# Species Richness Estimation for Benthic Data

Beth J. Norris

A thesis submitted for the degree of Doctor of Philosophy

School of Mathematics, Statistics and Actuarial Science

University of Kent

June 2012

# ABSTRACT

This thesis addresses species richness estimation for benthic data by describing the clustering of individuals within a species using a Neyman Type A distribution, and incorporating this into species richness estimates.

A review of current species richness estimation methods is included. The maximum-likelihood approach to species richness estimation is extended to incorporate the Neyman Type A model, with a gamma mixing distribution on the mean abundance of individuals within a species. Species richness estimates of this model are compared to those of the simpler negative binomial and Poisson models. The use of a penalised-likelihood is applied to avoid spuriously large estimates of species richness that can be associated with the 'boundary problem'.

The Bayesian approach to species richness is considered, using uninformative and informative priors. Informative priors are elicited using expert opinion obtained from a number of benthic ecologists at the Centre for Environment, Fisheries and Aquaculture Science. These are incorporated into species richness estimation in the form of priors, and also converted into penalties for use in the frequentist approach. Several benthic data sets are analysed throughout, along with a Lepidoptera data set, and a data set from a common bird census carried out in the USA. In addition, several simulation studies are undertaken to illustrate the performance of the estimators.

The research culminates in the application of species richness estimators to estimate species mortality due to dredging carried out off the Norfolk coast. Several estimators can be considered to gain a picture of the effect of dredging, and I recommend that species richness estimators should reflect the underlying distribution of the data. I also recommend that a precautionary approach should be taken when using these estimators in practical applications.

# ACKNOWLEDGEMENTS

I wish to thank my supervisors, Byron Morgan, Martin Ridout and Jon Barry for their advice and support throughout this research. I am very grateful to the Engineering and Physical Sciences Research Council and the Centre for Environment, Fisheries and Aquaculture Science for the project funding, along with the School of Mathematics, Statistics and Actuarial Science at the University of Kent and the National Centre for Statistical Ecology.

A special thanks to all the scientists at CEFAS who have assisted in developing my ecological understanding of benthic organisms, and for contributing data sets for analysis, and expert knowledge for elicitation. In addition, Tom Webb of the Department of Animal and Plant Sciences at the University of Sheffield who gave me an upper limit for my spuriously large species richness estimates.

Finally, I thank my family and friends who supported me during the highs and lows of the past few years, and Peter for always being there.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# GLOSSARY

The following notation will be used throughout this thesis:

$N$ total number of species in a population.

$x_i$ number of times the $i$th species is observed in the sample, $i = 0, 1, 2, ..., D$ (Only those species with $x_i > 0$ are observable in the sample).

$f_k$ number of species that are represented exactly $k$ times in the sample, $k = 0, 1, ..., r$.

$f_0$ number of species unobserved in the sample but present in the population, $f_0 = N - D$.

$n$ total number of individuals in the sample, $n = \sum_{i=1}^{D} x_i = \sum_{k=1}^{r} k f_k$.

$\boldsymbol{I}(A)$ indicator function, $\boldsymbol{I}(A) = 1$ if the event $A$ occurs, 0 otherwise.

$D$ number of distinct species discovered in the sample, $D = \sum_{i=1}^{N} \boldsymbol{I}(x_i > 0) = \sum_{k \geq 1} f_k$.

$g$ number of samples/grabs.

$a$ area of the grab.

$A$ area of the region of interest.

$\hat{N}_{C1}$ lower bound estimator of Chao (1984), given by $\hat{N}_{C1} = D + f_1^2/(2f_2)$.

$\hat{N}_{Ji}$ $i$th order jackknife estimator.

$\hat{N}_B$ bootstrap species richness estimator.

$\hat{N}_{ACE}$ Abundance-based Coverage Estimator.

$\boldsymbol{\theta}$ vector of parameters describing the abundance distribution of the data.

$L(\boldsymbol{\theta})$ likelihood function for parameter vector $\boldsymbol{\theta}$.

$l(\boldsymbol{\theta})$ log-likelihood function for parameter vector $\boldsymbol{\theta}$.

$l_p(\boldsymbol{\theta})$ profile log-likelihood function for parameter vector $\boldsymbol{\theta}$.

$d$ dimension of parameter vector.

$\chi^2_{d:\alpha}$ $\alpha\%$ point for the chi-squared distribution.

$z_{\alpha/2}$ $\alpha\%$ point for the standard normal distribution.

$\tau$ truncation point of the data.

$\gamma$ penalty parameter.

$h(\theta)$ penalty function for parameter $\theta$.

$\psi(\theta)$ odds function, $\psi(\theta) = p_0(\theta)/(1 - p_0(\theta))$.

$p(\boldsymbol{\theta})$ Bayesian prior on $\boldsymbol{\theta}$.

$\pi(\boldsymbol{\theta}|\boldsymbol{x})$ posterior distribution of $\boldsymbol{\theta}$ given data $\boldsymbol{x}$.

$N_S$ number of species in a data augmented super-population.

$\tau_\theta$ coefficient of overlap of the prior and posterior.

# 1. INTRODUCTION

## *1.1 Background*

This research is undertaken in collaboration with the Centre for Environment, Fisheries and Aquaculture Science (CEFAS), an agency of the UK Department for Environment, Food and Rural Affairs (DEFRA). CEFAS' primary purpose is to provide advice and support to the UK government and its agencies on a wide spectrum of issues, including climate change impacts and adaptation, marine planning and environmental licensing, sustainable fisheries, and marine biodiversity and habitats.

The world's oceans are an indispensable resource and must be protected and managed for future generations. The negative impacts of changes in marine biodiversity should be fully considered and minimised. Ecologists work to assess critical threats to marine systems and develop management strategies to mitigate them.

Ecology has evolved from a descriptive discipline to a highly quantitative field (Ludwig and Reynolds, 1988), and the use of statistics can play a vital role in assessing the impacts of threats to marine systems, such as climate change or marine aggregate extraction. There are various statistical methods that can be utilised to aid biodiversity conservation. Whole ecosystems can be modelled, exploring complex interactions between organisms and the environment. However, due to these complexities it can be advantageous to use key species as indicators of the health of the ecosystem.

Marine benthic organisms live in, on, or near the seabed. The ocean floor habitats in which they live constitute the largest single ecosystem on earth in terms of spatial

coverage (Snelgrove, 1997). Benthic organisms not only have value as an indicator species in the assessment of human impacts at sea, but also significantly effect major ecological processes including the regulation of carbon, nitrogen and sulphur cycling, water column processes, and pollutant distribution and fate (Snelgrove, 1997). Therefore, benthic organisms are frequently used for assessing biodiversity change as they may indicate impacts on the wider marine ecosystem.

Benthic organisms are largely sessile and so will give a good indication of locally induced changes. When impacts of activities such as dredging are considered, both the initial impact and the predicted rates of recovery of marine benthos are important (Kenny and Rees, 1996). In this research I concentrate on modelling initial impacts, which can include a reduction in species diversity, abundance and biomass.

The variety and abundance of benthos vary with latitude, depth, water temperature, pH and salinity, and also locally determined conditions such as the nature of the substrate (Britannica, 2012). Ecological factors such as predation and competition also have an effect on the community structure.

Benthic organisms vary in size, and in this research I concentrate on analysing data for Macrofauna, animals that are one centimetre or longer, and Megabenthos, which includes large crustaceans and molluscs. In benthic environments some organisms are colonial, that is individuals cannot always be separated and counted, but these species will be excluded from the analysis.

The effects of activities such as dredging depend not only on the magnitude and intensity of impact but also on the composition and spatial variability of benthic assemblages in different areas. There are many statistical aspects of the analysis and sampling of benthic data that are not well understood, and these need to be further developed in order that robust conclusions can be drawn from scientific research.

## 1.2 Statistical analysis of benthic data

The study and analysis of benthic organisms presents a number of statistical challenges. The data available for analysis are not ideal, because they are often limited to few samples because of the expense and difficulty of sampling. The complexity of benthic communities complicates modelling, because individuals of species are often found in clusters. Therefore many species can be missed during a sampling programme.

Modelling the spatial distribution of benthic organisms as cluster processes has been shown to work well and provides answers to important practical questions, for example regarding species loss (Boyd et al., 2006; Barry et al., 2010). The research of this thesis will build on this previous work.

The benthic abundance data analysed in this thesis consist of samples collected from the seabed by Hamon grab. The grab is a sample bucket attached to a pivoted arm, supported by a frame (Figure 1.1). The Hamon grab is activated by releasing tension in a wire and the sample bucket is driven through the sediment of the seabed (Boyd et al., 2006).

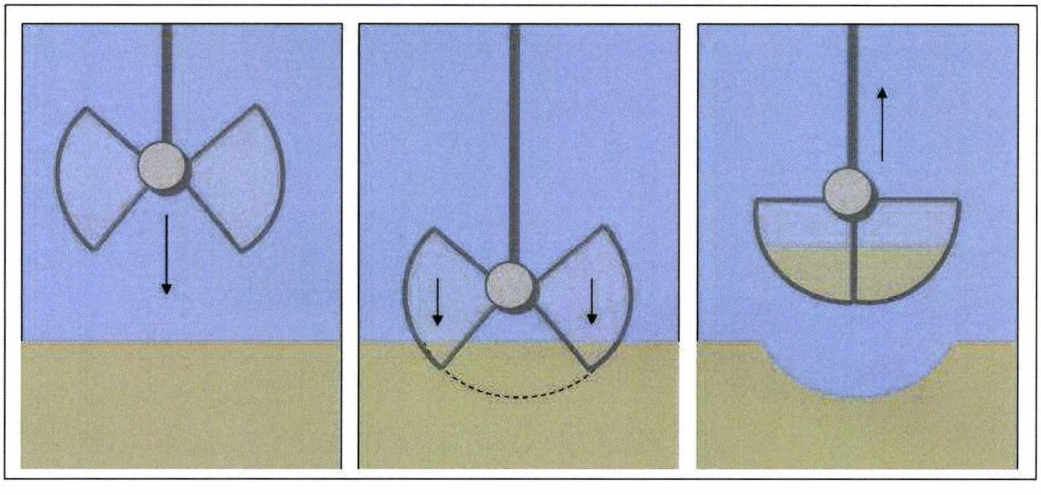


*Figure 1.1:* How a Hamon grab sampler works.

Under ideal conditions the Hamon grab should take a sample of known volume and surface area. Two grab sizes are used, and trials have indicated that the average sample surface area for the larger grab is $0.25m^2$, and for the smaller grab is $0.1m^2$ (Kenny and Rees, 1996). There are potential sources of sample error associated with the Hamon grab; for example the tendency of the grab to push itself away from the seabed when sampling coarse sediments results in obtaining only a scrape of sediment. However, for the purpose of this research I assume that surface area sampled is without error.

Replicate samples are collected, located randomly within the study region to avoid bias. After collection the whole sample is washed over square mesh sieves to remove the fine sediment and to separate the fauna into size-based fractions for later sorting and identification of species in the laboratory. This can be time consuming and a high level of expertise is required. This puts a constraint on the number of samples that can be processed.

Additional assumptions include that the community is 'closed' within a sampling period, and that sampling is carried out 'with replacement', in the sense that the community structure remains unchanged by sampling. This is a reasonable assumption to make, since the sampling fraction is small.

The specific benthic data sets analysed in this thesis have been collected around the UK coast during sampling programmes aboard CEFAS' research vessel, the CEFAS Endeavour. I will refer to them throughout by the name of the area. These areas include the coast off Norfolk, the Isle of Wight, Hastings, and the Eastern Channel. The Norfolk data contain five replicates per survey period, and the Isle of Wight and Hastings data contain ten replicates each. This level of replication is typical of benthic surveys. The Eastern Channel data set, however, is much larger, arising from an extensive survey programme of 225 replicate samples. For further details on the data sets see Kenny and Rees (1996); Boyd et al. (2006); Cooper et al. (2007).

During the thesis I also analyse the well-known Lepidoptera data used in Fisher et al. (1943) and the data set from the National Audubon Society Christmas Bird Count (CBC) of 1989 at Fort Myers, Florida (available from http://birdsource.tc.cornell.edu/cbcdata/). These data sets have been used in the literature to illustrate methods of species richness estimation and will be used for comparison of our methods.

## 1.3 Outline of the thesis

The study and analysis of benthic organisms presents a number of statistical challenges, and this thesis addresses some of these issues, focussing mainly on modelling the spatial distribution of benthic organisms as cluster processes. The aim is to answer critically important ecological questions relating to benthic organisms through the formulation and application of robust statistical procedures based on a range of appropriate stochastic models.

The specific objectives of the research are categorised into five parts: the first is to review the current state of the art and identify a gap in the knowledge, the following three focus on methodological development, and the last part examines the links between methods and the application of the models to assessing human impact on marine biodiversity.

Firstly, there are many diversity measures that can be utilised to look at differences in biodiversity between areas or years. However these are not adequate when analysing benthic data. The thesis focuses on estimating one aspect of biodiversity, species richness, and Chapter 2 reviews the appropriate literature in this field. The current state of the art is outlined, and the need for the research is clarified. This chapter also introduces a clustering process that is used to model the spatial distribution of benthic organisms, and with the aim of improving species richness estimation for benthic communities. It is shown that clustering needs to be accounted for when

estimating species richness.

Chapters 3 and 4 develop multinomial models for estimating species richness using contagious distributions, from a frequentist and a Bayesian approach respectively, incorporating a spatial aspect in the form of a cluster process. Chapter 3 highlights problems associated with parametric species richness estimation, including the choice of method for constructing confidence intervals, and how to deal with spuriously large estimates, specifically using penalties.

Chapter 4 investigates the Bayesian equivalent approach to species richness estimation, using uninformative priors, and extends current methods to account for the spatial clustering of benthic organisms. Since some species are not seen in the sample, alternative approaches used for handling missing data are considered, including reversible jump MCMC and data augmentation in a hierarchical Bayes framework.

Fourthly, I consider the use of informative priors, formed by eliciting expert information. Chapter 5 describes a process of elicitation of information from experts and incorporates this into priors. This is a particularly interesting aspect of the research, which as far as I am aware has not been done previously for benthic data. This chapter highlights difficulties within the elicitation process and applies the priors in estimating species richness for a number of benthic data sets.

Chapter 6 considers the link between the various methods of species richness estimation, by considering how elicited priors can be converted into penalties within a frequentist approach. The results of the frequentist approach are compared to those of the Bayesian approach and to the non-parametric estimators currently used for species richness estimation. I make a recommendation of the best method to use to estimate species richness for benthic data, and use this method to investigate the impacts of dredging on the Norfolk coast. These results are compared to those found

by Barry et al. (2010), who utilised a clustering model to describe the spatial pattern of each species and modelled the impact of dredging directly.

Chapter 7 emphasises the main contributions of this thesis and highlights future directions for the development of statistical models for benthic data. Supporting information is provided in the Appendix. The findings of this research can help refine current guidelines with regard to dredging and also be used to assess the impact of changes in biodiversity as a result of direct and indirect human impact.

An additional aim of this research is to construct a library of appropriate computer software, written in R, which will be made freely available, and which will link naturally with other statistical ecology software within the National Centre for Statistical Ecology (NCSE).

# 2. METHODS OF SPECIES RICHNESS ESTIMATION

## *2.1 Introduction*

To monitor the impact on benthic organisms from activities such as marine dredging, we need to be able to measure changes in the community. This can be done by measuring biodiversity and how it changes over time and between sites. There are several levels of biological diversity one could evaluate, including ecosystem diversity and genetic diversity, but I concentrate on the most commonly measured aspect, species diversity.

Biodiversity can be defined as *'the variety and abundance of species in a defined unit of study'* (Magurran, 2004). This definition specifies two components to be measured, variety and abundance of a species. Abundance can be defined as *'the total number of individuals or the density of individuals within an area'* (Buckland et al., 2005), but variety is less easily defined.

There are numerous indices that are used as measures of biodiversity, outlined in great detail in Magurran (1991, 2004) and Gotelli and Colwell (2010). Most of these indices assume that individuals are randomly sampled from an indefinitely large population and that all species are represented in the sample (Magurran, 1991).

However, the small sampling fraction and clustering nature of benthic organisms is such that we will not sample all species present in the area. Therefore, diversity indices may not behave as we would expect for benthic data, and may need to be adapted. A paper that I have contributed to, outlining the behaviour of diversity indices when applied to benthic data, is to be submitted to Ecological Indicators

shortly. However this thesis concentrates on how we can estimate the number of benthic species, namely 'species richness'.

This chapter reviews species richness estimators, and highlights the failure of some of these when applied to spatially clustered data. A clustering model is introduced to describe the spatial distribution of benthic organisms, and a simulation study shows the inadequacies of current species richness estimators in analysing clustered data. I also apply several non-parametric estimators applied to benthic data collected off the Isle of Wight coast.

## 2.2 Species richness estimation

Species richness is one of the earliest and most intuitive measures of biodiversity (McIntosh, 1967), and can be defined as *'the number of species present in the area of study'*. The need for the continued development of species richness estimators and the importance of knowledge transfer to biologists is encapsulated in this quote from (Kéry and Royle, 2008, p591)

> *'The most common approach to species-richness estimation is really no estimation at all: mere use of raw totals of detected species.'*

However, sampling from populations will rarely give a complete inventory of species due to constraints such as time and cost, and in marine ecosystems there is also the added constraint of the requirement of a high level of taxonomic expertise to avoid misidentification (Foggo et al., 2003).

The question of how to estimate the number of species in a population has been of interest for over six decades, and there are over 20 different techniques described that will produce an estimate of total species richness from sample data (Foggo et al., 2003; Chao, 2004; Magurran, 2004; Gotelli and Colwell, 2010). Some of these estimators are summarised here, and others are described in more depth in Chapters 3 and 4.

In some cases organisms may not be defined to species level. However within a particular sampling program organisms are identified to the same level, such as genus, and the same estimators can be used to estimate the richness at that particular level. I also concentrate on abundance-based species richness estimators, since that is the type of data that I am working with for benthic organisms. Some mention is made of incidence-based estimators, developed for data that are a measure of presence only.

### 2.2.1 Early approaches to species richness estimation

Early approaches to solve the species richness problem considered the relationship between species richness and the area sampled (Fisher et al., 1943), presenting this as a species accumulation curve. Samples are randomised and the cumulative number of species across a number of samples is plotted. A species richness estimate is obtained by fitting models to the curve, such as the Michaelis-Menten model or hyperbolic model (Keating and Quinn, 1998; Colwell and Coddington, 1994). A species richness estimate is extrapolated for the area of interest at the point where the curve reaches an asymptote.

A more recent adaptation of this approach is the Total-Species accumulation method (T-S), which incorporates spatial heterogeneity of samples into the estimate of species richness for large areas, by grouping samples into subsets based on shared environmental characteristics (O'Dea et al., 2006). First a species-accumulation curve is obtained for randomised samples of all the single subareas, and then the species-accumulation curve for all combinations of two subareas is calculated, and so on up to the combinations of all subareas. From these curves a total species curve is obtained, which can then be extrapolated to estimate the probable total number of species in the whole study area (Ugland et al., 2003).

