, where

$$z_i = \begin{cases} 1 & \text{if species } i \text{ exists in the region of interest} \\ 0 & \text{if not} \end{cases}$$

so that $S_\Omega = \sum_{i=1}^{M} z_i$. We let each of the M species in the super-community have probability ψ of independently occurring in our region of interest so that $z_i \overset{iid}{\sim} \text{Bernoulli}(\psi)$. We assume a labelling of all M species in the super-community such that $i = 1, ..., S_{obs}$ correspond to those observed during sampling and $i = S_{obs} + 1, \ldots, M$ indexes those species not observed. Given this, we have $z_i = 1$ for $i = 1, ..., S_{obs}$; whereas, z_i, for $i = S_{obs} + 1, \ldots, M$, are unknown and will be treated as latent variables in the model. Note also that $S_\Omega = S_{obs} + \sum_{i=S_{obs}+1}^{M} z_i$, so that inference on z_i, for $i > S_{obs}$, permits inference on S_Ω.

Our observed data are $\boldsymbol{n_i}$, for $i = 1, \ldots, S_{obs}$, and, given M, we augment these data with $\boldsymbol{n_i} = \mathbf{0}$ for $i > S_{obs}$. For $i \leq S_{obs}$, we assume

$$n_{i,q} \mid \mu_i, k \overset{iid}{\sim} \text{Negative Binomial}(\mu_i, k),$$

for sampling unit $q = 1, \ldots, Q$. For $i > S_{obs}$, we assume

$$n_{i,q} \mid \mu_i, z_i, k \overset{iid}{\sim} \begin{cases} \text{Negative Binomial}(\mu_i, k) & \text{if } z_i = 1 \\ I\{n_{i,q} = 0\} & \text{if } z_i = 0 \end{cases}$$

for $q = 1, \ldots, Q$. That is, the value $\boldsymbol{n_i} = \mathbf{0}$ arises either from a negative binomial distribution (if species i exists in the region of interest) or $\boldsymbol{n_i} = \mathbf{0}$ with probability 1 if species i does not exist in our region of interest.

Variability in the means μ_i arises from a mixing distribution. In the preceding chapter, this variability was modelled using an exponential distribution; however, we have found that more flexibility in the distribution improves estimation of S_Ω considerably. Gamma and log-normal distributions were considered as alternatives but led to computational problems in initial investigations in WinBUGS (e.g., trap errors and messages stating "cannot bracket slice for node" μ_i). Norris and Pollock (1998) considered finite mixtures of point masses on μ. In an empirical study of microbial diversity, Bunge and Barger (2008) found finite mixtures of exponential distributions were flexible and performed favourably compared with other distributions (i.e., gamma, lognormal, inverse Gaussian and Pareto distributions).

As a natural extension of the exponential distribution from Chapter 3, we assume the means μ_i arise as exchangeable draws from a two-component mixture of exponential distributions, G, having probability density function

$$g(\mu\,; \beta_1, \beta_2, \pi) = \pi\beta_1 e^{-\beta_1 \mu} + (1 - \pi)\beta_2 e^{-\beta_2 \mu}\,, \tag{4.1}$$

for $\mu > 0$, where $0 < \pi < 1$ is the mixing probability, and the exponential rate parameters are $\beta_1 > 0$ and $\beta_2 > 0$. The exponential distribution with rate β_1 has

weight π, and the exponential distribution with rate β_2 has weight $1-\pi$ in the mixture distribution. Without loss of generality, we assume $\beta_1 < \beta_2$, so that the model is identifiable to permit inference on β_1, β_2 and π. We can interpret the mixture G as representing two groups of species within our taxonomic group of interest. One group of species occur frequently in the sampling units; the expected number (μ) of con-specific individuals in a sampling unit is an exponential random variable with rate β_1. The species in the second group are rare, and appear less frequently in the sampling units, with $\mu \sim \text{Exponential}(\beta_2)$.

Letting $\Delta\beta = \beta_2 - \beta_1$, we assume mutually independent prior distributions for $k, \beta_1, \Delta\beta, \pi$ and ψ. Uniform(0,1) priors are placed on π and ψ. Diffuse gamma priors are placed on k, β_1 and $\Delta\beta$ where we assign the value 0.001 to the shape parameter and the rate parameter of each gamma prior (so the mean is 1 and the variance is 1000).

Following the model of Royle *et al.* (2007), we have placed a prior distribution on the total S_Ω in two stages. Recall that $S_\Omega = \sum_{i=1}^{M} z_i$, a priori. In the first stage, given ψ, S_Ω has a binomial prior distribution with index M and probability ψ. In the second stage, we have assigned a flat Uniform(0,1) prior on ψ. Integrating yields the marginal prior

$$\begin{aligned} P(S_\Omega \mid M) &= \int_0^1 P(S_\Omega \mid M, \psi) P(\psi) d\psi \\ &= \int_0^1 \binom{M}{S_\Omega} \psi^{S_\Omega} (1-\psi)^{M-S_\Omega} \, d\psi \\ &= \frac{1}{M+1}, \end{aligned}$$

and the result is a discrete uniform prior, $P(S_\Omega \mid M)$, on S_Ω with support over $\{0, 1, \ldots, M\}$.

The assumptions stated in Section 3.3 for our likelihood-based approach are also required in our Bayesian approach:

1. A uniform sampling effort is applied to each spatial sampling unit.
2. The spatial autocorrelation between spatial sampling units is assumed to be negligible so that sampling units can be treated as independent.
3. All individuals that occur in a sampling unit are detected and correctly identified to their species type.
4. The region of interest does not gain or lose species during the sampling period.
5. The sample abundances of different species in the sampling units are mutually independent.
6. The spatial distributions of species are stationary.

Under these assumptions, the hierarchical Bayesian model is concisely specified as follows:

Level I: *Likelihood*

- For $i = 1, \ldots, M$ and $q = 1, \ldots, Q$,

$$n_{i,q} \mid \mu_i, z_i, k \overset{iid}{\sim} \begin{cases} \text{Negative Binomial}(\mu_i, k) & \text{if } z_i = 1 \\ I\{n_{i,q} = 0\} & \text{if } z_i = 0 \end{cases}$$

- For $i = 1, \ldots, M$, $\quad z_i \mid \psi \overset{iid}{\sim} \text{Bernoulli}(\psi)$

Level II: *Priors*

- For $i = 1, \ldots, M$, $\quad \mu_i \mid \beta_1, \beta_2, \pi \overset{iid}{\sim} G(\mu; \beta_1, \beta_2, \pi)$

  with $g(\mu\,; \beta_1, \beta_2, \pi) = \pi\beta_1 e^{-\beta_1 \mu} + (1-\pi)\beta_2 e^{-\beta_2 \mu}$ and $\beta_1 < \beta_2$.

- $k \sim \text{Gamma}(0.001,\, 0.001)$

Level III: *Hyper-Priors*

- $\psi \sim \text{Uniform}(0,\, 1)$

- $\beta_1 \sim \text{Gamma}(0.001,\, 0.001)$

- $\Delta\beta = \beta_2 - \beta_1 \sim \text{Gamma}(0.001,\, 0.001)$

- $\pi \sim \text{Uniform}(0,\, 1)$

The latent variables and model parameters are collected in the set $\boldsymbol{\Theta} = \{z_{S_{obs}+1}, \ldots, z_M,\ \boldsymbol{\mu}, k, \psi, \pi, \beta_1, \Delta\beta\}$ with $\boldsymbol{\mu} = \langle \mu_1, \ldots, \mu_M \rangle$ and $S_\Omega = S_{obs} + \sum_{i=S_{obs}+1}^{M} z_i$, and inference proceeds through the posterior distribution $P(\boldsymbol{\Theta} \mid \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}})$, which we compute using MCMC.

4.3 Likelihood and Posterior

Before collecting data, given ψ, the indicator variable z_i has probability mass function

$$P\left(z_i \mid \psi\right) = \psi^{z_i}(1-\psi)^{1-z_i}\,,$$

for $i = 1, \ldots, M$. Conditional on z_i, μ_i and k,

$$P\left(\boldsymbol{n_i} \mid z_i, \mu_i, k\right) = z_i \cdot \left\{ \prod_{q=1}^{Q} P_{NB}\left(n_{i,q}\,; k, \mu_i\right) \right\} + (1 - z_i) \cdot I\left\{\boldsymbol{n_i} = \mathbf{0}\right\} \ , \tag{4.2}$$

where $P_{NB}\left(n_{i,q}\,; k, \mu_i\right)$ denotes the p.m.f. of the negative binomial distribution presented in Equation 3.1. The joint probability of $\boldsymbol{n_i}$ and z_i, given ψ, μ_i and k, is then

$$\begin{aligned} P\left(\boldsymbol{n_i}, z_i \mid \psi, \mu_i, k\right) &= P\left(\boldsymbol{n_i} \mid z_i, \mu_i, k\right) \cdot P\left(z_i \mid \psi\right) \\ &= \left[z_i \cdot \left\{ \prod_{q=1}^{Q} P_{NB}\left(n_{i,q}\,; k, \mu_i\right) \right\} + (1 - z_i) \cdot I\left\{\boldsymbol{n_i} = \mathbf{0}\right\} \right] \psi^{z_i} (1 - \psi)^{1 - z_i} \ . \end{aligned}$$

When $z_i = 1$ is known,

$$P\left(\boldsymbol{n_i}, z_i = 1 \mid \psi, \mu_i, k\right) = \psi \prod_{q=1}^{Q} P_{NB}\left(n_{i,q}\,; k, \mu_i\right) \ . \tag{4.3}$$

The observed data are $\boldsymbol{n_1}, \ldots, \boldsymbol{n_M}$ and $z_1 = z_2 = \ldots = z_{S_{obs}} = 1$. Using probability expressions of the forms in Equations 4.2 and 4.3, the likelihood function is

$$\begin{aligned} &L(\boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}} \mid \boldsymbol{\Theta}) \\ &\qquad = \prod_{i=1}^{S_{obs}} P\left(\boldsymbol{n_i}, z_i = 1 \mid \psi, \mu_i, k\right) \ \times \prod_{i=S_{obs}+1}^{M} P\left(\boldsymbol{n_i} \mid z_i, \mu_i, k\right) \\ &\qquad = \prod_{i=1}^{S_{obs}} \left\{ \psi \prod_{q=1}^{Q} P_{NB}\left(n_{i,q}\,; k, \mu_i\right) \right\} \ \times \\ &\qquad\quad \prod_{i=S_{obs}+1}^{M} \left[z_i \cdot \left\{ \prod_{q=1}^{Q} P_{NB}\left(n_{i,q}\,; k, \mu_i\right) \right\} + (1 - z_i) \cdot I\left\{\boldsymbol{n_i} = \mathbf{0}\right\} \right] \ , \end{aligned}$$

assuming species are mutually independent.

The joint posterior distribution of $\boldsymbol{\Theta}$, up to a normalizing constant, is then,

$$\begin{aligned} &P(\boldsymbol{\Theta} \mid \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}}) \\ &\quad \propto L(\boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}} \mid \boldsymbol{\Theta}) \times \left[\prod_{i=S_{obs}+1}^{M} \psi^{z_i} (1-\psi)^{1-z_i} \right] \\ &\qquad \times \left[\prod_{i=1}^{M} g(\mu_i \,; \beta_1, \beta_2, \pi) \right] \times P(k) P(\psi) P(\pi) P(\beta_1) P(\Delta\beta) \,, \end{aligned}$$

where $P(k)$ denotes the prior on k, and so forth. For $i = S_{obs} + 1, \ldots, M$, $\boldsymbol{n_i} = \mathbf{0}$ and the full conditional posterior distribution of z_i is Bernoulli with probability mass function

$$\begin{aligned} &P(z_i \mid \boldsymbol{\Theta}_{-z_i}, \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}}) \\ &\quad = \frac{\left[z_i \cdot \left\{ \prod_{q=1}^{Q} P_{NB}\left(n_{i,q} \,; k, \mu_i\right) \right\} + (1 - z_i) \right] \psi^{z_i} (1-\psi)^{1-z_i}}{\left\{ \prod_{q=1}^{Q} P_{NB}\left(n_{i,q} \,; k, \mu_i\right) \right\} \psi + (1-\psi)} \\ &\quad = \frac{z_i P_{NB}\left(0 \,; k, \mu_i\right)^Q \psi + (1 - z_i)(1-\psi)}{P_{NB}\left(0 \,; k, \mu_i\right)^Q \psi + (1-\psi)} \,, \end{aligned}$$

where $\boldsymbol{\Theta}_{-z_i}$ is $\boldsymbol{\Theta}$ excluding z_i.

The normalizing constant of the joint posterior distribution of $\boldsymbol{\Theta}$ is analytically intractable. Therefore, Bayesian inference is based upon drawing samples from the posterior, and then using the sample to derive Monte Carlo estimates of posterior quantities. To do this, we employ MCMC algorithms implemented in the Bayesian software program JAGS Version 2.2.0 (Plummer, 2003). The MCMC simulates draws from an ergodic Markov chain having $P(\boldsymbol{\Theta} \mid \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}})$ as its stationary distribution (Gelman *et al.*, 2004).

In the MCMC implementation, two or more simulation chains are run simultaneously. The initial values of the parameters $\boldsymbol{\Theta}$ in the chains are chosen to be dispersed

(i.e., in the tails of their marginal posterior distributions – this is accomplished after some initial MCMC simulations are run to get an idea of the posterior distributions of the parameters). The first phase of the MCMC simulation is considered a *burn-in* period during which the chains are allowed to reach steady state behaviour (i.e., the stationary distribution). We assume convergence has occurred when the Gelman and Rubin diagnostic statistic $\hat{R}$ is less than 1.1 for all monitored parameters, as recommended by Gelman *et al.* (2004).

Discarding the iterations corresponding to the burn-in phase, the MCMC simulation continues in a second phase called the *production run.* Simulated values in the production run are treated as samples from the joint posterior distribution $P(\boldsymbol{\Theta} \mid \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}})$. However, we note that the simulated values are not independent, and so autocorrelation is present between consecutive samples of the chain. To reduce the autocorrelation, the chain is 'thinned', and parameter values from every j^{th} iteration are recorded. The collection of recorded iterations of the production run constitute a sample from the joint posterior distribution $P(\boldsymbol{\Theta} \mid \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}})$. A large sample of values of $S_\Omega = S_{obs} + \sum_{i=S_{obs}+1}^{M} z_i$ from the production run are used to approximate the marginal posterior distribution of S_Ω.

Before the MCMC simulation is started, the value of M must be chosen as sufficiently large such that the upper tail of the posterior distribution of S_Ω will be well below M. So long as M is well above the largest value in the "effective" support of the posterior distribution of S_Ω, the results appear to be robust with respect to the

choice of M. We have run MCMC simulations with different values of M on the same data set and received essentially identical results for the posterior distribution of S_Ω. Given a data set, the time required to complete a specified number of iterations of the MCMC depends on the value of M.

In our simulations, the time to run the MCMC algorithm on a single data set varied from one day to more than six days, depending on the data set and the value of M. For the simulation studies presented in Chapter 5, it was necessary to fit the hierarchical Bayesian model to a large number of simulated data sets. This was not feasible on a single desktop PC running WinBUGS in a Windows operating system. Alternatively, we ran multiple sessions of JAGS on a Linux cluster of 12 nodes, each with a dual quad-core Xeon processor. Using the R2jags package, script files executed in the R software environment handled the MCMC in JAGS. Using this computational set-up, we were able to fit the hierarchical Bayesian model to up to sixty-four data sets simultaneously.

4.3.1 Hierarchical Bayesian Model specification in JAGS and WinBUGS

The specification of the hierarchical Bayesian model for JAGS and WinBUGS is shown below. The variable names in this model specification coincide with the notation in Section 4.2 (e.g., "psi" = ψ, "beta1" = β_1, "beta2.minus.beta1" = $\beta_2 - \beta_1$, "mu[i]" = μ_i, "n[i,q]" = $n_{i,q}$, "S.Omega" = S_Ω, etc.).

```
model {
    #Likelihood:
    for (i in 1:M) {
        z[i] ~ dbern(psi)
        p[i] <- k / (z[i]*mu[i] + k) # Use this line only in WinBUGS.
          # Use the next line only in JAGS because it requires 0 < p[i] < 1.
        p[i] <- k / (z[i]*mu[i] + k) - (1-z[i])*0.000000001
        for (q in 1:Q) {
            n[i,q] ~ dnegbin(p[i], k)
        }
    }
    # Priors:
    for (i in 1:M) {
        delta[i] ~ dbern(pi)
        beta[i] <- delta[i]*beta1 + (1 - delta[i])*beta2
        mu[i] ~ dexp(beta[i])
    }
    pi ~ dunif(0, 1)
    psi ~ dunif(0, 1)
    beta1 ~ dgamma(0.001, 0.001)
    beta2.minus.beta1 ~ dgamma(0.001, 0.001)
```

```
    beta2 <- beta1 + beta2.minus.beta1
    k ~ dgamma(0.001, 0.001)
    # Interested in the posterior distribution of S.Omega
    S.Omega <- sum(z[ ])
}
```

4.4 Bayesian Inference on S_Ω

The posterior samples of S_Ω from the MCMC algorithm typically produce a unimodal histogram with positive skewness. The lower tail of the distribution ends at or above S_{obs}, as it should be. Using the posterior distribution, we consider three point estimates of S_Ω: the posterior mean, median and mode. From the MCMC sampler, we are also able to compute the posterior variance of S_Ω.

In addition, we consider two interval estimates of the local species richness. First, we use the 95% highest posterior density (HPD) interval of S_Ω, which is the shortest interval having a specified coverage probability. Second, we construct the equal-tail 95% credible interval using the 2.5^{th} and 97.5^{th} percentiles from the estimated posterior of S_Ω. With the typically positive skewness of the posterior distribution, both interval estimates of S_Ω will be asymmetric, with their upper limits farther from the mode of the distribution than their lower limits.

4.5 Model Checking

The choice of the mixing distribution on the mean abundances $\boldsymbol{\mu}$ has an effect on the estimates of species richness (Barger & Bunge, 2008). If we have multiple models to fit for a given data set (e.g., different distributions on the means $\boldsymbol{\mu}$), model selection becomes an important issue. To choose between competing models, we employ the deviance information criteria (Spiegelhalter *et al.*, 2002). Based on the deviance statistic

$$D(\boldsymbol{\Theta}) = -2\, ln\, L(\boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}} \mid \boldsymbol{\Theta}) \,,$$

the deviance information criteria (DIC) is defined as

$$DIC = \overline{D(\boldsymbol{\Theta})} + p_D \,,$$

where $\overline{D(\boldsymbol{\Theta})} = \mathrm{E}[D(\boldsymbol{\Theta}) \mid \boldsymbol{n_1}, \ldots, \boldsymbol{n_M}, z_1, \ldots, z_{S_{obs}}]$ is the posterior mean of the deviance, estimated from all recorded iterations of the production run of the MCMC simulation. The term p_D is defined as

$$p_D = \overline{D(\boldsymbol{\Theta})} - D(\bar{\boldsymbol{\Theta}}) \,,$$

where $D(\bar{\boldsymbol{\Theta}})$ is the deviance evaluated at the posterior mean of $\boldsymbol{\Theta}$. While the deviance will decrease as the number of parameters in a model increases, the penalty term p_D favours models with fewer parameters. Lower values of DIC correspond to models that have a better fit and are more parsimonious.

4.6 Remarks

In this chapter, we have extended our abundance-based multiple-sampling-unit model from the previous chapter in three key areas. First, we have developed a data augmentation scheme based on the notion of a super-community and the incorporation of binary latent variables z_i, $i = S_{obs} + 1, \ldots, M$. Second, we have extended the distribution modelling variability in the expected contributions $(\mu_1, \mu_2, \ldots, \mu_M)$ from an exponential distribution to a mixture of two exponential distributions. Third, we have developed Bayesian inference via MCMC simulation rather than using a classical likelihood-based approach. The resulting inference is exact (up to Monte Carlo error), and does not rely on large sample approximations. We have also developed a computational implementation using the JAGS software program on a high performance Linux cluster. The computational set-up on the Linux cluster facilitates the simulation studies presented in Chapter 5. The change in the mixing distribution on $\boldsymbol{\mu}$ improves estimator performance, as will be demonstrated in the simulation studies.

It is important to emphasize that, like our likelihood-based approach, our Bayesian approach works with abundance-based data from multiple spatial sampling units. Kéry and Royle (2008) have also introduced a hierarchical Bayesian model to estimate species richness by using information from multiple spatial sites; however, their approach only uses detection histories, and requires repeated visits to each site. Our proposed methods are the only species richness methods that are designed to handle abundance-based data from multiple sites situated in a single region of interest.

Chapter 5

Simulation Studies

In this chapter, we study the performance of the methods proposed in Chapters 3 and 4 using simulation experiments. We focus primarily on the bias and precision of the resulting estimators of species richness in different scenarios corresponding to different levels of sampling effort. In addition, we compare our estimators with three commonly used non-parametric estimators.

The simulation studies are based on synthetic data generated under two settings. In the first, data are generated from a negative binomial mixture model. In the second setting, realistic sample data are simulated by sub-sampling from a census of woody plants in a 50-hectare plot of a Panamanian tropical rain forest.

5.1 Simulation Study 1

The first simulation setting uses samples generated from a negative binomial mixture model. Each set of samples for this setting is simulated as follows:

- **Step 1:** Fix values for the model parameters: S_Ω, the number Q of sampling units, the negative binomial dispersion parameter k, the rates β_1 and β_2 and the mixture weight π.

- **Step 2:** Generate independent values $\mu_1, \mu_2, \ldots, \mu_{S_\Omega}$ from the two-component exponential mixture model $G(\mu; \beta_1, \beta_2, \pi)$. Recall that μ_i represents the expected number of individuals that species i contributes to a sampling unit. Given μ_i, a vector $\boldsymbol{n_i}$ of Q independent counts for species i is generated from the negative binomial distribution with mean μ_i and variance $\mu_i + \mu_i/k$.

- **Step 3:** Include the vector of counts for a species in the sample if at least one of the counts is positive. That is, a species must contribute at least one individual to the sample; otherwise, the species is absent from the sample and therefore not recorded.

- **Step 4:** With the values in Step 1 fixed, repeat Steps 2 and 3, and generate 100 samples.

The simulation scheme is implemented in the R programming language. Based on the steps above, we draw two sets of 100 samples. The two sets are constructed

using identical values of the model parameters, but with Q, the number of sampling units, set to $Q = 8$ in the first set, and $Q = 16$ in the second set.

The fixed values of the parameters in Step 1 are obtained from a realistic setting by fitting our hierarchical Bayesian model to an oribatid mite data set collected from a Panamanian forest floor (data used with permission of Dr. Neville Winchester, University of Victoria). Chapter 6 considers a detailed case study examining these data; however, we use the data here to suggest realistic values of $S_\Omega, Q, k, \beta_1, \beta_2$ and π for the simulation study.

Table 5.1 lists the parameter values used in Step 1 for generating the two sets of 100 samples. Both sets of samples are constructed from artificial communities containing $S_\Omega = 148$ species. Each species has probability $\pi = 0.479$ of being in the group of relatively common species where their expected contributions (μ_i's) to the sampling units are random variables drawn from an exponential distribution with rate $\beta_1 = 0.0489$; whereas, each species has probability $1-\pi = 0.521$ of being in the group of relatively uncommon species where their mean contributions to the sampling units are random variables drawn from an exponential distribution with rate $\beta_2 = 1.991$. The dispersion parameter is $k = 0.558$, implying that the sample variance of the simulated con-specific abundances will tend to be approximately twice as large as the sample mean. Within the context of species richness, this overdispersion means fewer species will tend to occur in the samples than if there was no overdispersion (i.e., inference on S_Ω will tend to rely on a smaller number, S_{obs}, of species observed).

Table 5.1: Values of Parameters for Simulating Data

Parameter	Value for Set 1	Value for Set 2
S_Ω	148	148
k	0.558	0.558
β_1	0.0489	0.0489
β_2	1.991	1.991
π	0.479	0.479
Q	8	16

5.2 Simulation Study 2

In the second simulation setting, we create samples by sub-sampling from a *census* of woody plants where S_Ω is known (Hwang & Shen, 2010). A 500-metre by 1000-metre rectangular plot in the tropical lowland rainforest on Barro Colorado Island (BCI) has been surveyed for woody plant species seven times over the last 30 years (Hubbell *et al.*, 2005). We use the most recent *census* of the 50-hectare BCI plot from the year 2010, which included all woody plant specimens with diameters at breast height of at least 1 cm. We draw samples from the census data. The estimates of woody plant species richness produced from the samples can be compared with the known number of woody plant species, from the census, in the 50-hectare plot. For the purposes of this simulation setting, the 50-hectare plot is our region of interest.

At the time of the 2010 census, observers found a total of 221,757 specimens in the 50-hectare plot and identified them as belonging to $S_\Omega = 305$ woody plant species. The distribution of specimens among the species is quite uneven. Twenty-two species

only have one specimen in the BCI plot. The 153 least common species have a combined number of specimens equal to 2.1% of the total number of specimens in the 50-hectare plot. The single most abundant species has 30,130 specimens (13.6% of all specimens in the plot).

To create a sample from these census data, we simulate the placement of a fixed number of quadrats inside the 50-hectare plot. The quadrats are squares of uniform size. Using the known spatial coordinates of all specimens from the census of the plot, we can generate the observations within each quadrat. A sample will consist of the observations from the quadrats.

In this second simulation setting, we create four sampling scenarios and generate a total of 100 samples for each sampling scenario. The following steps are used to generate 100 samples:

- **Step 1:** Choose the number, Q, of quadrats to include in the sample, and fix the area of each quadrat.

- **Step 2:** Randomly place squares of the desired area within the perimeter of the 500-metre by 1000-metre BCI plot. Squares are added iteratively until Q non-intersecting squares have been established. These Q non-intersecting squares become the quadrats associated with one sample.

- **Step 3:** Given the quadrats, the observations from the census data are extracted for all specimens that reside in the quadrats. For each species, the result is a vector of Q sample abundances – one sample abundance for each quadrat. Note

that the sample obtained in this manner consists of the collection of vectors of sample abundances, but is restricted to include only those species that have at least one specimen in at least one simulated quadrat. A species that has zero specimens in all quadrats will be absent from the sample.

- **Step 4:** Using the specifications of the quadrats established in Step 1, Steps 2 and 3 are repeated until 100 independent sample data sets have been created.

Our hierarchical models are developed under the assumption that the sampling units are taken *with* replacement. In practice, however, the quadrats that constitute a sample are not permitted to intersect. Therefore, a sample of quadrats is taken *without* replacement. Fortunately, this discrepancy should have a negligible effect as at most 0.8% of the 50-hectare plot is included in a sample.

Under this simulation setting, we use four scenarios to investigate the impact of changing the number and size of the quadrats on the performance of estimators. The four scenarios considered are:

- *Scenario 1*: A sample consists of 40 quadrats, each quadrat being 5 metres by 5 metres.

- *Scenario 2*: A sample consists of 10 quadrats, each quadrat being 10 metres by 10 metres.

- *Scenario 3*: A sample consists of 40 quadrats, each quadrat being 10 metres by 10 metres.

- *Scenario 4*: A sample consists of 10 quadrats, each quadrat being 20 metres by 20 metres.

The four scenarios use samples containing a small fraction of the BCI plot in order for the sampling fractions to be comparable to the sampling fractions of real field surveys (Chiarucci *et al.*, 2003). Scenarios 1 and 2 have the smallest sampling fraction considered, with each sample occupying 1000 m^2, 0.2% of the 50-hectare plot. In Scenarios 3 and 4, a sample occupies 4000 m^2, 0.8% of the 50-hectare plot.

The impact of the sampling fraction on estimator performance will be examined in Section 5.7 by comparing the results of Scenarios 1 and 2 with the results of Scenarios 3 and 4. When the sampling fraction is increased from 0.2% in the first two scenarios to 0.8% in the last two scenarios, the quantity of quadrats or the size of the quadrats must change.