Another estimator of species richness, related to species accumulation curves, is $S_{\infty}$ proposed by Karakassis (1995). This method calculates a number that is theoretically the upper limit of the asymptote. A number of random permutations of the samples

are produced, and the cumulative number of species for a certain number of samples, $\bar{N}_g$, is estimated by averaging over the random permutations. The estimator is calculated by extrapolation, plotting the cumulative number of species in $g$ samples against the cumulative number of species in $g+1$ samples, and obtaining the regression line of these points. The estimator is the intercept between this line and the line $\bar{N}_g = \bar{N}_{g+1}$, i.e. the point where two successive samples are expected to present the same cumulative number of species (Figure 2.1) (Karakassis, 1995).

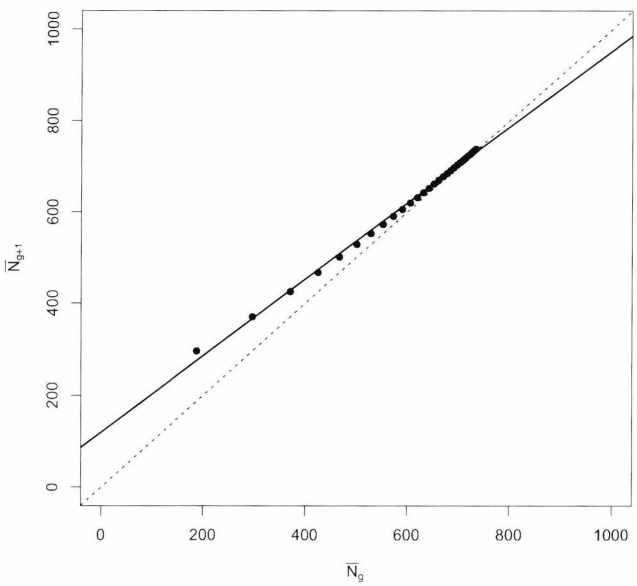



*Figure 2.1:* Illustration of the method of the $S_\infty$ estimator. Abundances of 1000 species over 100 grabs were simulated from a negative binomial distribution, $\alpha = 0.5$, $\beta = 0.5$. The points are the cumulative number of species in $g$ samples plotted against the cumulative number of species in $g + 1$ samples and the solid line is the regression line of these points. The dashed line is the line $\bar{N}_g = \bar{N}_{g+1}$. Using 30 permutations, $S_\infty = 717$.

### 2.2.2 Parametric approaches

The earliest parametric approaches to the species richness problem by Fisher et al. (1943) and Craig (1953) introduced the approach of modelling species abundance data using a Poisson model. In this model the distribution of the number of observed

individuals depends on only one parameter: the mean abundance, $\lambda$. This assumes that the abundance is equal for all species ($\lambda_1 = \lambda_2 = ... = \lambda_N$, where there are $N$ species in a population), which is unlikely in practical applications.

If species are not equally abundant, a mixture of distributions corresponding to different values of $\lambda$ can be used. Parametric estimators model the mean abundances $(\lambda_1, \lambda_2, ..., \lambda_N)$ as a random sample from a mixing distribution with density $f(\lambda, \theta)$, where $\theta$ is a low-dimensional parameter vector. This density has been modelled by many researchers as a gamma distribution, which, when combined with the Poisson distribution, leads to a negative binomial distribution for the number of individuals per species (Fisher et al., 1943).

Other parametric models that have been used to estimate species richness include the broken stick, the log normal, the inverse Gaussian and the generalized inverse Gaussian (Chao, 2004). However, when using parametric models we must make assumptions regarding species abundances. In addition, two models may fit the data equally well but give very different estimates and a good-fitting model does not necessarily give a satisfactory species richness estimate (Chao, 2004).

The parametric approach to species richness estimation will be considered in further detail in Chapter 3.

### 2.2.3 *Non-parametric approaches*

Concerns about making assumptions regarding the species abundances in a parametric approach led to the development of non-parametric approaches to species richness estimation. Here I outline some of the widely used and best performing of the non-parametric species richness estimators for abundance data.

The bootstrap method was developed for estimating species richness from quadrat sampling (Chao, 2004). Given the $n$ individuals observed, a random sample of size

$n$ is drawn from these with replacement. Assuming the proportion of the individuals for the $i$th species in the generated sample is $\hat{p}_i$, and $D$ is the number of observed species, then a bootstrap estimate of the total species richness is calculated using the formula

$$\hat{N_B} = D + \sum_{i=1}^{D}(1 - \hat{p}_i)^n. \tag{2.1}$$

The mean of a number of bootstrap estimates is taken as the final species richness estimate (Smith and van Belle, 1984). This estimate is considered as a conservative lower bound for $N$ (Mao, 2007).

The $Chao_1$ estimator is based on the concept that rare species carry the most information about the number of missing species, using only singletons and doubletons (those species represented by only one or two individuals respectively) to estimate the number of missing species (Chao, 2004). These estimators are based on Alan Turings frequency formulae used in cryptanalysis in World War II (Good, 1953; Chao et al., 2009). The original form of the estimator is

$$\hat{N} = D + \frac{f_1^2}{2f_2}, \tag{2.2}$$

where $f_1$ and $f_2$ are the number of singletons and doubletons. This formula breaks down if $f_2 = 0$, but a modified bias-corrected version of the estimator that is always obtainable is

$$\hat{N_{C1}} = D + \frac{f_1(f_1 - 1)}{2(f_2 + 1)} \tag{2.3}$$

(Chao, 2004).

The variance of the $Chao_1$ estimator is calculated as

$$\text{var}(\hat{N}_{C1}) = f_2 \left[ \frac{1}{2}\left(\frac{f_1}{f_2}\right)^2 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{4}\left(\frac{f_1}{f_2}\right)^4 \right], \tag{2.4}$$

for $f_1 > 0$ and $f_2 > 0$.

$Chao_2$ applies the same approach as the $Chao_1$ estimator, but looks at species that occur in only one or two samples, and is applied to replicated incidence data (Chao, 1987). These estimators, however, have been found to be robust estimators

of minimum species richness (Shen et al., 2003).

These non-parametric estimators are based on the assumption that sampling units are sampled with replacement. However, if sampling is carried out such that no sampled unit can be repeatedly observed, these estimators tend to overestimate richness for relatively high sampling fractions and do not converge to the true species richness (Chao and Lin, 2012). Therefore, Chao and Lin (2012) proposed a non-parametric lower bound for species richness

$$\hat{N_{CL}} = D + \frac{f_1^2}{\frac{n}{n-1}2f_2 + \frac{q}{1-q}f_1} \tag{2.5}$$

where $q$ is the ratio of sample size to population size. To use this method it is assumed that the total number of individuals in the population is known. When a small portion of individuals are taken from the entire population of individuals in the community, so that $q$ approaches zero, this lower bound approaches the $Chao_1$ estimator (Chao and Lin, 2012).

The Abundance-based Coverage Estimator (ACE) is based on the estimated sample coverage (the sum of the cell probabilities of the observed classes)

$$\hat{CO} = 1 - f_1 / \sum_{k \geq 1} i f_k.$$

It is a function of the rare species' frequencies, calculated as

$$\hat{N_{ACE}} = D_{abun} + \frac{D_{rare} + f_1 \hat{\gamma}^2}{\hat{CO}}, \tag{2.6}$$

where $D_{rare}$ are the number of rare species, and $D_{abun}$ are the number of abundant species observed in the sample. Here,

$$\hat{CV}^2 = \max \{N_{rare} \sum_{k \geq 1} k(k-1) f_k / [\hat{CO}(\sum_{k \geq} k f_k)^2] - 1, 0\} \tag{2.7}$$

denotes the estimated squared coefficient of variation (Chao, 2004). It is assumed that all species seen fewer than ten times in the sample are rare, and all those seen ten or more times are abundant and that the species abundances are well described by their mean and coefficient of variation (Chao, 2004). The Incidence-based Coverage

Estimator (ICE) yields similar estimates to ACE for presence only data.

Jackknife estimators were developed to reduce the bias of a biased estimator, and for species richness estimation this biased estimator would be the number of observed species. The idea behind the jackknife estimators is to average estimates calculated for subsets of the data by successively deleting a number of individuals from the original data. The first-order jackknife is calculated as

$$\hat{N}_{J1} = D + (n-1)f_1/n,$$

so only the number of singletons is used to estimate the number of unseen species. This is the most commonly used of the jackknife estimators, along with the second-order jackknife defined as

$$\hat{N}_{J2} = D + 2f_1 - f_2.$$

## 2.3 *Inadequacies of current species richness estimation methods applied to benthic data*

Studies have suggested that species richness estimation is dependent on spatial patterns (Baltanas, 1992; Fager, 1972) and Walther and Moore (2005) explained that performance is dependent on total species richness, sample size, and aggregation of species within samples. We would hope for a species richness estimator that is unbiased, precise and efficient. None of the commonly used estimators described here adequately account for spatial heterogeneity in species distributions, with the possible exception of the Total-Species accumulation method, the accuracy of which is yet to be fully investigated (Ugland and Gray, 2004).

When using species accumulation curves, various models may fit the data well, but give drastically different estimates (Chao, 2004). The method cannot be used on sparsely sampled communities because there will not be sufficient data to construct the accumulation shape, and the curve may not approach an asymptote. Alternatively, it has been seen that the number of individuals that need to be sampled

before the curve reaches an asymptote can be very large (Chao et al., 2009). Heck et al. (1975) noted that in a patchy environment, the species accumulation curve could even appear to approach an asymptote before many species were sampled (Figure 2.2). In most of the benthic data sets to be considered there are very few grabs samples, so this method would not be effective.

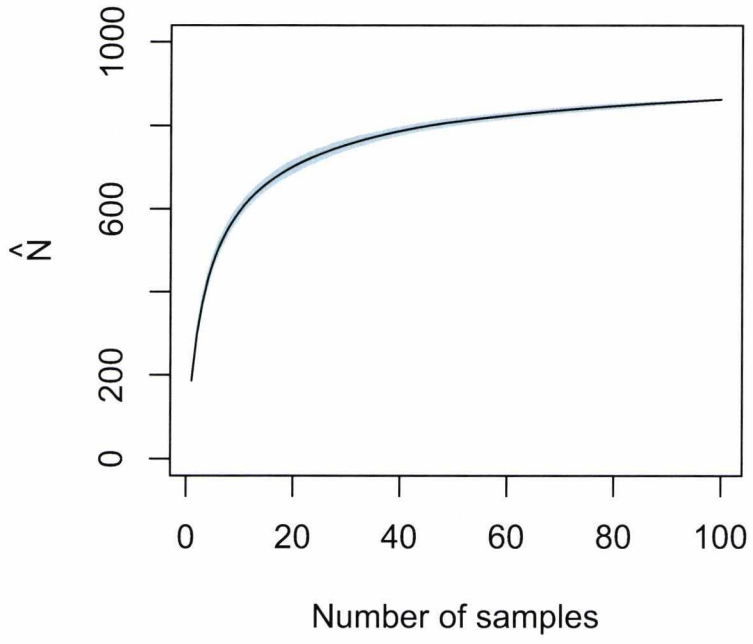



*Figure 2.2:* Species accumulation curve of negative binomial simulated abundance data for 1000 species. The shaded area shows the 95% confidence interval from the standard deviation produced from 100 permutations of the data.

Baltanas (1992) found that the performance of a species-area method was strongly affected by spatial aggregation of species, giving poor estimates that underestimate total species richness. Therefore this method is not recommended for benthic data sets, where spatial clustering of individuals within a species is anticipated.

In the study by Ugland and Gray (2004) the Total-species (T-S) accumulation method

gave a reasonable estimate for a benthic data set. However, when used on data from bird surveys taking into account habitat type and altitude, it overestimated total species richness (O'Dea et al., 2006). Further investigation into the strengths and limitations of this approach are required. This estimator also requires additional information to construct the subareas used within the estimator, which may not always be available.

The $S_{\infty}$ estimator is not suitable for benthic data sets, because with few samples this estimator tends to underestimate and could possibly even produce negative species richness estimates (Karakassis, 1995). Moreover, an assumption of this model, pointed out by Ugland and Gray (2004), is that the probability of being caught for any individual is proportional to sampling effort, and that 'catchability' is independent of the effort and is the same for all individuals. However, for large benthic communities, the catchability of rare species may not be the same as the catchability of more common species (Ugland and Gray, 2004).

Non-parametric estimators have been shown to be less biased and more precise than species accumulation curve extrapolation methods in general (Brose et al., 2003). However, when the performance of a number of estimators was ranked in a study on marine data sets, the results were in contrast to previous studies (Foggo et al., 2003; Walther and Morand, 1998; Hellmann and Fowler, 1999; Condit et al., 1996); in the study by Foggo et al. (2003) using three marine data sets, the first-order jackknife produced low variability, but showed decreasing precision with increasing numbers of unique species in the data set. The $Chao_1$ estimator also produced low precision when applied to a low-richness, low-abundance data set (Foggo et al., 2003).

Ugland and Gray (2004) analysed the performance of the $Chao_1$ estimator on a benthic data set and found that it was significantly underestimating the true species richness with limited sampling effort. The $Chao_1$ estimator was originally proposed to be a lower bound, however it has recently been justified for use as a point estimate

in extreme heterogeneous cases when the singletons in the sample have the same relative abundances in the community (Shen et al., 2003).

Studies suggest that the patchy spatial distribution of macrobenthic assemblages represents a hurdle to incidence-based estimators such as the $Chao_2$ estimator and $ICE$ (Chao, 2004). The $Chao_2$ estimator overestimated total species richness and displayed poor precision and accuracy when applied to marine data (Foggo et al., 2003). Using $ACE$, estimation of the number of missing species is based entirely on the rare species' frequencies, and as previously mentioned the assumption that rare species carry most information about unobserved species is not likely to be satisfied for benthic data sets. The bootstrap estimator has been shown to underestimate the number of species if there are a large number of rare species and the number of samples is small (Smith and van Belle, 1984).

Brose et al. (2003) found that no single estimator performed best at estimating species richness in all cases. However, most perform considerably better than taking the observed number of species as an estimate. Walther and Moore (2005) pointed out that there are no estimators that are suitable for all situations, or especially effective for a particular taxon, unless their performance is tied to the species-abundance distribution and sampling protocol used for that taxon. Therefore we may wish to incorporate the distribution of the species within an estimator.

## 2.4 Clustering of benthic organisms

In the natural world, clustered patterns are very common. For example, trees in natural forests have a clustered distribution, the pattern of which depends on the seed dispersal mode (Li et al., 2009). Populations of most benthic marine invertebrate species also have a clustered distribution (Heip, 1975), and estimates of species richness can be biased because rarer species are often missed during sampling, and clustering of individuals within a species means that a species is more likely to be missed during sampling.

Spatial clustering can be caused by a number of different processes, but I assume that the pattern occurring in benthic organisms can be adequately modelled by a mechanism involving 'parent points' and 'daughter points'; where the daughter points are scattered around the parent points. Such patterns are best described by the classical cluster model below.

Every point, $y$, in a given initial point process, $C_p$, is replaced by a cluster of $C^y$ points, not including the original point $y$ (Illian et al., 2008). The clusters, $C^y$, are finite point processes, and their set-theoretic union is the cluster point process

$$B = \bigcup_{y \in C_p} C^y. \tag{2.8}$$

Where the parent points form a Poisson process, Neyman-Scott cluster models can be utilised (Neyman and Scott, 1958). These model the location of cluster centres as a homogeneous Poisson process; the number of individuals in each cluster is modelled as a Poisson variable and these are located these around the cluster centre according to an isotropic spatial process.

A particular type of Neyman-Scott process, the Matérn process (Matérn, 1986) is used to model the spatial distribution of benthic organisms.

### *2.4.1 Matérn process*

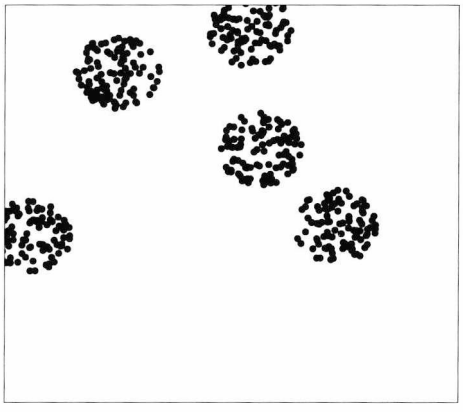
The Matérn process has three components:

1. Parent events form a Poisson process with intensity $\lambda$ (i.e. 'parents' are randomly distributed over the area with mean $\lambda$ per unit area),

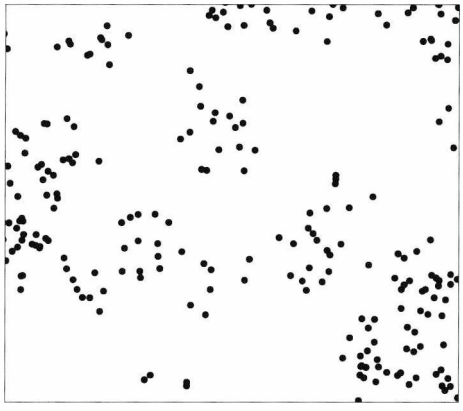
2. Each parent produces a Poisson number of 'daughters' with mean $\phi$,

3. The positions of the 'daughters' relative to their 'parents' are randomly chosen within a circle of radius $R$.

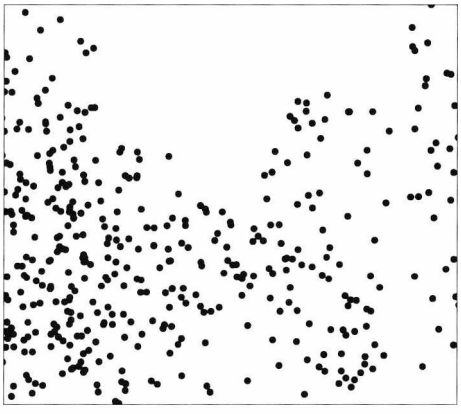
We are interested in the number of individuals (daughters) that fall into a given area representing the area sampled by a single grab in a benthic survey. If $\lambda$ is low and $\phi$ is high, then the Matérn process is very clustered when the radius is small (Figure 2.3a). If $\lambda$ is high and $\phi$ is low, the process looks more random (Figure 2.3b). As the radius parameter, $R$, increases, the realisations move further towards a random distribution (Figures 2.3c and 2.3d). The Matérn process assumes that clusters are independent of each other.



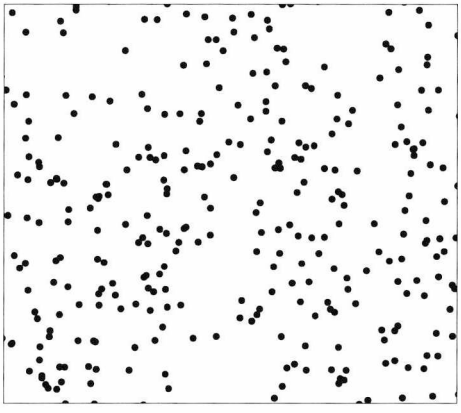
*(a)* $\lambda$=10, $\phi$=100, R=0.05.



*(b)* $\lambda$=100, $\phi$=10, R=0.05.



*(c)* $\lambda$=10, $\phi$=100, R=0.2.



*(d)* $\lambda$=100, $\phi$=10, R=0.2.

*Figure 2.3:* Realisations of the Matérn process.