To examine the effect of the *size* of the quadrats on estimator performance, one may wish to hold both the sampling fraction and the quantity of quadrats fixed, allowing only the size of the quadrats to vary. However, for a fixed number of quadrats, varying the size of the quadrats forces the sampling fraction to change, too. The relationship between the sampling fraction, the quantity of quadrats, and the size of the quadrats is simply

$$sampling\ fraction = \frac{(quantity\ of\ quadrats)(area\ of\ a\ quadrat)}{area\ of\ region}\ .$$

5.3 Comparison with Alternative Methods

We compare our maximum likelihood and Bayesian estimators with three commonly used species richness estimators. We also use S_{obs}, the number of species observed in the sample, as a baseline estimator of species richness.

Two of the competing estimators originate from the capture-recapture literature: the second-order jackknife estimator $\hat{S}_{Jack2}$ (Burnham & Overton, 1979), and an estimator $\hat{S}_{Chao2}$ from Chao (1987), both estimators being non-parametric. Both estimators are computed based on detection histories (also called incidence-based data); that is, they do not require the exact number of individuals of a species in a sampling unit, but are based simply on presence/absence data.

Our approaches use abundance-based data, and therefore, we also make comparisons to a commonly used abundance-based estimator, the non-parametric estimator of Chao and Lee (1992). We know of no other estimators of species richness based on abundance-based data, from multiple sampling units, sampled from within one community. Subsequently, we will label the Chao and Lee estimator, originally presented in their Equation 2.15, as $\hat{S}_{ACE}$. The estimator $\hat{S}_{ACE}$ was developed for use when the relative abundances of the species in the population are believed to be highly heterogeneous. Bunge and Fitzpatrick (1993) also recommended its use in the absence of information about the relative abundances of the species.

$\hat{S}_{ACE}$ is designed for use with abundance-based data from a single random sample of individuals. To apply $\hat{S}_{ACE}$ to a sample obtained from observations from multiple

sampling units in our simulation studies, we must reduce the data by combining the observations from all sampling units. More specifically, for $i = 1, 2, \ldots, S_{obs}$, we use $\sum_{q=1}^{Q} n_{i,q}$, the total number of individuals of species i that occur in the sample, rather than separate counts $\{n_{i,1}, n_{i,2}, \ldots, n_{i,Q}\}$.

We have selected the three non-parametric estimators $\hat{S}_{ACE}$, $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ for comparison as they have been used in several simulation studies (on real and simulated data sets – see Walther & Moore (2005) for a list of the studies) and have exhibited good overall performances relative to other estimators, including parametric estimators and estimators based on species-accumulation curves (Walther & Moore, 2005). In a comprehensive survey of 14 studies on species richness estimation, Walther and Moore (2005) ranked $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ as having the smallest overall biases, and second and third, respectively, in mean square error. The first-order jackknife estimator (Burnham and Overton, 1979) was ranked third in smallest bias and ranked first in mean square error; it uses presence/absence data. Since the two presence/absence estimators with the smallest biases are already considered, and with bias having a dominant role in estimator performance, the first-order jackknife estimator was not included in our comparison.

To introduce $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$, some notation is required. If a species is observed in the sample, then it is observed in at least one of the sampling units. We count the number of sampling units in which a species is observed, and define this as the *frequency of detection*. We let r_j represent the number of species with a frequency of detection equal to j, for $j = 1, 2, \ldots, Q$.

For example, suppose observations are made in 8 sampling units and a total of $S_{obs} = 10$ species are observed with the following frequencies of detection:

$$1, 1, 1, 2, 3, 3, 5, 5, 6, 8$$

Three species are each observed exactly once (i.e., $r_1 = 3$), and one species is observed in all 8 sampling units (i.e., $r_8 = 1$). In total, we have $r_1 = 3$, $r_2 = 1$, $r_3 = 2$, $r_5 = 2$, $r_6 = 1$, and $r_8 = 1$, while $r_4 = 0$ and $r_7 = 0$. Notice that $S_{obs} = \sum_{j=1}^{8} r_j$.

Chao's (1987) incidence-based estimator is then computed as

$$\hat{S}_{Chao2} = S_{obs} + \frac{r_1^2}{2r_2} .$$

Its estimated variance (see, e.g., Colwell, 2009) is

$$\widehat{Var}(\hat{S}_{Chao2}) = r_2 \cdot \left[\frac{1}{2} \left(\frac{r_1}{r_2} \right)^2 + \left(\frac{r_1}{r_2} \right)^3 + \frac{1}{4} \left(\frac{r_1}{r_2} \right)^4 \right].$$

Colwell (2009) also gives formulae for $\widehat{Var}(\hat{S}_{Chao2})$ when one or both of r_1 and r_2 are zero. Chao (1987) presented two methods for constructing confidence intervals based on this estimator. Assuming $\hat{S}_{Chao2}$ has an approximate normal distribution, then a symmetric 95% confidence interval is

$$\left(\hat{S}_{Chao2} - 1.96\sqrt{\widehat{Var}(\hat{S}_{Chao2})} \, , \; \hat{S}_{Chao2} + 1.96\sqrt{\widehat{Var}(\hat{S}_{Chao2})} \right) \tag{5.1}$$

Alternatively, Chao (1987) proposed an asymmetric log-linear confidence interval based on the assumption that $ln(\hat{S}_{Chao2} - S_{obs})$ is approximately normally distributed. This asymmetric log-linear confidence interval has the desirable feature that the lower

endpoint of the confidence interval will not be less than S_{obs}. The asymmetric log-linear 95% confidence interval is

$$\left(S_{obs} + (\hat{S}_{Chao2} - S_{obs})/K \ , \ S_{obs} + (\hat{S}_{Chao2} - S_{obs}) \cdot K\right), \tag{5.2}$$

where

$$K = exp\left\{1.96 \cdot \left[ln\left(1 + \frac{\widehat{Var}(\hat{S}_{Chao2})}{(\hat{S}_{Chao2} - S_{obs})^2}\right)\right]^{1/2}\right\}.$$

We examine the nominal coverage levels of both intervals in our simulation studies.

The second-order jackknife estimator (Burnham & Overton, 1979) is given by

$$\hat{S}_{Jack2} = S_{obs} + \frac{r_1(2Q-3)}{Q} - \frac{r_2(Q-2)^2}{Q(Q-1)} \ .$$

Burnham and Overton (1979) developed an estimator for the variance:

$$\begin{aligned}\widehat{Var}(\hat{S}_{Jack2}) = \ & (1 + \tfrac{2Q-3}{Q})^2 \cdot r_1 + (1 - \tfrac{(Q-2)^2}{Q(Q-1)})^2 \cdot r_2 \\ & + \textstyle\sum_{j=3}^{Q} r_j \ - \ \hat{S}_{Jack2} \ .\end{aligned}$$

They suggested constructing confidence intervals using a standard normal approximation, assuming this estimator has a small relative bias. The classic 95% confidence interval, based on $\hat{S}_{Jack2}$ and $\widehat{Var}(\hat{S}_{Jack2})$, has the same form as Equation 5.1. We also consider an asymmetric log-linear 95% confidence interval on $\hat{S}_{Jack2}$, treating $ln(\hat{S}_{Jack2} - S_{obs})$ as approximately normally distributed. The asymmetric log-linear confidence interval for $\hat{S}_{Jack2}$ is contructed in the same manner as that for $\hat{S}_{Chao2}$ in Equation 5.2.

For $\hat{S}_{ACE}$, we let n represent the total number of individuals in the sample, summed over all species. Let f_j denote the number of species that have exactly

j individuals in the sample, for $j = 1, 2, \ldots$. So, f_1 is the number of species that only contribute one individual to the sample. Chao, Ma, and Yang (1993) recommend partitioning the species into common and uncommon species based on a cut-off value τ. All species which contribute at most τ individuals to the sample are labeled uncommon species. Based on empirical analyses done by Chao, Ma, and Yang (1993), the cut-off τ is set equal to 10. Let $S_{uncommon}$ denote the number of uncommon species in the sample, and let S_{common} denote the number of species that each contribute more than τ individuals to the sample.

We use the formula for $\hat{S}_{ACE}$ intended for use in communities where the species have highly heterogeneous relative abundances, as this will be the case for the communities in Simulation Study 2 and the case study in the next chapter. Modifying Equation 2.15 in Chao and Lee (1992) to incorporate the cut-off value τ, we have

$$\hat{S}_{ACE} = S_{common} + \frac{S_{uncommon}}{\hat{C}} + \frac{f_1}{\hat{C}}\tilde{\gamma}^2 \,,$$

where $\hat{C} = 1 - f_1 / \sum_{j=1}^{\tau} f_j$, and the term $\tilde{\gamma}^2$ estimates the squared coefficient of variation of the relative abundances of the uncommon species in a community with

$$\tilde{\gamma}^2 = max \left\{ \hat{\gamma}^2 \left(1 + (1 - \hat{C}) \frac{\sum_{j=1}^{\tau} j(j-1) f_j}{\hat{C} \left(-1 + \sum_{j=1}^{\tau} j f_j \right)} \right), 0 \right\}$$

where

$$\hat{\gamma}^2 = max \left\{ \frac{S_{uncommon}}{\hat{C}} \frac{\sum_{j=1}^{\tau} j(j-1) f_j}{\left(\sum_{j=1}^{\tau} j f_j \right)^2 - 1} - 1 \,, 0 \right\}, \tag{5.3}$$

is the unadjusted estimate of the squared coefficient of variation. Chao and Lee (1992)

suggested the variance estimator

$$\widehat{Var}(\hat{S}_{ACE}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial \hat{S}_{ACE}}{\partial f_i}\frac{\partial \hat{S}_{ACE}}{\partial f_j} cov(f_i, f_j)\,, \tag{5.4}$$

where $cov(f_i, f_j) = f_i(1 - \frac{f_i}{\hat{S}_{ACE}})$ if $i = j$, and $cov(f_i, f_j) = -f_i f_j / \hat{S}_{ACE}$ if $i \neq j$. We will use $\hat{S}_{ACE}$ and $\widehat{Var}(\hat{S}_{ACE})$ to construct symmetric normal 95% confidence intervals and asymmetric log-linear 95% confidence intervals for species richness in the same manner as Equations 5.1 and 5.2.

5.4 Computing Bayesian Point Estimates and Credible Intervals of Species Richness

For each sample, we acquire our Bayesian point estimates of species richness – the posterior mean, median, and mode of S_Ω from the fit of our hierarchical Bayesian model. We run our MCMC simulation using two chains with a fixed known value for M, and dispersed initial values for the model parameters ψ, π, β_1, $\Delta\beta = (\beta_2 - \beta_1)$, and k. From our experience, we have found the autocorrelation in the MCMC chains is typically present for lags on the order of 500 to 5,000 iterations. Due to this persistent autocorrelation, the burn-in period of the MCMC simulation is taken to be at least 100,000 iterations for each sample. To help ensure the MCMC chains have approximately converged, the burn-in period continues until the Gelman and Rubin diagnostic statistic $\hat{R}$ is less than 1.1 for all monitored parameters. After the burn-in period, we use a production run of 500,000 iterations with some thinning –

keeping every 10^{th} observation in the MCMC chains to reduce the memory storage requirements. As a result, our posterior inference is based on two sets of 50,000 draws.

In some MCMC runs, the upper bound, M, in the super-community must be increased to ensure the maximum value of S_Ω is well below M. The posterior distribution of S_Ω is generally unaffected by the choice of M so long as M is sufficiently large.

From the posterior draws of an MCMC simulation, we acquire estimates of the posterior mean, median, and mode of S_Ω. We also record estimates of the posterior variance of S_Ω, the 95% highest posterior density (HPD) interval of S_Ω, and the equal-tail 95% credible interval. MCMC simulations were successfully completed on all samples from Simulation Study 1 and from Scenarios 3 and 4 in Simulation Study 2.

The MCMC simulations had problems with a handful of samples in the scenarios with the smallest sampling fractions (Scenarios 1 and 2) in Simulation Study 2. Three samples (out of 100) in Scenario 1 and five samples in Scenario 2 showed exceptionally large upward spikes in the MCMC trace plots for S_Ω with these spikes reaching rather large values. The occurrence of these extreme values in the sampler suggests a flat likelihood, with little information provided from the sparse data. These sample data sets were replaced with new simulated samples so that each scenario is based on 100 simulated samples.

5.5 Measures of Performance

To assess the performance of estimators, and to compare estimators, we focus on the bias, variance, and mean square error of the estimators computed across the simulated data sets.

Based on $N_{sim} = 100$ samples in a simulation scenario, the bias of an estimator $\hat{S}$ is estimated as

$$\widehat{bias}(\hat{S}) = \frac{1}{N_{sim}} \sum_{j=1}^{N_{sim}} \left(\hat{S}_{(j)} - S_{\Omega} \right) ,$$

where $\hat{S}_{(j)}$ is the species richness estimate based on the j^{th} sample, and S_{Ω} is the true species richness. The relative bias of an estimator in a simulation scenario is estimated as

$$\widehat{relative\, bias}(\hat{S}) = \frac{\widehat{bias}(\hat{S})}{S_{\Omega}} = \frac{1}{N_{sim}} \sum_{j=1}^{N_{sim}} \frac{\left(\hat{S}_{(j)} - S_{\Omega} \right)}{S_{\Omega}} .$$

As a measure of the precision of an estimator, within each scenario we compute the *sample variance* of the 100 estimates of species richness. For the estimators $\hat{S}_{Chao2}$, $\hat{S}_{Jack2}$, $\hat{S}_{ACE}$ and $\hat{S}_{\Omega}$, we compare this sample variance with the average of the 100 estimated variances of the estimator. For our Bayesian approach, the sample variance of the species richness estimates is not comparable to the average of the 100 posterior variance estimates.

As a combined measure of accuracy and precision, we compute the sample mean squared error (MSE). For each estimator in each scenario, we compute the sample MSE by adding the sample variance to the square of the estimated bias. In addition, we compute the coverage level of the 95% interval estimates as interval estimates of

species richness are often used in ecological management (Hwang & Shen, 2010). For each interval estimator, we report the percentage of the 100 interval estimates that include the true species richness.

5.6 Results from Simulation Study 1

Figure 5.1 displays box plots illustrating the empirical distribution of species richness estimates from the eight different estimators applied to 100 samples (eight sampling units per sample data set). The three Bayesian estimators (mean, median, and mode) have their distributions centered very close to $S_\Omega = 148$, the true number of species in the simulation. The five other estimators have most of their estimates below 148. The number, S_{obs}, of species observed and our MLE $\hat{S}_\Omega$ show the largest negative biases and the smallest ranges of values. The three non-parametric estimators have smaller negative biases than S_{obs} and $\hat{S}_\Omega$.

Figure 5.2 displays similar box plots for the simulations based on 16 sampling units per sample. We notice that with a doubling of the number of sampling units, all estimators have smaller ranges corresponding to a narrowing of their sampling distributions. In addition to the three Bayesian estimators, $\hat{S}_{Jack2}$ now has its distribution centered very close to $S_\Omega = 148$. While the other estimators (S_{obs}, $\hat{S}_{Chao2}$, $\hat{S}_{ACE}$ and $\hat{S}_\Omega$) have distributions centered below 148, these have been shifted upwards when a larger sampling effort is simulated.

Table 5.2 shows the estimated bias, variance, and MSE of the eight estimators

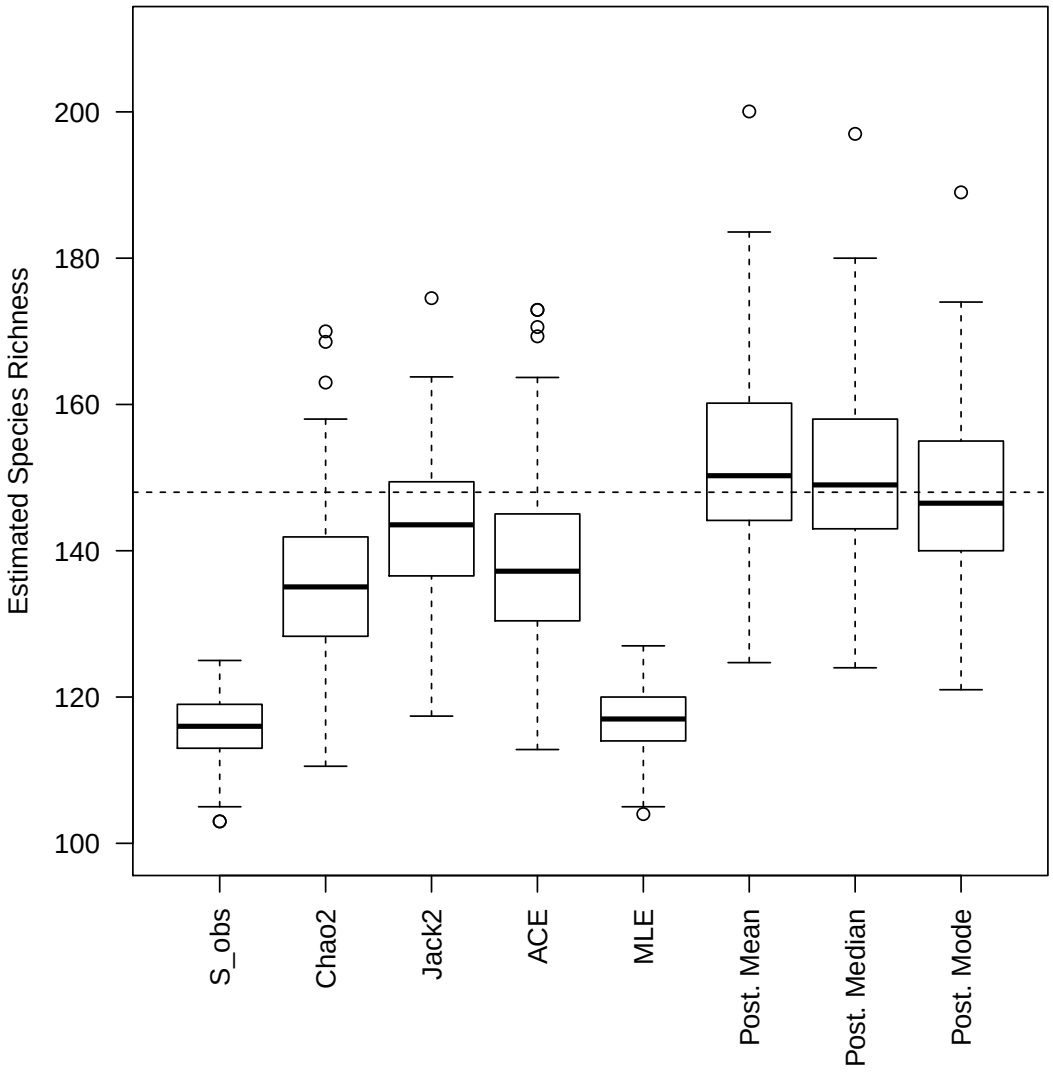

Figure 5.1: Side-by-side box plots of species richness estimates generated from samples of 8 sampling units in Simulation Study 1. The horizontal reference line is the actual species richness of 148. From left to right, the point estimators are S_{obs}, $\hat{S}_{Chao2}$, $\hat{S}_{Jack2}$, $\hat{S}_{ACE}$, MLE $\hat{S}_{\Omega}$ from Chapter 3, and the estimated posterior mean, posterior median, and posterior mode of S_{Ω} from Chapter 4.

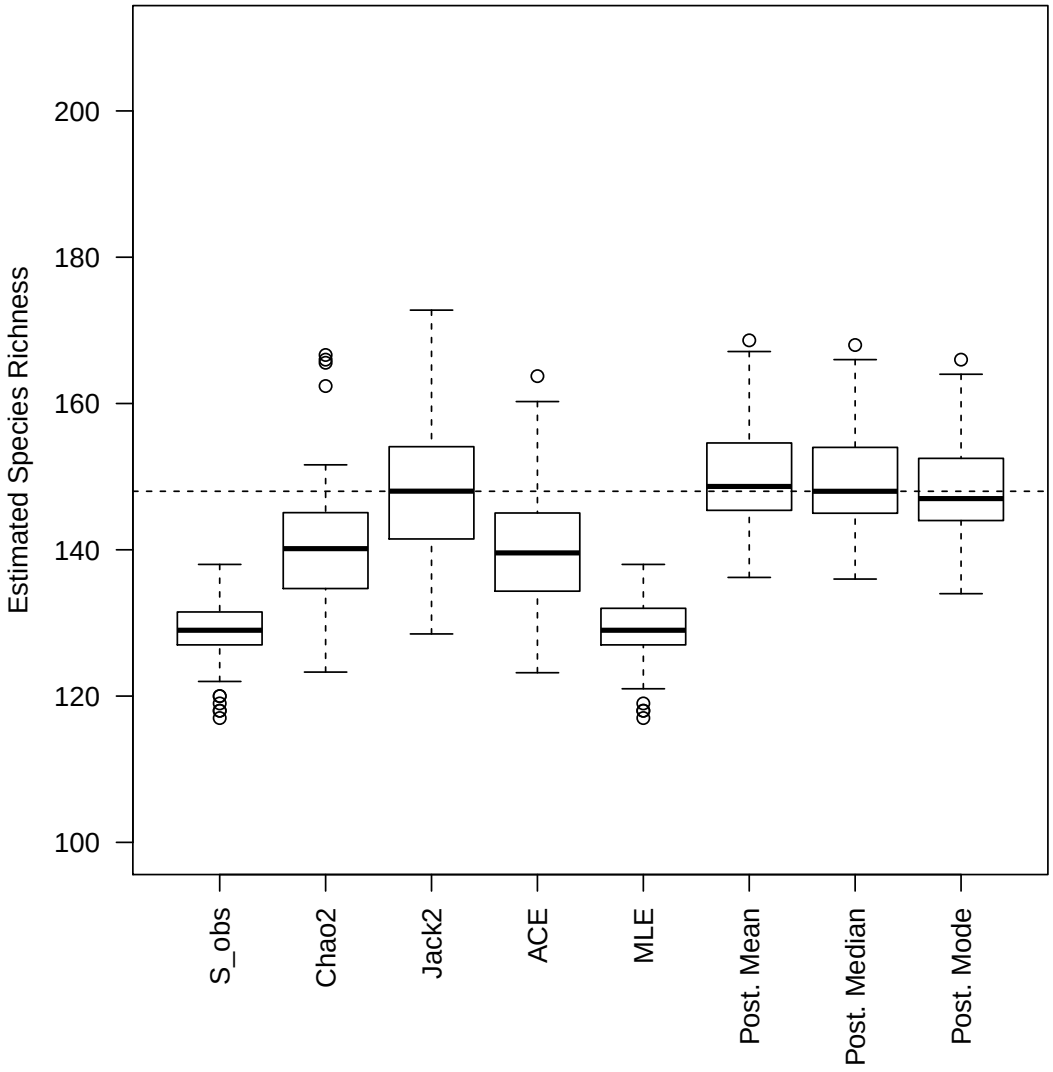

Figure 5.2: Side-by-side box plots of species richness estimates generated from samples of 16 sampling units in Simulation Study 1. The horizontal reference line is the actual species richness of 148.

applied to the two sets of 100 artificial samples. We see that the estimated posterior mode of S_Ω has the smallest bias at both levels of sampling effort. As expected, S_{obs} has the largest negative bias as a species richness estimator, but it is closely followed by the MLE $\hat{S}_\Omega$ at both levels of sampling effort. The sample variance of the estimators is smallest for S_{obs} and closely followed by the MLE $\hat{S}_\Omega$. The other six estimators have larger sample variances. The sample variances of all eight estimators decrease when the sampling effort is increased from 8 sampling units to 16 sampling units.

In Table 5.2, the second-order jackknife estimator $\hat{S}_{Jack2}$ has the smallest sample MSE when the samples use 8 sampling units, but the estimated posterior mode of S_Ω has a comparable sample MSE that is only 2.62% higher. When the simulations are based on 16 sampling units per sample, the posterior mode has the smallest sample MSE, and the posterior median and posterior mean follow; whereas, $\hat{S}_{Jack2}$ displays relatively poorer performance with 16 sampling units (relative increase in MSE is approximately 71% compared with the posterior mode).

Within each level of sampling effort, there is no single estimator that outperforms the other estimators in every performance measure. However, the posterior mode is quite competitive in terms of bias and sample MSE. The number, S_{obs}, of species observed and the MLE $\hat{S}_\Omega$ do have the smallest sample variances among the estimators; but, their estimates are always negatively biased, and their sample MSEs are more than three times larger than the sample MSEs of every other estimator. Among the Bayesian estimators, the posterior mode consistently has the smallest estimated bias

Table 5.2: Performance of Point Estimators in Simulation Study 1

Sampling Effort	Point Estimator	Sample Bias	Relative Difference from Smallest	Sample Variance	Relative Difference from Smallest	Sample MSE	Relative Difference from Smallest
8 sampling units	S_{obs}	-32.26	4,863.08%	**23.85**		1,064.56	710.97%
	$\hat{S}_{Chao2}$	-12.67	1,849.23%	128.94	440.63%	289.37	120.44%
	$\hat{S}_{Jack2}$	-4.99	667.69%	106.40	346.12%	**131.27**	
	$\hat{S}_{ACE}$	-9.73	1,396.92%	155.31	551.19%	250.01	90.45%
	MLE $\hat{S}_{\Omega}$	-30.99	4,667.69%	24.94	4.57%	985.32	650.61%
	Post. Mean	3.91	-701.54%	159.03	566.79%	174.30	32.78%
	Post. Median	2.34	-460.00%	150.41	530.65%	155.88	18.75%
	Post. Mode	**-0.65**		134.29	463.06%	134.71	2.62%
16 sampling units	S_{obs}	-19.31	-14,953.85%	**18.20**		391.07	698.10%
	$\hat{S}_{Chao2}$	-7.53	-5,892.31%	65.82	261.65%	122.49	149.98%
	$\hat{S}_{Jack2}$	-0.32	-346.15%	83.60	359.34%	83.70	70.82%
	$\hat{S}_{ACE}$	-7.56	-5,915.38%	56.86	212.42%	114.00	132.65%
	MLE $\hat{S}_{\Omega}$	-19.04	-14,746.15%	18.71	2.80%	381.23	678.02%
	Post. Mean	2.01	1,446.15%	52.83	190.27%	56.85	16.02%
	Post. Median	1.37	953.85%	51.14	180.99%	53.02	8.20%
	Post. Mode	**0.13**		48.98	169.12%	**49.00**	

and smallest sample MSE at both levels of sampling effort.

In Table 5.3, for a given level of sampling effort, we can compare the sample variance of the 100 estimates from a species richness estimator with the average of the 100 corresponding variance estimates. The relative difference in the rightmost column of Table 5.3 is computed by taking the difference between the sample variance (of the 100 estimates of species richness) and the average of the variance estimates and then dividing that difference by the sample variance of the 100 estimates of species richness. Ideally, the average of the 100 variance estimates will be close to the sample variance of the 100 species richness estimates and the relative difference will be close to zero.

Table 5.3: Performance of Variance Estimators for Simulation Study 1

Sampling Effort	Point Estimator	Sample Variance	Average of Variance Estimates	Relative Difference
8 sampling units	$\hat{S}_{Chao2}$	128.94	136.75	6.05%
	$\hat{S}_{Jack2}$	106.40	92.51	-13.05%
	$\hat{S}_{ACE}$	155.31	173.37	11.63%
	MLE $\hat{S}_{\Omega}$	24.94	1.88	-92.45%
16 sampling units	$\hat{S}_{Chao2}$	65.82	69.96	6.29%
	$\hat{S}_{Jack2}$	83.60	81.54	-2.46%
	$\hat{S}_{ACE}$	56.86	52.11	-8.37%
	MLE $\hat{S}_{\Omega}$	18.71	0.93	-95.04%

For the three non-parametric estimators ($\hat{S}_{Chao2}$, $\hat{S}_{Jack2}$ and $\hat{S}_{ACE}$), the average of the 100 variance estimates is within 13.05% of the corresponding sample variance of the 100 species richness estimates, when the samples are constructed from eight

sampling units. When the samples consist of observations from sixteen sampling units, the relative difference between the average of the variance estimates and the sample variance of the species richness estimates is less than 8.5% for the non-parametric estimators.

From Table 5.3, we see that the average of the asymptotic variance estimates associated with the MLE $\hat{S}_\Omega$ is much smaller than the sample variance of the $\hat{S}_\Omega$ estimates. This strong under-estimation of the variance of $\hat{S}_\Omega$ is present at both levels of sampling effort. In work not shown here, we applied the MLE $\hat{S}_\Omega$ and its asymptotic variance estimator to small artificial samples created from the same model that the MLE is based upon; satisfactory agreement was achieved between the sample variance of the $\hat{S}_\Omega$ estimates and the average of the asymptotic variance estimates. This suggests the poor performance of the asymptotic variance estimator for $\hat{S}_\Omega$ in Table 5.3 is primarily the result of model misspecification[1].

Table 5.4 presents the coverage probability of interval estimates. While the Bayesian intervals are not constructed to achieve 95% coverage in repeated samples (they are instead constructed based on posterior probability for a given data set) we see that these intervals do achieve a coverage level quite close to 95%, for both levels of sampling effort. This is not unexpected as Bayesian procedures typically exhibit good frequentist performance (see, e.g., Carlin & Louis, 2009).

[1]The MLE is based upon a negative binomial mixture model with an exponential distribution on $\boldsymbol{\mu}$; whereas, the sample data sets in Simulation Study 1 are simulated from a negative binomial mixture model with a two-component exponential mixture distribution on $\boldsymbol{\mu}$.

Table 5.4: Performance of 95% Interval Estimators for Simulation Study 1

	Coverage Percentage	
95% Interval Estimator	8 sampling units	16 sampling units
Chao2 Symmetric CI	63%	67%
Chao2 Asymmetric CI	84%	87%
Jack2 Symmetric CI	87%	92%
Jack2 Asymmetric CI	94%	97%
ACE Symmetric CI	73%	64%
ACE Asymmetric CI	91%	84%
ML Symmetric CI	0%	0%
ML Asymmetric CI	0%	0%
Likelihood Interval	0%	0%
Bayesian HPD Interval	94%	95%
Bayesian Equal-Tail CI	95%	97%

The asymmetric confidence interval constructed from $\hat{S}_{Jack2}$ also has coverage percentages close to the nominal level for the two levels of sampling effort. The symmetric confidence interval associated with $\hat{S}_{Jack2}$ has smaller coverage levels that are below the nominal 95% level. For $\hat{S}_{Chao2}$ and $\hat{S}_{ACE}$, the asymmetric confidence intervals have higher coverage percentages than their corresponding symmetric confidence intervals, though the coverage percentages are all below the nominal 95% level. The asymmetric confidence intervals have higher coverage levels because the non-parametric estimators tend to be negatively biased and the asymmetric confidence intervals tend to have higher upper limits than their corresponding symmetric confidence intervals.

For the maximum likelihood approach with an exponential distribution on $\boldsymbol{\mu}$, the

three 95% interval estimators[2] perform very poorly (Table 5.4). At both levels of sampling effort, all of the interval estimates have upper limits that are below the true S_Ω. As a result, these three interval estimators have coverage levels of 0%.

For the model-based samples generated in Simulation Study 1, the Bayesian estimators have performed well, with the posterior mode of S_Ω having the smallest estimated bias and sample MSE. The highest posterior density interval and the equal-tail credible interval have coverage levels that are very close to the nominal 95% level, so that these Bayesian intervals also exhibit good frequentist performance in this case.

5.7 Results from Simulation Study 2

In the preceding section, we simulated data under the ideal scenario for the Bayesian estimators, based on the model underlying these estimators. Here, we examine how the estimators fare when the data sets are simulated under a more realistic setting.

We consider four sampling scenarios in this simulation study, with the sampling scenarios differing in the quantity and size of the quadrats randomly placed in the 50-hectare tropical forest plot, for which we have a census.

Figures 5.3, 5.4, 5.5 and 5.6 display side-by-side box plots comparing sampling distributions of the eight point estimators for each of the four scenarios, respectively. Aside from two outliers, all of the species richness estimates summarized in the box plots are less than the actual number of species ($S_\Omega = 305$) in the 50-hectare plot.

[2]Using the MLE of S_Ω and the estimated asymptotic variance, symmetric and asymmetric 95% confidence intervals for S_Ω are constructed in the same manner as Equations 5.1 and 5.2, respectively.

This reflects the difficulty of the species richness estimation problem; all methods, including those estimators well established in the literature, perform poorly with all eight estimators being negatively biased. The negative bias decreases for each estimator when the sampling fraction increases (i.e., moving from Figures 5.3 and 5.4 to Figures 5.5 and 5.6).

Tables 5.5 and 5.6 show the estimated bias, variance and MSE of the eight estimators applied to the four sets of 100 samples in Simulation Study 2. As an estimator of species richness, S_{obs} has the largest estimated bias and the largest sample MSE in each of the four sampling scenarios, based on Figures 5.3 to 5.6 and Tables 5.5 and 5.6. The MLE $\hat{S}_{\Omega}$ follows next with the second largest bias and the second largest sample MSE. The estimators S_{obs} and $\hat{S}_{\Omega}$ have the smallest ranges and the smallest sample variances in every scenario. This mimics the observations of the previous simulation study.

In Scenarios 1 and 2 (i.e., samples that occupy 0.2% of the BCI plot), the posterior mean of S_{Ω} has the smallest bias and the smallest MSE. The posterior median and posterior mode of S_{Ω} follow closely behind, in terms of bias and MSE in Scenarios 1 and 2. Compared with the posterior mean, the non-parametric estimators have relative increases in MSE of at least 16% in Scenario 1 and at least 31% in Scenario 2.

The relative performances of the estimators change when the sampling fraction is increased. In Scenarios 3 and 4 (i.e., samples that occupy 0.8% of the BCI plot), $\hat{S}_{Jack2}$ has the smallest bias and the smallest sample MSE. In Scenario 3, $\hat{S}_{ACE}$ and

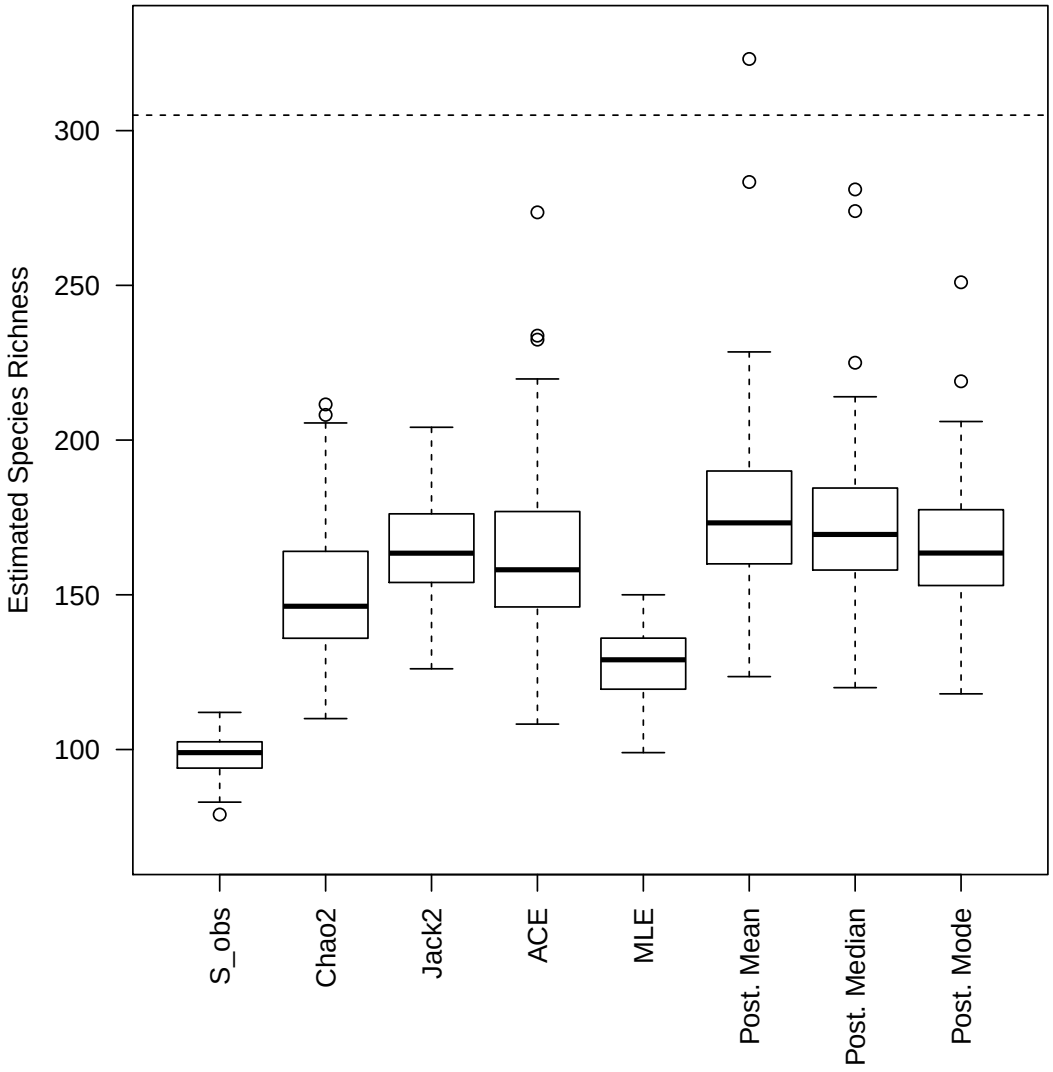

Figure 5.3: Side-by-side box plots of species richness estimates generated from samples of forty $5\,m \times 5\,m$ quadrats in Simulation Study 2. The horizontal reference line is the actual species richness of 305.

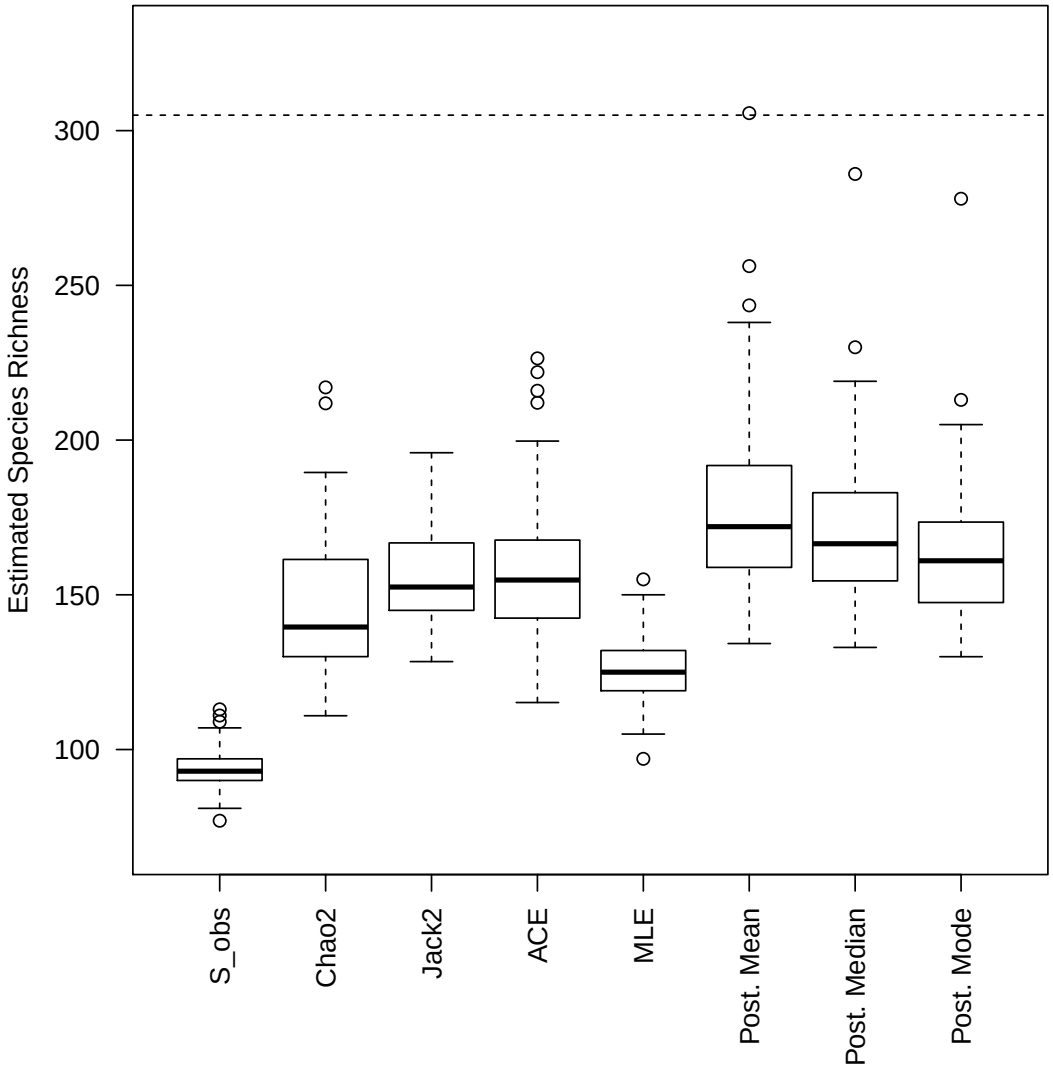

Figure 5.4: Side-by-side box plots of species richness estimates generated from samples of ten $10\,m \times 10\,m$ quadrats in Simulation Study 2. The horizontal reference line is the actual species richness of 305.

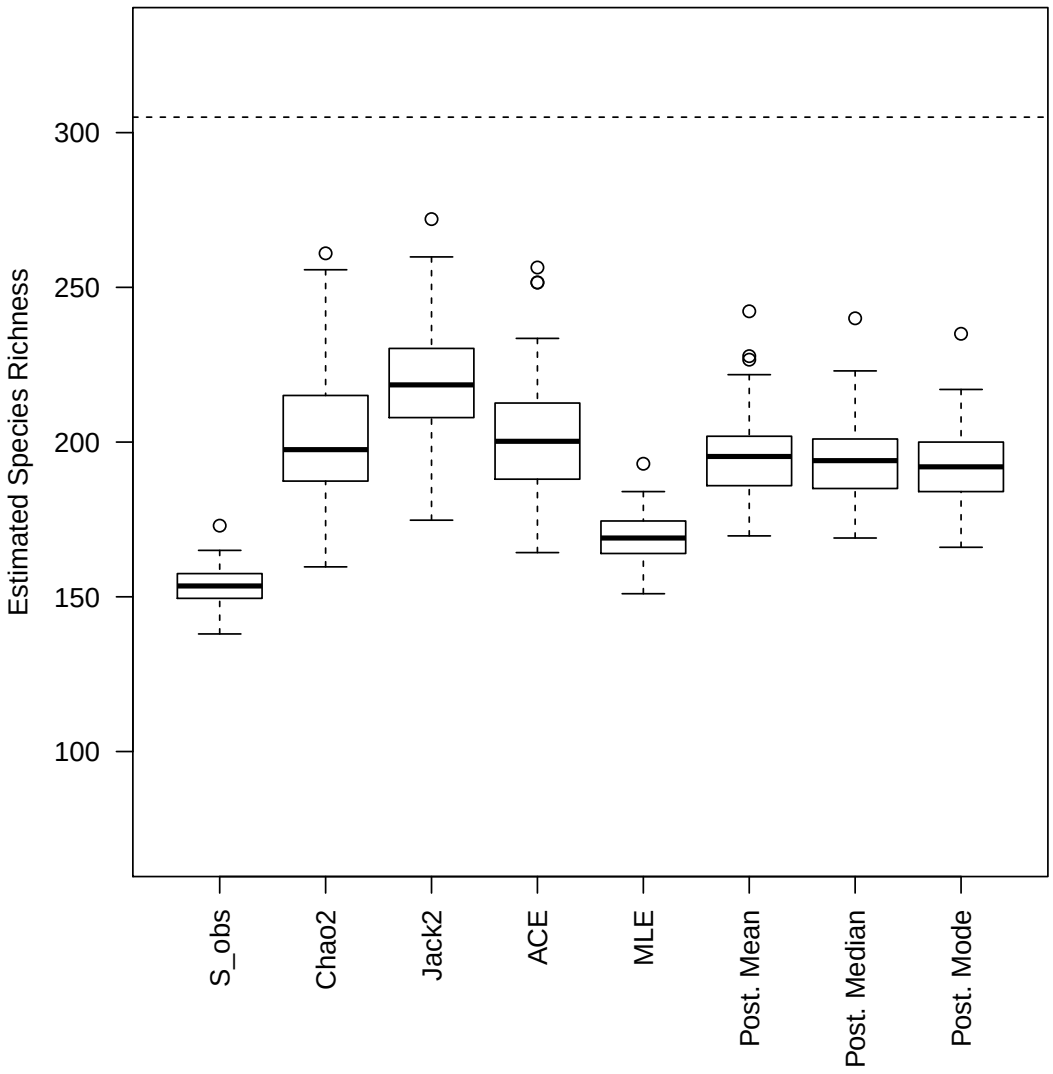

Figure 5.5: Side-by-side box plots of species richness estimates generated from samples of forty $10\,m \times 10\,m$ quadrats in Simulation Study 2. The horizontal reference line is the actual species richness of 305.

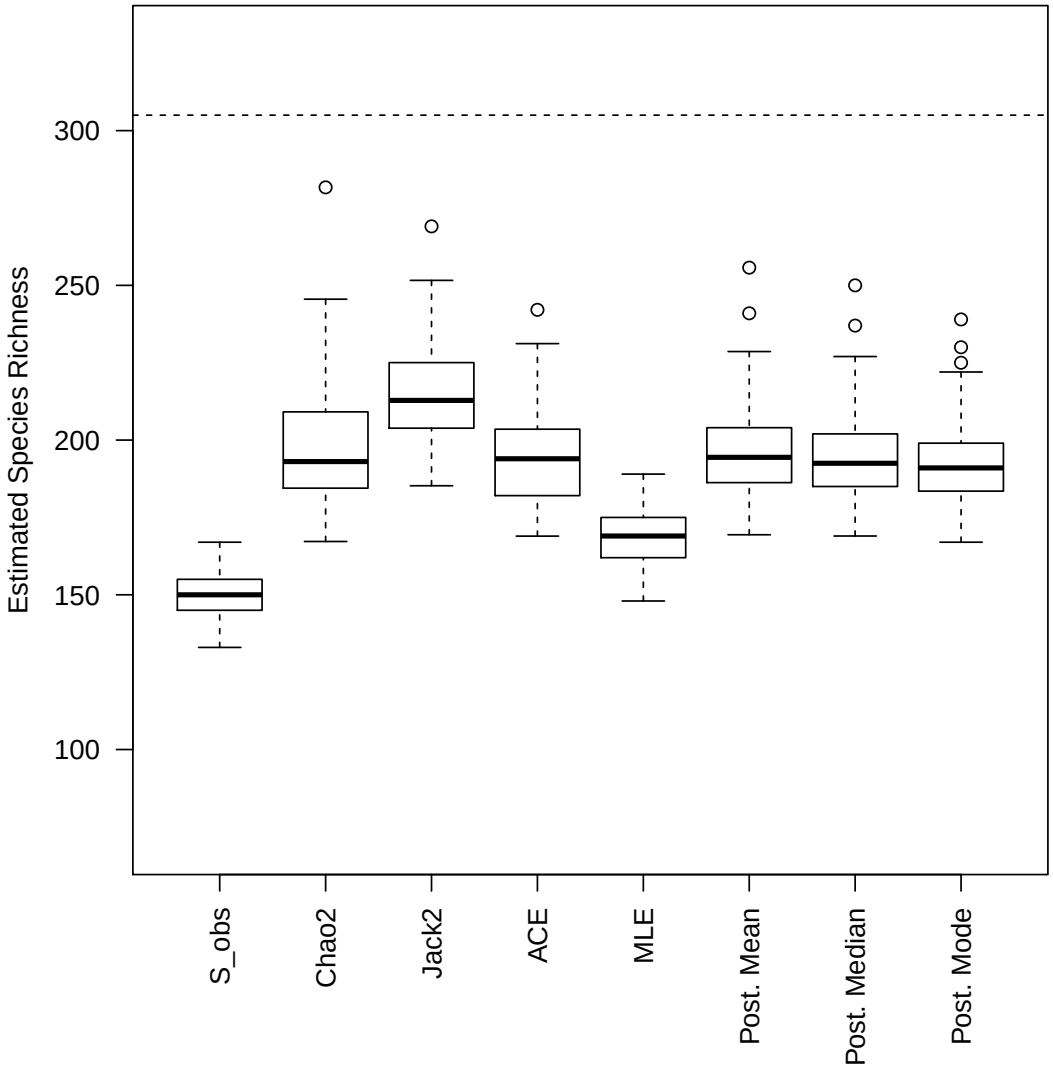

Figure 5.6: Side-by-side box plots of species richness estimates generated from samples of ten $20\,m \times 20\,m$ quadrats in Simulation Study 2. The horizontal reference line is the actual species richness of 305.

$\hat{S}_{Chao2}$ are the second and third best estimators, respectively, in terms of bias and sample MSE. In Scenario 4, $\hat{S}_{Chao2}$ has the second smallest bias and sample MSE; the performances of the three Bayesian estimators and $\hat{S}_{ACE}$ are close to $\hat{S}_{Chao2}$, in terms of bias and sample MSE.

When only considering the three Bayesian estimators, their relative performances are consistent across the four sampling scenarios. The posterior mean of S_Ω has the smallest bias and smallest sample MSE. The posterior mode of S_Ω has the smallest sample variance. The shape of the estimated posterior distribution of S_Ω helps explain the relatively better performance of the posterior mean. From examining the MCMC simulations, the posterior distribution of S_Ω is typically unimodal and positively skewed. The posterior mean of S_Ω is consistently larger than the posterior median and posterior mode. Therefore, keeping in mind that all estimators are negatively biased in Simulation Study 2, the posterior mean will have a smaller bias and a smaller MSE than the posterior median and posterior mode.

To examine the performance of estimators when the sampling fraction is held fixed but the quantity and size of the sampling units changes, we compare results from Scenarios 1 and 2. The initial inspection of the box plots in Figures 5.3 and 5.4 does not reveal notable differences in estimator performance across these scenarios. Table 5.5 indicates that the bias of all estimators are slightly smaller in Scenario 1 (samples consist of observations from forty $5\,m \times 5\,m$ quadrats). With the exception of the posterior mode, the estimators have smaller sample variance in Scenario 2 (samples consist of observations from ten $10\,m \times 10\,m$ quadrats); and, with the exception of

Table 5.5: Performance of Point Estimators in Scenarios 1 & 2 of Simulation Study 2

Sampling Effort	Point Estimator	Sample Bias	Relative Difference from Smallest	Sample Variance	Relative Difference from Smallest	Sample MSE	Relative Difference from Smallest
Scenario 1	S_{obs}	-206.77	60.29%	**44.18**		42,798.01	145.58%
Forty $5 \times 5\,m$	$\hat{S}_{Chao2}$	-154.01	19.39%	539.33	1,120.76%	24,257.62	39.20%
quadrats	$\hat{S}_{Jack2}$	-141.23	9.48%	285.51	546.24%	20,230.02	16.08%
	$\hat{S}_{ACE}$	-141.84	9.95%	801.92	1,715.12%	20,921.30	20.05%
	MLE $\hat{S}_{\Omega}$	-176.87	37.11%	117.25	165.39%	31,400.24	80.18%
	Post. Mean	**-129.00**		784.72	1,676.19%	**17,427.00**	
	Post. Median	-132.84	2.98%	628.44	1,322.45%	18,274.90	4.87%
	Post. Mode	-138.62	7.46%	439.79	895.45%	19,655.30	12.79%
Scenario 2	S_{obs}	-211.60	66.50%	**41.17**		44,815.73	165.19%
Ten $10 \times 10\,m$	$\hat{S}_{Chao2}$	-159.34	25.38%	422.97	927.37%	25,811.75	52.73%
quadrats	$\hat{S}_{Jack2}$	-149.84	17.90%	209.46	408.77%	22,660.55	34.09%
	$\hat{S}_{ACE}$	-147.29	15.89%	503.18	1,122.20%	22,196.69	31.34%
	MLE $\hat{S}_{\Omega}$	-179.49	41.23%	111.93	171.87%	32,328.59	91.30%
	Post. Mean	**-127.09**		747.41	1,715.42%	**16,899.70**	
	Post. Median	-134.41	5.76%	538.93	1,209.04%	18,604.98	10.09%
	Post. Mode	-142.21	11.90%	449.00	990.60%	20,672.68	22.33%

Table 5.6: Performance of Point Estimators in Scenarios 3 & 4 of Simulation Study 2

Sampling Effort	Point Estimator	Sample Bias	Relative Difference from Smallest	Sample Variance	Relative Difference from Smallest	Sample MSE	Relative Difference from Smallest
Scenario 3	S_{obs}	-151.64	76.67%	**38.86**		23,033.55	200.22%
Forty $10 \times 10\ m$	$\hat{S}_{Chao2}$	-103.86	21.01%	446.25	1048.35%	11,233.59	46.42%
quadrats	$\hat{S}_{Jack2}$	**-85.83**		305.70	686.67%	**7,672.20**	
	$\hat{S}_{ACE}$	-102.99	19.99%	332.26	755.02%	10,939.40	42.58%
	MLE $\hat{S}_{\Omega}$	-135.50	57.87%	57.10	46.94%	18,417.35	140.05%
	Post. Mean	-109.63	27.73%	174.58	349.25%	12,193.18	58.93%
	Post. Median	-110.85	29.15%	161.54	315.70%	12,449.27	62.26%
	Post. Mode	-112.74	31.35%	145.22	273.70%	12,855.53	67.56%
Scenario 4	S_{obs}	-154.96	70.68%	**52.06**		24,064.66	184.00%
Ten $20 \times 20\ m$	$\hat{S}_{Chao2}$	-107.28	18.16%	365.70	602.46%	11,875.39	40.15%
quadrats	$\hat{S}_{Jack2}$	**-90.79**		230.83	343.39%	**8,473.37**	
	$\hat{S}_{ACE}$	-110.78	22.02%	198.43	281.16%	12,470.21	47.17%
	MLE $\hat{S}_{\Omega}$	-136.34	50.17%	79.36	52.44%	18,667.95	120.31%
	Post. Mean	-108.89	19.94%	197.93	280.20%	12,055.40	42.27%
	Post. Median	-110.44	21.64%	184.43	254.26%	12,381.42	46.12%
	Post. Mode	-112.93	24.39%	162.71	212.54%	12,915.90	52.43%

the posterior mean, estimators have smaller sample MSEs in Scenario 1.

We also compare the estimator performance between Scenarios 3 and 4. The box plots in Figures 5.5 and 5.6 do not reveal notable differences in estimator performance between Scenarios 3 and 4. In Table 5.6, the bias and sample MSE of the estimators are smaller in Scenario 3 (samples consist of observations from forty $10\,m \times 10\,m$ quadrats) than the corresponding performance measures for Scenario 4, except for the posterior mean and posterior median. The sample variance of the non-parametric estimators $\hat{S}_{Chao2}$, $\hat{S}_{Jack2}$ and $\hat{S}_{ACE}$ are smaller in Scenario 4 (samples consist of ten $20\,m \times 20\,m$ quadrats), while the sample variance of the other five estimators are smaller in Scenario 3.

In summary, for a fixed sampling fraction of the 50-hectare BCI plot (i.e., 0.2% in Scenarios 1 and 2, and 0.8% in Scenarios 3 and 4), estimators applied to samples constructed from forty small quadrats tend to have slightly smaller bias and smaller MSE but larger variance than the same estimators when applied to samples constructed from ten larger quadrats. If these observations were to extend to other settings, investigators must balance the additional travel required to access more quadrats of smaller size with the potential reductions in bias and MSE.

Table 5.7 displays the sample variance of the species richness estimates from 100 samples and the average of the corresponding variance estimates of the species richness estimators. For the non-parametric estimators $\hat{S}_{Chao2}$, $\hat{S}_{Jack2}$ and $\hat{S}_{ACE}$, we observe wider ranges of relative differences between the sample variance of the species richness estimates and the average of the variance estimates than in the first simulation study.

In each sampling scenario, the average of the 100 asymptotic variance estimates associated with the MLE $\hat{S}_\Omega$ is at least twice as small as the sample variance of the $\hat{S}_\Omega$ estimates. As a result, confidence intervals constructed from our likelihood approach will be too narrow and will not be reliable when applied to the sample data sets in Simulation Study 2.