Data can be simulated using the three components of the Matérn process, and the number of points falling within a defined area can be summed to give a sample for a particular grab size. Data can be simulated by producing a Matérn realisation of points using parameters of choice, by first generating a number of parents with their locations randomly distributed within an area equal in size to the grab area plus a boundary of width $R$ to allow for any parents that may lie outside the grab area, but generate children within the grab area (See Figure 2.4). Then a Poisson number

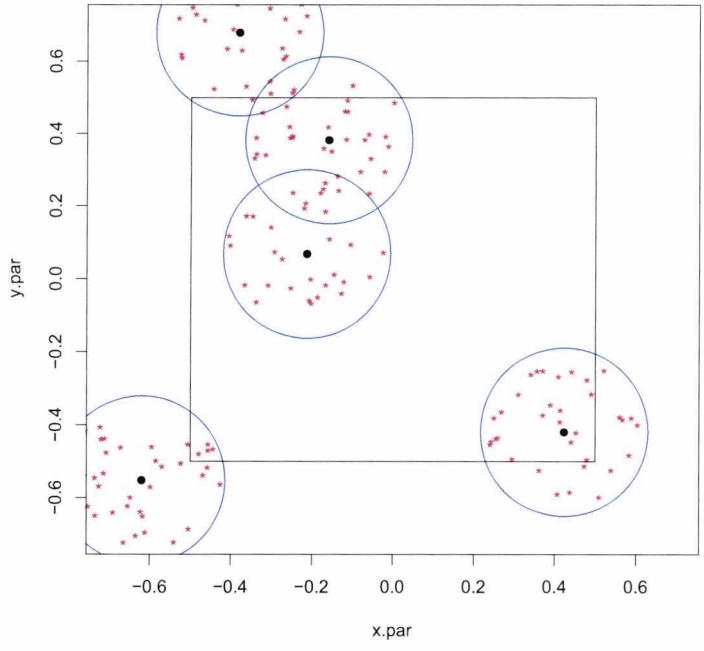



*Figure 2.4:* Example of a realisation of the Matérn process for $R = 0.2$, $\lambda = 5$ and $\phi = 30$. The square indicates the grab area of unit size. The closed circles represent the parents and the stars the children, with the open circles indicating the outer limits of the clusters. We can see that in some cases the parents lie outside the grab area, but can still produce children which lie within the grab area.

of children are allocated per parent, with intensity $\phi$, and located randomly within a circle of radius $R$ around the parent. The number of children located within the grab area are then output to the data set as the result of one grab sample. This was repeated $g$ times for each species, where $g$ is the number of grab samples.

## 2.5 Performance of species richness estimators

I investigated the performance of the non-parametric species richness estimators $Chao_1$, the first and second order jackknife, $ACE$ and the bootstrap when applied to clustered data. These estimators were chosen because they are commonly used to estimate species richness, and were found to perform well in some previous studies (Foggo et al., 2003; Brose et al., 2003).

I did not include species accumulation methods as in most of the benthic data sets to be considered there are very few grabs samples, so this method would not be effective. I also did not include incidence based estimators such as $ICE$ and $Chao_2$ as it has previously been shown that these estimators do not perform well for data exhibiting clustering, and also I have abundance data so it seemed wasteful to use an incidence based estimator which will throw away some of the available data.

Using the Matérn process I simulated populations of 1000 species, using various Matérn parameters. I compared the performance of the estimators over varying sample sizes, and would expect that as the number of grabs increases, the performance of estimators should improve. I also investigated the effect of increasing the radius, and varying the mean number of children per parent, that is clustering intensity.

During these simulations I not only considered the effect of varying each individual parameter, but also the simultaneous variation of the three parameters. A similar pattern emerged across the results, and it was clear to see the effect of changing each parameter, and so only the results of varying the individual parameters are reported.

I also considered how the estimators performed when applied to benthic data collected off the coast of the Isle of Wight.

### 2.5.1 Varying sample size

To investigate the effect of varying sample size I performed simulations for various number of grabs. I performed the simulations for various cluster intensities, however a similar pattern emerged in each case, so I show here only the results for $\mu = 10$, $\lambda = 5$, $R = 0.05$, where $\mu = \lambda\phi$.

We can see that for these parameter values, when the number of grabs increases to over 50 we are observing most of the species present in the survey area (Table 2.1). At 50 grabs the first order jackknife and bootstrap estimators are slightly over estimating the species richness on average, however the 95% central interval shows a tight range of estimates with low variation between them.

As we might expect, at 100 grabs all the estimators give a species richness of $\hat{N} = 1000$, which is equal to the observed number of species and the true species richness. The estimators should perform well when all species have been seen, because they are primarily based on the number of singletons and doubletons in the sample.

When our sample size is only five grabs, we see that the estimators are negatively biased, and underestimate the true species richness significantly at 10 grabs also. However, the second order jackknife gets relatively close to the true species richness within its 95% central interval over the 50 simulations.

### 2.5.2 Varying cluster radius

To investigate the effect of varying the cluster radius I performed simulations for various $R$ values. I performed the simulations for various clusterings, however a similar pattern emerged in each case so I show here only the results for $\mu = 10$, $\lambda = 5$.

Table 2.2 shows that as the radius increases, the non-parametric estimators improve in their performance. All of the estimators perform reasonably well when the radius

| No. grabs | Estimator | Mean | Median | SD | RMSE | MAE | 95% central interval |
|---|---|---|---|---|---|---|---|
| 5 | $D$ | 451 | 451 | 14 | 549 | 549 | (426, 478) |
| | $\hat{N}_{C1}$ | 530 | 529 | 25 | 471 | 470 | (485, 576) |
| | $\hat{N}_{J1}$ | 584 | 583 | 23 | 416 | 416 | (544, 625) |
| | $\hat{N}_{J2}$ | 603 | 606 | 34 | 399 | 397 | (531, 665) |
| | $\hat{N}_{ACE}$ | 506 | 506 | 18 | 494 | 494 | (475, 540) |
| | $\hat{N}_{B}$ | 521 | 521 | 18 | 479 | 479 | (491, 552) |
| 10 | $D$ | 699 | 696 | 14 | 301 | 301 | (673, 725) |
| | $\hat{N}_{C1}$ | 777 | 772 | 22 | 224 | 223 | (736, 822) |
| | $\hat{N}_{J1}$ | 846 | 844 | 21 | 155 | 154 | (809, 888) |
| | $\hat{N}_{J2}$ | 852 | 848 | 34 | 152 | 148 | (776, 917) |
| | $\hat{N}_{ACE}$ | 745 | 741 | 16 | 256 | 255 | (719, 777) |
| | $\hat{N}_{B}$ | 780 | 776 | 16 | 220 | 220 | (753, 812) |
| 50 | $D$ | 998 | 998 | 1 | 2 | 2 | (995, 1000) |
| | $\hat{N}_{C1}$ | 1000 | 1000 | 3 | 3 | 2 | (996, 1006) |
| | $\hat{N}_{J1}$ | 1004 | 1004 | 3 | 5 | 4 | (998, 1009) |
| | $\hat{N}_{J2}$ | 998 | 998 | 7 | 7 | 6 | (984, 1011) |
| | $\hat{N}_{ACE}$ | 999 | 999 | 1 | 2 | 1 | (996, 1001) |
| | $\hat{N}_{B}$ | 1004 | 1004 | 1 | 4 | 4 | (999, 1006) |
| 100 | $D$ | 1000 | 1000 | 0 | 0 | 0 | (1000, 1000) |
| | $\hat{N}_{C1}$ | 1000 | 1000 | 0 | 0 | 0 | (1000, 1000) |
| | $\hat{N}_{J1}$ | 1000 | 1000 | 0 | 0 | 0 | (1000, 1001) |
| | $\hat{N}_{J2}$ | 1000 | 1000 | 1 | 1 | 0 | (999, 1002) |
| | $\hat{N}_{ACE}$ | 1000 | 1000 | 0 | 0 | 0 | (1000, 1000) |
| | $\hat{N}_{B}$ | 1000 | 1000 | 0 | 0 | 0 | (1000, 1000) |

*Table 2.1:* Summary statistics of non-parametric species richness estimators over 50 simulated data sets from the Matérn process with $N = 1000$, $\mu = 10$, $\lambda = 5$, $R = 0.05$ and varying the sample size. Results shown are the mean, median, standard deviation, SD, root mean squared error, RMSE, mean absolute error, MAE, and 95% central interval across the 50 estimates.

| $R$ | Estimator | Mean | Median | SD | RMSE | MAE | 95% central interval |
|---|---|---|---|---|---|---|---|
| 0.01 | D | 666 | 664 | 12 | 334 | 334 | (640, 688) |
| | $\hat{N}_{C1}$ | 719 | 717 | 16 | 281 | 281 | (694, 747) |
| | $\hat{N}_{J1}$ | 786 | 784 | 17 | 215 | 214 | (752, 819) |
| | $\hat{N}_{J2}$ | 769 | 769 | 27 | 232 | 231 | (724, 816) |
| | $\hat{N}_{ACE}$ | 701 | 700 | 13 | 299 | 299 | (674, 725) |
| | $\hat{N}_{B}$ | 737 | 735 | 13 | 264 | 263 | (706, 761) |
| 0.05 | D | 701 | 702 | 14 | 300 | 299 | (663, 726) |
| | $\hat{N}_{C1}$ | 774 | 774 | 20 | 227 | 226 | (732, 826) |
| | $\hat{N}_{J1}$ | 845 | 846 | 20 | 156 | 155 | (795, 893) |
| | $\hat{N}_{J2}$ | 845 | 840 | 30 | 158 | 155 | (787, 922) |
| | $\hat{N}_{ACE}$ | 745 | 748 | 16 | 255 | 255 | (703, 778) |
| | $\hat{N}_{B}$ | 781 | 784 | 16 | 219 | 219 | (737, 813) |
| 0.1 | D | 732 | 732 | 15 | 268 | 268 | (694, 761) |
| | $\hat{N}_{C1}$ | 830 | 830 | 25 | 172 | 170 | (788, 882) |
| | $\hat{N}_{J1}$ | 902 | 902 | 22 | 100 | 98 | (864, 943) |
| | $\hat{N}_{J2}$ | 921 | 920 | 36 | 87 | 79 | (853, 996) |
| | $\hat{N}_{ACE}$ | 788 | 788 | 17 | 213 | 212 | (753, 821) |
| | $\hat{N}_{B}$ | 824 | 824 | 17 | 177 | 176 | (785, 857) |
| 0.5 | D | 870 | 868 | 10 | 130 | 130 | (849, 886) |
| | $\hat{N}_{C1}$ | 978 | 978 | 21 | 31 | 26 | (935, 1021) |
| | $\hat{N}_{J1}$ | 1087 | 1087 | 18 | 88 | 87 | (1047, 1122) |
| | $\hat{N}_{J2}$ | 1083 | 1081 | 36 | 90 | 83 | (1004, 1158) |
| | $\hat{N}_{ACE}$ | 954 | 954 | 12 | 48 | 46 | (928, 976) |
| | $\hat{N}_{B}$ | 991 | 990 | 11 | 15 | 12 | (967, 1009) |

*Table 2.2:* Summary statistics of non-parametric species richness estimators over 50 simulated data sets from the Matérn process with $N = 1000$, $\mu = 10$, $\lambda = 5$, and varying the Matérn radius. Results shown are the mean, median, standard deviation, SD, root mean squared error, RMSE, mean absolute error, MAE, and 95% central interval across the 50 estimates.

is large at 0.5, in comparison to the grab size of $0.25m^2$. However we can see that jackknife estimators overestimate the true species richness.

For all estimators and parameter values, the species richness estimators are performing better than the observed species richness. However when the radius is small the estimators are biased. The variation in the estimates decreases as I increase the radius.

Overall, the jackknife estimators seem to be performing the best in terms of bias. However the bootstrap estimator and the $Chao_1$ estimator show much less variance. ACE performs the worst, showing greater negative bias in its estimates than the other non-parametric estimators, even when the radius is larger. However, ACE does perform better than $D$, showing that any species richness estimator is better than just using the raw species count.

### 2.5.3 *Varying clustering intensity*

To investigate the effect of varying clustering intensity, I performed simulations for various $\lambda$ values, the mean number of parents. I performed the simulations for various abundances, however a similar pattern emerged for each so I show here only the results for $\mu = 10$. Therefore the corresponding clustering intensity is $10/\lambda = (10, 2, 1, 0.5)$.

Table 2.3 shows that as the clustering intensity decreases, the non-parametric estimators improve in their performance. All of the estimators perform reasonably well when clustering intensity is low at 0.5. However we can see that the first order jackknife seems to overestimate the true species richness.

Again the species richness estimators are performing better than the observed species richness. However, when clustering intensity is higher, they do not estimate much more than the observed species richness, and are highly biased. The variance in the estimates is reflected in the variance of the observed number of species between

| $\lambda$ | Estimator | Mean | Median | SD | RMSE | MAE | 95% central interval |
|---|---|---|---|---|---|---|---|
| 1 | $D$ | 276 | 278 | 14 | 724 | 724 | (245, 305) |
| | $\hat{N}_{C1}$ | 287 | 288 | 15 | 713 | 713 | (254, 318) |
| | $\hat{N}_{J1}$ | 293 | 291 | 15 | 708 | 707 | (258, 322) |
| | $\hat{N}_{J2}$ | 295 | 295 | 16 | 705 | 705 | (262, 328) |
| | $\hat{N}_{ACE}$ | 279 | 280 | 14 | 721 | 721 | (247, 308) |
| | $\hat{N}_{B}$ | 285 | 286 | 15 | 715 | 715 | (252, 315) |
| 5 | $D$ | 694 | 694 | 17 | 307 | 306 | (657, 726) |
| | $\hat{N}_{C1}$ | 767 | 766 | 21 | 234 | 233 | (733, 810) |
| | $\hat{N}_{J1}$ | 838 | 836 | 21 | 164 | 162 | (796, 880) |
| | $\hat{N}_{J2}$ | 837 | 833 | 31 | 166 | 163 | (790, 903) |
| | $\hat{N}_{ACE}$ | 739 | 740 | 17 | 262 | 261 | (700, 769) |
| | $\hat{N}_{B}$ | 774 | 776 | 18 | 226 | 226 | (733, 808) |
| 10 | $D$ | 812 | 812 | 13 | 189 | 188 | (772, 833) |
| | $\hat{N}_{C1}$ | 921 | 923 | 21 | 82 | 79 | (875, 956) |
| | $\hat{N}_{J1}$ | 1013 | 1018 | 19 | 23 | 19 | (965, 1043) |
| | $\hat{N}_{J2}$ | 1026 | 1029 | 34 | 42 | 35 | (955, 1078) |
| | $\hat{N}_{ACE}$ | 885 | 886 | 15 | 116 | 115 | (843, 908) |
| | $\hat{N}_{B}$ | 921 | 923 | 15 | 80 | 79 | (878, 947) |
| 20 | $D$ | 871 | 872 | 10 | 130 | 129 | (849, 888) |
| | $\hat{N}_{C1}$ | 977 | 981 | 19 | 30 | 24 | (934, 1012) |
| | $\hat{N}_{J1}$ | 1086 | 1087 | 19 | 88 | 86 | (1049, 1121) |
| | $\hat{N}_{J2}$ | 1081 | 1090 | 36 | 89 | 81 | (1000, 1141) |
| | $\hat{N}_{ACE}$ | 953 | 955 | 13 | 48 | 47 | (927, 977) |
| | $\hat{N}_{B}$ | 991 | 992 | 12 | 15 | 12 | (966, 1011) |

*Table 2.3:* Summary statistics of non-parametric species richness estimators over 50 simulated data sets from the Matérn process with $N = 1000$, $\mu = 10$, $R = 0.05$ and varying the clustering intensity. Results shown are the mean, median, standard deviation, SD, root mean squared error, RMSE, mean absolute error, MAE, and 95% central interval across the 50 estimates.

simulations which we can see by looking at the variance of $D$.

The jackknife estimators seem to be performing the best in terms of bias, and the bootstrap estimator and the $Chao_1$ estimator show much less variance. ACE performs the worst of the richness estimators used.

### 2.5.4 Estimators applied to Isle of Wight benthic data

Here I show the inadequacy of these estimators in estimating species richness for benthic data. I use the Isle of Wight data set as an example, where a survey collected data using two grab sizes, large $0.25m^2$ and small $0.1m^2$. Overall 273 species were recorded in the area, 240 using the large grabs, and 198 using the small grabs. Ten grabs were collected for each grab size.

For a species richness estimator to be adequate in modelling these data, the minimum number of species it should estimate is 273, the observed number of species seen in this area. An estimator that gives a value less than this is underestimating the species richness, and will not be useful for similar benthic data sets.

Table 2.4 shows the estimates given by the non-parametric estimators for the Isle of Wight data. There were not readily available variance formulae for all estimators, but the variance and confidence intervals have been calculated where possible. We can see that for the data collected using the small grabs, only the second order Jackknife gives an estimated species richness above the total observed number of species of 273. All the other estimators perform poorly, however the observed species richness does lie within the 95% confidence interval of the $Chao_1$ estimator.

For the data collected using the large grabs, the bootstrap estimator again does not reach the total observed number of species of 273. However, the other estimators do. Comparing the estimates from the first and second order jackknife, we see that including more data in our estimation increases the species richness estimate, and I

| Grab size | Estimator | Estimate | SD | 95% Confidence Interval |
|---|---|---|---|---|
| $0.25m^2$ | $D$ | 240 | - | - |
| | $\hat{N}_{C1}$ | 286 | 16.71 | (263, 332) |
| | $\hat{N}_{J1}$ | 293 | 10.26 | (273, 313) |
| | $\hat{N}_{J2}$ | 314 | - | - |
| | $\hat{N}_{ACE}$ | 285 | - | - |
| | $\hat{N}_{B}$ | 265 | - | - |
| $0.1m^2$ | $D$ | 198 | - | - |
| | $\hat{N}_{C1}$ | 241 | 15.6 | (219, 284) |
| | $\hat{N}_{J1}$ | 258 | 6.85 | (245, 271) |
| | $\hat{N}_{J2}$ | 287 | - | - |
| | $\hat{N}_{ACE}$ | 246 | - | - |
| | $\hat{N}_{B}$ | 226 | - | - |

*Table 2.4:* Non-parametric species richness estimates for Isle of Wight benthic data sets. Results shown are the species richness estimate, standard deviation, SD, and 95% confidence interval (where available) for the species richness estimators $D$: the number of observed species, $\hat{N}_{C1}$: the $Chao_1$ estimator, $\hat{N}_{J1}$: the first order jackknife, $\hat{N}_{J2}$: the second order jackknife, $\hat{N}_{ACE}$: the Abundance-based Coverage Estimator, $\hat{N}_{B}$: the bootstrap estimator.

would expect that the larger estimate is therefore closer to the true species richness. This estimate is substantially larger than that of the $Chao_1$ species richness estimator, however it does lie within the 95% confidence interval of the latter.

On the whole we can see that the non-parametric species richness estimators perform badly for this benthic data set, possibly as they were not designed for clustered data, and therefore do not account for this during their estimation.

## 2.6 Discussion

The aims of this chapter were to introduce the concept of species richness, and describe some of the methods from the literature which have been developed to estimate this measure of biodiversity. Some of these estimators were considered in more detail, with the aim to highlight that they are inadequate when applied to clustered data, and this was shown by a simulation study and also by application to a benthic data set. A clustering model was introduced, that might be used to model the spatial distribution of benthic organisms.

A review of the species richness literature showed that there is no estimator that is best in all cases, and this was supported by our results. In the simulation study I found that the jackknife estimators performed the best in terms of bias. However the bootstrap estimator and the $Chao_1$ estimator show much less variance.

I have shown that as the sample size increased, the estimates improved. However they did underestimate species richness significantly when there was a low number of grabs. A similar pattern emerged when increasing the cluster radius, and decreasing the clustering intensity of the individuals.

As the radius increases and the clustering intensity decreases, the spatial distribution of the individuals within a species becomes more like a Poisson process. Since the non-parametric estimators were based on the Poisson distribution, they should be able to estimate species richness well at this level.

All of the species richness estimators perform better than using the observed number of species, showing that any species richness estimator is better than just using the raw species count. However, when clustering intensity is higher, they do not estimate much more than the observed species richness, and are highly biased. The variance of the estimates is high in some cases, but this reflects the high variance in the observed number of species between simulations which is shown by the variance of $D$.

When applying the estimators to the benthic data all the estimators behaved poorly. They do not estimate species richness well for a benthic data set, probably because of spatial clustering, suggesting an alternative approach is required.

Some of the non-parametric estimators were developed as lower bound estimators, however we require an estimator that will accurately estimate the true species richness, and a confidence interval for the estimate.

This chapter has shown the importance of using a species richness estimator other than just taking the observed number of species. Although the non-parametric estimators considered are limited by their bias, I have shown they improve on the observed number of species, but are not suitable for our purpose. The review of the literature has also shown that species accumulation methods fail for clustered data. Therefore I will consider an alternative approach to species richness estimation, using a parametric approach.

Walther and Moore (2005) pointed out that there are no estimators that are especially effective for a particular taxon, unless their performance is tied to the species-abundance distribution and sampling protocol used for that taxon. Therefore, in Chapter 3 I introduce an estimator that is linked to the species abundance distribution of benthic organisms, thus making it suitable for the estimation of species richness for benthic organisms in many situations.

## *2.7 Conclusions*

This chapter shows that there is clearly some scope for the development of a new species richness estimator for use with benthic data. Current methods such as non-parametric estimators, do not deal well with spatially clustered data, and this has been confirmed by my simulation study. I have also shown that merely using the number of observed species is inadequate for biodiversity estimation.

I have introduced a clustering model that could be used to describe the spatial distribution of the benthic organisms, and Chapter 3 will consider how this can be built into a parametric species richness estimator.

# 3. SPECIES RICHNESS ESTIMATION - A MAXIMUM-LIKELIHOOD FREQUENTIST APPROACH

## *3.1 Introduction*

Chapter 2 described a range of species richness estimators. It is important that species richness is estimated accurately, as underestimation may mean that an important area for biodiversity is missed. Alternatively, overestimation might mean wasting conservation effort on areas that are not very diverse. Chapter 2 showed that current methods that do not account for spatial clustering are not adequate for analysing benthic data. Therefore, this chapter will present an approach to species richness estimation which incorporates the spatial pattern of the individuals of each species. As species spatial heterogeneity factors will be taken into account, this estimator should be more suitable for estimating total species richness for benthic data.

This chapter applies the frequentist approach to species richness estimation, and introduces a method using the Neyman Type A distribution to model the spatial pattern of individuals within a species. The chapter aims to present a model which can accurately estimate species richness when species are clustered, using a maximum-likelihood approach. Methods to calculate confidence intervals and goodness of fit measures are examined.

Problems which arose with the maximum-likelihood approach include the question of truncating data and the boundary problem, and I introduce an approach to dealing with the boundary problem using penalties. Several data sets are analysed using the maximum-likelihood approach.

## 3.2 Maximum-likelihood estimation

Barry (2009) proposed a parametric approach to estimate species richness by combining variable species abundance and the spatial pattern of the species' members. As species heterogeneity factors will be taken into account, this estimator should be suitable for estimating species richness for benthic data from grab samples. Models are fitted by the method of maximum likelihood, which was first developed in this context by Craig (1953), following Fisher et al. (1943) in fitting a parametric model to observed species abundances.

The aim is to estimate the total number of species in a survey area. Barry (2009) notes that the number of species that will be detected in $g$ grabs will be determined by the total number of species in the area, the number of individuals per species, and the spatial pattern of individuals of each species. Barry (2009) describes his approach as method-of-moments estimation, but the method can also be regarded as maximum-likelihood.

If we let $x_i$ be the observed number of individuals of species $i$ in our sample, for $i = 1, ..., D$, and $f_k$ be the number of species seen $k$ times, then of the $N$ species present in the population, only $D$ have been observed, and there are $N-D$ unobserved species in cell $f_0$.

Letting $p_k(\boldsymbol{\theta})$ be the probability of seeing a species $k$ times, for $k = 1, ..., r$, distributed according to some abundance distribution with parameters $\boldsymbol{\theta}$, we can write $p_0(\boldsymbol{\theta}) = 1 - \sum_{k \geq 1} p_k(\boldsymbol{\theta})$. Then $p_k(\boldsymbol{\theta})/(1 - p_0(\boldsymbol{\theta}))$ are the zero-truncated cell probabilities.

The likelihood for $N$ and $\boldsymbol{\theta}$ can be written as

$$L(N, \boldsymbol{\theta}) = \frac{N!}{(N-D)!} p_0(\boldsymbol{\theta})^{N-D} \{1 - p_0(\boldsymbol{\theta})\}^D \prod_{k \geq 1} \left( \frac{p_k(\boldsymbol{\theta})}{(1 - p_0(\boldsymbol{\theta}))} \right)^{f_k}. \qquad (3.1)$$

This can be factorised as $L(N, \boldsymbol{\theta}) = L_b(N, \boldsymbol{\theta}) L_c(\boldsymbol{\theta})$ (Sanathanan, 1977) where the first likelihood, $L_b(N, \boldsymbol{\theta})$, can be written as

$$L_b(N, \boldsymbol{\theta}) = \frac{N!}{(N-D)!} p_0(\boldsymbol{\theta})^{N-D} \{1 - p_0(\boldsymbol{\theta})\}^D, \tag{3.2}$$

and the second as

$$L_c(\boldsymbol{\theta}) = \prod_{k \geq 1} \left( \frac{p_k(\boldsymbol{\theta})}{(1 - p_0(\boldsymbol{\theta}))} \right)^{f_k}. \tag{3.3}$$

There are two widely used estimators of the parameters $(N, \boldsymbol{\theta})$ (Fewster and Jupp, 2009):

1. The maximum-likelihood estimator (MLE) $(\hat{N}, \hat{\boldsymbol{\theta}})$, which maximises the full likelihood (Equation 3.1) simultaneously with respect to $N$ and $\boldsymbol{\theta}$,

2. The conditional maximum-likelihood estimator $(\hat{N}_c, \hat{\boldsymbol{\theta}}_c)$, where $\hat{\boldsymbol{\theta}}_c$ is the value of $\boldsymbol{\theta}$ which maximises the conditional likelihood of $\boldsymbol{\theta}$ based on the conditional distribution of $f_1, ..., f_r$ (Equation 3.3), and $\hat{N}_c$ is the value of $N$ that maximises $L_b(N, \hat{\boldsymbol{\theta}}_c)$. This approach results in the conditional likelihood estimator for $N$ (Sanathanan, 1977):

$$\hat{N}_c = \frac{D}{1 - p_0(\hat{\boldsymbol{\theta}}_c)}. \tag{3.4}$$

Theorem 3 of Sanathanan (1972) states that the estimators $\hat{N}$ and $\hat{N}_c$ satisfy $\hat{N}_c \geq \hat{N}$. Despite the discrepancy between $\hat{N}_c$ and $\hat{N}$, which is often small in practice, in many cases it is simpler to calculate the conditional likelihood estimator, and this method most certainly had computational advantage when these estimators were considered in the 1970s.

Fewster and Jupp (2009) have shown that the difference between $\hat{N}_c$ and $\hat{N}$ is of order 1, and as $N \to \infty$, $\hat{N}_c / \hat{N} \to 1$ (Chao and Bunge, 2002).

## 3.3 Species abundance models

In the simple case that the population species abundances come from a Poisson distribution, with equal mean abundance of $\lambda$ for each species, then the probability of seeing a species $k$ times in a single grab is

$$p_k(\lambda) = \frac{(\lambda S)^k e^{-\lambda S}}{k!},$$

where $S$ is the sampling effort. The probability that a species is not found is

$$p_0(\lambda) = e^{-\lambda S}.$$

For a single grab of area $a$ from a total area $A$, $S = a/A$. Since grabs are independent, and in our case $a << A$ (Bolam, 2011), the possibility that grab samples overlap can be ignored in practice. I can extend this to get the probability that a species is not detected in $g$ grabs by letting $S = ga/A$. Within each survey, sampling effort is constant, and therefore I can incorporate $S$ into the abundance parameter $\lambda$.

We can extend the Poisson model by allowing the parameter $\lambda$ to vary between species. Specifically, the expected numbers of individuals per unit area, $\lambda$, are modelled as a random sample from a mixing distribution with density $f(\lambda; \boldsymbol{\theta})$. The marginal probability that a particular species is observed $k$ times in the sample is

$$p_k(\boldsymbol{\theta}) = \int_0^\infty p_k(\lambda) f(\lambda; \boldsymbol{\theta}) \mathrm{d}\lambda, \ k = 0, 1, ... \tag{3.5}$$

where $\boldsymbol{\theta}$ are the parameters describing the abundance distribution.

A density commonly adopted to describe the distribution of the species abundances is the gamma density:

$$f(\lambda; \alpha, \beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}. \tag{3.6}$$

The probability of observing a species $k$ times in the sample becomes

$$p_k(\boldsymbol{\theta}) = \int_0^\infty \frac{(\lambda)^k e^{-\lambda}}{k!} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \mathrm{d}\lambda. \tag{3.7}$$

This corresponds to a negative binomial distribution for the species counts, and $p_k(\boldsymbol{\theta})$ can be written as

$$p_k(\boldsymbol{\theta}) = \frac{\Gamma(\alpha + k)}{k!\Gamma(\alpha)} \left(\frac{1}{\beta + 1}\right)^{\alpha} \left(\frac{\beta}{\beta + 1}\right)^{k}. \tag{3.8}$$

For $\alpha = 1$, the gamma distribution simplifies to an exponential distribution and the number of individuals seen per species will follow a geometric distribution, where

$$p_k(\boldsymbol{\theta}) = \left(\frac{1}{\beta + 1}\right) \left(\frac{\beta}{\beta + 1}\right)^{k-1}. \tag{3.9}$$

### 3.3.1 *Incorporating clustering into the species abundance model*

I wish to assume the realistic situation for benthic species that densities vary between species, and also that individuals of a species are highly clustered. I therefore describe the species abundances using a Neyman-Scott clustering process (Neyman and Scott, 1958).

Barry (2009) proposed the use of a Matérn process (Matérn, 1986) to model the spatial distribution of benthic organisms. If we suppose that the radius of the clusters, $R$, is very small in comparison to the size of the grab sample, we will have very tight clusters and can assume to a good approximation that an individual will be located in the grab sample if and only if the cluster centre is located in the grab sample area. If $R = 0$ the Matérn process reduces to a Neyman Type A distribution (Johnson et al., 1997) with probability function

$$p_x(\lambda, \phi) = \frac{e^{-\lambda}\phi^x}{x!} \sum_{j=0}^{\infty} \frac{(\lambda e^{-\phi})^j j^x}{j!}. \tag{3.10}$$

This seems a plausible step to take as Skellam (1958) states that

> *'the type A distribution exhibits considerable robustness, and can be employed as an approximation in certain circumstances where the condition requiring compact clustering can be greatly relaxed'.*

In addition, the Neyman Type A distribution has been widely used to model spatial distributions for ecological data including plants, larvae and bacteria (Evans, 1953; Bliss and Fisher, 1953; Martin and Katti, 1962).

### 3.3.2 The Neyman Type A model

If the distribution of the individuals has a Neyman Type A distribution, then for a particular species

$$p_k(\lambda, \phi) = \frac{e^{-\lambda}\phi^k}{k!} \sum_{j=0}^{\infty} \frac{(\lambda e^{-\phi})^j j^k}{j!}, \tag{3.11}$$

where

$$E(Y) = \lambda\phi, \text{ and } \mathrm{Var}(Y) = \lambda\phi(1+\phi).$$

Given this Neyman Type A distribution for the spatial distribution of a particular species, we can assume that both parameters $\lambda$ and $\phi$ will vary between species. To estimate the Neyman Type A parameters we can assume that each is a random variable from some mixing distribution (or if the parameters are not thought to be independent, from a bivariate distribution).

To use the estimate $\hat{N}$ we find the expectation over the joint distribution of $\lambda$ and $\phi$. If I give the mean abundance, $\lambda\phi = \mu$, a gamma distribution, then the probability that a species is observed $k$ times in a sample is the marginal likelihood

$$p_k(\boldsymbol{\theta}) = \int_0^{\infty} \frac{e^{-\lambda}(\phi)^k}{k!} \sum_{j=0}^{\infty} \left\{ \frac{(\lambda e^{-\phi})^j j^k}{j!} \right\} \frac{\beta^\alpha (\lambda\phi)^{\alpha-1} e^{-\beta\lambda\phi}}{\Gamma(\alpha)} \mathrm{d}\lambda, \tag{3.12}$$

where $\boldsymbol{\theta} = (\phi, \alpha, \beta)$. This allows the species density to vary, but assumes that the mean number of individuals per cluster, $\phi$, does not vary between species. I term this the Neyman Type A-gamma distribution.

If we extend the model so that not only the mean abundance but also the clustering parameter, $\phi$, can vary between species, the probability that a species is observed $k$

times in a sample becomes

$$p_k(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty \frac{e^{-\lambda}\phi^k}{k!} \sum_{j=0}^{\infty} \left\{ \frac{(\lambda e^{-\phi})^j j^k}{j!} \right\} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \frac{b^a \phi^{a-1} e^{-b\phi}}{\Gamma(a)} \mathrm{d}\lambda \mathrm{d}\phi, \tag{3.13}$$

where $\phi$ comes from a gamma distribution with parameters $a$ and $b$, and $\boldsymbol{\theta} = (\alpha, \beta, a, b)$. This model links more closely to the actual behaviour we see in benthic organisms, however the double integral greatly complicates computation of these probabilities. Therefore, I chose to work with the simpler Neyman Type A-gamma distribution where $\phi$ does not vary between species, and this can be extended later.

I can use these probabilities to obtain the maximum-likelihood estimate for $N$. For the conditional likelihood estimator, I need to maximise only the likelihood of the zero-truncated probabilities of the abundance distribution. I can then use $\hat{\boldsymbol{\theta}}$ to obtain an estimate for $N$.

## 3.4 Confidence interval estimation

In addition to obtaining a point estimate for the species richness, $N$, I also need to calculate a confidence interval for the estimate of $N$.

### 3.4.1 Horvitz-Thompson interval estimation

The Horvitz-Thompson point estimate, defined as

$$\hat{N}_{HT} = \frac{1}{1 - p_0(\hat{\theta}_c)} \sum_{i=1}^{N} I_i, \tag{3.14}$$

where $I_i = 1$ if species $i$ is present in the survey, and $I_i = 0$ otherwise, is equivalent to the conditional likelihood estimate $\hat{N}_c$, where $\hat{\theta}_c$ maximises the log-likelihood of the zero truncated abundance distribution of the data and $D = \sum_{i=1}^{N} I_i$.

van der Heijden et al. (2003) show that the variance of $\hat{N}_c$ can be estimated using the law of total variance, by

$$\mathrm{var}(\hat{N}_c) = \mathbb{E}[\mathrm{var}(\hat{N}_c | I_1, ..., I_N)] + \mathrm{var}(\mathbb{E}[\hat{N}_c | I_1, ..., I_N)]), \tag{3.15}$$

where the first term on the right hand side is estimated using the $\delta$-method, and reflects sampling fluctuation in the abundance distribution conditional on the data (van der Heijden et al., 2003).

The second term on the right hand side of Equation 3.15 reflects variation in the obtained sample. $\hat{N}_c$ is assumed equal to $\mathbb{E}(\hat{N}_c|I_1, ..., I_N))$ for a sufficiently large sample size, and so the variance in this second term is estimated by

$$\widehat{\text{var}}[\mathbb{E}(\hat{N}_c|I_1, ..., I_N)] = \frac{p_0(\hat{\boldsymbol{\theta}})}{(1 - p_0(\hat{\boldsymbol{\theta}}))^2} \sum_{i=1}^{N} I_i. \tag{3.16}$$

So the variance of $\hat{N}_c$ in Equation 3.15 is

$$\text{var}(\hat{N}_c) = \boldsymbol{a}^T \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{a}|_{\hat{\boldsymbol{\theta}}} + \frac{p_0(\hat{\boldsymbol{\theta}})}{(1 - p_0(\hat{\boldsymbol{\theta}}))^2} \sum_{i=1}^{N} I_i, \tag{3.17}$$

where $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is the observed Fisher information matrix and $\boldsymbol{a}^T$ is the vector of derivatives

$$\boldsymbol{a}^T = \left( \frac{\partial \hat{N}_c}{\partial \boldsymbol{\theta}_1}, ..., \frac{\partial \hat{N}_c}{\partial \boldsymbol{\theta}_d} \right) \tag{3.18}$$

evaluated at $\hat{\boldsymbol{\theta}}_c$, where $d$ is the dimension of the parameter vector $\boldsymbol{\theta}_c$

For example, when using the Poisson distribution to model the abundance of the observed data, $X = x_1, ..., x_D$, we can use Equation 3.17 to calculate the variance of the estimator as

$$\text{var}(\hat{N}_c) = \left( D \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \right)^2 \left( \sum_{i=1}^{D} x_i \hat{\lambda}^{-2} - \frac{D \exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \right) + D \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2}, \tag{3.19}$$

Using this variance estimate we can construct a symmetric 95% confidence interval for $N_c$ as

$$\hat{N}_c \pm 1.96 \ \text{var}(\hat{N}_c)^{1/2}.$$

However, Cruyff and van der Heijden (2008) show through simulations that, especially for smaller samples, confidence intervals should be asymmetric. They suggest

development of a confidence interval for the logarithm of $\hat{N}$, which will allow for asymmetry, and such an interval is outlined in Section 3.4.3.