Table 5.7: Performance of Variance Estimators for Simulation Study 2

Sampling Scenario	Point Estimator	Sample Variance	Average of Variance Estimates	Relative Difference
Scenario 1	$\hat{S}_{Chao2}$	539.33	537.59	-0.32%
Forty $5 \times 5\, m$	$\hat{S}_{Jack2}$	285.51	243.31	-14.78%
Quadrats	$\hat{S}_{ACE}$	801.92	745.08	-7.09%
	MLE $\hat{S}_\Omega$	117.25	39.77	-66.08%
Scenario 2	$\hat{S}_{Chao2}$	422.97	478.81	13.20%
Ten $10 \times 10\, m$	$\hat{S}_{Jack2}$	209.46	201.31	-3.89%
Quadrats	$\hat{S}_{ACE}$	503.18	727.23	44.53%
	MLE $\hat{S}_\Omega$	111.93	44.08	-60.62%
Scenario 3	$\hat{S}_{Chao2}$	446.25	395.97	-11.27%
Forty $10 \times 10\, m$	$\hat{S}_{Jack2}$	305.70	253.53	-17.06%
Quadrats	$\hat{S}_{ACE}$	332.26	321.94	-3.11%
	MLE $\hat{S}_\Omega$	57.10	18.38	-67.81%
Scenario 4	$\hat{S}_{Chao2}$	365.70	352.43	-3.63%
Ten $20 \times 20\, m$	$\hat{S}_{Jack2}$	230.83	218.24	-5.45%
Quadrats	$\hat{S}_{ACE}$	198.43	269.63	35.88%
	MLE $\hat{S}_\Omega$	79.36	21.54	-72.85%

Table 5.8 displays the coverage levels of the interval estimators of species richness for the four sampling scenarios in Simulation Study 2. Unfortunately, all of

the interval estimators have very poor coverage levels with most interval estimates having upper endpoints below the true species richness of the 50-hectare plot. No interval estimators have coverage levels near the nominal 95% level. In fact, the highest observed coverage level is 24% (for the Bayesian equal-tail confidence interval in Scenario 2). The three interval estimators from our likelihood-based approach have coverage levels of 0% in all four scenarios. The coverage levels of the Bayesian interval estimators decrease when the sampling effort increases (either by increasing the number of quadrats, or by increasing the size of the quadrats); in the scenarios with the lowest sampling effort, the sparser data causes flatter likelihoods and, as a result, wider interval estimates are produced that are more likely to contain S_Ω. While an unfortunate result, it should be noted that all methods considered, including well established methods from the literature, perform poorly here. This reflects the difficulties of conducting inference on S_Ω in applications.

5.8 Comparing Results of the Simulation Studies

To compare the performance of the species richness estimators in the two simulation studies, we can look at the relative bias of the estimators. Table 5.9 presents the relative biases of the estimators in the different sampling scenarios for the two simulation studies. We see the relative bias of each estimator decreases when the sampling effort is increased in a simulation study. For each species richness estimator, its relative bias in the two scenarios of Simulation Study 1 are much less than the relative biases

Table 5.8: Performance of 95% Interval Estimators for Simulation Study 2

	Coverage Percentage			
95% Interval Estimator	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Chao2 Symmetric CI	0%	1%	6%	2%
Chao2 Asymmetric CI	6%	2%	8%	6%
Jack2 Symmetric CI	0%	0%	1%	0%
Jack2 Asymmetric CI	0%	0%	1%	1%
ACE Symmetric CI	3%	2%	3%	0%
ACE Asymmetric CI	13%	6%	3%	1%
ML Symmetric CI	0%	0%	0%	0%
ML Asymmetric CI	0%	0%	0%	0%
Likelihood Interval	0%	0%	0%	0%
Bayesian HPD Interval	4%	12%	0%	1%
Bayesian Equal-Tail CI	7%	24%	0%	1%

in the more realistic scenarios of Simulation Study 2.

For each column of Table 5.9, the smallest relative bias is highlighted in bold. In Simulation Study 1, the very small relative biases of the Bayesian estimators are to be expected as the samples are generated from our hierarchical Bayesian model. The second order jackknife estimator also has small relative biases in Simulation Study 1. In Simulation Study 2, all estimators in all four scenarios have negative relative biases. For samples from Scenarios 1 and 2 of Simulation Study 2, the Bayesian posterior estimators have the smallest relative biases. For the samples involving a larger sampling fraction in Scenarios 3 and 4 of Simulation Study 2, $\hat{S}_{Jack2}$ has the smallest relative bias.

Table 5.9: Relative Bias of Point Estimators in the Two Simulation Studies

Point	Simulation Study 1		Simulation Study 2			
Estimator	8 Sampling Units	16 Sampling Units	Scenario 1	Scenario 2	Scenario 3	Scenario 4
S_{obs}	-21.80%	-13.05%	-67.79%	-69.38%	-49.72%	-50.81%
$\hat{S}_{Chao2}$	-8.56%	-5.09%	-50.49%	-52.24%	-34.05%	-35.17%
$\hat{S}_{Jack2}$	-3.37%	-0.22%	-46.30%	-49.13%	**-28.14%**	**-29.77%**
$\hat{S}_{ACE}$	-6.58%	-5.11%	-46.51%	-48.29%	-33.77%	-36.32%
MLE $\hat{S}_{\Omega}$	-20.94%	-12.86%	-57.99%	-58.85%	-44.43%	-44.70%
Post. Mean	2.64%	1.35%	**-42.30%**	**-41.67%**	-35.94%	-35.70%
Post. Median	1.58%	0.93%	-43.55%	-44.07%	-36.34%	-36.21%
Post. Mode	**-0.44%**	**0.09%**	-45.45%	-46.63%	-36.96%	-37.03%

5.9 Remarks

The Bayesian estimators perform well under the ideal conditions of Simulation Study 1. These estimators have relatively small bias and small MSEs, in particular, the posterior mode. The second-order jackknife estimator $\hat{S}_{Jack2}$ is also competitive in Simulation Study 1. The two Bayesian 95% interval estimators and the asymmetric 95% confidence interval associated with $\hat{S}_{Jack2}$ achieve coverage levels very close to the nominal level. In both simulation studies, S_{obs} and the MLE $\hat{S}_{\Omega}$ consistently have the largest bias and the largest MSEs.

With the realistic samples of Simulation Study 2, the best overall estimator of the eight estimators examined depends on the amount of sampling effort. For the small samples in Scenarios 1 and 2, the posterior mean of S_{Ω} had the smallest bias and smallest MSE. Increasing the sampling effort by a factor of four, $\hat{S}_{Jack2}$ had the smallest bias and smallest sample MSE in Scenarios 3 and 4.

Based on these results, we recommend the use of our Bayesian method in settings where the sampling units represent a small fraction of the region of interest. In addition, we recommend expanding the Bayesian model to consider alternative mixing distributions on $\boldsymbol{\mu}$, in particular, distributions such as a three-component exponential mixture or a lognormal distribution that place more support near zero; such distributions would result in larger estimates of species richness (i.e., possibly reducing bias of the estimator and increasing the coverage level of corresponding interval estimates). For our likelihood-based method, changing the mixing distribution on $\boldsymbol{\mu}$ accordingly

would achieve similar outcomes.

Chapter 6

Case Study: Oribatid Mites in a Tropical Forest

6.1 Introduction

In 2003 and 2004, a large-scale investigation of arthropod diversity was conducted in the lowland tropical rainforest of the San Lorenzo Protected Area (SLPA) of Panama (Basset *et al.*, 2007). The main goal of this study was to estimate the arthropod species richness in a tropical rainforest. As arthropods make up a majority of eukaryotic organisms (Basset *et al.*, 2007), an estimate of arthropod species richness is useful in determining how much more sampling is required to inventory the species diversity of the tropical rainforest and make conclusions about arthropod species richness at a global spatial scale.

As part of this large-scale investigation of arthropods, oribatid mites were collected

to look for diversity patterns between the forest floor and canopy habitats in the SLPA lowland tropical rainforest. As part of the full spectrum of arthropods sampled in this international project, oribatid mite estimates can then be used in the estimation of the total species richness of arthropods in a tropical rainforest. Here, we analyse the oribatid mite data set and estimate the species richness of the oribatid mites for two forest habitats, the forest canopy and the forest floor.

6.2 Sampling Oribatid Mites in a Lowland Tropical Rainforest

Here, we briefly introduce the oribatid mites and describe the lowland tropical rainforest of the San Lorenzo Protection Area in the Colón Province of the Republic of Panama. We also describe how the field study of oribatid mite diversity was conducted.

Oribatid mites are an order of microarthropods in the class Arachnida and subclass Acari. They tend to be the numerically-dominant microarthropod fauna in most forest floor habitats (Petersen & Luxton, 1982) and forest canopy habitats (Behan-Pelletier & Walter, 2000). They occur on many parts of the tree surface including the bark, leaves, and suspended soil mats (Lindo & Winchester, 2007). Oribatid mites feed on living or dead plant and fungal material, and their feeding activity is important in the decomposition process in soils. Adult oribatid mites range in size from 0.2 to 1.4 mm in length, and are considered important functional components

of forest ecosystems (Lindo & Winchester, 2006).

The San Lorenzo Protected Area is a contiguous lowland evergreen forest in Panama, some 6,000 hectares in area. The rainforest has an average of 3676 woody plant stems (greater than 10 mm in diameter at breast height) per hectare with trees reaching heights up to 45 metres (Basset *et al.*, 2007). The climate is wet and warm all year round with an average rainfall of 3139 mm, and the average annual temperature was 26 degrees Celsius over the years 1998 to 2002 (Basset *et al.*, 2007).

The field study took place at eight sites in the SLPA in September and October of 2003. The eight sites were chosen to represent the variety of the forest environment while at the same time being relatively accessible to field observers (e.g., close to access roads). As a result, the sites were not randomly chosen, with seven sites on the slopes of a hill (altitude 130 metres) and one site on a floodplain. The sites measured 20 m $\times$ 20 m in area. Each site was located within two kilometres of all other sites. For more information on the SLPA and the study sites, we refer the reader to Basset *et al.* (2007).

At each of the eight sites, three trees large enough to safely climb were randomly chosen. On the ground around each selected tree, eight cores of soil and ground litter were collected – two replicate cores were collected at each cardinal direction (N, W, S, E) from the tree trunk. In addition, eight tree bark scrapings were collected at different heights on each selected tree. The heights of the tree bark scrapings varied from tree to tree. The average height of the tree bark scrapings was 20.1 m from the ground (minimum 5 m, maximum 32 m, standard deviation 4.9 m). The

average height of the tree bark scrapings was significantly different among the 24 trees (ANOVA F test = 5.90, df = 23 and 168, p-value < 0.001). The three trees with the highest average core heights differed significantly from the three trees with the lowest averages, while the 18 other trees as a group did not differ significantly in their average core heights (Tukey's method, family error rate of 5%). Each core of soil or tree bark scraping measured 3×5 cm in area. A total of 192 cores from the forest floor and 192 cores from the forest canopy were collected.

Through subsequent laboratory work, the weight of the fresh substrate in each of the cores was recorded. The microarthropods were then extracted from the substrate over a 48-hour period using a Berlese-Tullgren apparatus and stored in 75% ethanol. After the substrate from each core was dried, its weight was again measured. From each core extraction, all adult oribatid mites were identified to their species type under microscope, if possible, and counted. As common in the taxonomic study of arthropods (Basset *et al.*, 2007), mites in pre-adult stages of development were not recorded because of difficulties with identification.

6.3 Exploratory Data Analysis

For each species observed, the number of adult mites extracted from the substrate was recorded for each core. A total of 16,124 adult oribatid mites were collected from the 384 cores. A total of 141 distinct species were identified from the individual specimens. In the family *Galumnidae*, 1021 adult mites were collected, but these

individuals could not be identified to genus and species type because no taxonomist familiar with *Galumnidae* was available. As a result, this family is excluded from the analysis, except where noted.

The distribution of abundances of species in the data set is quite uneven. For fifteen species, only one adult mite is observed in the field study. The 71 species with the lowest abundances constitute a combined percentage of 3.4% of the 15,103 mites identified to species types. While the two most abundant species (*Scheloribatidae Scheloribates species 1* and *Haplozetidae Rostrozetes ovulum*) are represented by 12.7% and 13.0%, respectively, of the 15,103 identified adult mites.

6.3.1 Association between numbers of mites and weights of the substrate

During the field study, the substrate cores were approximately the same dimensions (3×5 cm). However, inspection of the data set reveals that the number of adult oribatid mites extracted from a core varies from 0 to 457 mites with a standard deviation of 51.75. The wet weights of the substrate from the 384 cores also vary considerably with a minimum of 5.5 grams and a maximum of 159.0 g. The dry weights have a minimum of 1.5 g and a maximum of 66.0 g. In the canopy, a weak positive association is observed between the number of adult mites in a core and the *wet* weight of the substrate from the core (Pearson's correlation coefficient $r = 0.337$, p-value < 0.001; Kendall's $\tau = 0.222$, p-value < 0.001). A weak positive association

is also observed between the number of adult mites in a core and the *dry* weight of the substrate from the canopy (Pearson's $r = 0.334$, p-value < 0.001; Kendall's $\tau = 0.253$, p-value < 0.001). For the cores from the forest floor, we observe a weak positive association between the number of adult mites in a core and the weight of the core (using wet weights: Pearson's $r = 0.094$ with p-value $= 0.195$, Kendall's $\tau = 0.087$ with p-value $= 0.076$; using dry weights: Pearson's $r = 0.092$ with p-value $= 0.206$, Kendall's $\tau = 0.082$ with p-value $= 0.093$).

For the canopy, when we combine cores from the same tree, the association between the total dry weight of the substrate and the number of mites collected from the tree is stronger (Pearson's $r = 0.754$ with p-value < 0.001, Kendall's $\tau = 0.531$ with p-value < 0.001). The association becomes stronger still when we combine canopy cores from the same site (Pearson's $r = 0.926$ with p-value $= 0.001$, Kendall's $\tau = 0.786$ with p-value $= 0.006$).

For the forest floor, the association between the dry weights of the substrate and the number of mites in the substrate is strongest when we combine cores within the same site (Pearson's $r = 0.205$ with p-value $= 0.626$; Kendall's $\tau = 0.286$ with p-value $= 0.399$); however, the association does not appear to be statistically significant.

For the canopy, the scatter plot shown in Figure 6.1 illustrates a positive association between the dry weights of the canopy substrate and the number of mites collected in the canopy of the sites. For the forest floor, there is little association shown in the scatter plot of Figure 6.2 between the dry weights of the soil substrate and the number of mites collected from the sites.

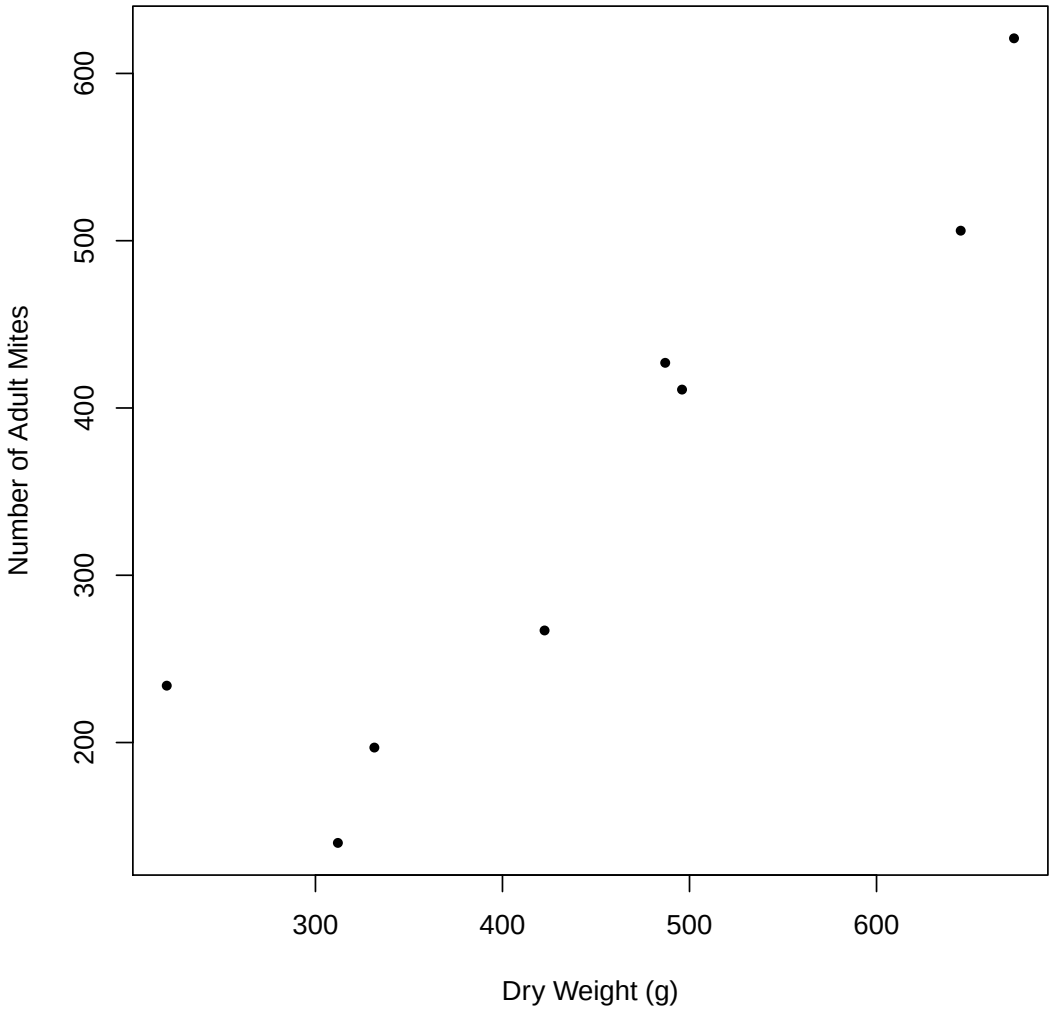

Figure 6.1: Scatter plot of number of adult mites collected from the canopy at eight sites versus the total dry weight of the sampled substrate.

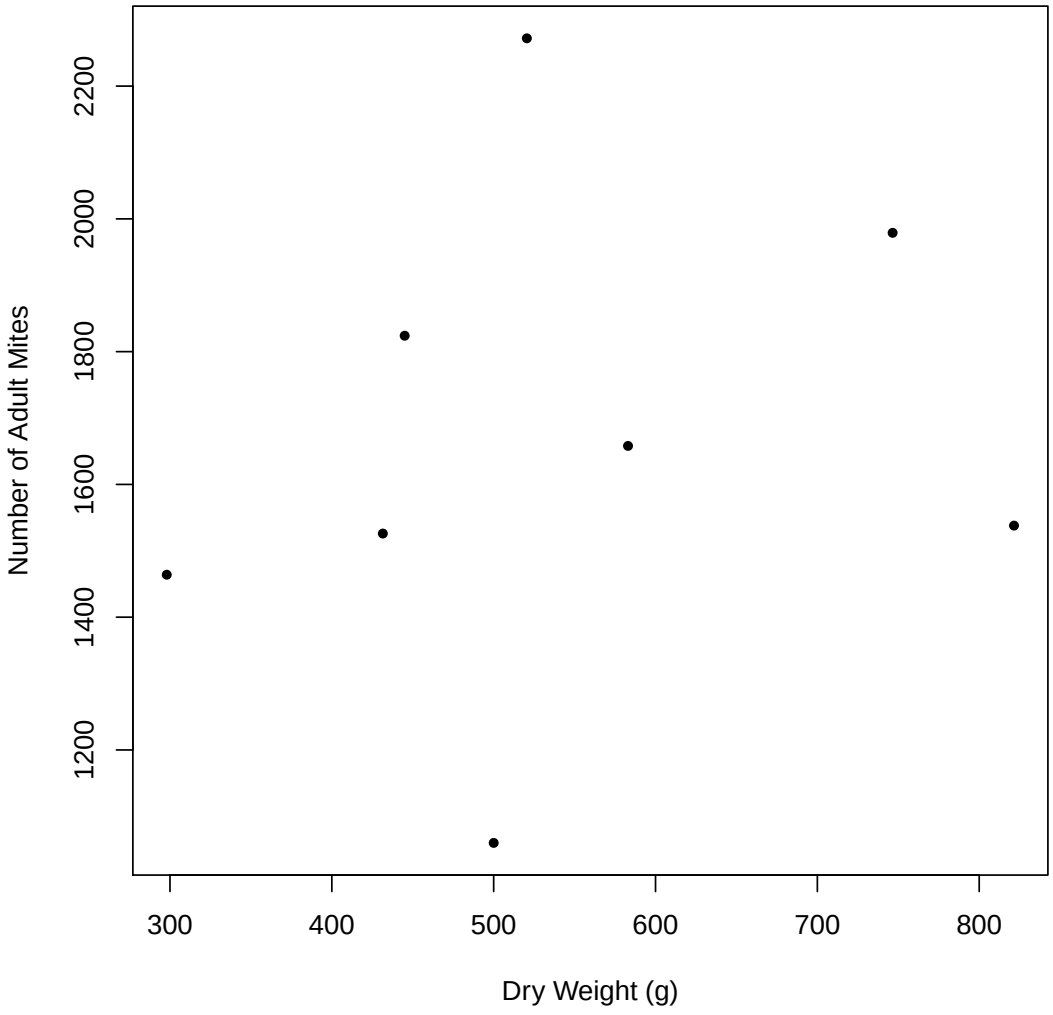

Figure 6.2: Scatter plot of number of adult mites collected from the forest floor at eight sites versus the total dry weight of the sampled substrate.

For the canopy data set, the number of mites collected from a core is positively associated with the weight of the substrate in the core. Therefore, the expected number of mites in a core will not be the same for all cores, rather, it will vary with the weight of the substrate in the core. This is contrary to our multi-site abundance-based models developed in Chapters 3 and 4. In our models, the sample abundances $n_{i,1}, n_{i,2}, \ldots, n_{i,Q}$ of species i in the Q sampling units are *identically distributed* (i.e. the expected values are equal: $E[n_{i,1}] = E[n_{i,2}] = \ldots = E[n_{i,Q}]$). To account for this, in Section 6.5, we propose an extension to our hierarchical Bayesian model to include a positive linear relationship between the dry weight of the substrate and the expected number of mites collected.

6.3.2 Comparing the Mite Composition of the Canopy and Forest Floor

To begin, we compare the composition of species in the forest floor and canopy. If the data suggest the forest floor and canopy differ significantly in the composition and relative abundances of the oribatid mite species, it is important to account for this difference in which case the mite communities on the forest floor and in the canopy can be analysed separately.

Table 6.1 presents totals of the variables measured from the cores for the canopy and the forest floor. Collectively, the cores from the forest floor contain a larger number of species and a larger number of mites than the canopy. In total, the ground cores

contain 4.7 times more mites than the canopy cores. The total wet and dry weights of the soil substrate collected from the ground are larger than the corresponding total weights from the tree bark scrapings. The overall percent moisture is the percentage by which the total wet weight exceeds the total dry weight, relative to the total dry weight. The substrate from the ground cores has a higher average moisture content than the substrate from the canopy cores. The canopy cores contain an average of 0.739 adults mites per gram dry weight whereas the ground cores have an average of 2.865 adult mites per gram dry weight.

Table 6.1: Summary statistics for canopy & ground, excluding *Galumnidae*

	Canopy	Forest Floor
Number of species observed	118	129
Total number of adult mites	2,651	12,452
Total wet weight of collected substrate (g)	9,383.5	13,677.1
Total dry weight of collected substrate (g)	3,588.0	4,346.0
Overall percent moisture (%)	161.5	214.7
Number of mites per gram dry weight	0.739	2.865

Several factors may explain why the ground measurements differ from those in the canopy. The quality and quantity of the substrate, the dispersal capabilities of the mites and differences in micro climate conditions may explain the lower abundances in the canopy (Nadkarni and Longino, 1990). In addition, mites may specialize on particular host-tree species in the canopy whereas the ground habitat may be more uniform (Basset *et al.*, 2007).

Out of 141 species identified, 106 (75.3%) were shared in common between the

forest floor and canopy. More species in the rainforest may be shared in common between the forest floor and canopy, but, by chance, they were absent from the canopy cores and/or the ground cores. A total of 12 species were observed in the canopy but not observed on the forest floor. In addition, 23 species were observed on the forest floor but not in the canopy.

As part of the comparison of the oribatid mite communities, we investigate if similar numbers of species are expected to be observed in the canopy and ground for the same amount of sampling effort. We can measure sampling effort by the number of cores, the weight of the sampled substrate, or the number of adult mites collected.

For the oribatid mite data, it is inappropriate to compare the canopy and ground communities with sampling effort measured by either the number of cores or the weight of the sampled substrate. Suppose that the ground mite community and canopy mite community contain the same species with identical relative abundances. The average number of adult mites collected from the forest floor of a site is significantly greater than the average number collected from the canopy of a site (matched pairs test on the eight sites: $t = 12.52$, p-value < 0.001). Therefore, we would expect more individuals and hence more species to be observed in a given number of ground cores versus the same number of canopy cores. Using the fact that the average number of mites per gram dry weight is 2.865 for the ground substrate and 0.739 for the canopy substrate (from Table 6.1), a similar argument can be made against measuring sampling effort using the weight of the sampled substrate.

We now consider measuring the sampling effort using the number of adult mites.

We use rarefaction curves (Heck *et al.*, 1975) to compare the expected number of species observed in samples of individuals of the same size collected from the forest floor and canopy. This requires assuming the mites in the 192 cores from the forest floor constitute a random sample of individuals, and similarly, for the canopy. Suppose a total of S_{obs} species are observed in an original sample of N individuals with x_i individuals of species i, for $i = 1, 2, \ldots, S_{obs}$, so that $N = \sum_{i=1}^{S_{obs}} x_i$. Assuming a subsample of m individuals is randomly sampled without replacement from the original sample, the expected number $E[S_m]$ of species in a subsample of size m is

$$E[S_m] = S_{obs} - \sum_{i=1}^{S_{obs}} \frac{\binom{N-x_i}{m}}{\binom{N}{m}},$$

for $m = 0, 1, \ldots, N$. Note that $E[S_0] = 0$ and $E[S_N] = S_{obs}$. A rarefaction curve is produced by plotting $E[S_m]$ as a function of m.

Figure 6.3 shows the two rarefaction curves for the ground and canopy. The rarefaction curves lie quite close together initially, for subsamples of 200 or fewer individuals. As the size m of the subsamples reaches the size of the original canopy sample (2,651 individuals), the expected number of species in the canopy subsamples approaches 118, which exceeds the expected number of species (101) in ground subsamples by 17. The dissimilarity of the two rarefaction curves suggests large samples of equal size from the canopy and the ground may contain different numbers of species.

A rarefaction curve is based on the assumption that a random sample of individuals is used, or equivalently, that the individuals are randomly distributed. The

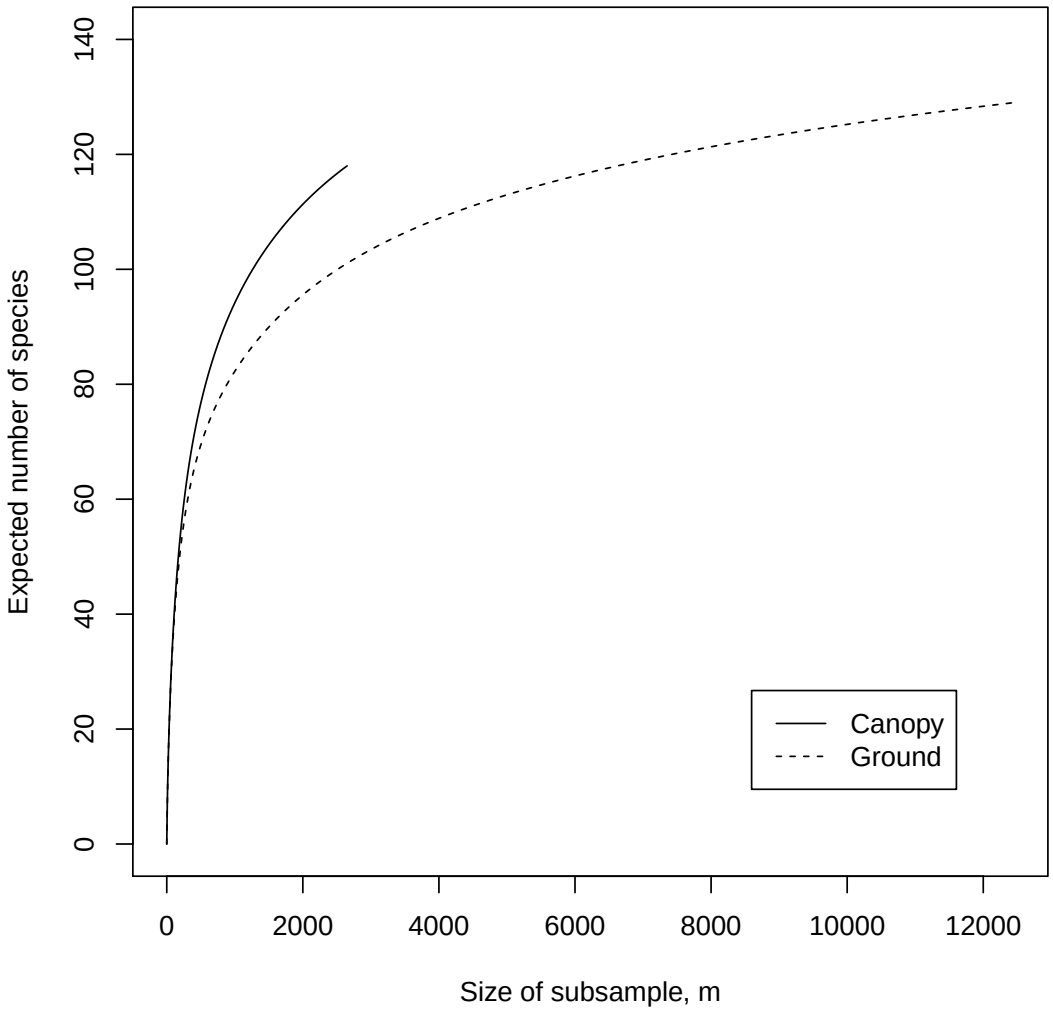

Figure 6.3: Rarefaction curves for the oribatid mites collected from the forest canopy (solid line) and from the forest floor (dashed line) of the SLPA.

oribatid mites collected from the forest floor do not constitute a random sample of individuals, and the individuals are not randomly distributed among the cores taken from the ground (Index of Dispersion = 10,052.68, p-value < 0.001). Similarly, the oribatid mites collected from the forest canopy do not constitute a random sample of individuals and are not randomly distributed among the cores (Index of Dispersion = 2738.29, p-value < 0.001). Thus, while the rarefaction curves lend some insight, we cannot rely solely on these curves.

We can compare the composition of the oribatid mites in the canopy and floor using an index of the dissimilarity (Legendre & Legendre, 1998). There are many such indices available (e.g., see Section 7.3 of Legendre & Legendre, 1998).

Let S_{obs} denote the total number of oribatid mite species observed in the rainforest and $n_{i,q}$ denote the number of adult mites of species i in the q^{th} sample for $i = 1, 2, \ldots, S_{obs}$. The *sample relative abundance* of species i in the q^{th} sample is $n_{i,q}/N_q$, where $N_q = \sum_{j=1}^{S_{obs}} n_{j,q}$ is the total number of mites in the q^{th} sample. The complement of Whittaker's index of association (Legendre & Legendre, 1998) measures the dissimilarity or distance between two samples based on the sample relative abundances of the species. For two samples labelled 1 and 2, the complement of Whittaker's index of association is

$$D = \frac{1}{2} \sum_{i=1}^{S_{obs}} \left| \frac{n_{i,1}}{N_1} - \frac{n_{i,2}}{N_2} \right| . \tag{6.1}$$

The difference D varies between 0 and 1, and can be expressed as a percentage between 0% and 100%. When the two samples do not share any species in common,

we have $D = 100\%$. When the relative abundances of a species in the two samples are identical, it contributes zero to the index D, so that if all species have identical proportions in the two samples, we have $D = 0\%$.

We note that a common index, the Bray-Curtis dissimilarity index[1] (Legendre & Legendre, 1998), is equivalent to Equation 6.1 when the raw abundances ($n_{i,q}$) in the Bray-Curtis dissimilarity index are replaced by the sample relative abundances of the species.

Observations from a single core can be quite sparse, with many species being absent. As a result, we consider combining the observations from the eight canopy cores of a tree to form a single sample, and this is done for the ground, as well. With 24 trees in total, we have two samples associated with each tree: a ground sample and a canopy sample. Considering all pairs of canopy samples, the average dissimilarity D is 73.9% (SD = 12.5%); whereas, for pairs of ground samples, the average dissimilarity $D = 58.2\%$ (SD = 12.0%) is lower. Higher average dissimilarity among the canopy observations suggests greater variability in oribatid mite composition within the canopy versus within the forest floor.

Based on the differences described above between the oribatid mite observations in the canopy and forest floor, we will treat the canopy and forest floor as separate

[1]The Bray-Curtis dissimilarity index is a measure of percent dissimilarity between two samples. The equation for the Bray-Curtis dissimilarity index is

$$BC = \frac{\sum_{i=1}^{S_{obs}} |n_{i,1} - n_{i,2}|}{\sum_{i=1}^{S_{obs}} n_{i,1} + \sum_{i=1}^{S_{obs}} n_{i,2}}.$$

communities in the remaining sections of this chapter. As a result, we compute estimates of oribatid mite species richness for the canopy and forest floor separately.

6.4 Remarks Regarding Assumptions for the Methods

Appendix A assesses the plausibility of the assumptions (described in Section 3.3) of our estimation methods. Some assumptions are likely to be violated with the canopy and ground data sets. However, these assumptions underly not only our methods, but the other methods as well. The estimator $\hat{S}_{ACE}$ assumes a simple random sample of individuals has been collected – this is not the case for the observations collected from the canopy and forest floor. The estimators $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ assume a species has the same probability of occurring in each sampling unit, but this is violated in the canopy when a strong association is observed between the number of mites collected and the weight of the substrate. This assumption is also violated if the spatial distributions of the species are not stationary across the rainforest.

The assumption of 100% probability of detection of all mites in the cores is unlikely to be true as the immatures cannot be identified and therefore were not recorded. Furthermore, all estimators are affected by the inability to identify the 1,021 *Galumnidae* mites to their species types.

Species are highly unlikely to be mutually independent, however this is a common assumption made by all the species richness estimators considered here. In the re-

mainder of this chapter, we will combine the observations from the 24 canopy cores of a site to have one canopy sampling unit for each site. With little dependence introduced by the host tree species and the distances between sites, the sets of observations in the canopy sampling units will be treated as being independent. We do the same for the forest floor.

The violations of the assumptions reduce the reliability of the species richness estimates and their associated measures of uncertainty. In an attempt to address violations of the first assumption regarding uniform sampling effort, the next section proposes an adjustment to our hierarchical Bayesian model to incorporate the dry weights of the substrate.

6.5 Modelling the Association between Mite Abundances and Substrate Weights

In the exploratory data analysis, the number of mites collected from the canopy of a site was found to have a strong positive correlation with the dry weight of the sampled substrate. In this section, we propose a modification of our hierarchical Bayesian model to account for the dry weights of the substrate.

Consider either the forest floor or the canopy, and let W_q denote the sum of the dry weights of the substrate collected from the sampling unit in the q^{th} site, for sites $q = 1, 2, \ldots, 8$. The dry weights will be measured in units of 500 grams. Let $\boldsymbol{W} = \langle W_1, W_2, \ldots, W_8 \rangle$ denote the vector of total dry weights of the substrate

sampled from the eight sites.

In our hierarchical model, a species contributes individuals to a sampling unit according to a negative binomial distribution. We adjust the mean of the negative binomial distribution to be proportional to the weight of the substrate, so that $E[n_{i,q}] = W_q \mu_i$.

With S_{obs} species observed in the layer of interest (canopy or ground) in the rainforest, the data is augmented to include a fixed known number $(M - S_{obs})$ of zero vectors (i.e., $\boldsymbol{n_i} = \mathbf{0}$ for $i = S_{obs} + 1, \ldots, M$). As before, the binary variable $z_i = 1$ indicates species i exists in the layer of interest, and $z_i = 0$ otherwise, for $i = 1, \ldots, M$. For $i \leq S_{obs}$, we assume

$$n_{i,q} \mid W_q, \mu_i, k \overset{ind}{\sim} \text{Negative Binomial}(W_q \mu_i, k),$$

for $q = 1, \ldots, 8$. For $i > S_{obs}$, we assume

$$n_{i,q} \mid W_q, \mu_i, z_i, k \overset{ind}{\sim} \begin{cases} \text{Negative Binomial}(W_q \mu_i, k) & \text{if } z_i = 1 \\ I\{n_{i,q} = 0\} & \text{if } z_i = 0 \end{cases}$$

for $q = 1, \ldots, 8$.

As previously in Chapter 4, $\boldsymbol{\mu} = \langle \mu_1, \mu_2, \ldots, \mu_M \rangle$ is treated as a vector of independent and identically distributed random variables from a mixture of two exponential distributions with exponential rate parameters β_1 and β_2 $(\beta_1 < \beta_2)$, and mixture weight π. After incorporating the substrate weights in the negative binomial sampling model, we proceed with the Bayesian inference as described in Chapter 4.

The dispersion parameter k has been fixed at a common value for all species. Realistically, k should change with the weight of the sampled substrate if the expected

number of individuals in a sampling unit is proportional to the weight of the sampled substrate. For species i, suppose $n_{i,q} \mid \mu_i, k \sim$ Negative Binomial$(W_q\mu_i, k)$ when 500 grams ($W_q = 1$) of substrate are examined. As the weight increases, the expected number $W_q\mu_i$ of individuals increases, and the distribution of $n_{i,q}$ may approach a Poisson distribution as the size of the sampling unit expands beyond the scale of local clusters of individuals. In these circumstances, when the weight of substrate is increased, we expect k to increase and tend towards positive infinity in the limit.

6.6 Inference on Oribatid Mite Species Richness

For the canopy, we combine all the observations from the cores of a site, and treat the eight sites as the sampling units. With little dependence introduced by the host tree species and the distances between sites (see Section A.2 in Appendix A), the sets of observations in the canopy sampling units will be treated as being independent. Likewise, for the forest floor, we also combine the observations from all cores within a site, and treat the sets of observations from different sites as independent.

6.6.1 Estimating oribatid mite species richness within the canopy of the forest

To estimate the number of oribatid mite species in the canopy of the rainforest, we use the likelihood-based approach from Chapter 3, the Bayesian approach from Chapter 4, and the substrate-weight-adjusted version of the Bayesian approach proposed in

Section 6.5. For Bayesian inference on the species richness, the posterior mean is used as a point estimate. For each site, the observations from the 24 canopy cores within the site are pooled to form one sample of raw abundances, and our methods use this abundance data from each of the eight sites. We also apply the non-parametric estimators introduced in Chapter 5.

Table 6.2 presents the estimates of oribatid mite species richness in the canopy along with measures of variability in parentheses and 95% interval estimates. When our Bayesian model is fit to the data, the posterior mean is slightly smaller when the model incorporates the weights of the collected substrate; the posterior mean is 145.8 with the substrate weights versus 146.2 without them. The 95% credible intervals are also slightly narrower when the Bayesian model incorporates the substrate weights; for example, the HPD interval is (129, 167) with the substrate weights and (127, 168) without the substrate weights.

The MLE $\hat{S}_{\Omega}$ is the lowest estimate (127) and has the narrowest 95% interval estimates, which is consistent with its behaviour in the simulation studies in Chapter 5. The second-order jackknife estimator $\hat{S}_{Jack2}$ produces the largest estimate (170.4). Among most estimators, the 95% interval estimates do overlap. Although, the interval estimates associated with the MLE $\hat{S}_{\Omega}$ do not intersect with the interval estimates associated with $\hat{S}_{Jack2}$.

In each MCMC simulation for the Bayesian methods, two chains were run based on dispersed initial values for the model parameters (ψ, π, β_1, $(\beta_2 - \beta_1)$, k). For the data augmentation, we set $M = 350$. The posterior values of S_{Ω} were all safely well below

Table 6.2: Estimates of Oribatid Mite Species Richness in Canopy

Estimator	Point Estimate (ESE)	95% Asymmetric Confidence Interval	95% Symmetric Confidence Interval
S_{obs}	118		
$\hat{S}_{Chao2}$	162.7 (19.2)	(138.0, 218.1)	(125.2, 200.3)
$\hat{S}_{Jack2}$	170.4 (12.7)	(150.8, 201.9)	(145.5, 195.4)
$\hat{S}_{ACE}$	140.6 (10.5)	(127.5, 171.7)	(120.1, 161.2)
		95% Asymmetric CI	Likelihood Interval
MLE $\hat{S}_{\Omega}$	127 (3.3)	(122.5, 135.9)	(122, 135)
		95% HPD Interval	95% Equal-Tail Interval
Posterior Mean of S_{Ω}	146.2	(127, 168)	(130, 174)
Post. Mean adjusted with substrate weights	145.8	(129, 167)	(130, 171)

M. A burn-in period of at least 100,000 iterations was employed to ensure that the Gelman and Rubin diagnostic statistic was less than 1.1 for all monitored parameters. An additional 500,000 iterations was then run, thinning every 10^{th} observation. This resulted in two sets of 50,000 posterior draws for a total of 100,000 posterior draws used for inference.

A total of 152 adult *Galumnidae* mites were observed in the substrate collected from the forest canopy. Treating the mites from the *Galumnidae* family as one 'species' and including them in the analysis had a negligible effect; the estimates of species richness typically increased by one, and the lower and upper limits of the interval estimates also increased by one.

In addition to the negative binomial model, we also consider a Poisson sampling model in our Bayesian approach. Given that species i exists in the rainforest canopy, its abundances $n_{i,1}, n_{i,2}, \ldots, n_{i,8}$ in the canopy sampling units can be treated as independent and identically distributed Poisson random variables with mean μ_i, for $i = 1, 2, \ldots, M$. Under this Poisson sampling model, the total number $\sum_{q=1}^{8} n_{i,q}$ of adult mites of species i collected from the canopy is a Poisson random variable with mean $8\mu_i$. This Poisson sampling model is equivalent to pooling the observations from the eight sites together, essentially working with a single sample.

The models can also be altered to use a different probability distribution on the mean abundances, $\boldsymbol{\mu}$, of the species in the sampling units. In addition to the two-component exponential mixture distribution, we consider a lognormal distribution on $\boldsymbol{\mu}$. As mentioned in Section 3.2.2, the lognormal distribution is often used to model the distribution of expected abundances in sampling units, or equivalently, the distribution of the regional abundances (i.e., population numbers) of the species.

Finally, we consider one last alteration to the hierarchical Bayesian model. As described in Section 6.5, we can incorporate the total weight of the canopy substrate collected from a site into the expected abundance of a species in the sampling unit. For the Poisson sampling model, if species i exists in the canopy, then $n_{i,q}$ is modelled as a Poisson random variable with mean $W_q\mu_i$, for $i = 1, 2, \ldots, M$. The dry weight W_q of the substrate collected from the q^{th} sampling unit is measured in units of 500 grams, for $q = 1, 2, \ldots, 8$.

In total, we consider eight variations of the Bayesian model for these data, based

on all possible combinations of sampling model (negative binomial or Poisson), probability distribution on $\boldsymbol{\mu}$ (two-component exponential mixture or lognormal), and whether or not to incorporate the substrate weights. For models with a lognormal distribution on $\boldsymbol{\mu}$, M was increased to 800 when needed to ensure the upper tail of the posterior distribution of S_Ω was well below M.

Table 6.3 displays the posterior mean and interval estimates of species richness from the eight Bayesian models fitted to the canopy data. The posterior standard deviation of species richness S_Ω is in parentheses next to the posterior mean. The last column of Table 6.3 displays the deviance information criterion (DIC) of each model. DIC penalizes models for a poor fit to the data (i.e., high estimated expected deviance) and for the number of effective parameters. A lower DIC indicates a preferred model (in terms of fit and parsimony). The DIC values suggest the negative binomial sampling models are more appropriate than the Poisson sampling models. For each combination of sampling model and distribution on $\boldsymbol{\mu}$, the weight-adjusted model has a lower DIC than the model that does not incorporate the weights of the substrate, indicating that the incorporation of the substrate weights is an appropriate adjustment. For each combination of sampling model and distribution on $\boldsymbol{\mu}$, the point and interval estimates of species richness from the weight-adjusted model are very similar to the corresponding model without the adjustment for substrate weights.

The weight-adjusted negative binomial sampling model with a two-component exponential mixture on $\boldsymbol{\mu}$ has the lowest DIC (highlighted in bold in Table 6.3). For this model, the estimated number of oribatid mite species in the canopy is 145.8

Table 6.3: Bayesian Inference on Oribatid Mite Species Richness in Canopy

Sampling Model	Distribution on $\boldsymbol{\mu}$	Adjust for Substrate Weights?	Posterior Mean of S_Ω (SD)	95% HPD Credible Interval	95% Equal-Tail Credible Interval	DIC
Poisson	Two-component Exponential	No	136.6 (6.7)	(124, 149)	(126, 152)	4,776.0
Poisson	Two-component Exponential	Yes	136.6 (6.8)	(124, 149)	(126, 153)	4,376.8
Poisson	Lognormal	No	166.0 (25.0)	(129, 215)	(135, 233)	4,845.3
Poisson	Lognormal	Yes	165.2 (24.0)	(129, 211)	(135, 228)	4,441.2
Negative Binomial	Two-component Exponential	No	146.2 (11.6)	(127, 168)	(130, 174)	3,155.2
Negative Binomial	Two-component Exponential	Yes	145.8 (10.6)	(129, 167)	(130, 171)	**3,102.4**
Negative Binomial	Lognormal	No	164.3 (22.7)	(131, 208)	(135, 221)	3,156.9
Negative Binomial	Lognormal	Yes	164.7 (22.9)	(130, 207)	(136, 222)	3,109.4

(posterior mean) with a 95% highest posterior density interval of (129, 167) and a 95% equal-tail credible interval of (130, 171). This model has a smaller posterior mean of species richness S_Ω, a smaller posterior standard deviation of S_Ω and smaller upper limits in its interval estimates than the model with the second smallest DIC (the weight-adjusted negative binomial sampling model with a lognormal distribution on $\boldsymbol{\mu}$).

6.6.2 Estimating oribatid mite species richness on the forest floor

For the forest floor data, Table 6.4 presents the point estimates with measures of variability in parentheses and 95% interval estimates of the number of oribatid mite species. The MLE $\hat{S}_\Omega$ produces the smallest estimate (131) of species richness. In fact, the lower limits of the interval estimates associated with the likelihood-based approach are very close to the observed number of species (129).

The estimator $\hat{S}_{Jack2}$ produces the largest estimate (160.3). The symmetric 95% confidence intervals associated with $\hat{S}_{Chao2}$ and $\hat{S}_{ACE}$ have unrealistic lower limits because they are less than the number of species observed in the canopy. The last two rows of Table 6.4 refer to Bayesian estimates based on the negative binomial sampling model with a two-component exponential mixture distribution on $\boldsymbol{\mu}$. The interval estimates from our Bayesian approaches overlap substantially with the asymmetric 95% confidence intervals from the three non-parametric estimators. The Bayesian

approach that incorporates the substrate weights has a slightly larger posterior mean of S_Ω (148.7 versus 148.0) and slightly larger upper endpoints of its interval estimates than the Bayesian approach that does not incorporate the substrate weights.

Table 6.4: Estimates of Oribatid Mite Species Richness on Forest Floor

Estimator	Point Estimate (ESE)	95% Asymmetric Confidence Interval	95% Symmetric Confidence Interval
S_{obs}	129		
$\hat{S}_{Chao2}$	148.9 (10.2)	(136.7, 180.3)	(128.9, 168.9)
$\hat{S}_{Jack2}$	160.3 (10.3)	(145.7, 187.8)	(140.0, 180.6)
$\hat{S}_{ACE}$	144.8 (8.7)	(134.8, 172.3)	(127.7, 161.8)
		95% Asymmetric CI	Likelihood Interval
MLE $\hat{S}_\Omega$	131 (1.6)	(129.5, 136.9)	(129, 134)
		95% HPD Interval	95% Equal-Tail Interval
Posterior Mean of S_Ω	148.0	(136, 159)	(138, 162)
Post. Mean adjusted with substrate weights	148.7	(136, 161)	(138, 164)

A total of 869 adult mites from the family *Galumnidae* were observed in the substrate extracted from the forest floor. As with the canopy, if we treat the *Galumnidae* mites on the forest floor as belonging to a single species, then the point estimates of species richness typically increased by one and the endpoints of the interval estimates increased by one, too.

We fit the eight variations of the hierarchical Bayesian model described in Section 6.6.1 to the ground data, and the results are presented in Table 6.5. The DIC is

smallest for the four negative binomial sampling models. This suggests that the Poisson sampling models are not appropriate for the mite data from the forest floor.

For each combination of sampling model and distribution on $\boldsymbol{\mu}$ in Table 6.5, the weight-adjusted model has a higher DIC than the corresponding model that does not incorporate the weights of the substrate. Therefore, incorporating the weights of the sampled ground substrate does not appear to improve each model's fit to the data. This is in-line with the exploratory analysis. In addition, incorporating the weights of the sampled ground substrate produces minimal change in the posterior mean and interval estimates of species richness.

The negative binomial sampling model with a lognormal distribution on $\boldsymbol{\mu}$ with no adjustments for the substrate weights has the lowest DIC, overall. For this model, the estimated number of oribatid mite species in the canopy is 155.4 (posterior mean) with a 95% highest posterior density interval of (135, 180) and a 95% equal-tail credible interval of (138, 187).

6.7 Discussion

The MLE $\hat{S}_{\Omega}$ using an exponential distribution on $\boldsymbol{\mu}$ produces the lowest estimate of species richness for both the canopy and the forest floor. The highest estimate of species richness for each data set is from the non-parametric estimator $\hat{S}_{Jack2}$. For the canopy, the point estimates of the number of species in the canopy range from $\hat{S}_{\Omega} = 127$ (ESE = 3.3) to $\hat{S}_{Jack2} = 170.4$ (12.7). For the ground, a similar

Table 6.5: Bayesian Inference on Oribatid Mite Species Richness on Forest Floor

Sampling Model	Distribution on $\boldsymbol{\mu}$	Adjust for Substrate Weights?	Posterior Mean of S_Ω (SD)	95% HPD Credible Interval	95% Equal-Tail Credible Interval	DIC
Poisson	Two-component Exponential	No	143.9 (5.0)	(134, 153)	(135, 155)	14,204.8
Poisson	Two-component Exponential	Yes	143.9 (5.0)	(134, 153)	(135, 155)	14,911.5
Poisson	Lognormal	No	156.3 (13.4)	(134, 183)	(138, 190)	14,224.2
Poisson	Lognormal	Yes	155.8 (12.1)	(136, 179)	(139, 185)	14,928.2
Negative Binomial	Two-component Exponential	No	148.0 (6.1)	(136, 159)	(138, 162)	5,006.4
Negative Binomial	Two-component Exponential	Yes	148.7 (6.7)	(136, 161)	(138, 164)	5,021.0
Negative Binomial	Lognormal	No	155.4 (12.9)	(135, 180)	(138, 187)	**4,993.7**
Negative Binomial	Lognormal	Yes	155.5 (12.6)	(136, 180)	(138, 187)	4,997.2

but more narrow range of estimates is produced with the lowest estimate $\hat{S}_{\Omega} = 131$ (1.6) and the highest estimate $\hat{S}_{Jack2} = 160.3$ (10.3). Several times more individuals were collected from the ground than the canopy of the eight sites. As a result, the estimated standard deviations associated with the species richness estimates and the widths of the interval estimates are smaller for the ground data set.

In Chapter 5, all estimators were negatively biased when applied to the realistic data sets in the second simulation study. Although we are working with different data sets now, this still gives us a cause for concern, especially considering the violations of the assumptions discussed in Appendix A. For species-rich taxa like the oribatid mites, it is common for estimators to actually provide lower-bound estimates on species richness (Gotelli & Colwell, 2001).

If one wishes to err on the side of overestimation, then the lognormal distribution on $\boldsymbol{\mu}$ in the hierarchical Bayesian models produces the interval estimates with the largest upper limits. One may consider other probability distributions for $\boldsymbol{\mu}$. Based on our experience (see also O'Hara, 2005), distributions that place more support on small positive values of μ near zero will produce larger species richness estimates, but also introduce larger measures of uncertainty (i.e., larger standard deviations and higher upper limits of interval estimates). Our likelihood-based approach would produce higher species richness estimates if, instead of the exponential distribution, a distribution on $\boldsymbol{\mu}$ with more support near zero was used, such as the lognormal distribution.

In the past, it has been common to use a Poisson sampling model with a probability

distribution on $\boldsymbol{\mu}$, albeit for modelling the abundances of species in one sample (Fisher *et al.*, 1943; O'Hara, 2005) rather than in multiple samples from one region. However, among the eight variations of hierarchical Bayesian models applied to the canopy and forest floor data sets, the negative binomial sampling models consistently provided a better fit to the two data sets.

For the canopy, a strong positive association exists between the dry weights of the sampled substrate and the number of adult mites observed. Incorporating the substrate weights into the hierarchical Bayesian models was beneficial because it improved the fit of the models to the data (as measured from the DIC values in Table 6.3). However, given a sampling model and a distribution on $\boldsymbol{\mu}$, it is interesting to note that Bayesian inference on species richness was very similar between the models with and without the adjustments for the substrate weights. The optimal Bayesian model was the negative binomial sampling model adjusted for the substrate weights with a two-component exponential mixture for $\boldsymbol{\mu}$, giving a 95% HPD interval of (129, 167) for the oribatid mite species richness in the canopy of the rainforest.

For the forest floor, the exploratory analysis did not reveal a significant association between the substrate weights and the numbers of mites observed. Consequently, the weight-adjusted Bayesian models did not offer any improvement in fit to the ground data against the corresponding models that did not adjust for the ground substrate weights. The optimal Bayesian model was the negative binomial sampling model with a lognormal distribution for $\boldsymbol{\mu}$ and with no adjustment for the substrate weights, resulting in a 95% HPD interval of (135, 180) for the species richness on the forest

floor.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Species richness estimation based on multiple samples from one community is an open area of research (Bunge & Barger, 2009) and a challenging problem. In this dissertation, we have presented hierarchical models for abundance-based data from multiple sampling locations in a region. These models are unique among models for species richness estimation as they use the sample abundances of species in multiple samples from the *same* community; whereas, previously proposed estimators use multiple samples that must be drawn from distinct communities (e.g., Mao & Lindsay, 2004).

Our statistical models and estimators are appropriate for populations where conspecific individuals tend to occur in clusters. In the presence of this clustering, individuals in a spatial-based sampling unit are not a random sample, and therefore a negative binomial sampling model seems more appropriate than a Poisson sampling

model. This was true for the hierarchical Bayesian models applied to the oribatid mite data sets.

In Simulation Study 2 in Chapter 5, all of the estimators considered produced negatively-biased estimates, and the bias shrunk with increased sampling effort. The equations for the estimators $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ (shown in Chapter 5) principally involve S_{obs}, r_1 and r_2. With samples representing a tiny fraction of the region, r_1 is generally larger than r_2. As the sampling effort increases modestly, r_1 will continue to be larger than r_2 (Mao & Colwell, 2005). The number, S_{obs}, of species observed during sampling increases with additional sampling effort. Thus, increases in sampling effort (though still dealing with tiny fractions of the population) imply a general tendency of increasing estimates of species richness for $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$.

For $\hat{S}_{ACE}$, the coefficient of variation of the relative abundances of the species is estimated, but the estimated coefficient of variation tends to be negatively biased, especially with small sample sizes (Chao & Lee, 1992). This results in negatively-biased estimates of species richness. As the sample size increases, the estimate of the coefficient of variation and, consequently, $\hat{S}_{ACE}$ will increase.