### 3.4.2 Confidence regions from profile likelihoods

A standard method of interval estimation uses the approximate $100(1-\alpha)\%$ confidence limits satisfying

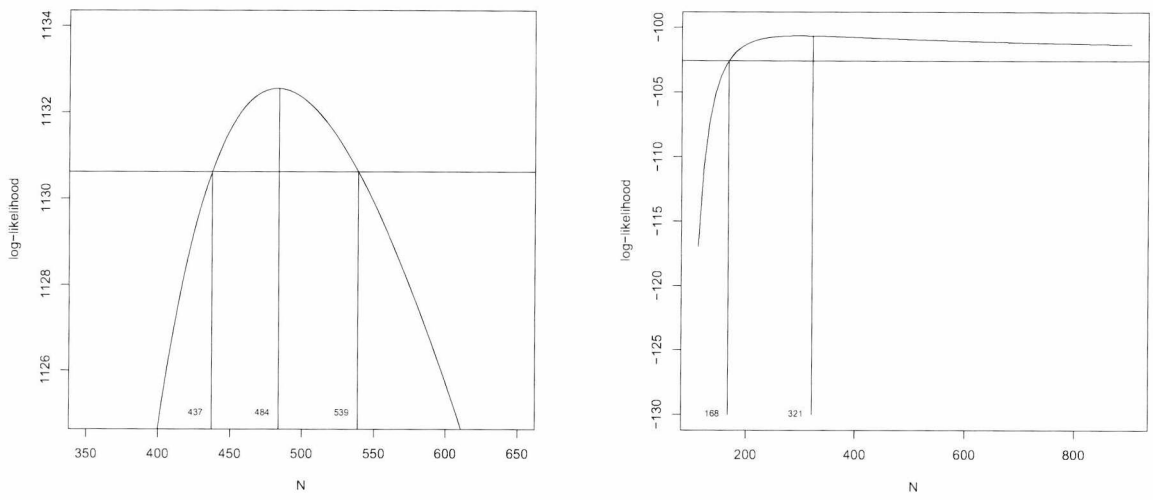
$$2\left\{l_p(\hat{N},\boldsymbol{\theta}) - l_p(N,\boldsymbol{\theta})\right\} = \chi^2_{d:\alpha} \tag{3.20}$$

where $l_p(N,\boldsymbol{\theta})$ denotes the profile log-likelihood of $N$, and $d$ is the dimension of the parameter vector (Fewster and Jupp, 2009). Therefore, the confidence set consists of all values of $N$ for which the log-likelihood lies within $\frac{1}{2}\chi^2_{d:\alpha}$ of the maximum value of $l_p(\hat{N},\boldsymbol{\theta})$.

Since I am interested in estimating $N$, I can treat the abundance parameter vector $\boldsymbol{\theta}$ as a vector of nuisance parameters, and fix the values at their maximum for each possible $\hat{N}$. Therefore, I let $d = 1$ and use $\chi^2_{1:5}$ to obtain 95% confidence limits for $\hat{N}$ (Morgan, 2009, p90). An example of this profile confidence interval is shown in Figure 3.1a using the Poisson to model species' abundances. Here I use the full profile log-likelihood to find the species richness estimates.

In theory this approach can be applied when using any abundance distribution, but when using the negative binomial I have shown that it is possible to get an almost flat log-likelihood profile (Figure 3.1b). It might be an advantage where this may arise to consider a profile confidence interval for $\log(N)$, calculated using the same method, to reduce the flatness of the profile.

However, it is possible that the upper confidence limit for $N$ could be infinite, and this possibility was shown by Morgan and Ridout (2009) for the beta-binomial model. They showed that the overall log-likelihood tends to a constant as $N \to \infty$, but depending on the value of this constant in relation to the log-likelihood at the




*(a)* Poisson *(b)* negative binomial

*Figure 3.1:* Profile confidence intervals for $\hat{N}$ using (a) the Poisson and (b) the negative binomial distributions to model species' abundances. Abundance data were simulated from the Poisson and negative binomial distributions respectively, with $N = 500$. The values of $N$ corresponding to log-likelihood values above the horizontal line lie within the 95% confidence limit of $\hat{N}$, and the vertical line illustrate the upper and lower limit of the 95% confidence interval for $N$. We can see that there appears to be no upper limit for $N$ for the negative binomial example.

maximum-likelihood estimate it could lead to an infinite upper confidence limit for $N$. This possibility is shown in the following example, for the negative binomial model.

*Infinite upper confidence limit for N*

As $N \to \infty$, Morgan and Ridout (2009) showed that the value of the full likelihood depends only on the conditional term of the likelihood, and that this conditional likelihood should approach a constant. The mean of the abundance distribution, in this case the negative binomial, can be approximated by equating the conditional mean number of times a species is seen to its expectation giving

$$\frac{\sum_{k=1}^{r} k f_k}{D} \approx \frac{\tilde{\mu}}{1-p_0} \approx \frac{N\tilde{\mu}}{D}, \text{ so that} \tag{3.21}$$

$$\hat{\mu} \approx \frac{1}{N}\sum_{k=1}^{r} k f_k. \tag{3.22}$$

I present an example where a data set has been simulated from the negative binomial distribution with $\alpha = 2$ and $\beta = 2$. The number of observed species was $D = 103$ and the conditional mean number of times a species is seen is $453/103 = 4.398$. Table 3.1 shows the contributions to the log-likelihood from the first and second components in the log-likelihood , $l_b$ and $l_c$, the overall profile log-likelihood, $l_p$, and the maximum-likelihood estimate of $\mu$, the mean of the negative binomial distribution, conditional on $N$, and the corresponding value from Equation 3.22, $\tilde{\mu}$. The estimate of the negative binomial parameter, $\alpha$, is also shown.

The first row of the table corresponds to the number of species observed in the sample, $D$, with the second and third corresponding to the approximate chi-squared lower confidence limit and the maximum-likelihood estimate respectively. This table shows that the likelihood values stabilise as $N \to \infty$ as we would expect, and the approximate value of $\tilde{\mu}$ is accurate for all values of $N$ considered. In addition,

$$l_b \approx D(\log D - 1) = 374.3771 \tag{3.23}$$

| $N$ | $l_b$ | $l_c$ | $l_p$ | $\hat{\mu}$ | $\tilde{\mu}$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|
| 103 | 364.0682 | -249.7406 | 114.3276 | 4.3989 | 4.3981 | 1.5649 |
| 147 | 374.7233 | -238.6929 | 134.9750 | 3.0820 | 3.0816 | 0.6000 |
| 320 | 374.5712 | -237.6117 | 136.9590 | 1.4155 | 1.4155 | 0.1766 |
| 1,000 | 374.4312 | -237.8920 | 136.5390 | 0.4350 | 0.4530 | 0.0455 |
| 10,000 | 374.3822 | -238.1181 | 136.2641 | 0.0452 | 0.0453 | 0.0042 |
| 100,000 | 374.3772 | -238.1430 | 136.2342 | 0.0045 | 0.0045 | 0.0004 |
| 1,000,000 | 374.3750 | -238.1504 | 136.2246 | 0.0005 | 0.0005 | 0.0001 |

*Table 3.1:* The behaviour of the profile log-likelihood for $N$ for a data set of 200 species simulated from the negative binomial distribution with parameters $\alpha = 2$ and $\beta = 2$. The number of species observed in the simulated data set was $D = 103$.

is approximate as $N \to \infty$ and very accurate for $N \geq 100,000$.

So we see that the profile log-likelihood tends to a limit as $N \to \infty$. However, the 95% confidence set consists of all values of $N$ for which the log-likelihood lies within $\frac{1}{2}\chi^2_{1:5}$ of the maximum value of $l_p(\hat{N}, \boldsymbol{\theta})$. For this example the limit is at the log-likelihood value of 135.0484, and so it appears that the upper 95% profile confidence limit for $N$ is infinite.

Morgan and Ridout (2009) surmise that this result is not particular to the beta-binomial model, and I have shown that this can also occur for the negative binomial model.

### 3.4.3 Confidence intervals for $\log\hat{N}$

If it is difficult to evaluate the profile log-likelihood, $l_p$, then Fewster and Jupp (2009) suggest use of the intervals

$$\log\hat{N} \pm z_{\alpha/2} \left\{ \frac{q(\hat{\boldsymbol{\theta}})^T i_c(\hat{\boldsymbol{\theta}})^{-1} q(\hat{\boldsymbol{\theta}}) + p_0(\hat{\boldsymbol{\theta}})}{\hat{N}(1 - p_0(\hat{\boldsymbol{\theta}}))} \right\}^{1/2} \tag{3.24}$$

based on the asymptotic distribution of $\hat{N}$, or the analogous expression based on $(\hat{N}_c, \hat{\boldsymbol{\theta}}_c)$ when using the conditional MLE. This gives asymptotic $100(1-\alpha)\%$ confidence intervals for $\log N$, where $i_c(\boldsymbol{\theta})$ is the Fisher information on $\boldsymbol{\theta}$ based on a single observation from the zero-truncated abundance distribution, and $q(\boldsymbol{\theta})$ is a column vector defined as

$$q(\boldsymbol{\theta})^T = \frac{\mathrm{d}\log(1-p_0(\boldsymbol{\theta}))}{\mathrm{d}\boldsymbol{\theta}}. \tag{3.25}$$

As $\log\hat{N}$ is asymptotically normally distributed, we form a 95% confidence interval for $\log\hat{N}$ using $z_{2.5} = 1.96$.

This method should be an improvement on the Horvitz-Thompson interval as it is able to produce an asymmetric confidence interval for $N$.

If we assume the abundances of the observed species, $x_1, ..., x_D$, come from a Poisson distribution we will have $1 - p_0(\boldsymbol{\theta}) = 1 - e^{-\lambda}$. The probability mass function for the zero-truncated Poisson is

$$f(y|y>0;\lambda) = \frac{f(y|\lambda)}{f(y>0|\lambda)} = \frac{e^{-\lambda}\lambda^y}{y!(1-e^{-\lambda})}, y = 1, 2, ... \tag{3.26}$$

and so the conditional log-likelihood can be written as

$$l_c(\lambda) = \sum_{i=1}^{D} \left(x_i \log(\lambda) - \lambda - \log(1-e^{-\lambda}) - \log(x_i!)\right). \tag{3.27}$$

Therefore we can calculate $q(\boldsymbol{\theta})$ as

$$q(\boldsymbol{\theta})^T = \frac{\mathrm{d}\log(1-p_0(\boldsymbol{\theta}))}{\mathrm{d}\boldsymbol{\theta}} = \frac{e^{-\lambda}}{1-e^{-\lambda}}, \tag{3.28}$$

and the observed Fisher information on $\hat{\lambda}$ as

$$i_c(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 l_c}{\partial\lambda^2} = \sum_{i=1}^{D} x_i\lambda^{-2} - \frac{De^{-\lambda}}{(1-e^{-\lambda})^2}, \tag{3.29}$$

and form the confidence interval for $\log(\hat{N})$ using Equation 3.24.

We can use the same method to construct confidence intervals when the species

abundances are described by a negative binomial distribution. The conditional log-likelihood for the zero-truncated negative binomial can be written as

$$\begin{aligned} l_c(\boldsymbol{\theta}) = & \sum_{i=1}^{D} \log(\Gamma(\alpha + x_i)) - \log(x_i!) - \log(\Gamma(\alpha)) \\ & + \alpha \log(\tfrac{1}{1+\beta}) + x_i \log(\tfrac{\beta}{1+\beta}) - \log(1 - \left(\tfrac{1}{1+\beta}\right)^{\alpha}). \end{aligned}$$

Now we have

$$1 - p_0(\boldsymbol{\theta}) = 1 - \left(\frac{1}{1+\beta}\right)^{\alpha}$$

and we can calculate

$$q(\boldsymbol{\theta})^T = \left(\frac{\mathrm{dlog}(1 - p_0(\boldsymbol{\theta})}{\mathrm{d}\alpha}, \frac{\mathrm{dlog}(1 - p_0(\boldsymbol{\theta}))}{\mathrm{d}\beta}\right) = \left(\frac{-\frac{1}{1+\beta}^{\alpha} \log(\frac{1}{1+\beta})}{1 - \frac{1}{1+\beta}^{\alpha}}, \frac{-\frac{1}{1+\beta}^{\alpha} \alpha}{\frac{1}{1+\beta}(1 - \frac{1}{1+\beta}^{\alpha})}\right).$$

The derivatives for $i_c(\hat{\boldsymbol{\theta}})$ can be found in Appendix A. However, instead of using these equations directly I could use the Hessian matrix which is produced when maximising the conditional log-likelihood for $\boldsymbol{\theta}$ within `R`. This would be helpful when extending to more complicated distributions such as the Neyman Type A-gamma model.

For the Neyman Type A-Gamma distribution I can calculate a value for $p_0(\hat{\boldsymbol{\theta}})$ within `R` using the `integrate` function, and also obtain a Hessian matrix during optimisation of the conditional log-likelihood. However I also need to calculate $q(\boldsymbol{\theta})$, which must be done numerically.

### *3.4.4 Bootstrap confidence intervals*

We can use one of several bootstrap approaches to construct confidence intervals. The first is an abundance-model-based bootstrap, as described by Wang and Lindsay (2005), in which we plug our fitted parameters into our abundance model and use it to simulate $\hat{N}$ new observations, $x_i$, for our population. Species with $x_i = 0$ are omitted, because they would not be present in our sample, and we have a new data set on which to apply the estimator. An estimate is computed, and the whole process is repeated to generate $S$ bootstrap samples for $N$.

95% confidence limits are then obtained using Efron's percentile method (Efron, 1981), by using the 2.5 and the 97.5 percentiles of the bootstrap distribution as the limits of the confidence interval. This method can be refined by using the bias-corrected and accelerated (BCa) bootstrap, which adjusts for both bias and skewness in the bootstrap distribution (Efron, 1987).

However, as Wang and Lindsay (2005) state, it is individuals which are sampled from the population, not the species counts, which come from aggregating individuals. Therefore, we can consider a multinomial-based bootstrap, proposed by Wang and Lindsay (2005), in which we create an estimated population and simulate draws of individuals from it.

Wang and Lindsay (2005) do this by creating $\hat{N}$ cells, with each cell corresponding to a species. The cells are then divided into $r$ groups, corresponding to the species abundance levels, and weighted. The multinomial parameter for each cell in the $j$th group is then calculated as $p_j$, and bootstrap samples of fixed size $S$ are generated by drawing individuals from a multinomial distribution, of $S$ trials and event probabilities $p_j$. Again, the estimator can be computed for each sample, and a confidence interval constructed using the percentile method.

Further bootstrap methods suggested by Wang and Lindsay (2005) include a more non-parametric bootstrap, which generates $n$ non-zero observations of $X$ from the multinomial corresponding to the empirical distribution of the non-zero counts, where $n$ is the number of individuals in the sample, and a hybrid bootstrap, which involves sampling $n$ from a binomial and then drawing $n$ times from the empirical multinomial.

Wang and Lindsay (2005) found that the multinomial-based bootstrap tended to be more reliable than the others. However as the bootstrap sample size is fixed, this method best matches data collected by sampling a fixed number of individuals, which

is not the kind of data I have. van der Heijden et al. (2003) used an abundance-model based parametric bootstrap and found that coverage probabilities (the percentage of confidence intervals that contain the true value of the parameter) could be low when the population size and density were small.

## 3.5 Goodness of fit and model selection

To quantify the fit of a model, I can use the Pearson $\chi^2$ test, which makes a comparison between the observed and fitted values. The Pearson goodness-of-fit statistic is given by:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}, \tag{3.30}$$

where $r$ is the number of categories comparing observed and expected values over. Since we cannot observe $f_0$, the zero category is excluded from the calculation. When more than 20% of the categories have a frequency less than five, the Pearson goodness-of-fit statistic breaks down, and therefore in this case I group the data, and $r$ is the number of groups. This should ensure the $\chi^2$ test is accurate.

We then calculate the degrees of freedom, $df = r - 1 - d$, where $d$ is the number of parameters estimated, and look at the corresponding chi-squared percentage point, $\chi_{df:\alpha}$, where $\alpha$ indicates the $\alpha/100$ probability we reject the fit.

When using this method we may expect some degree of mismatch between model and data. However, as suggested by Morgan (2009), we can use this goodness-of-fit measure as a guideline.

To choose between models we can select the model with the smallest values of an information criterion, such as the Akaike Information Criterion (AIC). The AIC is defined as

$$AIC = -2\log L + 2d \tag{3.31}$$

where $L$ is the likelihood of the fitted model, and $d$ is the number of parameters estimated (Akaike, 1973).

This is a simple way to choose between competing models, that does not require models to be nested. The AIC can be thought of as a measure of lack of model fit plus a penalty for estimating $d$ parameters (Burnham et al., 1995).

To compare the fit of models, we can look at the difference between each model and that with the lowest AIC, termed the $\Delta$AIC, and (Burnham and Anderson, 2002, p 70) suggest that models with $\Delta\text{AIC} < 2$ are plausible, $4 < \Delta\text{AIC} < 7$ are considerably less plausible, and $\Delta\text{AIC} > 10$ means that the model is unlikely.

In fitting the full likelihood to the data it is possible to obtain positive log-likelihood values and therefore negative AIC values, as the likelihood is not itself the probability of observing the data, but just proportional to it. However, using the conditional likelihood only negative values are obtained as we are just calculating the probability of observing the data. Where negative AIC values occur, we still take the model with the lowest AIC as the best of the set.

## 3.6 Excluding highly abundant species

If we have a species in our sample that is highly abundant, its large numbers of individuals may skew the estimator by having a significant effect on the model fitting. Therefore excluding these abundant species from the estimation of the number of missing species may give a more accurate estimate of species richness.

To illustrate the importance of excluding abundant species, I use the microbial organisms sample data presented in Barger and Bunge (2008). The non-zero observed frequencies for the full data set, listed as $(k, f_k)$, are: (1, 15),(2, 6),(3, 7),(4, 2),(5, 1),(6, 1),(7, 1),(8, 1),(9, 1),(12, 1),(15, 1),(20, 1),(164, 1). The observed number of species is 36 and the observed number of individual organisms is 303. If we include all

of the observed data, and use the conditional log-likelihood estimator and a Poisson model for the species abundances, the fitted model corresponds very badly to the observed data, as seen in Figure 3.2.

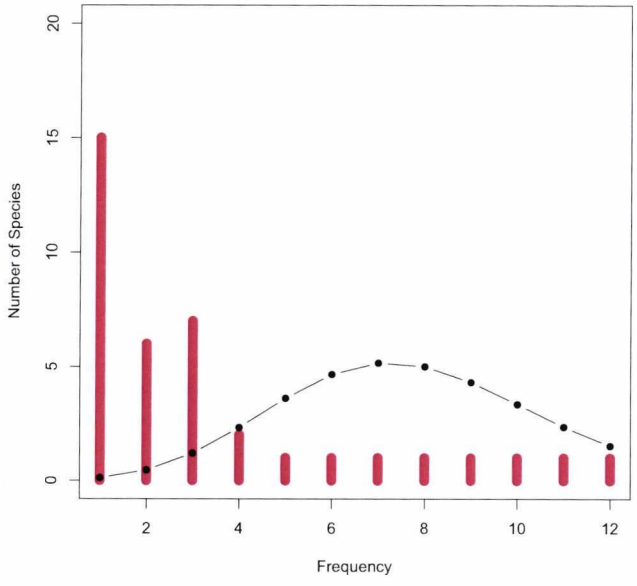



*Figure 3.2:* Observed (red) and expected (black) frequencies evaluated at $(\hat{N}_c, \hat{\theta}_c)$ for the Poisson model fitted to the microbial data without using truncation (data point at 164 not shown).

We need to decide on the criteria to define a species as abundant before we can proceed with data truncation, and there are several ways that this could be done. Firstly, an abundant species could be defined as a species that is seen in every single grab. However, although not likely, there is the possibility that a species may fall into this category even if there is only a single individual of the species in each grab.

Alternatively, we could define a species as abundant if there is more than a certain number of individuals of that species present over all the grabs. Under this definition, an abundant species may be seen only in a single grab, but it would be more likely to skew the overall species richness estimate.