In addition to the challenges associated with small samples, two additional factors cause our maximum likelihood and Bayesian methods to produce negatively-biased estimates in Simulation Study 2. First, the distribution of the abundances of species from the census of the 50-hectare plot on Barro Colorado Island (BCI) is very skewed, with many rare species only having one or two individuals in the plot. The abundances are not adequately modelled by an exponential distribution or a two-component mix-

ture of exponential distributions, but rather by the lognormal distribution or distributions developed under neutral theories of biodiversity (Volkov *et al.*, 2003). As a consequence, our methods do not use an appropriate mixing distribution on $\boldsymbol{\mu}$ for the sub-samples generated from the census data. Without sufficient support near zero for the mixing distribution on $\boldsymbol{\mu}$, the resulting estimators underestimate the species richness. A second factor that contributes to the negative bias of our estimators is our simplifying assumption that the negative binomial dispersion parameter k is fixed at a common value for all species. Condit *et al.* (2000) observed that species with lower abundances had higher degrees of spatial aggregation in the BCI plot. If one were to allow each species to have its own negative binomial dispersion parameter, the dispersion parameters associated with the rare spatially-aggregated species in BCI will be smaller, resulting in higher estimated probabilities of absence from samples and higher estimates of species richness.

The interval estimators from all methods had very poor coverage levels in Simulation Study 2. A significant reason for this is the negative bias of the estimators. For our maximum likelihood and Bayesian methods, the choices for the parametric distribution on $\boldsymbol{\mu}$ and a common value of k for all species contribute to the poor performance of the interval estimators. In addition, large samples are needed in order for the confidence intervals based on asymptotics (i.e., interval estimates associated with our MLE and the non-parametric estimators) to be appropriate. Furthermore, all estimation methods make the simplifying assumption that species are mutually independent. Therefore, the interval estimators do not account for uncertainty introduced

through inter-species interactions.

In the case study on oribatid mites, we considered various extensions of the Bayesian model and incorporated the weight of the substrate from each sampling unit. In other scenarios, covariates associated with sampling units could be incorporated as a fairly straight-forward extension to the modelling framework proposed here.

7.2 Future Work

One statistical model and associated species richness estimator will not be best for all data sets and scenarios. Based on the our studies, we recommend our Bayesian model in settings where con-specific clustering is suspected, the sampling units are spaced apart (to reduce spatial autocorrelation), and the sampling units constitute a small fraction of the region (i.e., less than 1% of the region of interest is sampled).

We have found that species richness estimates are sensitive to the mixing distribution placed on the expected numbers (μ's) of con-specific individuals in the sampling units. Using more positively skewed distributions with more support near zero, such as the lognormal and Pareto distributions, may produce more reasonable estimates of species richness in some settings; although, these distributions will result in higher estimated variances and wider confidence intervals. Thus, the choice of mixing distribution for $\boldsymbol{\mu}$ is guided, essentially, by the classical trade-off between bias and variance. In addition to investigating other mixing distributions, developing

techniques for model checking and goodness of fit for abundance-based data from multiple samples is an important area for future research.

It would be desirable to consider mixing distributions on the dispersion parameter k of the negative binomial distribution. Going even further, bivariate probability distributions on μ and k could be used, relaxing the independence assumption. In particular, a non-parametric approach based on Dirichlet process mixtures (Antoniak, 1974; Ferguson, 1973) could be used to choose the optimal mixing distribution on (μ, k). Empirical observations may influence the modelled association between μ and k. For example, Condit *et al.* (2000) found a negative association between the degree of spatial aggregation and the observed abundance of woody plant species in five of six large study plots in different tropical forests in five countries.

While sampling units may be randomly distributed in a region, spatial autocorrelation among observations is still present. Future work will investigate the incorporation of spatial autocorrelation into the hierarchical model (e.g., using multivariate conditional autoregressive models; Besag, 1974), and these spatial models would be a useful extension to the developments made in this dissertation.

Bibliography

C.E. Antoniak, 1974. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems", *The Annals of Statistics*, **2**: 1152-1174.

O. Arrhenius, 1921. "Species and area", *Journal of Ecology*, **9**: 95-99.

S. Banerjee, P.B. Carlin and E.A. Gelfand, 2004. *Hierarchical Modeling and Analysis for Spatial Data*, CRC, New York, NY, USA.

K. Barger and J. Bunge, 2008. "Bayesian estimation of the number of species using noninformative priors", *Biometrical Journal*, **50**: 1064-1076.

Y. Basset, B. Corbara, H. Barrios, P. Cuénoud, M. Leponce, H.-P. Aberlenc, J. Bail, D. Bito, J.R. Bridle, G. Castaño-Meneses, L. Cizek, A. Cornejo, G. Curletti, J.H.C. Delabie, A. Dejean, R.K. Didham, M. Dufrêne, L.L. Fagan, A. Floren, D.M. Frame, F. Hallé, O.J. Hardy, A. Hernandez, R.L. Kitching, T.M. Lewinsohn, O.T. Lewis, M. Manumbor, E. Medianero, O. Missa, A.W. Mitchell, M. Mogia, V. Novotny, F. Odegaard, E.G. de Oliveira, J. Orivel, C.M.P. Ozanne, O. Pascal, S. Pinzón, M. Rapp, S.P. Ribeiro, Y. Roisin, T. Roslin, D.W. Roubik, M. Samaniego, J. Schmidl, L.L. Sorensen, A. Tisheckkin, C. Van Osselaer, and N.N. Winchester, 2007. "IBISCA-Panama, a large-scale study of arthropod beta-diversity and vertical stratification in a lowland rainforest: rationale, study sites and field protocols", *Entomologie*, **77**: 39-69.

V.M. Behan-Pelletier and D.E. Walter, 2000. "Biodiversity of oribatid mites (Acari: Oribatida) in tree canopies and litter", In: D.C. Coleman and P.F. Hendrix (Eds.), *Invertebrates as Webmasters in Ecosystems*. CAB International, Wallingford, pp. 187-202.

J. Besag, 1974. "Spatial interaction and the statistical analysis of lattice systems", *Journal of the Royal Statistical Society, B*, **36**: 192-236.

J. Besag, P. Green, D. Higdon and K. Mengersen, 1995. "Bayesian computation and stochastic systems", *Statistical Science*, **10**: 3-66.

T. Boulinier, J.D. Nichols, J.R. Sauer, J.E. Hines and K.H. Pollock, 1998. "Estimating species richness - the importance of heterogeneity in species detectability", *Ecology*, **79**: 1018-1028.

M.G. Bulmer, 1974. "On fitting the Poisson lognormal distribution to species-abundance data", *Biometrics*, **30**: 101-110.

J. Bunge and K. Barger, 2008. "Parametric models for estimating the number of classes", *Biometrical Journal*, **50**: 971-982.

J. Bunge and K. Barger, 2009. *Estimating the Number of Classes in a Population*, Cornell University, New York. `http://www.stat.cornell.edu/~bunge/research.html` Last accessed on April 16, 2012.

J. Bunge and M. Fitzpatrick, 1993. "Estimating the number of species: a review", *Journal of the American Statistical Association*, **88**: 364-373.

K.P. Burnham and W.S. Overton, 1979. "Robust estimation of population size when capture probabilities vary among animals", *Ecology*, **60**: 927-936.

R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu, 1995. "A limited memory algorithm for bound constrained optimization", *SIAM Journal of Scientific Computing*, **16**: 1190-1208.

B.P. Carlin and T.A. Louis, 2009. *Bayesian Methods for Data Analysis*, 3^{rd} edition, CRC Press, Boca Raton, FL, USA.

G. Casella and E.I. George, 1992. "Explaining the Gibbs sampler", *The American Statistician*, **46**: 167-174.

R.M. Cassie, 1962. "Frequency distribution models in the ecology of plankton and other organisms", *Journal of Animal Ecology*, **31**: 65-92.

A. Chao, 1984. "Nonparametric estimation of the number of classes in a population", *Scandinavian Journal of Statistics*, **11**: 265-270.

A. Chao, 1987. "Estimating the population size for capture-recapture data with unequal catchability", *Biometrics*, **43**: 783-791.

A. Chao, 2005. "Species richness estimation", *Encyclopedia of Statistical Sciences*, S. Kotz ed., p. 7909-7916, Wiley, Hoboken, N.J.

A. Chao, R.K. Colwell, C.-W. Lin, and N.J. Gotelli, 2009. "Sufficient sampling for asymptotic minimum species richness estimators", *Ecology* **90**: 1125-1133.

A. Chao, W.-H. Hwang, Y.-C. Chen, and C.-Y. Kuo, 2000. "Estimating the number of shared species in two communities." *Statistics Sinica*, **10**: 227-246.

A. Chao and S.-M. Lee, 1992. "Estimating the number of classes via sample coverage", *Journal of the American Statistical Association*, **87**: 210-217.

A. Chao, M.-C. Ma and M.C.K. Yang, 1993. "Stopping rules and estimation for recapture debugging with unequal failure rates", *Biometrika*, **80**: 192-201.

A. Chao, T.-J. Shen and W.-H. Hwang, 2006. "Application of Laplaces Boundary-mode approximations to estimate species and shared species richness", *Australian and New Zealand Journal of Statistics*, **48**: 117-128.

A. Chiarucci, N.J. Enright, G.L.W. Perry, B.P. Miller and B.B. Lamont, 2003. "Performance of nonparametric species richness estimators in a high diversity plant community", *Diversity and Distributions*, **9**: 283-295.

S. Chib and E. Greenberg, 1994. "Bayes inference for regression models with ARMA(p, q) errors", *Journal of Econometrics*, **64**: 183-206.

S. Chib and E. Greenberg, 1995. "Understanding the Metropolis-Hastings algorithm", *The American Statistician*, **49**: 327-335.

R.K. Colwell, 2009. *EstimateS 8.2 User's Guide*. University of Connecticut, Storrs, Connecticut. `http://viceroy.eeb.uconn.edu/EstimateSPages/EstSUsersGuide/EstimateSUsersGuide.htm` Last accessed on April 16, 2012.

R.K. Colwell, and J.A. Coddington, 1994. "Estimating terrestrial biodiversity through extrapolation", *Philosophical Transactions: Biological Sciences*, **345**: 101-118.

R. Condit, 1998. *Tropical Forest Census Plots*, Springer-Verlag, New York, NY, USA.

R. Condit, P.S. Ashton, P. Baker, S. Bunyavejchewin, S. Gunatilleke, N. Gunatilleke, S.P. Hubbell, R.B. Foster, A. Itoh, J.V. LaFrankie, H.S. Lee, E. Losos, N. Manokaran, R. Sukumar and T. Yamakura, 2000. "Spatial patterns in

the distribution of tropical tree species", *Science*, **288**: 1414-1418.

D.R. Cox and D.V. Hinkley, 1974. *Theoretical Statistics*, Chapman & Hall, London, UK.

N.A.C. Cressie, 1993. *Statistics for Spatial Data*, Wiley, New York, NY, USA.

J.N. Darroch, 1958. "The multiple-recapture census, I. Estimation of a closed population", *Biometrika*, **45**: 343-359.

J.N. Darroch and D. Ratcliff, 1980. "A note on capture-recapture estimation", *Biometrics*, **36**: 149-153.

O.H. Diserud and S. Engen, 2000. "A general and dynamic species abundance model, embracing the lognormal and the gamma models", *The American Naturalist*, **155**: 497-511.

R.M. Dorazio and J.A. Royle, 2005. "Estimating size and composition of biological communities by modeling the occurrence of species", *Journal of the American Statistical Association*, **100**: 389-398.

R.M. Dorazio, J.A. Royle, B. Soderstrom, and A. Glimskar, 2006. "Estimating species richness and accumulation by modeling species occurrence and detectability", *Ecology*, **87**: 842-854.

C.C.Y. Dorea and S.A. Mingoti, 2006. "Estimating the total number of distinct species using quadrat sampling and under-dependence structure", *Journal of Applied Statistics*, **33**: 497-512.

C.F. Dormann, J.M. McPherson, M.B. Araújo, R. Bivand, J. Bolliger, G. Carl, R.G. Davies, A. Hirzel, W. Jetz, W.D. Kissling, I. Kühn, R. Ohlemüller, P.R. Peres-Neto, B. Reineking, B. Schröder, F.M. Schurr, and R. Wilson, 2007. "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review", *Ecography*, **30**: 609-628.

P.K. Dunn and G.K. Smyth, 2005. "Series evaluation of Tweedie exponential dispersion models densities", *Statistics and Computing*, **15**: 267-280.

J.A. Dupuis and J. Joachim, 2006. "Bayesian estimation of species richness from quadrat sampling data in the presence of prior information", *Biometrics*, **62**: 706-712.

B. Efron and R. Thisted, 1976. "Estimating the number of unseen species: How many words did Shakespeare know?", *Biometrika*, **63**: 435-447.

J.M. Elliott, 1977. *Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates*, Freshwater Biological Station Association, Ambleside, UK.

S. Engen, 1978. *Stochastic Abundance Models*, Chapman & Hall, London, UK.

S. Engen and R. Lande, 1996a. "Population dynamic models generating the log-normal species abundance distribution", *Mathematical Biosciences*, **132**: 169-183.

S. Engen and R. Lande, 1996b. "Population dynamic models generating species abundance distributions of the gamma type", *Journal of Theoretical Biology*, **178**: 325-331.

T.S. Ferguson, 1973. "A Bayesian analysis of some nonparametric problems", *The Annals of Statistics*, **1**: 209-230.

R.A. Fisher, A.S. Corbet, and C.B. Williams, 1943. "The relation between the number of species and the number of individuals in a random sample of an animal population", *Journal of Animal Ecology*, **12**: 42-58.

C.H. Flather, 1996. "Fitting species-accumulation functions and assessing regional land use impacts on avian diversity", *Journal of Biogeography*, **23**: 155-168.

D. Gamerman and H.F. Lopes, 2006. *Markov Chain Monte Carlo*, 2^{nd} edn. CRC, New York, NY, USA.

K.J. Gaston, P.A.V. Borges, F. He and C. Gaspar, 2006. "Abundance, spatial variance and occupancy: arthropod species distribution in the Azores", *Journal of Animal Ecology*, **75**: 646-656.

A.E. Gelfand and A.F.M. Smith, 1990. "Sampling-based approaches to calculating marginal densities", *Journal of the American Statistical Association*, **85**: 398-409.

A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, 2004. *Bayesian Data Analysis*, 2^{nd} edition, CRC Press, New York, NY, USA.

A. Gelman and D.B. Rubin, 1992. "Inference from iterative simulation using multiple sequences", *Statistical Science*, **7**: 457-472.

S. Geman and D. Geman, 1984. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**: 721-724.

C.J. Geyer, 1992. "Practical Markov chain Monte Carlo (with discussion)", *Statistical Science*, **7**: 473-511.

H.A. Gleason, 1922. "On the relation between species and area", *Ecology*, **3**: 158-162.

L.A. Goodman, 1949. "On the estimation of the number of classes in a population", *Annals of Mathematical Statistics*, **20**: 572-579.

N.J. Gotelli and R.K. Colwell, 2001. "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness", *Ecology Letters*, **4**: 379-391.

J.F. Grassle and N.J. Maciolek, 1992. "Deep-sea species richness: Regional and local diversity estimates from Quantitative Bottom Samples", *The American Naturalist*, **139**: 313-341.

P. Greig-Smith, 1952. "The use of random and contiguous quadrats in the study of the structure of plant communities", *Annals of Botany*, **16**: 293-316.

P. Greig-Smith, 1983. *Quantitative Plant Ecology*, 3^{rd} edition, University of California Press, Berkeley, CA, USA.

J. Gurevitch, S.M. Scheiner, and G.A. Fox, 2002. *The Ecology of Plants*, Sinauer Associates Inc., Sunderland, MA, USA.

N.G. Hairston, 1959. "Species abundances and community organization", *Ecology*, **40**: 404-416.

B. Harris, 1959. "Determining bounds on integrals with applications to cataloging problems", *Annals of Mathematical Statistics*, **30**: 521-548.

J. Harte, A. Kinzig, and J. Green, 1999. "Self-similarity in the distribution and abundance of species", *Science*, **284**: 334-336.

M. Harwit and R. Hildebrand, 1986. "How many more discoveries in the Universe?", *Nature*, **320**: 724-726.

W.K. Hastings, 1970. "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, **57**: 97-109.

F. He and P. Legendre, 1996. "On species-area relations", *The American Naturalist*, **148**: 719-737.

F. He and P. Legendre, 2002. "Species diversity patterns derived from species-area models", *Ecology*, **83**: 1185-1198.

K.L. Heck, G. van Belle, and D. Simberloff, 1975. "Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size", *Ecology*, **56**: 1459-1461.

P.G. Hoel, S.C. Port, and C.J. Stone, 1972. *Introduction to stochastic processes*, Houghton Mifflin, Boston, MA, USA.

S.-P. Huang and B.S. Weir, 2001. "Estimating total number of alleles using a sample coverage method", *Genetics*, **159**: 1365-1373.

S.P. Hubbell, 2001. *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press, Princeton, NJ, USA.

S.P. Hubbell, R. Condit, and R.B. Foster, 2005. *Barro Colorado Forest Census Plot Data*, Smithsonian Tropical Research Institute, Centre for Tropical Forest Science. `https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci` Last accessed on April 16, 2012.

R.G. Hughes, 1986. "Theories and models of species abundance", *The American Naturalist*, **128**: 879-899.

W.-H. Hwang, and T.-J. Shen, 2010. "Small-sample estimation of species richness applied to forest communities", *Biometrics*, **66**: 1052-1060.

N.L. Johnson, S. Kotz, and A.W. Kemp, 1993. *Univariate Discrete Distributions 2nd Edition*, John Wiley & Sons, Toronto, Canada.

J.G. Kalbfleisch, 1985. *Probability and Statistical Inference*, 2^{nd} edition, Volume 2, Springer-Verlag, New York, NY, USA.

R.A. Kempton, and L.R. Taylor, 1974. "Log-Series and log-normal pa-

rameters as diversity discriminants for the Lepidoptera", *Journal of Animal Ecology*, **43**: 381-399.

R.A. Kempton and R.W.M. Wedderburn, 1978. "A comparison of three measures of species diversity", *Biometrics*, **34**: 25-37.

M. Kéry and J.A. Royle, 2008. "Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys", *Journal of Applied Ecology*, **45**: 589-598.

A.R. Kiester, 2001. "Species diversity, overview", *Encyclopedia of Biodiversity*, S.A. Levin, editor. Volume 5, pages 441-451, Academic Press, San Diego, CA, USA.

R. King, B.J.T. Morgan, O. Gimenez, and S.P. Brooks, 2010. *Bayesian Analysis for Population Ecology*, CRC Press, New York, USA.

C.J. Krebs, 1999. *Ecological Methodology*, 2^{nd} edition, Benjamin/Cummings, Menlo Park, CA, USA.

P. Legendre and L. Legendre, 1998. *Numerical Ecology*, Elsevier, New York, NY, USA.

W.A. Lewins and D.N. Joanes, 1984. "Bayesian estimation of the number of species", *Biometrics*, **40**: 323-328.

H. Li, 2008. *Bayesian Hierarchical Models for Spatial Count Data with Application to Fire Frequency in British Columbia*, M.S. thesis, University of Victoria, Victoria, BC, Canada.

Z. Lindo and N.N. Winchester, 2006. "A comparison of microarthropod assemblages with emphasis on oribatid mites in canopy suspended soils and forest floors associated with ancient western redcedar trees", *Pedobiologia*, **50**: 31-41.

Z. Lindo and N.N. Winchester, 2007. "Resident corticolous oribatid mites (Acari: Oribatida): Decay in community similarity with vertical distance from the ground", *Écoscience*, **14**: 223-229.

W.A. Link, 2003. "Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities", *Biometrics*, **59**: 1123-1130.

J.S. Liu, W.H. Wong, and A. Kong, 1994. "Correlation structure and

convergence rate of the Gibbs sampler: applications to the comparison of estimators and augmentation schemes", *Biometrika*, **81**: 27-40.

D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, 2000. "WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility", *Statistics and Computing*, **10**: 325-337.

M. Luoto, R. Virkkala, R.K. Heikkinen and K. Rainio, 2004. "Predicting bird species richness using remote sensing in boreal agricultural-forest mosaics", *Ecological Applications*, **14**: 1946-1962.

G. Mace, H. Masundire, J. Baillie, 2005. Chapter 4 of *Ecosystems and Human Well-being: Current State and Trends*, Volume 1, R. Hassan, R. Scholes and N. Ash, editors. Island Press, Washington, DC. `http://www.millenniumassessment.org/en/index.aspx`
Last accessed on April 16, 2012.

C.X. Mao and R.K. Colwell, 2005. "Estimation of species richness: mixture models, the role of rare species, and inferential challenges", *Ecology*, **86**: 1143-1153.

C.X. Mao, R.K. Colwell, and J. Chang, 2005. "Estimating the species accumulation curve using mixtures", *Biometrics*, **61**: 433-441.

C.X. Mao and B.G. Lindsay, 2004. "Estimating the number of classes in multiple populations: a geometric analysis", *The Canadian Journal of Statistics*, **32**: 303-314.

C.X. Mao and B.G. Lindsay, 2007. "Estimating the number of classes", *The Annals of Statistics*, **35**: 917-930.

D.C. Martin and S.K. Katti, 1965. "Fitting of certain contagious distributions to some available data by the maximum likelihood method", *Biometrics*, **21**: 34-48.

R.M. May, 1975. "Patterns of species abundance and diversity", *Ecology and evolution of communities*, pp 81-120, M.L. Cody and J.M. Diamon, editors. Belknap Press of Harvard University Press, Cambridge, MA, USA.

R.M. May, 1992. "How many species inhabit the earth?", *Scientific American*, **267**: 42-48.

R.M. May, J.H. Lawton, and N.E. Stork, 1995. Chapter 1 of *Extinction Rates*, J.H. Lawton and R. M. May, editors. Oxford University Press, New York,

NY, USA.

B.J. McGill, B.A. Maurer, and M.D. Weiser, 2006. "Empirical evaluation of neutral theory", *Ecology*, **87**: 1411-1423.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, 1953. "Equation of state calculations by fast computing machine", *Journal of Chemical Physics*, **21**: 1087-1091.

P. Müller, 1993. *A generic approach to posterior integration and gibbs sampling*, Technical Report No. 91-09, Purdue University, West Lafayette, IN, USA. `http://people.ee.duke.edu/~lcarin/tr91-09.pdf` Last accessed on April 16, 2012.

N.M. Nadkarni and J.T. Longino, 1990. "Invertebrates in canopy and ground organic matter in a neotropical montane forest, Costa Rica", *Biotropica*, **22**: 286-289.

T.K. Nayak, 1991. "Estimating the number of component processes of a superimposed process", *Biometrika*, **78**: 75-81.

J. Neyman, 1939. "On a new class of 'contagious' distributions applicable in entomology and bacteriology", *Annals of Mathematical Statistics*, **10**: 35-57.

J.L. Norris and K.H. Pollock, 1998. "Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species", *Environmental and Ecological Statistics*, **5**: 391-402.

R.B. O'Hara, 2005. "Species richness estimators: how many species can dance on the head of a pin?", *Journal of Animal Ecology*, **74**: 375-386.

J.K. Ord and G.A. Whitmore, 1986. "The Poisson-inverse Gaussian distribution as a model for species abundance", *Communications in Statistics - Theory and Methods*, **15**: 853-871.

D.L. Otis, K.P. Burnham, G.C. White, and D.R. Anderson, 1978. *Statistical inference from capture data on closed animal populations*, Wildlife Society, Washington, USA.

M.W. Palmer, 1990. "The estimation of species richness by extrapolation", *Ecology*, **71**: 1195-1198.

H.-Y. Pan, A. Chao, and W. Foissner, 2009. "A nonparametric lower bound for

the mumber of species shared by multiple communities", *Journal of Agricultural, Biological, and Environmental Statistics*, **14**: 452-468.

N. Picard, M. Karembé, and P. Birnbaum, 2004. "Species-area curve and spatial pattern", *Écoscience*, **11**: 45-54.

H. Petersen and M. Luxton, 1982. "A comparative analysis of soil fauna populations and their role in decomposition processes", *Oikos*, **39**: 287-388.

E.C. Pielou, 1957. "The effect of quadrat size on the estimation of the parameters of Neyman's and Thomas's distributions", *Journal of Ecology*, **45**: 31-47.

J.B. Plotkin and H.C. Muller-Landau, 2002. "Sampling the species composition of a landscape", *Ecology*, **83**: 3344-3356.

J. Plotkin, M. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. Ashton, 2000. "Species-area curves, spatial aggregation, and habitat specialization in tropical forests", *Journal of Theoretical Biology*, **207**: 81-99.

M. Plummer, 2003. "JAGS: A program for analysis of bayesian graphical models using Gibbs sampling", *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Technische Universität Wien, Vienna, Austria.

G. Pólya, 1930. "Sur quelques points de la theorie des probabilites", *Annales de l'institut Henri Poincaré*, **1**: 117-162.

F.W. Preston, 1948. "The commonness and rarity of species", *Ecology*, **29**: 254-283.

R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org` Last accessed on April 16, 2012.

A.E. Raftery and S. Lewis, 1992. "How many iterations in the Gibbs sampler?", pages 763-773 in *Bayesian Statistics 4*, editors J.M. Bernardo *et al.*, Oxford University Press, Oxford, UK.

B.D. Ripley, 1981. *Spatial Statistics*, Wiley, New York, NY, USA.

C.P. Robert, 2001. *The Bayesian Choice*, 2^{nd} edition, Springer-Verlag, New York, NY, USA.

C.R. Robert and G. Casella, 2004. *Monte Carlo Statistical Methods*, 2^{nd} edition, Springer, New York, NY, USA.

G.O. Roberts, A. Gelman, and W.R. Gilks, 1994. *Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms*, Technical Report, University of Cambridge, Cambridge, UK.

J. Rodrigues, L.A. Milan, and J.G. Leite, 2001. "Hierarchical Bayesian estimation for the number of species", *Biometrical Journal*, **43**: 737-746.

J.A. Royle, R.M. Dorazio, and W.A. Link, 2007. "Analysis of multinomial models with unknown index using data augmentation", *Journal of Computational and Graphical Statistics*, **16**: 67-85.

L. Sanathanan, 1977. "Estimating the size of a truncated sample", *Journal of the American Statistical Association*, **72**: 669-672.

H.S. Sichel, 1975. "On a distribution law for word frequencies", *Journal of the American Statistical Association*, **70**: 542-547.

H.S. Sichel, 1986. "Parameter estimation for a word frequency distribution based on occupancy theory", *Communications in Statistics - Theory and Methods*, **15**: 935-949.

H.S. Sichel, 1997. "Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution", *South African Statististical Journal*, **31**: 13-37.

A.R. Solow, 1994. "On the Bayesian estimation of the number of the species in a community", *Ecology*, **75**: 2139-2142.

A.R. Solow, 1996. "Estimating the size of the source population from a matched sample of parts", *Mathematical Geology*, **28**: 783-789.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, 2002. "Bayesian measures of model complexity and fit (with discussion)", *Journal of the Royal Statistical Society, B*, **64**: 583-639.

A.J. Stam, 1987. "Statistical problems in ancient numismatics", *Statistica Neerlandica*, **41**: 151-173.

D. Stoyan and H. Stoyan, 1994. *Fractals, Random Shapes and Point Fields*,

John Wiley & Sons, Chichester, USA.

V. St-Louis, A.M. Pidgeon, M.K. Clayton, B.A. Locke, D. Bash, and V.C. Radeloff, 2009. "Satellite image texture and a vegetation index predict avian biodiversity in the Chihuahuan Desert of New Mexico", *Ecography*, **32**: 468-480.

L.R. Taylor, 1961. "Aggregation, variance and the mean", *Nature*, **189**: 732-735.

L.R. Taylor, I.P. Woiwod, and J.N. Perry, 1978. "The density-dependence of spatial behaviour and the rarity of randomness", *Journal of Animal Ecology*, **47**: 383-406.

R. Thisted and B. Efron, 1987. "Did Shakespeare write a newly-discovered poem?", *Biometrika*, **74**: 445-455.

M. Thomas, 1949. "A generalization of Poisson's binomial limit for use in ecology", *Biometrika*, **36**: 18-25.

L. Tierney, 1994. "Markov chains for exploring posterior distributions (with discussion)", *Annuals of Statistics*, **22**: 1701-1762.

K.I. Ugland, J.S. Gray, and K.E. Ellingsen, 2003. "The speciesaccumulation curve and estimation of species richness", *Journal of Animal Ecology*, **72**: 888-897.

USGS Patuxent Wildlife Research Center, 2010. *North American Breeding Bird Survey* website, Laurel, MD, USA. `http://www.pwrc.usgs.gov/bbs/` Last accessed on April 16, 2012.

I. Volkov, J.R. Banavar, S.P. Hubbell, and A. Maritan, 2003. "Neutral theory and relative species abundance in ecology", *Nature*, **424**: 1035-1037.