If we were considering a single grab, or pooling data over the grabs, we could split the sample into 'rare' species and 'abundant' species by partitioning the data into those species seen up to $\tau$ times in the sample, and those seen more than $\tau$ times, where $\tau \leq r$ (Chao et al., 2000). An estimate, $\hat{N}_\tau$, is obtained by maximising the likelihood based completely on the 'rare' species data, and then the final estimate of species richness is calculated as

$$\hat{N} = \hat{N}_\tau + E_\tau,$$

where $E_\tau = \sum_{i=\tau+1}^{r} f_i$, the number of truncated species.

Previous studies have chosen $\tau$ by applying a goodness of fit method (Behnke et al., 2006; Barger and Bunge, 2008). Behnke et al. (2006) considered data subsets consisting the observed frequency counts from one up to some maximum value, which they called the right truncation point. They then fitted all models at all right truncation points and selected the 'best of the best' model, based on goodness of fit (defined by two chi-square statistics), the minimal standard error (among the fitted models), and the maximal data usage (highest right truncation point).

This method would be computationally intensive, because every model is calculated for every possible $\tau$. This would not be a practical approach when fitting the Neyman Type A-gamma distribution to the abundance data, as because of the summation within the Neyman Type A probability calculations, and the need to integrate the marginal likelihood, fitting this model by maximum-likelihood can be computationally expensive. Also, if the estimator is sensitive to the choice of $\tau$, then we are assuming that certain species hold most of the information on the missing species, namely the 'rare' ones (Chao et al., 2000).

Using this truncation method discards some of the data from the maximum-likelihood estimator. To avoid throwing away these data, I truncated the data at a point $\tau$, but instead of using $\hat{N} = \hat{N}_\tau + E_\tau$, I include $E_\tau$ within the likelihood calculation.

So instead of maximising the conditional log-likelihood

$$l_c(\boldsymbol{\theta}) = \sum_{k\geq 1} f_k \log(p_k(\boldsymbol{\theta}) - (1 - p_0(\boldsymbol{\theta}))), \tag{3.32}$$

I let $q_\tau = 1 - \left(\sum_{k=1}^{\tau} p_k(\boldsymbol{\theta})/(1 - p_0(\boldsymbol{\theta}))\right)$ and maximise

$$l_c(\boldsymbol{\theta}) = \sum_{k=1}^{\tau} f_k \log(p_k(\boldsymbol{\theta}) - (1 - p_0(\boldsymbol{\theta}))) + E_\tau \log(q_\tau). \tag{3.33}$$

When using the full likelihood I adjust in the same way.

Using this approach, I am still required to select $\tau$, although we might expect that the choice of $\tau$ in this approach may be less influential on the species richness estimate because no data is discarded. The 'best' choice of truncation point may vary between models, and so I follow the approach of using the $\tau$ suggested by the goodness of fit method, if this value is already available in the literature for a particular data set.

For some data sets, when the number of individuals for a particular species exceeds approximately 150, an error is produced within R when fitting the Neyman Type A-gamma model. Therefore, in these cases it is necessary to truncate the data at $\tau = 150$ to obtain a species richness estimate, and the likelihood of Equation 3.33 is used in maximisation. In all other cases, unless otherwise stated, the data are not truncated.

## 3.7 Computational difficulties in fitting the Neyman Type A - gamma model

Computational difficulties arise in R when fitting the Neyman Type A-gamma model to the data using the maximum-likelihood approach.

The marginal likelihood for the abundance probabilities, (Equation 3.12)

$$p_k(\boldsymbol{\theta}) = \int_0^\infty \frac{e^{-\lambda}\phi^k}{k!} \sum_{j=0}^{\infty} \left\{ \frac{(\lambda e^{-\phi})^j j^k}{j!} \right\} \frac{\beta^\alpha (\lambda\phi)^{\alpha-1} e^{-\beta\lambda\phi}}{\Gamma(\alpha)} \mathrm{d}\lambda \tag{3.34}$$

cannot be easily computed. Firstly, the probability density function of the Neyman Type A distribution contains an infinite summation, which must be

truncated. I used a truncation point of 100 for this summation, which was shown to estimate the probabilities well in most cases. However, when $\lambda \to 0$ and $\phi \to \infty$, even when the terms up to ten million in this summation were calculated, the probability estimates showed an error of 0.01. However, since such extreme parameter values were not used for simulations, I did not increase the truncation point.

When using function `integrate` to calculate the marginal probabilities, computation could be slow, and further investigation is required into how this could be improved. In addition, when the number of individuals for a particular species exceeded 150, the marginal probabilities could not be calculated and it was necessary to truncate the data at this point to obtain a species richness estimate.

## 3.8 Boundary problem

The 'boundary problem' is a serious issue that can occur when estimating species richness parametrically through maximum-likelihood estimation, often occurring when heterogeneity is severe (Pledger and Phillpot, 2008). It arises when a great deal of positive mass is located near the origin of the abundance distribution (Wang and Lindsay, 2005; Kuhnert et al., 2008; Böhning, 2009). This increases the influence of that part of the abundance distribution, which can give a spuriously large species richness estimate. The concept of the 'boundary problem' is complex, and more detail, in terms of non-parametric mixtures, is outlined in Wang and Lindsay (2008).

The boundary problem presents a hurdle to several estimators, such as the nonparametric maximum-likelihood estimator of Wang and Lindsay (2008), for which they found it caused a severe instability problem. It has also been demonstrated that mixtures of several exponential family distributions suffer under the boundary problem (Wang and Lindsay, 2008; Kuhnert et al., 2008), and it is stated by Böhning (2009) that it is so far impossible to detect which data set has this problem and which does not.

### 3.8.1 Proposed solutions to the boundary problem

Penalising the likelihood has been found to be useful where the likelihood function is relatively flat, to combat the boundary problem (Wang and Lindsay, 2005; Moreno and Lele, 2010). However, Moreno and Lele (2010) point out that there is no unique specification of the penalty term to be used, or theoretical basis to choose one.

Wang and Lindsay (2005) proposed a class of nonparametric maximum-likelihood estimators (NPMLE) for the species richness problem, using a penalty term on the log-likelihood to eliminate the instability problem. The estimators are constructed using the conditional likelihood, and Wang and Lindsay (2005) hoped that the penalty functions considered would attain high stability whilst retaining sensitivity.

There are several properties that we would like a penalty to have. As the sample size increases, Moreno and Lele (2010) state that the penalty should tend to zero to maintain the asymptotic properties of the MLE. Also, if sampling depth, defined as the proportion of the species in the population which have been observed in the sample, is high, the penalty should tend to zero, because the data should be well behaved.

As the probability of not seeing a species, $p_0$, approaches 1, the species richness estimate, $\hat{N}$, tends to infinity (Figure 3.3). This is when we would require a penalty, to obtain a realistic species richness estimate. However, if the true value of $p_0$ is low, then a species richness estimate should not be penalised, as the boundary problem should not occur. If we do apply a penalty in this case, then we risk bias in the species richness estimate.

Penalising the likelihood can be seen as an attempt to incorporate prior information from a naive estimator (Wang and Lindsay, 2005), such as the number of observed species. The goal is then to penalise estimates that are too far from this naive estimate. However, it would be important to use a suitable naive estimate, because

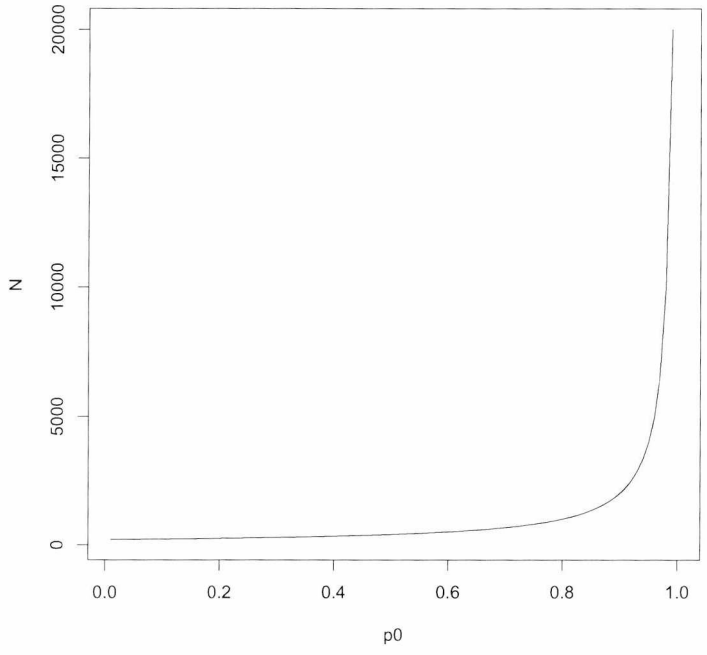



*Figure 3.3:* Graph showing how the maximum-likelihood species richness estimate $\hat{N} \to \infty$ as the probability of not seeing a species $p_0(\theta) \to 1$.

if this initial estimate is too low in comparison to the true species richness, then the penalty could be too harsh and negatively bias the final species richness estimate.

Böhning (2009) proposed an alternative method to combat the boundary problem in species richness estimation, using a non-parametric empirical Bayes approach. This method uses information from the sample to inform the model, and does not suffer under the boundary problem (Böhning, 2009). However, if the sample size is small, there will be large variation, and Böhning (2009) suggest using smoothing probabilities to combat this.

### *Penalised log-likelihood*

The objective of Wang and Lindsay (2005) was to see if there was a way to penalise the nonparametric likelihood such that the resulting maximum-likelihood estimator retained much of its flexibility, but behaved more stably. Parametric maximum-likelihood estimates have a tendency to become spuriously large, and it is hoped that

by adding a penalty term to the likelihood we will obtain estimates of species richness that are accurate and stable.

If $l(N, \boldsymbol{\theta})$ is a log-likelihood function then the penalised log-likelihood corresponding to penalty parameter $\gamma$ and penalty function $h(N, \boldsymbol{\theta})$ is defined as (Wang and Lindsay, 2005)

$$l^{\gamma}(N, \boldsymbol{\theta}) = l(\theta) - \gamma h(N, \boldsymbol{\theta}). \tag{3.35}$$

We can consider maximising the full penalised likelihood of Equation 3.35; or to estimate $N$ using the conditional penalised log-likelihood where we first find the penalised MLE of $\boldsymbol{\theta}$ from the penalised conditional log-likelihood corresponding to the conditional likelihood $L_c$ in Equation (3.3) of

$$l_c^{\gamma}(\boldsymbol{\theta}) = l_c(\boldsymbol{\theta}) - \gamma h(\boldsymbol{\theta}), \tag{3.36}$$

and then use Equation 3.4 to find the penalised conditional MLE of $N$

$$\hat{N}_c^{\gamma} = D/\{1 - p_0(\hat{\boldsymbol{\theta}}_c^{\gamma})\}. \tag{3.37}$$

Wang and Lindsay (2005) remark that for $\gamma > 0$ and $h(\boldsymbol{\theta}) > 0$, a maximiser of the penalised conditional log-likelihood in Equation (3.36) tends to avoid $\boldsymbol{\theta}$ with large values of $h(\boldsymbol{\theta})$, that is with a large penalty function.

Wang and Lindsay (2005) considered three penalty functions. The first of these was

$$h_1(\boldsymbol{\theta}) = \log\left(p_0(\boldsymbol{\theta})\right), \ \gamma_1 = 0.5, \tag{3.38}$$

which did not eliminate the boundary problem. As $p_0$ increases, the penalty term becomes flat rapidly, and therefore ignorable (Wang and Lindsay, 2005) (Figure 3.4). This means that there will not be enough penalty where it is required at large values of $p_0$, and too much penalty elsewhere.

The second choice of penalty function by Wang and Lindsay (2005) was the odds

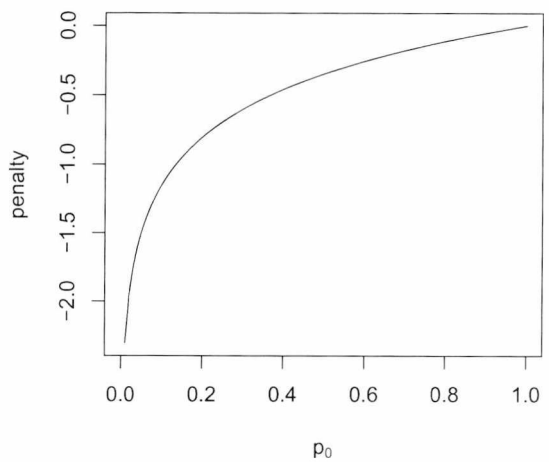



*Figure 3.4:* Graph showing how penalty 1, $\gamma_1 h_1 = 0.5\log p_0$, becomes flat rapidly as $p_0$ increases.

function, $\psi(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta})/(1 - p_0(\boldsymbol{\theta}))$. As $\hat{N}_c = D/\{1 - p_0(\hat{\boldsymbol{\theta}}_c)\}$, then $\hat{N}_c = D(1 + \psi)$; so imposing a penalty on $\psi$ reduces the magnitude of $\hat{N}_c$.

(Wang and Lindsay, 2005) highlight that the penalised likelihood

$$l_2(\boldsymbol{\theta}) = l_c(\boldsymbol{\theta}) - \gamma_2\psi(\boldsymbol{\theta}), \quad \gamma_2 > 0$$

cannot have its maximum at $\psi(\boldsymbol{\theta}) = \infty$, so extreme estimates due to the boundary problem cannot occur (Figure 3.5).

The penalty parameter $\gamma_2$ can be tuned to control the variability of $\hat{N}_c$. However, the optimal choice of $\gamma_2$ depends strongly on the value of the odds function (Wang and Lindsay, 2005). So far there is no known method devised to choose the best penalty for a particular problem.

The final penalty to be considered by Wang and Lindsay (2005) was a more severe penalty,

$$l_3(\boldsymbol{\theta}) = l_c(\boldsymbol{\theta}) - \gamma_3(\psi - \eta)^2\boldsymbol{I}(\psi > \eta), \quad \gamma_3, \eta > 0,$$

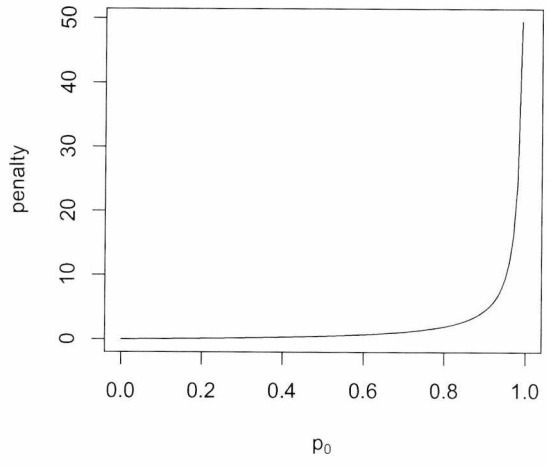



*Figure 3.5:* The behaviour of penalty 2, $\gamma_2 h_2 = 0.5\psi$, as $p_0$ increases, where $\psi$ is the odds function $\psi = p_0/(1 - p_0)$.

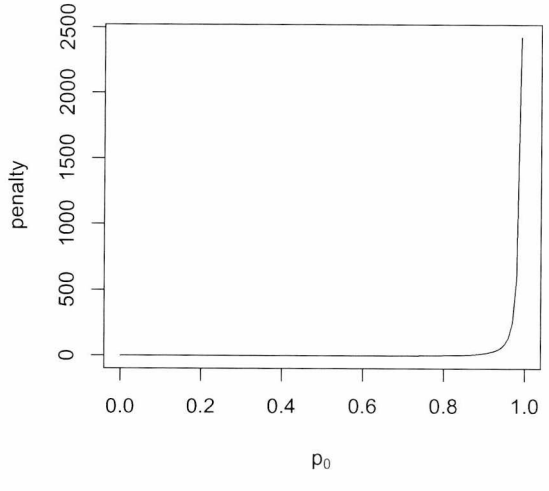



*Figure 3.6:* Graph plotting the behaviour of penalty 3, $\gamma_3(\psi - \eta)^2 \boldsymbol{I}(\psi > \eta)$, as $p_0$ increases, where $\psi$ is the odds function, $\gamma_3 = 1/2\eta$ and $\eta = \hat{N}_{C1}/D - 1$. For this example, $D = 200$ and $\hat{N}_{C1} = 300$.

where $\boldsymbol{I}$ is the indicator function. This penalises the quadratic distance between $\psi$ and $\eta$ when $\psi$ exceeds threshold $\eta$, and offers no penalty for small $\psi$, but as $\psi$ becomes larger, the penalty becomes much larger (Figure 3.6) (Wang and Lindsay, 2005). Comparing the two graphs in Figures 3.5 and 3.6 we see that this is indeed a much harsher penalty than penalty 2.

This penalty requires the choice of two parameters, $\gamma_3$ and $\eta$. In an attempt to avoid under- or over-penalising, Wang and Lindsay (2005) used adaptive values for the penalty parameters and let

$$\eta = \frac{\hat{N}_{C1}}{D} - 1 \tag{3.39}$$

and

$$\gamma_3 = \frac{1}{2\eta}, \tag{3.40}$$

where $\hat{N}_{C1}$ is the lower bound estimator of Chao (1984), given by $\hat{N}_{C1} = D + f_1^2/(2f_2)$.

This penalty term shrinks the MLE towards the naive estimate $\hat{N}_{C1}$, which is likely to cause negative bias in some cases, especially when applied to benthic data, as we have seen in Chapter 2 that the $Chao_1$ estimator $\hat{N}_{C1}$ can severely underestimate the true species richness when dealing with clustered individuals.

Therefore we might consider an alternative naive estimator to $\hat{N}_{C1}$. However none of the nonparametric species richness estimators mentioned in Chapter 2 are suitable because they were unable to give an adequate estimate for clustered data. Instead I could use an iterative approach, to avoid this negative bias. We could obtain an initial species richness estimate using this penalty with the Chao estimate as a naive species richness estimate. This would enable us to get an estimate for $N$, which we might call $\hat{N}_0$. Then using this estimate in place of $\hat{N}_{C1}$ in $\eta$, I could repeat the maximisation of the likelihood, and obtain another estimate of $N$, $\hat{N}_1$. If I repeated this process, I would hope the estimates $\hat{N}_0, \hat{N}_1, \hat{N}_2, ...$ would converge to some final value, which I would take as my final species richness estimate.

I would hope that using this approach would penalise the log-likelihood, but not as harshly as using the Chao estimate in the penalty. However, this would require a high computation time when applied to more complicated distributions, depending on the number of iterations required for convergence (if they converge at all). Therefore, we might use a one-step iteration in the hope that this would still give a more realistic estimate than merely using the $Chao_1$ estimator within the penalty.

## 3.9 Analysis of methods via simulation

As well as assessing the performance of the maximum-likelihood method for various models, there are several aspects that have been highlighted that require further investigation. These include:

- the calculation of confidence intervals,
- the sampling depth required to avoid the boundary problem,
- how we might combat the boundary problem using penalties.

I did this using simulated data. In all of these simulations unless stated otherwise I applied the negative binomial MLE and no penalty. I simulated data sets $\boldsymbol{Y}$ such that $y_i \sim \text{Poisson}(\lambda_i)$, and $\lambda_i \sim \text{Gamma}(\alpha, \beta)$. The data were generated for 10 grabs, and an example of the data when pooled is shown in Table 3.2.