B.A. Walther and J.L. Moore, 2005. "The concepts of bias, and precision and accuracy, and their use in performance of species richness estimators, with a literature review of estimator performance", *Ecography*, **28**: 815-829.

M. Whiteside and M. Eakin, 2004. "A better estimate for the number of signatures on a petition", *Proceedings of the American Statistical Association*, Toronto, Canada.

B.K. Williams, J.D. Nichols, and M.J. Conroy, 2002. *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA, USA.

M. Williamson and K.J. Gaston, 2005. "Lognormal distribution is not appropriate null hypothesis for the species-abundance distribution", *Journal of Animal Ecology*, **74**: 409-422.

P. Yip, 1991. "A method of inference for a capture-recapture experiment in discrete time with variable capture probabilities", *Communications in Statistics – Stochastic Models*, **7**: 343-362.

H. Zhang and H. Stern, 2005. "Investigation of a generalized multinomial model for species data", *Journal of Statistical Computation and Simulation*, **75**: 347-362.

Appendix A

Checking Assumptions for Species Richness Estimation in the Case Study

Our multi-site abundance-based models from Chapters 3 and 4 are developed based on several assumptions. The estimators $\hat{S}_{ACE}$, $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ from the simulation studies of Chapter 5 make some of the same assumptions. Here, we examine the plausibility of these assumptions for the oribatid mite data used in Chapter 6.

A.1 Uniform sampling effort

The cores have a uniform area of 15 cm^2 each. However, the weights of the substrate in the cores vary widely (e.g., dry weights: mean = 20.6 g, SD = 10.3 g, min = 1.5

g, max = 66 g). In Section 6.3.1, a positive association was identified between the substrate weights and the number of adult mites in the canopy cores. Therefore, we cannot expect the numbers of individuals in the canopy cores to be close in value if the substrate weights are very different. We have modified our Bayesian model to account for this in Section 6.5.

Estimators $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ assume that a species has the same probability of occurrence in each sampling unit. This assumption is unlikely to be true for the canopy data set because the probability of occurrence of a species will tend to be positively associated with the weight of the substrate. Consequently, $\hat{S}_{Chao2}$ and $\hat{S}_{Jack2}$ should be interpreted with care as these estimates may not be reliable in this context.

A.2 Sampling units contain independent sets of observations

Our methods assume that the set of observations in a sampling unit is *independent* of the observations in other sampling units. As mentioned earlier, the locations of the eight sites were not randomly chosen. However, they were chosen to represent the variety of the rainforest environment within the SPLA (Basset *et al.*, 2007). The eight sites each have a land area of 20 m $\times$ 20 m. So, the total ground surface area occupied by the sites is 0.0004 km^2 which is approximately 5.46×10^{-6}% of the 73.20 km^2 of lowland tropical rainforest in the SLPA. We can surmise that the total

surface area of the canopy cores from the eight sites is also a very small fraction of the total surface area of the rainforest's canopy. Therefore, although the substrate is sampled without replacement, this will not hinder us from treating the observations from different sites as being independent as only a very small fraction of the region is sampled without replacement.

For the eight cores of ground substrate collected near the base of a tree, it is not appropriate to treat the sets of observations in the cores as independent due to their close proximity. For the same reason, it is not appropriate to treat the sets of observations in the eight cores from the canopy of a tree as independent. We thus pool the ground observations from a tree, and treat the eight ground cores together as one sampling unit. This is also done for the canopy cores in a tree. As a result, in the canopy, we have one sampling unit for each tree giving us a total of 24 canopy sampling units. Likewise, we have 24 ground sampling units from the forest floor.

We investigate two factors that may introduce some dependence between sets of observations from sampling units.

First, there are instances where the same host species of tree is sampled more than once. Of the 24 trees in the field study, there are 18 distinct tree species, with two species each sampled twice and two species each sampled three times. We investigate if trees of the same species have a lower degree of dissimilarity in mite composition than observations taken from trees of different species.

For the forest floor, a lower average dissimilarity in mite composition is observed among pairs of ground sampling units from trees of the same species: the average

value of D is 48.6% (SD = 8.2%); whereas, for pairs of sampling units from trees of different species, the average value of D is 58.5% (SD = 10.4%); however, there are only eight pairs involving trees of the same species for this comparison. For the canopy, the average dissimilarity in mite composition among sampling units from trees of the same species is comparable to the average dissimilarity among sampling units from trees of different species: for pairs from the same tree species, the average of D is 75.1% (SD = 11.8%); for pairs from different tree species, the average of D is 73.9% (SD = 8.1%). We acknowledge that the species types of the host trees may introduce a dependence among observations in the sampling units, in particular, for the forest floor.

The second factor we investigate is spatial autocorrelation. Combining all 24 cores from the canopy of a site, the dissimilarity index D is computed for each pair of sites. No significant association is found between the values of D and the distance between sites (Mantel statistic $r = 0.019$, p-value = 0.448). Doing the same for the ground data, no significant spatial autocorrelation is found for D among the forest floor observations of the sites (Mantel statistic $r = 0.056$, p-value = 0.384). With no significant spatial autocorrelation detected among sites, we will treat the eight sites as containing independent sets of observations, for both the canopy and the forest floor.

A.3 Closed Population

The 384 cores were sampled from the lowland tropical rainforest over a period of two months (September and October) in 2003. Given the relatively limited dispersal range of oribatid mites and the relatively brief period of sampling that was free of major disturbances to the rainforest, it is plausible that few, if any, new species entered the rainforest, and few, if any, existing species were extirpated during the period of sampling. We will assume the rainforest did not gain or lose mite species during the period of sampling.

A.4 100% Probability of Detection and Correct Identification

Estimators used in the case study of Chapter 6 assume all individuals in the cores are detected and correctly identified. As described below, there are indeed violations of these assumptions for the oribatid mite data set.

The size of the adult mites did not prevent the identification of specimens to their species type. However, the unavailability of a taxonomist familiar with the *Galumnidae* family of oribatid mites meant 1,021 adult mites were not classified to their genus and species type. It is believed that the 1,021 adult mites belonged to at least eight species in the *Galumnidae* family. In Section 6.6, we consider a way to include the observations from the *Galumnidae* family in the species richness estimates

for the canopy and forest floor.

Immature mites (mites in stages of development before the adult stage) were excluded from consideration. Although this is standard practice due to the difficulty classifying the immatures to their species type (Basset *et al.*, 2007), it means some species may have been present in the sampled substrate, but were not recorded as their representatives were immature specimens. However, based on the life history of the oribatid mites, (Dr. Neville Winchester, 2012, personal communication) immature mites are likely represented by adults in the sampled substrate, and it is unlikely a significant number of unrecorded species may be in the immature specimens.

Nevertheless, it may be the case that some species have gone undetected as they only had immatures in the sampled substrate. Failure to detect and/or correctly identify species would introduce bias.

A.5 Species are mutually independent

The assumption of mutual independence is unlikely to hold true, as factors such as inter-species interactions and characteristics of the sites and micro-habitats can influence the spatial distributions and abundances of multiple species. The lack of mutual independence among the species will introduce bias in the species richness estimation.

A.6 Spatial distributions of species are stationary

In our models, the number of individuals of a species in a sampling unit has the same probability distribution for all sampling units randomly placed in the region.

This assumption of stationarity in the distribution of a species across the region is unrealistic for the mite communities in the canopy and forest floor. In particular, during the investigation into the dissimilarity of the mite composition among sites, we noticed higher dissimilarity index values for pairs of sites that involved the site located on a floodplain (the other seven sites were located on a hillside). This suggests the spatial distribution of species may be different in the site located on the floodplain.

Appendix B

Bayesian Inference and Computation

Statistical inference concerns the learning of some unknown aspect of the population from which the data were drawn. Bayesian inference fits a probability model to observed data and summarizes the result through a probability distribution on the unknown parameters $\boldsymbol{\theta}$ or unobserved data $\tilde{y}$ we are interested in. In other words, Bayesian inferences are made in terms of probability statements conditional on the observed data $\boldsymbol{y}$. The Bayesian method offers potentially attractive advantages over the frequentist statistical approach for modeling spatial data (Banerjee *et al.*, 2004). The maximum likelihood method can be complex in multidimensional and constrained settings even though some numerical procedures such as EM algorithms have been introduced to handle this. Under the Bayesian setting, computational challenges associated with computing posterior distributions can be overcome by applying Markov

Chain Monte Carlo methods which will be introduced in Section B.2.

B.1 Framework for Bayesian Inference

A Bayesian statistical model is made of a sampling distribution (likelihood function), $P(\boldsymbol{y} \mid \boldsymbol{\theta})$, for observed data conditional on unknowns $\boldsymbol{\theta}$, and a prior distribution, $p(\boldsymbol{\theta})$, that reflects various degrees of belief on the likely values of unknowns (Robert, 2001). Given these two distributions, the joint distribution or full probability model can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = P(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$$

and the posterior distribution is obtained via Bayes' rule (Gelman *et al.*, 2004)

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{P(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{B.1}$$

where $p(\boldsymbol{y}) = \sum_{\boldsymbol{\theta}} P(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ in the case of discrete $\boldsymbol{\theta}$ and $p(\boldsymbol{y})$
$= \int L(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ in the case of continuous $\boldsymbol{\theta}$. Since $p(\boldsymbol{y})$ does not depend on $\boldsymbol{\theta}$, it can be considered as a constant with fixed $\boldsymbol{y}$. Therefore, (B.1) can be obtained up to normalizing constant as,

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto P(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{B.2}$$

that is proportional to the likelihood function times the prior. Using numerical methods described in the next section, we can work with (B.2) for model estimation and avoid computing the normalizing constant which is not easily obtained.

Complex models can be built through the specification of several simple stages under a Bayesian hierarchical framework. A hierarchical Bayes model (Robert, 2001)

is a Bayesian statistical model where the prior distribution $p(\boldsymbol{\theta})$ is decomposed into several conditional levels of distributions

$$p_1(\boldsymbol{\theta} \mid \boldsymbol{\theta_1}),\ p_2(\boldsymbol{\theta_1} \mid \boldsymbol{\theta_2}),\ ...,\ p_n(\boldsymbol{\theta_{n-1}} \mid \boldsymbol{\theta_n})$$

and a marginal distribution

$$p_{n+1}(\boldsymbol{\theta_n})$$

such that

$$p(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta_1} \times ... \times \boldsymbol{\Theta_n}} p_1(\boldsymbol{\theta} \mid \boldsymbol{\theta_1}) \cdots p_n(\boldsymbol{\theta_{n-1}} \mid \boldsymbol{\theta_n}) p_{n+1}(\boldsymbol{\theta_n}) d\boldsymbol{\theta_1} \cdots \boldsymbol{\theta_n}.$$

The parameters $\boldsymbol{\theta_i}$ are called hyper-parameters of level i, for $1 \leq i \leq n$. The most common hierarchical Bayesian model is the case when $n = 2$. At the first stage, a distribution for the data given parameters is specified. At the second stage, prior distributions for parameters given hyper-parameters are specified and distributions for hyper-parameters are specified at the third stage. The prior distributions at the first level may correspond to the structural information about the model such as uncertain linear restrictions on the parameters of a regression model, whereas the prior distributions at the second level correspond to the more subjective information that accounts for the imprecision of these restrictions. The hierarchical modeling improves the robustness of the resulting Bayes estimators, since uncertainty regarding the model structure can be incorporated into additional prior distributions. In addition, the decomposition of a prior distribution into its components in the hierarchical Bayes approach simplifies Bayesian calculations and allows for an easier approximation of some posterior quantities by simulation.

The choice of prior distribution is critical for Bayesian inference (Gelman *et al.*, 2004). If prior information is available from external knowledge, this information can be used to construct a prior distribution for unknowns. The mechanism for converting prior information to prior probability distributions is often unclear; moreover, prior information will typically not induce a unique prior distribution. Often, there is little prior information regarding model unknowns, in which case a noninformative or vague prior distribution can be employed. Such priors typically arise in the form of a parametric distribution with large or infinite variance. For large data sets, this approach is reasonable as the likelihood will dominate the prior, and inference will be primarily data-driven. For small data sets, this approach is not reasonable and inference will be sensitive to prior choice.

When the posterior distribution follows the same parametric form as the prior distribution, the prior is called a conjugate prior. Probability distributions that belong to the exponential family of distributions always have conjugate prior distributions (Robert, 2001). Suppose $f(y_i \mid \theta)$'s are from the exponential family of distributions and have form

$$f(y_i \mid \theta) = s(y_i)t(\theta)e^{a(y_i)b(\theta)} \tag{B.3}$$

for $i = 1, ..., n$. The likelihood function for a random sample is then

$$L(\boldsymbol{y}; \theta) = \phi(\boldsymbol{y})t(\theta)^n \exp(w(\boldsymbol{y})b(\theta)), \tag{B.4}$$

where

$$\phi(\boldsymbol{y}) = \prod_{i=1}^{n} s(y_i) \qquad \text{and} \qquad w(\boldsymbol{y}) = \sum_{i=1}^{n} a(y_i).$$

If the prior distribution of θ is specified as

$$p(\theta) \propto t(\theta)^{\eta}\exp(b(\theta)v),$$

then the posterior distribution is

$$p(\theta \mid \boldsymbol{y}) \propto t(\theta)^{n+\eta}\exp(b(\theta)(w(\boldsymbol{y}) + v))$$

which has the same density form as the prior distribution. This choice of prior density is conjugate and is often called the natural conjugate prior. For example, the Gamma(α, β) distribution with p.d.f.

$$\begin{aligned} p(\theta) &= \tfrac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \\ &\propto \theta^{\alpha-1}\exp(-\beta\theta) \end{aligned}$$

is a natural conjugate prior for a Poisson distribution in the form

$$f(x \mid \theta) = \frac{\theta^{x}\exp(-\theta)}{x!}$$

In this case, the posterior distribution can be written as

$$\begin{aligned} p(\theta \mid x) &\propto f(x \mid \theta) \times p(\theta) \\ &\propto \theta^{x}\exp(-\theta)\theta^{\alpha-1}\exp(-\beta\theta) \\ &\propto \theta^{x+\alpha-1}\exp(-\theta(1+\beta)), \end{aligned}$$

which is the kernel of Gamma$(x + \alpha, 1 + \beta)$ distribution. Table B.1 lists natural conjugate priors for some common exponential families. Conjugate priors can be convenient to work with and can simplify computation as we shall see in the next section which discusses Bayesian computation through the Metropolis-Hastings algorithm.

Table B.1: Natural conjugate priors for some common exponential families

$f(x \mid \theta)$	$p(\theta)$	$p(\theta \mid x)$
Normal(θ, σ^2)	Normal(μ, τ^2)	Normal$(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
Poisson(θ)	Gamma(α, β)	Gamma$(\alpha + x, \beta + 1)$
Gamma(ν, θ)	Gamma(α, β)	Gamma$(\alpha + \nu, \beta + x)$
Binomial(n, θ)	Beta(α, β)	Beta$(\alpha + x, \beta + n - x)$
NegBin(m, θ)	Beta(α, β)	Beta$(\alpha + m, \beta + x)$
Multinomial$(\theta_1, ..., \theta_k)$	Dirichlet$(\alpha_1, ..., \alpha_k)$	Dirichlet$(\alpha_1 + x_1, ..., \alpha_k + x_k)$
Normal$(\mu, 1/\theta)$	Gamma(α, β)	Gamma$(\alpha + 0.5, \beta + (\mu - x)^2/2)$

B.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a general numerical framework for generating dependent realizations from possibly high dimensional probability distributions. This framework is used for Bayesian inference where it is of interest to summarize the posterior distribution (B.1) and the corresponding normalizing constant can not be obtained analytically. In what follows we review some basic theory related to Markov chains, in particular the limit theorems which justify the use of MCMC in practice. We will then discuss the Gibbs sampler and Metropolis-Hastings algorithms, illustrating these algorithms with some simple examples. Finally, practical implementation issues are discussed.

In general, a stochastic process (Gamerman and Lopes, 2006) is defined as a

collection of random quantities denoted $\theta^{(t)} \in S$ where $t \in T$. The index set T takes nonnegative integers and S is known as the state space. For simplicity of presentation, we will start by assuming S is discrete. We will then briefly discuss results for general state spaces. A Markov chain is a special type of stochastic process where the past and future states are conditionally independent given the current state. This property can be stated as

$$Pr(\theta^{(n+1)} \in A \mid \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, ..., \theta^{(0)} \in A_0) = Pr(\theta^{(n+1)} \in A \mid \theta^{(n)} = x) \tag{B.5}$$

for sets $A, A_{n-1}, ..., A_0 \subset S$ and $x \in S$. In the case of homogeneous Markov chain where (B.5) does not depend on n, a transition kernel $P(x, A)$ can be defined as:

1. $P(x, .)$ is a probability distribution over S for all $x \in S$;

2. the function $x \mapsto P(x, A)$ can be evaluated for all $A \subset S$.

When dealing with discrete state space, we often work with the set A of the form $A = \{y\}$ and the transition probability $P(x, \{y\}) = P(x, y)$ is defined as:

1. $P(x, y) \geq 0$ for $\forall x, y \in S$;

2. $\sum_{y \in S} P(x, y) = 1$ for $\forall x \in S$.

For a discrete state space S with r elements, a $r \times r$ transition matrix P can be

established as

$$P = \begin{bmatrix} P(x_1, x_1) & \cdots & P(x_1, x_r) \\ \vdots & & \vdots \\ P(x_r, x_1) & \cdots & P(x_r, x_r) \end{bmatrix}.$$

The transition probability from state x to state y over m steps can be obtained as the matrix product of P m-times denoted as P^m (Gamerman and Lopes, 2006). We let $\pi^{(n)}$ with components $\pi_n(x_i) = Pr(\theta^{(n)} = x_i)$ denote a row vector containing marginal probabilities associated with $\theta^{(n)}$. The recursive relationship between successive marginal distributions of the chain can be written as $\pi^{(n)} = \pi^{(0)}P^{n-1}P = \pi^{(n-1)}P$. We let ρ_{xy} denote the probability of the chain, starting from state x, eventually reaching state y. A state $y \in S$ is said to be recurrent if $\rho_{yy} = 1$ and is said to be transient if $\rho_{yy} < 1$. For a recurrent state y, if $\mathrm{E}[T_y \mid \theta^{(0)}] < \infty$ where $T_y = \min\{n \geq 1 : \theta^n = y\}$ denotes the hitting time of y, the state y is said to be positive recurrent which is an important property for establishing limiting results. Within the context of iterative simulation algorithms, asymptotic behavior of the chain as the number of iterations $n \to \infty$ is the most important area of the Markov chain theory. A distribution π is said to be a stationary distribution of a chain with transition probabilities $P(x, y)$ if

$$\sum_{x \in S} \pi(x)P(x, y) = \pi(y) \qquad \text{for} \qquad \forall y \in S \tag{B.6}$$

which in matrix form is $\pi = \pi P$. If the stationary distribution π exists and

$$\lim_{n\to\infty} P^{(n)}(x, y) = \pi(y),$$

then the sequence of marginal distributions $\pi^{(n)}$ will approach π as $n \to \infty$, independently of the initial distribution of the chain. In this sense, π is also referred to

as the limiting distribution. There are situations where stationary distributions exist but limiting distributions do not (Gamerman and Lopes, 2006). In order to establish limiting results, we will introduce the notion of periodicity. The period of a state x is the largest common divisor of the set $\{n \geq 1 : P^{(n)}(x, x) > 0\}$ denoted by d_x. A state is said to be ergodic if the state is positive recurrent and aperiodic ($d_x = 1$). In addition, a chain is ergodic if all its states are ergodic. Given this, an important limiting theorem, the ergodic theorem, can be stated based on the ergodicity of the chain. Suppose $\theta^{(n)}$ is ergodic with stationary distribution π and $t(\theta)$ is a real valued function with $\mathrm{E}_\pi[t(\theta)] < \infty$. Then the ergodic average

$$\bar{t}_n = (1/n)\sum_{i=1}^{n} t(\theta^{(i)}) \xrightarrow{a.s} \mathrm{E}_\pi[t(\theta)] \qquad \text{as} \qquad n \to \infty.$$

This is the Markov chain equivalent to the strong law of large numbers and it is this theorem that justifies the use of MCMC for estimating expectations taken with respect to the posterior distribution for Bayesian inference.

In practice, when using Markov Chain simulation to fit statistical models in a Bayesian framework, the state space S corresponds to a parameter space that in general will not be a discrete set. Nevertheless, the ergodic theorem described above can be extended and applied to more general state spaces. Assuming S is a continuous state space, the transition kernel is defined through a conditional probability density function

$$p(x, y) = \frac{\partial P(x, y)}{\partial y}$$

where

$$P(x,y) = Pr(\theta^{(n+1)} \leq y \mid \theta^{(n)} = x) = Pr(\theta^{(1)} \leq y \mid \theta^{(0)} = x), \qquad \text{for } x, y \in S.$$

Then the continuous version of (B.6) can be written as

$$\pi(y) = \int_{-\infty}^{\infty} \pi(x)p(x,y)dx \tag{B.7}$$

where π is the stationary distribution of the chain. With these definitions, the limiting results considered in the discrete case will carry over to the continuous case, though, a thorough technical presentation of these general results is beyond the scope of this thesis.

The key to MCMC simulation for Bayesian inference is to simulate realizations $\boldsymbol{\theta^{(0)}}$, $\boldsymbol{\theta^{(1)}}$, $\boldsymbol{\theta^{(2)}}$, ... from an ergodic Markov chain whose stationary distribution is the posterior distribution of interest. Starting from an initial state $\boldsymbol{\theta^{(0)}}$, realizations of the chain are produced successively until the chain 'forgets' this initial state and begins to exhibit its steady state behavior. If the chain reaches approximate stationarity at iteration T, the set of sampled values, $\boldsymbol{\theta^{(0)}}$, ..., $\boldsymbol{\theta^{(T)}}$, is discarded as a 'burn-in' period and successive realizations $\boldsymbol{\theta^{(T+1)}}$, $\boldsymbol{\theta^{(T+2)}}$, $\boldsymbol{\theta^{(T+3)}}$, ... are approximate draws from the posterior distribution (which is the stationary distribution of the Markov chain). Bayes inference can be based on summarizing the posterior distribution using a Monte Carlo sample of size J draws after the burn-in period. For a given Bayesian inference problem, there are many ways to construct the required Markov chain. We will introduce the two most widely used MCMC algorithms, the Gibbs sampler and the Metropolis-Hastings algorithm, in the next two sections. We will then, in Section

B.3, discuss convergence diagnostics for Markov chain simulation.

B.2.1 Gibbs Sampling

Gibbs sampling is a useful MCMC simulation scheme that samples iteratively from the full conditional distribution of each parameter, given all the other parameters and the data. The original work is presented by Geman and Geman (1984) within the context of image analysis but has since been applied in far more general contexts. Gelfand and Smith (1990), in a landmark paper, present many ways of applying the Gibbs sampler to a wide variety of Bayesian inference problems. Suppose our posterior distribution $[\boldsymbol{\theta}|\boldsymbol{y}]$ is k-dimensional, where $\boldsymbol{y}$ denotes the observed data. For any component θ_i of $\boldsymbol{\theta}$, the full conditional distribution is defined as $[\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k, \boldsymbol{y}] = [\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{y}]$. Given this, Gibbs sampling is preformed according to the following iterative scheme:

1. Set iteration counter t to 1 and choose a starting point for $\boldsymbol{\theta}$ in the parameter space; that is, set

$$\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_k^{(0)})^T;$$

2. At iteration t, draw a new value $\boldsymbol{\theta}^{(t)}$ by successively generating each component of $\boldsymbol{\theta}$ from its corresponding full conditional distribution:

$$[\theta_i^{(t)}|\theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, ..., \theta_k^{(t-1)}, \boldsymbol{y}];$$

   A single 'sweep' of the Gibbs sampler consists of a cycle through all k components of $\boldsymbol{\theta}$.

3. Change iteration counter from t to $t+1$ and repeat steps 2 and 3, to produce successive values of the Markov chain.

Note that Gibbs sampling reduces the problem of simulating from a high dimensional distribution to the problem of simulating from a sequence of lower dimensional (usually scalar) distributions.

To illustrate the workings of the Gibbs sampler, we demonstrate using a simple case of two Bernoulli random variables X and Y (Casella and George, 1992). Suppose the joint probability function is

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix} \quad \text{with } p_1+p_2+p_3+p_4 = 1.$$

Then the conditional distributions of $f(X \mid Y = y)$ and $f(Y \mid X = x)$ can be easily calculated and are summarized with two matrices

$$A_{y|x} = \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix}$$

and

$$A_{x|y} = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix},$$

which, in the context of the Gibbs sampler, can be viewed as the transition matrices that give the probabilities of getting to y state from x and vice versa. Suppose we are interested in the marginal distribution of X which is given by

$$f_x = [f_x(0)\ f_x(1)] = [p_1+p_3,\ p_2+p_4]. \tag{B.8}$$

Instead of generating samples from f_x, the Gibbs sampler generates a sequence of random variables

$$Y_0',\ X_0',\ Y_1',\ X_1',\ ...,\ Y_k',\ X_k' \tag{B.9}$$

where $X_j' \sim f(x \mid Y_j' = y_j')$ and $Y_{j+1}' \sim f(y \mid X_j' = x_j')$. The transition matrix for the X' sequence in (B.9) is $A_{x|x} = A_{y|x}A_{x|y}$. Furthermore, if we denote the marginal distribution of X_k' as

$$f_k = [f_k(0),\ f_k(1)],$$

then for any k,

$$f_k = f_{k-1}A_{x|x} = f_0 A_{x|x}^k$$

where f_0 is the initial probability. Hoel *et al.* (1972) showed that as long as all the entries of $A_{x|x}$ are positive, f_k will converge to the unique stationary distribution f that satisfies

$$fA_{x|x} = f \tag{B.10}$$

regardless of f_0. The marginal distribution f_x defined as (B.8) satisfies (B.10), that is

$$f_x A_{x|x} = f_x A_{y|x} A_{x|y} = f_x.$$

Therefore, for large k, the distribution f_k, is approximately f_x, and thus we see, through this simple example, how the Gibbs sampler generates approximate, dependent realizations from a pre-specified target distribution f_x. A general proof that the Gibbs sampler produces the required ergodic Markov chain under very general conditions can be found in Robert and Casella (2004).

Once a Gibbs sampler has been implemented, there are several ways to form a sample size M from the desired distribution (posterior distribution). The first approach is the independent sampling procedure (Gelfand and Smith, 1990) that generates M independent Gibbs sequences until convergence, say after k iterations, and uses the final value X'_k in each of the sequences to form the sample. This approach requires Mk generations, but provides a sample with independent values, provided the M chains are initialized independently. The second approach advocated by Geyer (1992) considers a long single chain. A sample of size M can be formed by M successive values from the chain after reaching convergence at iteration k. This generation method requires $k + M$ iterations; however, if the chain's autocorrelation is too high, it may take too long for a single chain to adequately cover the entire parameter space appropriately. That is, the effective sample size may be far less than M due to the high autocorrelation. A third approach takes every l^{th} iteration after the burn-in period (Raftery and Lewis, 1992), thereby thinning the chain and reducing autocorrelation between those values of the chain that are recorded. This approach requires $k + lM$ iterations. It reduces the autocorrelation between sampled values and is advantageous if computer storage of values is limited. Combinations of these approaches are also adapted and methods of assessing convergence will be discussed in detail in Section B.3. Examples 1 and 2 serve to further illustrate how the Gibbs sampler works, again, considering only simple settings for now.

Example 1 (Casella and George, 1992): Suppose the joint distribution of X and Y is given by

$$f(x,y) \propto \left(\begin{array}{c} n \\ x \end{array} \right) y^{x+\alpha-1}(1-y)^{n-x+\beta-1},\ x = 0,1,...,n \text{ and } 0 \leq y \leq 1.$$

This is a non-standard bivariate distribution; however, the full conditional distributions are easily recovered as

$$f(x \mid y) \sim \text{Binomial}(n, y) \tag{B.11}$$

$$f(y \mid x) \sim \text{Beta}(x+\alpha, n-x+\beta). \tag{B.12}$$

If we are interested in the marginal distribution $f(x)$ of X, the Gibbs sequence (B.9) can be iteratively generated by simulations from (B.11) and (B.12). In this example, $M = 500$ parallel chains are produced and the 10^{th} value from each chain is used to form the sample. Figure B.1 displays the histogram obtained from the Gibbs sampling output with $n = 16$, $\alpha = 2$ and $\beta = 4$. The solid line represents the density of the true marginal distribution of $f(x) = \int f(x,y)dy$, which can be shown to have a Beta-Binomial form

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)} \qquad x = 0,1,...,n.$$

It is apparent that samples generated using the Gibbs sampler recover the properties of the true marginal distribution very well in this example.

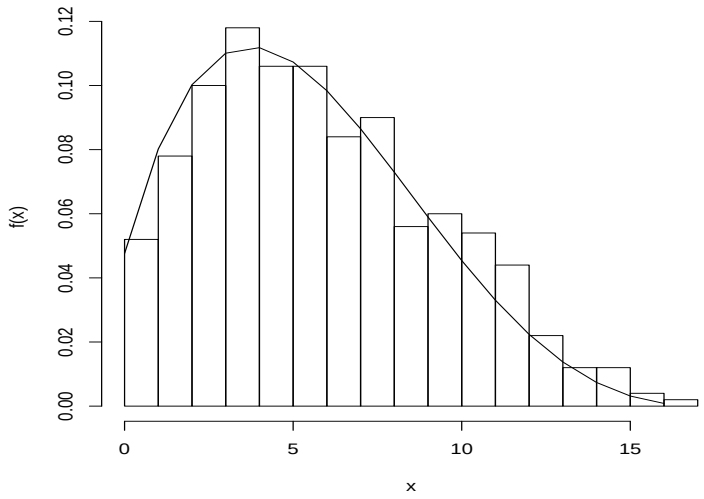

Figure B.1: Histogram of Gibbs sampling for $f(x)$ in Example 1

Example 2 (Casella and George, 1992): Suppose the conditional distributions $X \mid Y$ and $Y \mid X$ are exponential distributions restricted to the interval $(0, B)$, that is

$$f_{X|Y}(x \mid y) \propto y\mathrm{e}^{-yx},\ 0 < x < B < \infty \tag{B.13}$$

and

$$f_{Y|X}(y \mid x) \propto x\mathrm{e}{-}xy,\ 0 < y < B < \infty. \tag{B.14}$$

Similar to Example 1, Gibbs sampling is applied based on (B.13) and (B.14) and $M = 500$ parallel chains are produced, and the 15^{th} value from each chain is used to form the sampleFigure B.2 displays a histogram of the simulated X values with $B = 5$. The restriction that $B < \infty$ ensures that the marginal distribution $f(x)$ exists. When we employ the restriction $B < \infty$,

$$f_{X|Y}(x \mid y) = \frac{y\mathrm{e}^{-yx}}{1 - y\mathrm{e}^{-yB}} \qquad \text{and} \qquad f_{Y|X}(y \mid x) = \frac{x\mathrm{e}^{-xy}}{1 - x\mathrm{e}^{-xB}}.$$

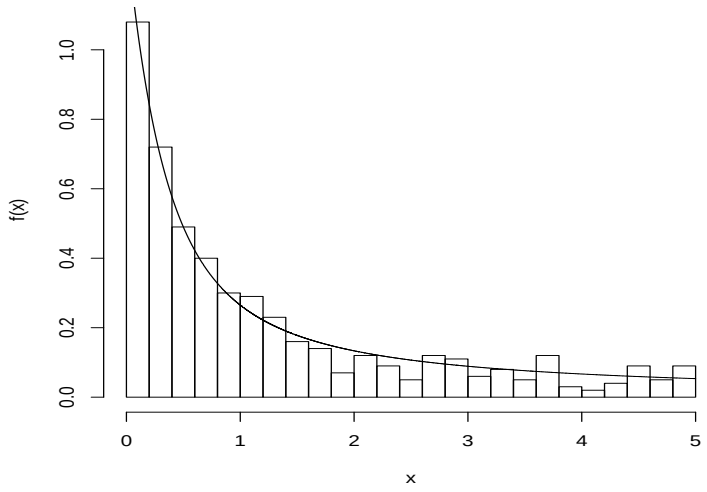

Figure B.2: Histogram of Gibbs sampling for $f_X(x)$ in Example 2

Moreover,

$$f_X(x) \propto \frac{1 - x\mathrm{e}^{-xB}}{x} \tag{B.15}$$

is the true marginal distribution for X in this example. For the purpose of comparison, the density (B.15) after proper normalization is the solid line in Figure B.2. Once again, samples generated using the Gibbs sampling recover the true density fairly well.

To employ Gibbs sampling for a given problem, it is clear that we must be able to simulate from the required low-dimensional full conditional distributions. If we recognize the full conditional density as some standard distribution, such as normal or gamma, we can simulate from it directly. Otherwise, we will need to employ techniques that enable sampling from arbitrary one (or low) dimensional distributions. There are many ways to complete this task, for example, through rejection sampling,

inverse-probability sampling (based on the probability integral transform), the ratio-of-uniforms method just to name a few. In this thesis we will employ the Metropolis-Hastings algorithm for this task.

B.2.2 The Metropolis-Hastings Algorithm

The Metropolis-Hastings Algorithm is another MCMC algorithm that can be used to generate simulations from an arbitrary distribution. The objective here is to generate samples from a target distribution $\pi(x) = f(x)/K$ where K is the (possibly unknown) normalizing constant. Suppose $h(x)$ is a known density and for a known constant $c = \sup_x \frac{f(x)}{h(x)}$, $f(x) \leq ch(x)$ for all x. To obtain a random value from $\pi(.)$, a classical simulation method named acceptance-rejection sampling can be performed as follows:

1. generate a candidate value z from $h(.)$;
2. calculate the ratio $r = f(z)/[ch(z)]$;
3. generate a value $u \sim \mathrm{U}(0, 1)$;
4. if $u \leq r$, return z; otherwise, goto step 1.

It is easily shown that the accepted value z comes from $\pi(.)$; however, finding a constant c that does the trick may be difficult in many applications and this method may result in an undesirably large number of rejections (Chib and Greenberg, 1995).

As in acceptance-rejection sampling, the Metropolis-Hastings is a rejection algorithm that allows us to generate samples from a non-standard distribution, $\pi(x)$, and

requires only knowledge of the corresponding density up to a normalizing constant. The original idea is presented in papers written by Metropolis *et al.* (1953) and Hastings (1970). This algorithm generates a proposed value x^* from some candidate distribution, $q(x^*|x^{(t-1)})$, conditional on the previous value $x^{(t-1)}$, and either accepts or rejects this proposed value with a certain probability. More precisely, given a candidate distribution, the Metropolis-Hastings algorithm proceeds as follows:

1. Set iteration counter t to 1 and choose a initial value $x^{(0)}$ in the parameter space;

2. At iteration t, generate x^* from the candidate distribution $q(x^*|x^{(t-1)})$;

3. Compute the acceptance ratio r as:

$$r = \frac{\pi(x^*)q(x^{(t-1)}|x^*)}{\pi(x^{(t-1)})q(x^*|x^{(t-1)})};$$

4. Set

$$x^{(t)} = \begin{cases} x^* & \text{with probability } \min(1, r) \\ x^{(t-1)} & \text{with probability } 1 - \min(1, r) \end{cases}$$

5. Change iteration counter from t to $t + 1$;

6. Repeat steps 2-5 until convergence is reached.

Notice that we don't need to know the normalizing constant associated with the distribution function of $\pi(x)$, since the normalizing constant is canceled in the calculation of the acceptance ratio.

To implement the Metropolis-Hastings algorithm, it is necessary to specify a suitable candidate distribution (Chib and Greenberg, 1995). Typically, candidate distributions are selected from a family of distributions that require the specification of tuning parameters such as the location and scale. The first method given by Metropolis *et al.* (1953) produces a random walk chain. The candidate value $\boldsymbol{x}^*$ is drawn according to the process

$$\boldsymbol{x}^* = \boldsymbol{x} + \boldsymbol{z}$$

where $\boldsymbol{x}$ is the current value of the chain and $\boldsymbol{z}$ is called the increment random variable that is generated from a multivariate density $q_1(.)$. In the single variable case, the most commonly used candidate distribution is a normal distribution (Gamerman and Lopes, 2006) centered at the current value, $x^* \sim \mathrm{N}(x, \sigma^2)$. Here, σ^2 is a prechosen constant, commonly referred to as a tuning parameter, that is chosen so that the algorithm performs adequately. Note that when the candidate density $q_1(.)$ is symmetric, that is $q_1(\boldsymbol{z}) = q_1(-\boldsymbol{z})$, the acceptance ratio can be reduced to

$$\boldsymbol{r} = \min\Big\{\frac{\pi(\boldsymbol{x}^*)}{\pi(\boldsymbol{x})}, 1\Big\}.$$

The second method suggested by Tierney (1994) is represented by a vector autoregressive process of order 1. In this case,

$$\boldsymbol{x}^* = \boldsymbol{a} + B(\boldsymbol{x} - \boldsymbol{a}) + \boldsymbol{z}$$

where $\boldsymbol{z}$ is generated from some distribution $q_2(.)$. The vector $\boldsymbol{a}$ and matrix B are both conformable with $\boldsymbol{x}$. Setting $B = -I$ where I is the identity matrix produces an autoregressive chain that has values reflected about the point $\boldsymbol{a}$ and induces negative

correlation between successive elements of the chain. There are also other possible methods to choose candidate distributions such as using independent candidate functions (Hastings, 1970), exploiting the form of $\pi(.)$ (Chib and Greenberg, 1994) and using the acceptance-rejection sampling method with a pseudo-dominating density (Tierney, 1994). To illustrate the Metropolis-Hastings algorithm, we present an example based on simulating the bivariate normal distribution and consider three different candidate distributions.

Example 3 (Chib and Greenberg, 1995): Consider the bivariate normal distribution $\mathrm{N}_2(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} = (1, 2)^T$ is the mean vector and

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

is the 2×2 covariance matrix. The density function can be written as

$$\pi(\boldsymbol{x}) = \frac{1}{2\pi|\Sigma|^{-1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}.$$

Therefore, the acceptance ratio for a symmetric candidate distribution is

$$r = \min\Big\{ \frac{\exp\{-\frac{1}{2}(\boldsymbol{x}^* - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}^* - \boldsymbol{\mu})\}}{\exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}},\ 1 \Big\}.$$

Three candidate distributions are proposed. The first one is a random walk generating density

$$\boldsymbol{x}^* = \boldsymbol{x} + \boldsymbol{z}, \tag{B.16}$$

where the i^{th} component of $\boldsymbol{z}$ is uniformly distributed on the interval $(-\delta_i, \delta_i)$ for $i = 1,\ 2$. We set $\delta_1 = 0.75$ and $\delta_2 = 1$. The second candidate distribution is also a

random walk generating density (B.16) but $\boldsymbol{z}$ is distributed as $\mathrm{N}_2(0, D)$ where

$$D = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.4 \end{pmatrix}.$$

Therefore, each component of $\boldsymbol{z}$ can be easily generated independently from a normal distribution. The third candidate distribution is an autoregressive generating density

$$\boldsymbol{x}^* = \boldsymbol{\mu} - (\boldsymbol{x} - \boldsymbol{\mu}) + \boldsymbol{z},$$

where components of $\boldsymbol{z}$ are distributed independently as $\mathrm{Unif}(-1, 1)$. To illustrate the characteristics of the output, Figure B.3 (panels (b), (c) and (d)) are the scatter plots of 6000 simulated values using the three candidate densities. The top left panel of Figure B.3 (panel (a)) is the scatter plot of 4000 values simulated directly from the desired bivariate normal distribution. It is clear that the Metropolis-Hastings algorithm based on either of the three candidate distributions reproduces the shape of target bivariate distribution very well in this simple example.

Choosing the tuning parameters which represent the spread and scale of the candidate distribution is an important matter. Choice of tuning parameter will effect the acceptance rate of the chain. Consider the simple situation where the candidate distribution is a Normal distribution centered at the current value that has variance σ^2 which is the tuning parameter. A large value for the variance will allow the candidate to propose moves that are distant from the current value, but it is at the likely cost of having a very small acceptance rate. On the other hand, a small value for the variance allows only a close move around the current value, which may give a high

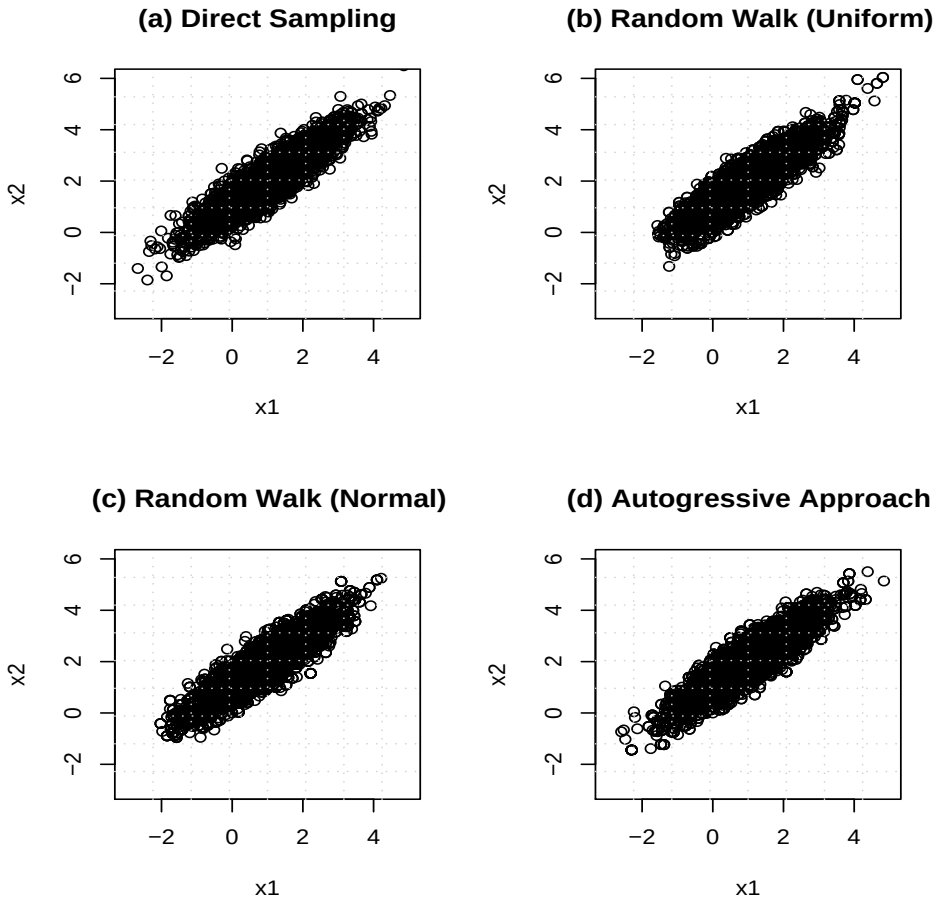

Figure B.3: Scatter Plots of Simulated Draws for Example 3

acceptance rate but at the expense of higher chain autocorrelation. In this case, the chain will take longer to traverse the support of the density. Autocorrelations across sample values are likely to be high in both situations. Müller (1993) recommended that the acceptance rate for a random walk chain should be around 0.5. The work by Roberts, Gelman and Gilks (1994) suggested that if the candidate distributions are normal, the ideal acceptance rate is approximately 0.45 in one-dimensional case and approximately 0.23 as the number of dimensions approaches infinity. In general, the ideal candidate distribution will yield an acceptance rate between 0.2 to 0.5 (Besag *et al.*, 1995). In practice, we can run the algorithm for several iterations in a preliminary tuning phase in order to achieve a reasonable acceptance rate. If the acceptance rate is too low, we decrease the value for the variance of the candidate distribution. If the acceptance rate is too high, we increase the value for the variance of the candidate distribution. Once a reasonable acceptance rate is achieved, the tuning parameter and hence the candidate distribution is held fixed to produce successive realizations from the Markov chain.

Our overall strategy for Markov chain simulation will be based on a combination of the Gibbs sampler and the Metropolis-Hastings algorithm. Working within a Gibbs sampling framework, we will successively draw realizations from full conditional distributions. If, at a given step in the algorithm, the corresponding full conditional distribution is of standard form, this will be a simple task, which we shall refer to as a 'Gibbs step'. If, on the other hand, the full conditional distribution is not of standard form we could apply the Metropolis-Hastings algorithm to obtain the

required draw. One obvious disadvantage of this approach is that it requires simulating a Markov chain *within* another Markov chain; moreover, this would be required at every sweep of the Gibbs sampler which seems computationally too intensive for practical application. Fortunately, this is not required. Upon encountering a non-standard full conditional distribution within a Gibbs sampler, applying only a single iteration of the Metropolis-Hastings algorithm (termed a 'Metropolis-Hastings step'), and subsequently moving to the next component of the Gibbs sampler is sufficient for producing an ergodic Markov chain with the required stationary distribution (see, for example, Robert and Casella (2004) for detailed discussions and technical results).

B.3 Diagnosing Convergence

MCMC algorithms provide a way to generate realizations from a distribution without knowing the normalizing constant of that distribution. The next question that arises is how one can assess convergence of the MCMC algorithm. That is, we need to decipher at which iteration the MCMC sequence has reached approximate stationarity. Further to this we need to consider the number of subsequent iterations required so that the chain will exhibit all the features of the target distribution. The theoretical foundations of MCMC algorithms show that the chains constructed by MCMC algorithms are ergodic under fairly general conditions. Therefore, the ergodic theorem guarantees the estimation consistency of various aspects of the target distribution. For practical implementation, we must of course settle for a finite Monte Carlo sam-

ple size. Thus we must decide a suitable number of post burn-in iterations at which point it safe to stop these algorithms. Monte Carlo sample size calculations based on a pre-specified error bounds are very difficult or impossible to perform in most practical situations. Thus, empirical methods based on MCMC output $\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^t, ...$ are typically used for convergence assessment, to determine when the chain has reached approximate stationarity and has exhibited all the features of the posterior distribution.

For typical applications, two different types of convergence need to be assessed. The first type is convergence to the stationary distribution. This can be done through a trace plot (time series plot) of the sampled values, which can indicate when the Markov chain has forgotten its initial state and has begun to exhibit its steady state behavior. For most MCMC algorithms, convergence to the stationary distribution is not the major issue. Instead, the speed of exploration of the target distribution and the degree of correlation between sampled values within the chain are most important. Therefore, it is crucial to examine the autocorrelation plot of the chain. A useful technique is to monitor convergence of ergodic averages, $\frac{1}{T}\sum_{t=1}^{T} h(\boldsymbol{\theta}^t)$, to their asymptotic values for a function $h(.)$ such that $\mathrm{E}_\pi[h(\boldsymbol{\theta})] < \infty$. A plot of ergodic averages after each iteration can be used to check convergence. When $\boldsymbol{\theta}$ is high dimensional, it will not be possible to examine trace plots, ergodic average plots and autocorrelation plots for each component of $\boldsymbol{\theta}$. In this case, a representative subset of model parameters, in addition to certain functions of model parameters is monitored. One useful summary of all model parameters is the logarithm of the

posterior distribution at each state of the chain (up to an additive constant not depending on $\boldsymbol{\theta}$) which can always be examined as an overall summary of the chain.

Instead of diagnosing convergence based on a single chain, multiple chains initialized from different starting values are often used. By simulating several independent parallel chains, the variability and dependence on initial values are reduced and convergence is easier to assess by plotting multiple chains onto the same axis; however, there are dangers in a naive implementation of the multiple chains principle. The slowest chain will govern convergence and it is extremely important that the different chains are initialized at points that are well dispersed over the parameter space.

Aside from the examination of trace plots, the Gelman-Rubin diagnostic statistic (Gelman and Rubin, 1992) is another useful tool for deciding how long the Markov chain should run. In this context, we begin by running $2N$ iterations of M parallel chains each initialized at dispersed points in the target distribution. After discarding the first N iterations, we compute the between and within sequence variances, denoted as B and W respectively. For the parameter of interest ϕ,

$$B = \frac{N}{M-1}\sum_{j=1}^{M}(\bar{\phi}_{\cdot j} - \bar{\phi}_{\cdot\cdot})^2$$

and

$$W = \frac{1}{M}\sum_{j=1}^{M} S_j^2 \text{ where } S_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\phi_{ij} - \bar{\phi}_{\cdot j})^2.$$

In these formulas, ϕ_{ij} denotes the i^{th} value from j^{th} chain, $\bar{\phi}_{\cdot j} = \frac{1}{N}\sum_{i=1}^{N}\phi_{ij}$ and $\bar{\phi}_{\cdot\cdot} = \frac{1}{M}\sum_{j=1}^{M}\bar{\phi}_{\cdot j}$. The marginal posterior variance $Var[\phi \mid y]$, where y is the data,

can be estimated using a weighted average of B and W:

$$\hat{Var}^{+}(\phi \mid y) = \frac{N-1}{N}W + \frac{1}{N}B.$$

This estimator is biased; however, it is asymptotically unbiased under stationarity. For finite N, $\hat{Var}^{+}(\phi \mid y)$ overestimates $Var[\phi \mid y]$ while W underestimating $Var[\phi \mid y]$ since the individual sequences have not had time to explore the target. Therefore, the Gelman and Rubin diagnostic statistic

$$\hat{R} = [\frac{\hat{Var}^{+}(\phi \mid y)}{W}]^{1/2}$$

is always greater than 1 and declines to 1 as $N \to \infty$. Further simulations are required when $\hat{R}$ is high. It is typically recommended that simulations continue until $\hat{R} < 1.1$. Note that this diagnostic is univariate, and hence must be applied to each component of the parameter vector separately and monitored for all parameters of interest.

It is generally safe to conclude convergence if the autocorrelation function drops quickly, the trace plot and ergodic plot indicate stability of Monte Carlo estimates and the Gelman and Rubin diagnostic statistics close to 1 for all parameters of interest. If this is not the case, running a longer chain may solve the problem but techniques based on alternations to the model or algorithm can be more efficient. Reparameterizations such as centering covariates and hierarchical centering can improve the algorithm (Gamerman and Lopes, 2006). When high correlations exist between components, blocking techniques that update groups of parameters (based on their joint full conditional distribution) can be very beneficial in improving computational performance (in the sense of lowering chain autocorrelations). In this case,

slow componentwise moves are replaced by moves dictated by the joint full conditional distribution for the block of parameters considered (Liu *et al.*, 1994). Carefully considering the identifiability of parameters in the model, and seeking a better proposal distributions can also be helpful. Having discussed methods for posterior computation, we complete the literature review by discussing Bayesian methods for model selection and goodness-of-fit. We focus on methods that are applicable to hierarchical models and that are easily implemented via MCMC.

B.4 Bayesian Model Selection using the Deviance Information Criterion

For a given problem in data analysis, there will usually be several models under consideration. In general, a larger model has more flexibility and therefore has the advantage of fitting the data better; however, these large models become more difficult to compute and interpret. Choosing between competing, perhaps non-nested models is thus an important issue in any data analysis.

The standard Bayesian approach to model selection arises through the Bayes factor (BF) which, given a prior over a set of competing models, is obtained as the ratio of the posterior and prior model odds. The BF can also be written as a ratio of the marginal likelihoods of observed data $\boldsymbol{y}$ under each model defined as

$$BF = \frac{p_1(\boldsymbol{y})}{p_2(\boldsymbol{y})} = \frac{\int L_1(\boldsymbol{y} \mid \boldsymbol{\theta_1})\pi_1(\boldsymbol{\theta_1})d\boldsymbol{\theta_1}}{\int L_2(\boldsymbol{y} \mid \boldsymbol{\theta_2})\pi_2(\boldsymbol{\theta_2})d\boldsymbol{\theta_2}},$$

where $L_i(\boldsymbol{y} \mid \boldsymbol{\theta_i})$ is the likelihood function under model i and $\pi_i(\boldsymbol{\theta_i})$ is the prior specification assigned to model i, $i = 1, 2$. the Bayes factor has a nice interpretation in terms of posterior model probabilities; however, the calculation of the marginal likelihoods is difficult for complex models. The Akaike Information Criterion (AIC) is another technique for model comparison taking the form

$$AIC = -2l_{M_i}(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}) + 2p$$

and the Bayesian Information Criterion (BIC) is yet another criteria having the form

$$BIC = -2l_{M_i}(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}) + \log(n)p.$$

Both are easily computable alternatives to the BF. Here, $l_{M_i}(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$ is the log-likelihood for model M_i, $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ is the MLE of $\boldsymbol{\theta}$ under model M_i, n is the number of observations and p is the number of parameters. Models with lower AIC or BIC values are preferred. The BIC is particulary relevant for Bayesian model selection as it can be asymptotically related to posterior model probabilities derived under a uniform prior over the model space. Both the AIC and BIC impose penalties for model complexity based on the number of model parameters p. Unfortunately, these methods are not appropriate for hierarchical spatial model where parameters include correlated random effects. In this context, correlation between parameters becomes an issue and hence the *effective* number of parameters is not entirely clear. As a result, penalizing model complexity becomes a more subtle issue.

To address this concern, the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) is an extension of the AIC approach that may be applied

for choosing between hierarchical spatial models. The criteria is based on the deviance statistic which is defined as

$$D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{y} \mid \boldsymbol{\theta}),$$

where $L(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood function of the data given parameters under the model. Then the DIC is defined as

$$DIC = \overline{D(\boldsymbol{\theta})} + p_D,$$

where $\overline{D(\boldsymbol{\theta})} = \mathrm{E}[D(\boldsymbol{\theta}) \mid \boldsymbol{y}]$ is the posterior mean of the deviance, a measure of fit with lower values indicating superior fit to the data. The quantity p_D is a penalty term that measures the complexity of the model and is defined by

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}),$$

where $D(\bar{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior mean of $\boldsymbol{\theta}$. Rather than simply penalizing the model depending on the total 'raw' number of parameters appearing in the model, the p_D penalty accounts for spatial correlation or shrinkage among correlated parameters and can be interpreted as an estimate of the effective number of model parameters. Spiegelhalter et al. (2002) present a detailed justification for the definition of p_D and illustrate its use with several examples. There, it is also shown that in the special case of non-hierarchical generalized linear models (models without random effects), p_D is approximately equal to the raw parameter count p and that this approximation is exact in the case of the normal linear model. With models incorporating correlated parameters, p_D will be smaller than the raw number

of parameters. Overall, just as with the AIC and BIC, the model with lower DIC score is preferred since it reaches the best combination of fit and parsimony.

B.5 Goodness-of-fit for Bayesian Models using Posterior Predictive Model Checking

A fundamental aspect of any model based data analysis involves checking the fit of the proposed model to the data. Conclusions drawn from any analysis are of course conditional on such checks of model adequacy. As with model fitting, the Bayesian approach to goodness-of-fit relies heavily on simulation based methods. We focus here on methods based on the posterior predictive distribution. The posterior predictive checking technique (Gelman et al., 2004) is based on an examination of the fit of a model to the observed data by drawing replicated data values from the posterior predictive distribution defined as

$$p(\boldsymbol{y}^{rep} \mid \boldsymbol{y}) = \int p(\boldsymbol{y}^{rep} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta},$$

where $\boldsymbol{y}$ is a vector of the observed data, $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{y}^{rep}$ represents a hypothetical replicate data set, which is assumed to be drawn under the same conditions as the observed data, and also assumed to be conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{\theta}$.

In practice, the posterior predictive distribution is computed using simulation. Having obtain samples from the posterior distribution for a given model, simulation

from the posterior predictive distribution is straightforward using one-for-one composition sampling. The replicated data, drawn from the predictive distribution induced by the model, should look similar to the observed data when the model fits. Any obvious departures from the observed data indicate potential failings of the model (in a predictive sense). To quantify the lack of fit, the posterior predictive p-value measures the probability that the replicated data could be more extreme than the observed data (where the probability is computed W.R.T the posterior predictive distribution). To illustrate further, let $T(.)$ be a 'checking function' (a statistic) used to summarize some aspect of the data. This is usually chosen to be a specific feature of the data, that is of interest. The posterior predictive p-value is calculated as:

$$p_B = \Pr(T(y^{rep}) \geq T(\boldsymbol{y}) \mid \boldsymbol{y}).$$

Using carefully constructed checking functions, one can search for specific discrepancies between observed and simulated data.