### 3.9.1 Confidence intervals

I wished to investigate the coverage properties of the confidence intervals, because I hoped that when using 95% confidence intervals, the true species richness would lie

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_k$ | 54 | 33 | 34 | 10 | 10 | 3 | 2 | 1 | 2 | 1 |

*Table 3.2:* Example abundance data simulated for a population of $N = 200$ from the negative binomial distribution with parameters $\alpha = 2$ and $\beta = 1$. $D = 150$.

within the interval 95% of the time. van der Heijden et al. (2003) reported a simulation study that assesses the stability of the Horvitz-Thompson variance estimators, and also evaluate the coverage probability of confidence intervals obtained by abundance-model-based parametric bootstrapping. I repeated this study using the Poisson model and the conditional maximum-likelihood estimator, and also included the coverage results for the other confidence interval estimators mentioned in Section 3.4, the profile confidence intervals for $N$, and the confidence interval for $\log(N)$. Although van der Heijden et al. (2003) found the coverage of the abundance-model-based parametric bootstrap low when the population size and density were small, I chose to use the same method for comparison with their results.

I simulated data from a Poisson distribution, with parameters shown in Table 3.3. As far as I am aware, a comparison of the coverage of these confidence interval methods has not been published before. Coverage was calculated as the proportion of confidence interval estimates for which the true species richness fell between the upper and lower confidence limit.

The results given by van der Heijden et al. (2003) indicated that the Horvitz-Thompson confidence interval had a higher coverage probability than that of the bootstrap confidence interval when both $\lambda$ and $N$ were small. However for other values they were comparable. My results for the bootstrap confidence intervals were slightly different (See Table 3.3). However, looking at the mean widths of the intervals in Table 3.4, I can see that the bootstrap intervals are very wide for the low values of $\lambda$ and $N$ due to spuriously large estimates arising in the bootstrap process.

Table 3.3 shows good coverage probabilities for the profile confidence intervals, performing well in all cases, and outperforming the Horvitz-Thompson confidence intervals in some cases.

van der Heijden et al. (2003) suggested the use of confidence intervals which can

| | | $N$ | | | |
|---|---|---|---|---|---|
| $\lambda$ | Method | 100 | 250 | 500 | 1000 |
| 0.5 | HT | 0.948 | 0.944 | 0.960 | 0.952 |
| | PR | 0.936 | 0.946 | 0.954 | 0.952 |
| | BOOT | 0.942 | 0.916 | 0.912 | 0.930 |
| | LN | 0.684 | 0.668 | 0.684 | 0.644 |
| 1 | HT | 0.958 | 0.948 | 0.948 | 0.946 |
| | PR | 0.960 | 0.954 | 0.942 | 0.956 |
| | BOOT | 0.932 | 0.914 | 0.902 | 0.916 |
| | LN | 0.832 | 0.792 | 0.778 | 0.792 |
| 1.5 | HT | 0.950 | 0.960 | 0.946 | 0.936 |
| | PR | 0.944 | 0.962 | 0.952 | 0.938 |
| | BOOT | 0.914 | 0.908 | 0.910 | 0.888 |
| | LN | 0.864 | 0.846 | 0.852 | 0.830 |
| 2 | HT | 0.924 | 0.948 | 0.950 | 0.948 |
| | PR | 0.938 | 0.944 | 0.944 | 0.952 |
| | BOOT | 0.896 | 0.914 | 0.912 | 0.918 |
| | LN | 0.876 | 0.888 | 0.880 | 0.890 |
| 2.5 | HT | 0.960 | 0.952 | 0.964 | 0.948 |
| | PR | 0.966 | 0.962 | 0.964 | 0.944 |
| | BOOT | 0.944 | 0.936 | 0.946 | 0.930 |
| | LN | 0.938 | 0.926 | 0.926 | 0.918 |

*Table 3.3:* Coverage of 95% confidence intervals calculated using HT: Horvitz-Thompson, PR: Profile log-likelihood for $\hat{N}$, BOOT: abundance-based Bootstrap, and LN: $\log\hat{N}$ following Fewster and Jupp (2009), for the Poisson MLE applied to Poisson simulated data sets over 100 simulations.

| | | $N$ | | | |
|---|---|---|---|---|---|
| $\lambda$ | Method | 100 | 250 | 500 | 1000 |
| 0.5 | HT | 121.0 | 175.0 | 235.5 | 325.0 |
| | PR | 109.0 | 183.9 | 240.9 | 328.7 |
| | BOOT | 6890.8 | 181.4 | 224.5 | 301.1 |
| | LN | 56.5 | 82.6 | 112.2 | 155.5 |
| 1 | HT | 48.4 | 75.2 | 105.5 | 146.2 |
| | PR | 49.8 | 76.1 | 106.0 | 146.6 |
| | BOOT | 46.1 | 67.5 | 92.7 | 126.9 |
| | LN | 31.1 | 48.4 | 67.9 | 94.6 |
| 1.5 | HT | 28.5 | 44.7 | 62.6 | 88.4 |
| | PR | 28.9 | 45.0 | 62.8 | 88.5 |
| | BOOT | 26.0 | 39.8 | 55.5 | 78.0 |
| | LN | 21.4 | 33.7 | 47.2 | 66.7 |
| 2 | HT | 19.0 | 29.8 | 42.0 | 59.2 |
| | PR | 19.2 | 30.0 | 42.1 | 59.3 |
| | BOOT | 17.6 | 27.2 | 38.4 | 54.0 |
| | LN | 15.7 | 24.7 | 34.8 | 49.1 |
| 2.5 | HT | 13.5 | 21.2 | 29.9 | 42.0 |
| | PR | 14.1 | 21.4 | 30.0 | 42.1 |
| | BOOT | 12.8 | 19.9 | 28.0 | 39.5 |
| | LN | 11.9 | 18.7 | 26.3 | 37.1 |

*Table 3.4:* Width of 95% confidence intervals, HT: Horvitz-Thompson, PR: Profile likelihood for $\hat{N}$, BOOT: Bootstrap, LN: $\log\hat{N}$, for the Poisson MLE applied to Poisson simulated data sets over 100 simulations.

be asymmetric, such as those of $\log(\hat{N})$. However, the confidence intervals for $\log(\hat{N})$ suggested by Fewster and Jupp (2009) surprisingly were much narrower than those from the other methods, and this was the likely cause of the much lower coverage probabilities associated with this method. When I increased the value of $\lambda$, the Fewster and Jupp (2009) $\log(\hat{N})$ method did give excellent coverage. For example, for $N = 1000$, $\lambda = 5$ I obtained coverage of 95.4% over 500 simulations. However in reality we are more likely to have low sampling depth (the proportion of the species in the population which have been observed in the sample), not 0.99 as in this example. As sampling depth decreases with the decrease in $\lambda$, the coverage probability also decreases.

The poor performance of the $\log(\hat{N})$ confidence intervals warranted further investigation, but since the other methods were performing well I did not investigate further at this time.

I wished to run the same investigation for the negative binomial case, however I experienced a number of problems which arose due to the boundary problem. Therefore I needed to investigate the use of these confidence intervals with penalties. Since Table 3.3 shows good coverage probabilities for the profile intervals, I investigated their performance for the penalised MLE and used them for further confidence interval estimation when applying the maximum-likelihood estimator.

### *3.9.2 Sampling depth required to avoid the boundary problem*

To investigate the boundary problem, I looked at the homogeneous case where data were simulated from the Poisson distribution. As defined previously, sampling depth is the proportion of the species in the population which have been observed. Table 3.5 shows that in most cases the conditional maximum-likelihood estimator gave a value close to the true $N$ for the Poisson model, and when applying a Poisson model to Poisson simulated data, there was no significant evidence of the boundary problem occurring when sampling depth was more than 10%. The maximum estimate over

these simulations was 1784, and the minimum was 593.

However, when I decreased sampling depth we see clear signs of the boundary problem occurring. When sampling depth decreased to less than 10% we see that the standard deviation of the species richness estimates increased significantly. The median of the sample remained respectable until the sampling depth decreased to 2.5%, but the central 95% interval clearly showed that the boundary problem is occurring in several cases when sampling depth was 5% or smaller.

I repeated the simulation for the heterogenous case and simulated data from the negative binomial distribution, for populations of various gamma parameter values, and $N = 1000$. For each simulation setting, I generated 200 data sets, and the summary statistics of the sampling distributions of the estimators based on 200 samples are presented in Table 3.6, following Wang and Lindsay (2005).

The results show that when sampling depth fell below 50% that there was evidence of the boundary problem. The mean estimate became large, and compared to the median estimate, we see that the distribution of the estimates is skewed, with some very large estimates contributing to the large mean estimate. This is highlighted in the 95% confidence interval, which is extremely wide and as the sampling depth decreased, the standard deviation increased.

Compared to the results for the Neyman Type A-gamma model, (Table 3.7) signs of the boundary problem were now evident at higher sampling depths, and when sampling depth was at 20% we see some extreme species richness estimates. I would expect the standard deviation to increase because the number of parameters in the model increase, less information is used to estimate each parameter, so there will be more variance in the estimates. However, the spuriously large estimates can be attributed to the boundary problem.

| Depth | $\lambda$ | Mean | Median | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|
| 0.025 | 0.03 | 419,622 | 511,924 | 284,056 | 506,327 | 419,015 | (501, 633,810) |
| 0.05 | 0.05 | 317,227 | 1089 | 541,478 | 626,391 | 316,575 | (633, 999,470) |
| 0.075 | 0.08 | 154,855 | 1148 | 509,875 | 531,651 | 154,083 | (813, 2223) |
| 0.1 | 0.11 | 1223 | 1047 | 770 | 1444 | 438 | (812, 1410) |
| 0.2 | 0.20 | 1049 | 1015 | 211 | 213 | 166 | (882, 1182) |
| 0.3 | 0.40 | 1018 | 1004 | 141 | 142 | 108 | (923, 1104) |
| 0.4 | 0.50 | 1007 | 1011 | 82 | 83 | 65 | (948, 1053) |
| 0.5 | 0.70 | 1004 | 1001 | 59 | 59 | 48 | (961, 1044) |
| 0.6 | 0.90 | 999 | 1000 | 40 | 40 | 32 | (970, 1025) |
| 0.75 | 1.40 | 1001 | 998 | 24 | 24 | 19 | (984, 1017) |
| 0.9 | 2.30 | 1001 | 1000 | 11 | 11 | 9 | (993, 1007) |

*Table 3.5:* Sample mean, median, standard deviation, SD, root mean squared error, RMSE, mean absolute error, MAE, and a central 95% interval of the Poisson MLE, calculated for 200 data sets of 1000 species simulated using a Poisson distribution for each sampling depth/Poisson parameter combination.

| Depth | $(\alpha, \beta)$ | Mean | Median | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|
| 0.2 | (4,0.06) | 2490 | 1078 | 2850 | 3216 | 1677 | (608, 9553) |
| | (2,0.11) | 2217 | 893 | 2709 | 2970 | 1592 | (194, 9288) |
| | (1,0.25) | 1844 | 850 | 2053 | 2219 | 1252 | (195, 7827) |
| | (0.5,0.56) | 1966 | 889 | 1959 | 2184 | 1357 | (407, 7575) |
| 0.3 | (4, 0.09) | 1404 | 961 | 1657 | 1706 | 688 | (284, 7689) |
| | (2,0.20) | 1373 | 983 | 1280 | 1333 | 571 | (659, 6170) |
| | (1,0.43) | 1500 | 928 | 1491 | 1573 | 748 | (559, 6169) |
| | (0.5,1.04) | 1215 | 944 | 827 | 854 | 444 | (598, 4274) |
| 0.4 | (4,0.13) | 1106 | 990 | 484 | 496 | 221 | (770, 2064) |
| | (2,0.29) | 1057 | 977 | 345 | 350 | 208 | (716, 1918) |
| | (1,0.67) | 1132 | 1005 | 524 | 540 | 273 | (701, 2411) |
| | (0.5,1.79) | 1176 | 1000 | 572 | 598 | 305 | (708, 2761) |
| 0.5 | (4,0.19) | 1017 | 990 | 139 | 140 | 105 | (823, 1402) |
| | (2,0.41) | 1033 | 1005 | 155 | 158 | 115 | (821, 1393) |
| | (1,1) | 1029 | 980 | 320 | 321 | 135 | (753, 1459) |
| | (0.5,3) | 1040 | 996 | 202 | 206 | 138 | (795, 1559) |
| 0.75 | (4,0.41) | 1004 | 1000 | 42 | 43 | 34 | (934, 1103) |
| | (2,1) | 1002 | 999 | 46 | 46 | 36 | (921, 1108) |
| | (1,3) | 1001 | 994 | 48 | 48 | 36 | (920, 1119) |
| | (0.5,15) | 1005 | 1000 | 58 | 58 | 44 | (905, 1161) |
| 0.9 | (4,0.78) | 1001 | 1000 | 17 | 17 | 13 | (968, 1039) |
| | (2,2.17) | 1002 | 1002 | 18 | 18 | 14 | (967, 1039) |
| | (1,9) | 1000 | 1000 | 16 | 16 | 13 | (971, 1032) |
| | (0.5,100) | 1002 | 1001 | 21 | 21 | 16 | (966, 1050) |

*Table 3.6:* Sample mean, median, standard deviation, SD, root mean squared error, RMSE, mean absolute error, MAE, and a central 95% interval of the negative binomial MLE, calculated for 200 data sets of 1000 species simulated using a negative binomial distribution for each sampling depth/ negative binomial parameter combination.

| Depth | Mean | Median | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|
| 0.2 | 2715 | 984 | 5656 | 5910 | 2080 | (188, 13,303) |
| | 2564 | 942 | 3607 | 3932 | 1974 | (192, 14,336) |
| | 2094 | 701 | 3080 | 3269 | 1696 | (194, 11,661) |
| | 2952 | 833 | 3810 | 4281 | 2430 | (324, 13,109) |
| 0.3 | 2180 | 974 | 3201 | 3411 | 1479 | (285, 14,296) |
| | 2256 | 929 | 3321 | 3551 | 1570 | (286, 12,603) |
| | 2280 | 870 | 3502 | 3729 | 1646 | (308, 14,761) |
| | 1643 | 900 | 2335 | 2422 | 978 | (469, 9563) |
| 0.4 | 1157 | 964 | 971 | 984 | 410 | (382, 4369) |
| | 1290 | 950 | 1489 | 1517 | 522 | (418, 8485) |
| | 1423 | 983 | 1683 | 1735 | 621 | (607, 5438) |
| | 1338 | 999 | 1363 | 1404 | 521 | (644, 3832) |
| 0.5 | 1095 | 986 | 414 | 425 | 232 | (511, 2099) |
| | 1058 | 964 | 336 | 341 | 212 | (741, 1864) |
| | 1079 | 986 | 328 | 338 | 222 | (709, 2096) |
| | 1109 | 1016 | 322 | 340 | 225 | (739, 2011) |
| 0.75 | 1011 | 1002 | 91 | 91 | 70 | (866, 1219) |
| | 1008 | 994 | 91 | 91 | 70 | (862, 1236) |
| | 1020 | 1016 | 88 | 90 | 69 | (861, 1223) |
| | 1057 | 1005 | 277 | 283 | 128 | (756, 1953) |
| 0.9 | 1002 | 1001 | 27 | 27 | 21 | (948, 1070) |
| | 1002 | 1000 | 28 | 28 | 23 | (945, 1068) |
| | 1028 | 996 | 193 | 195 | 55 | (891, 1007) |
| | 998 | 996 | 32 | 32 | 24 | (954, 1059) |

*Table 3.7:* Sample mean, median, standard deviation, SD, root mean squared error, RMSE, mean absolute error, MAE, and a central 95% interval of the Neyman Type A-gamma MLE, calculated for 200 data sets of 1000 species simulated using a Neyman Type A-gamma distribution for each sampling depth/ Neyman Type A-gamma parameter combination.

### 3.9.3 Combatting the boundary problem using penalties

To combat the boundary problem, I opted to use penalties, however before implementing the penalised maximum-likelihood estimators, further investigation was required in these areas:

- how to choose the penalty parameter $\gamma_2$ for the second Wang and Lindsay (2005) penalty,
- bias for penalty 3 when using clustered data,
- an iterative approach to penalty 3.
- using confidence intervals with penalties.

#### *Choice of penalty parameter $\gamma_2$*

A value of $\gamma_2 = 0.5$ was selected by Wang and Lindsay (2005). However it was shown by simulation in Wang and Lindsay (2005) that the optimal choice of $\gamma_2$ depended strongly on the true value of the odds function $\psi$. A larger true value of $\psi$ required more bias correction, and vice versa. I performed a simulation study to examine the performance of penalty 3 using one-step iteration, again simulating data sets from the negative binomial and Neyman Type A-gamma distributions for 1000 species.

To investigate which values of $\gamma_2$ were effective for benthic data sets, I set up simulations as before from the negative binomial model (Tables 3.8 and 3.9). We can see from the results that as the sampling depth decreased, the variance of the estimates increased as we would expect. As $\alpha$ decreased, the estimates had less bias overall, but as the penalty parameter $\gamma_2$ increased, there was too much penalty in some cases. This suggests that as $\alpha$ gets smaller the estimates are more sensitive to the penalty parameter used.

I repeated the simulations using the Neyman Type A-gamma species richness estimator and again there was an increase in variance as the sampling depth decreased

| Depth | $\alpha$ | $\gamma_2$ | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|---|
| 0.2 | 4 | 0.25 | 1168 | 959 | 624 | 646 | 374 | (628, 3343) |
| | | 0.5 | 1001 | 913 | 339 | 339 | 240 | (620, 2129) |
| | | 1 | 889 | 863 | 191 | 221 | 184 | (612, 1450) |
| | | 1.5 | 841 | 836 | 150 | 219 | 188 | (600, 1173) |
| | 2 | 0.25 | 1229 | 857 | 1027 | 1052 | 537 | (537, 4481) |
| | | 0.5 | 989 | 824 | 533 | 533 | 345 | (534, 2705) |
| | | 1 | 830 | 758 | 275 | 323 | 274 | (527, 1741) |
| | | 1.5 | 762 | 722 | 189 | 303 | 277 | (522, 1384) |
| | 1 | 0.25 | 1116 | 775 | 922 | 929 | 533 | (474, 4228) |
| | | 0.5 | 892 | 720 | 494 | 506 | 381 | (471, 2508) |
| | | 1 | 734 | 664 | 268 | 377 | 343 | (458, 1585) |
| | | 1.5 | 662 | 614 | 187 | 387 | 363 | (450, 1244) |
| | 0.5 | 0.25 | 1028 | 725 | 802 | 803 | 534 | (386, 3506) |
| | | 0.5 | 815 | 664 | 440 | 477 | 403 | (384, 2085) |
| | | 1 | 660 | 589 | 245 | 419 | 385 | (382, 1317) |
| | | 1.5 | 586 | 545 | 174 | 449 | 422 | (372, 1044) |
| 0.3 | 4 | 0.25 | 1130 | 983 | 505 | 522 | 280 | (718, 2501) |
| | | 0.5 | 1046 | 944 | 339 | 342 | 216 | (716, 1931) |
| | | 1 | 960 | 900 | 220 | 223 | 171 | (712, 1533) |
| | | 1.5 | 910 | 864 | 166 | 189 | 158 | (701, 1337) |
| | 2 | 0.25 | 1258 | 941 | 982 | 1015 | 472 | (670, 3760) |
| | | 0.5 | 1101 | 909 | 575 | 584 | 342 | (660, 2675) |
| | | 1 | 968 | 854 | 345 | 346 | 257 | (643, 1914) |
| | | 1.5 | 896 | 812 | 254 | 275 | 229 | (632, 1546) |
| | 1 | 0.25 | 1067 | 915 | 500 | 504 | 310 | (567, 2567) |
| | | 0.5 | 977 | 881 | 346 | 347 | 252 | (559, 1945) |
| | | 1 | 880 | 827 | 235 | 264 | 223 | (547, 1472) |
| | | 1.5 | 820 | 785 | 184 | 258 | 227 | (538, 1252) |
| | 0.5 | 0.25 | 1135 | 901 | 702 | 715 | 427 | (524, 3643) |
| | | 0.5 | 1001 | 861 | 453 | 453 | 326 | (520, 2440) |
| | | 1 | 877 | 801 | 290 | 315 | 268 | (510, 1700) |
| | | 1.5 | 807 | 754 | 221 | 294 | 262 | (502, 1398) |

*Table 3.8:* Summary statistics of the negative binomial penalised MLE using penalty 2 and a range of penalty parameter values, calculated for 200 data sets of 1000 species simulated using a negative binomial distribution for each sampling depth/ negative binomial parameter combination.

| Depth | $\alpha$ | $\gamma_2$ | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|---|
| 0.4 | 4 | 0.25 | 1033 | 975 | 227 | 229 | 161 | (762, 1785) |
| | | 0.5 | 1010 | 959 | 201 | 201 | 148 | (760, 1631) |
| | | 1 | 973 | 935 | 166 | 168 | 131 | (758, 1438) |
| | | 1.5 | 945 | 918 | 142 | 153 | 125 | (756, 1315) |
| | 2 | 0.25 | 1017 | 934 | 297 | 298 | 202 | (703, 1874) |
| | | 0.5 | 989 | 919 | 256 | 256 | 188 | (702, 1742) |
| | | 1 | 948 | 895 | 208 | 214 | 172 | (700, 1565) |
| | | 1.5 | 917 | 873 | 179 | 197 | 165 | (698, 1442) |
| | 1 | 0.25 | 1046 | 958 | 355 | 358 | 214 | (683, 1937) |
| | | 0.5 | 1009 | 940 | 284 | 285 | 192 | (676, 1746) |
| | | 1 | 958 | 904 | 220 | 224 | 170 | (667, 1518) |
| | | 1.5 | 921 | 879 | 186 | 202 | 163 | (657, 1377) |
| | 0.5 | 0.25 | 1041 | 980 | 268 | 271 | 193 | (696, 1691) |
| | | 0.5 | 1006 | 959 | 230 | 230 | 175 | (691, 1543) |
| | | 1 | 954 | 920 | 186 | 191 | 157 | (684, 1362) |
| | | 1.5 | 914 | 888 | 158 | 180 | 151 | (677, 1249) |
| 0.5 | 4 | 0.25 | 1010 | 980 | 162 | 162 | 110 | (815, 1500) |
| | | 0.5 | 1001 | 972 | 152 | 152 | 107 | (811, 1452) |
| | | 1 | 984 | 958 | 138 | 139 | 102 | (805, 1375) |
| | | 1.5 | 969 | 945 | 127 | 130 | 99 | (803, 1312) |
| | 2 | 0.25 | 998 | 969 | 141 | 141 | 104 | (799, 1373) |
| | | 0.5 | 988 | 960 | 134 | 135 | 101 | (795, 1345) |
| | | 1 | 970 | 946 | 123 | 127 | 100 | (788, 1297) |
| | | 1.5 | 954 | 934 | 114 | 123 | 101 | (781, 1256) |
| | 1 | 0.25 | 1018 | 983 | 148 | 149 | 109 | (809, 1389) |
| | | 0.5 | 1006 | 974 | 141 | 141 | 106 | (804, 1353) |
| | | 1 | 985 | 958 | 128 | 129 | 101 | (796, 1296) |
| | | 1.5 | 967 | 943 | 118 | 123 | 100 | (787, 1248) |
| | 0.5 | 0.25 | 1012 | 963 | 191 | 192 | 134 | (760, 1509) |
| | | 0.5 | 998 | 953 | 178 | 178 | 129 | (757, 1450) |
| | | 1 | 974 | 934 | 158 | 161 | 123 | (751, 1360) |
| | | 1.5 | 953 | 920 | 144 | 151 | 121 | (745, 1292) |

*Table 3.9:* Summary statistics of the negative binomial penalised MLE using penalty 2 and a range of penalty parameter values, calculated for 200 data sets of 1000 species simulated using a negative binomial distribution for each sampling depth/ negative binomial parameter combination. cont.

(Tables 3.10 to 3.12). We also see a pattern of less bias as $\alpha$ increased, and less sensitivity to the penalty parameter used.

A penalty parameter between 0.25 and 1 performed best, although the exact value did vary between depths and with the parameter values of the negative binomial and Neyman Type A-gamma used. Therefore we see that the choice of $\gamma_2 = 0.5$ was fairly effective for the negative binomial and gave a reasonable estimate of $N = 1000$ in all cases. However, to avoid too much negative bias, I may want to consider a smaller penalty parameter of $\gamma_2 = 0.25$ when using the Neyman Type A-gamma model when sampling depth is low.

*Penalty 3*

I investigated the bias of penalty 3 for negative binomial and Neyman Type A-gamma estimators. I performed a simulation study to examine the performance of penalty 3 using one-step iteration, again simulating data sets from the negative binomial and Neyman Type A-gamma distributions for 1000 species. The results showed that this penalty caused negative bias in the richness estimate, particulary for smaller values of $\alpha$ (Tables 3.13 and 3.14). When the depth was greater than 0.5 there was no need to apply a penalty, and the results showed that the estimates here were close to the true species richness of $N = 1000$. However, overall this penalty was too harsh, and when $\alpha = 0.5$ and sampling depth was 0.2 or 0.3 we can see a very low mean estimate of $N$ of 616 and 761 respectively, and the true species richness did not lie within the 95% central interval of the sample estimates.

When using the Neyman Type A-gamma model, at every depth the penalised MLE over-penalised the likelihood, and the bias increased significantly as the sampling depth decreased (Table 3.14). The true species richness was only included within the 95% central interval when sampling depth reached 90%, despite the estimates without penalty giving good results for sampling depths of 0.75 and over. This penalty was clearly too harsh for data with a spatial aspect, and introduced too much bias towards

| Depth | $\alpha$ | $\gamma_2$ | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|---|
| 0.2 | 4 | 0.25 | 1092 | 945 | 590 | 597 | 397 | (203, 3053) |
| | | 0.5 | 914 | 859 | 394 | 404 | 302 | (190, 2006) |
| | | 1 | 739 | 734 | 334 | 424 | 342 | (188, 1582) |
| | | 1.5 | 652 | 666 | 312 | 467 | 393 | (188, 1412) |
| | 2 | 0.25 | 932 | 760 | 636 | 639 | 448 | (190, 2645) |
| | | 0.5 | 725 | 665 | 384 | 472 | 401 | (184, 1736) |
| | | 1 | 610 | 600 | 263 | 470 | 419 | (190, 1262) |
| | | 1.5 | 536 | 543 | 251 | 527 | 481 | (182, 1167) |
| | 1 | 0.25 | 912 | 672 | 648 | 654 | 469 | (219, 2635) |
| | | 0.5 | 679 | 604 | 358 | 481 | 411 | (187, 1570) |
| | | 1 | 534 | 521 | 209 | 511 | 478 | (194, 1006) |
| | | 1.5 | 453 | 464 | 176 | 574 | 549 | (188, 798) |
| | 0.5 | 0.25 | 974 | 669 | 835 | 835 | 590 | (315, 3589) |
| | | 0.5 | 715 | 587 | 440 | 524 | 451 | (310, 2005) |
| | | 1 | 532 | 476 | 238 | 525 | 491 | (195, 1150) |
| | | 1.5 | 458 | 428 | 172 | 568 | 546 | (194, 870) |
| 0.3 | 4 | 0.25 | 1139 | 942 | 613 | 628 | 382 | (375, 2805) |
| | | 0.5 | 908 | 868 | 424 | 434 | 331 | (291, 1850) |
| | | 1 | 781 | 803 | 278 | 354 | 285 | (292, 1318) |
| | | 1.5 | 719 | 749 | 255 | 379 | 306 | (280, 1175) |
| | 2 | 0.25 | 1115 | 927 | 578 | 590 | 394 | (510, 2914) |
| | | 0.5 | 913 | 889 | 336 | 347 | 279 | (308, 1818) |
| | | 1 | 735 | 728 | 264 | 374 | 310 | (291, 1328) |
| | | 1.5 | 657 | 664 | 206 | 400 | 350 | (285, 1068) |
| | 1 | 0.25 | 1065 | 856 | 648 | 651 | 385 | (475, 3220) |
| | | 0.5 | 869 | 776 | 347 | 371 | 281 | (468, 1805) |
| | | 1 | 737 | 713 | 239 | 355 | 308 | (283, 1352) |
| | | 1.5 | 674 | 657 | 161 | 364 | 333 | (285, 1041) |
| | 0.5 | 0.25 | 1026 | 750 | 794 | 794 | 471 | (464, 4088) |
| | | 0.5 | 874 | 713 | 446 | 464 | 361 | (449, 2444) |
| | | 1 | 743 | 630 | 282 | 382 | 336 | (421, 1590) |
| | | 1.5 | 669 | 590 | 205 | 390 | 358 | (399, 1236) |

*Table 3.10:* Summary statistics of the Neyman Type A-gamma penalised MLE using penalty 2 and a range of penalty parameter values, calculated for 200 data sets of 1000 species simulated using a Neyman Type A-gamma distribution for each sampling depth/ Neyman Type A-gamma parameter combination.

| Depth | $\alpha$ | $\gamma_2$ | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|---|
| 0.4 | 4 | 0.25 | 1065 | 938 | 608 | 612 | 295 | (397, 2091) |
| | | 0.5 | 955 | 913 | 408 | 410 | 255 | (385, 1683) |
| | | 1 | 871 | 874 | 285 | 312 | 230 | (380, 1376) |
| | | 1.5 | 798 | 821 | 250 | 321 | 253 | (379, 1227) |
| | 2 | 0.25 | 1140 | 982 | 727 | 740 | 356 | (404, 2803) |
| | | 0.5 | 1006 | 923 | 465 | 465 | 283 | (394, 2014) |
| | | 1 | 900 | 874 | 310 | 326 | 240 | (381, 1591) |
| | | 1.5 | 818 | 817 | 267 | 323 | 256 | (379, 1330) |
| | 1 | 0.25 | 1079 | 923 | 691 | 695 | 329 | (405, 2803) |
| | | 0.5 | 966 | 884 | 449 | 450 | 279 | (394, 1934) |
| | | 1 | 875 | 849 | 302 | 326 | 247 | (381, 1591) |
| | | 1.5 | 800 | 797 | 251 | 321 | 263 | (379, 1245) |
| | 0.5 | 0.25 | 1080 | 953 | 619 | 624 | 313 | (405, 2091) |
| | | 0.5 | 974 | 913 | 418 | 418 | 269 | (394, 1778) |
| | | 1 | 885 | 866 | 290 | 312 | 236 | (381, 1485) |
| | | 1.5 | 811 | 813 | 249 | 313 | 251 | (379, 1310) |
| 0.5 | 4 | 0.25 | 1062 | 978 | 363 | 368 | 217 | (509, 2219) |
| | | 0.5 | 1004 | 964 | 329 | 329 | 219 | (499, 2030) |
| | | 1 | 941 | 922 | 283 | 289 | 207 | (489, 1719) |
| | | 1.5 | 893 | 901 | 243 | 265 | 200 | (488, 1533) |
| | 2 | 0.25 | 1035 | 974 | 225 | 228 | 168 | (719, 1618) |
| | | 0.5 | 1007 | 956 | 205 | 205 | 156 | (713, 1570) |
| | | 1 | 963 | 923 | 173 | 177 | 146 | (702, 1363) |
| | | 1.5 | 925 | 889 | 151 | 168 | 144 | (693, 1304) |
| | 1 | 0.25 | 990 | 937 | 210 | 211 | 167 | (716, 1502) |
| | | 0.5 | 965 | 916 | 190 | 193 | 156 | (719, 1420) |
| | | 1 | 923 | 878 | 162 | 180 | 153 | (692, 1299) |
| | | 1.5 | 890 | 859 | 135 | 174 | 148 | (677, 1204) |
| | 0.5 | 0.25 | 1045 | 967 | 277 | 281 | 191 | (730, 1824) |
| | | 0.5 | 1014 | 940 | 243 | 243 | 177 | (725, 1676) |
| | | 1 | 965 | 922 | 197 | 201 | 157 | (714, 1467) |
| | | 1.5 | 927 | 885 | 169 | 184 | 155 | (705, 1356) |

*Table 3.11:* Summary statistics of the Neyman Type A-gamma penalised MLE using penalty 2 and a range of penalty parameter values, calculated for 200 data sets of 1000 species simulated using a Neyman Type A-gamma distribution for each sampling depth/ Neyman Type A-gamma parameter combination. cont.

| Depth | $\alpha$ | $\gamma_2$ | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|---|
| 0.75 | 4 | 0.25 | 1000 | 979 | 94 | 94 | 66 | (865, 1292) |
| | | 0.5 | 997 | 978 | 93 | 93 | 67 | (862, 1250) |
| | | 1 | 989 | 968 | 93 | 94 | 71 | (855, 1243) |
| | | 1.5 | 969 | 956 | 74 | 80 | 64 | (848, 1177) |
| | 2 | 0.25 | 995 | 989 | 109 | 110 | 82 | (828, 1252) |
| | | 0.5 | 991 | 985 | 104 | 105 | 79 | (826, 1233) |
| | | 1 | 981 | 976 | 102 | 103 | 79 | (819, 1216) |
| | | 1.5 | 972 | 971 | 98 | 102 | 79 | (814, 1196) |
| | 1 | 0.25 | 996 | 984 | 87 | 87 | 71 | (856, 1197) |
| | | 0.5 | 991 | 979 | 85 | 85 | 70 | (853, 1189) |
| | | 1 | 984 | 975 | 83 | 84 | 68 | (846, 1172) |
| | | 1.5 | 973 | 965 | 81 | 86 | 70 | (840, 1168) |
| | 0.5 | 0.25 | 1083 | 1020 | 272 | 285 | 150 | (740, 1961) |
| | | 0.5 | 1085 | 1018 | 269 | 282 | 144 | (769, 1961) |
| | | 1 | 1087 | 1003 | 278 | 292 | 156 | (772, 1961) |
| | | 1.5 | 1060 | 993 | 280 | 286 | 151 | (737, 1961) |
| 0.9 | 4 | 0.25 | 1000 | 1003 | 31 | 31 | 25 | (914, 1063) |
| | | 0.5 | 999 | 1001 | 30 | 30 | 24 | (930, 1061) |
| | | 1 | 995 | 997 | 30 | 30 | 24 | (920, 1057) |
| | | 1.5 | 990 | 994 | 34 | 36 | 26 | (900, 1056) |
| | 2 | 0.25 | 1002 | 995 | 29 | 29 | 24 | (952, 1069) |
| | | 0.5 | 1002 | 995 | 32 | 32 | 24 | (951, 1089) |
| | | 1 | 997 | 992 | 28 | 28 | 23 | (950, 1064) |
| | | 1.5 | 992 | 988 | 29 | 30 | 24 | (927, 1061) |
| | 1 | 0.25 | 1014 | 1003 | 141 | 142 | 48 | (916, 1285) |
| | | 0.5 | 1013 | 1001 | 142 | 142 | 47 | (916, 1284) |
| | | 1 | 1011 | 999 | 142 | 142 | 47 | (916, 1281) |
| | | 1.5 | 1011 | 996 | 141 | 141 | 44 | (916, 1280) |
| | 0.5 | 0.25 | 990 | 985 | 34 | 36 | 29 | (916, 1062) |
| | | 0.5 | 988 | 983 | 35 | 37 | 30 | (909, 1058) |
| | | 1 | 984 | 980 | 34 | 37 | 30 | (909, 1052) |
| | | 1.5 | 979 | 976 | 33 | 39 | 32 | (915, 1046) |

*Table 3.12:* Summary statistics of the Neyman Type A-gamma penalised MLE using penalty 2 and a range of penalty parameter values, calculated for 200 data sets of 1000 species simulated using a Neyman Type A-gamma distribution for each sampling depth/ Neyman Type A-gamma parameter combination. cont.

| Depth | $(\alpha, \beta)$ | Mean | Median | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|
| 0.2 | (4,0.06) | 958 | 950 | 239 | 243 | 190 | (210, 1377) |
| | (2,0.11) | 796 | 800 | 253 | 325 | 258 | (193, 1291) |
| | (1,0.25) | 684 | 696 | 179 | 363 | 320 | (195, 1008) |
| | (0.5,0.56) | 616 | 613 | 127 | 405 | 384 | (393, 885) |
| 0.3 | (4, 0.09) | 894 | 928 | 281 | 301 | 211 | (286, 1406) |
| | (2,0.20) | 918 | 886 | 174 | 192 | 159 | (660, 1275) |
| | (1,0.43) | 840 | 816 | 162 | 228 | 194 | (557, 1185) |
| | (0.5,1.04) | 761 | 759 | 109 | 263 | 242 | (571, 963) |
| 0.4 | (4,0.13) | 1007 | 967 | 167 | 167 | 129 | (770, 1443) |
| | (2,0.29) | 957 | 943 | 152 | 158 | 129 | (715, 1277) |
| | (1,0.67) | 943 | 930 | 148 | 159 | 131 | (697, 1265) |
| | (0.5,1.79) | 894 | 870 | 120 | 160 | 136 | (695, 1185) |
| 0.5 | (4,0.19) | 1002 | 985 | 119 | 119 | 93 | (822, 1289) |
| | (2,0.41) | 1003 | 986 | 117 | 117 | 93 | (821, 1264) |
| | (1,1) | 968 | 954 | 121 | 125 | 97 | (750, 1237) |
| | (0.5,3) | 952 | 941 | 106 | 116 | 96 | (773, 1217) |
| 0.75 | (4,0.41) | 1003 | 1000 | 42 | 42 | 33 | (934, 1102) |
| | (2,1) | 1001 | 997 | 45 | 45 | 35 | (920, 1106) |
| | (1,3) | 999 | 992 | 46 | 46 | 36 | (919, 1115) |
| | (0.5,15) | 999 | 994 | 54 | 54 | 42 | (902, 1134) |
| 0.9 | (4,0.78) | 1001 | 1000 | 17 | 17 | 13 | (968, 1039) |
| | (2,2.17) | 1002 | 1002 | 17 | 18 | 14 | (967, 1039) |
| | (1,9) | 1000 | 1000 | 16 | 16 | 13 | (971, 1032) |
| | (0.5,100) | 1001 | 1001 | 20 | 20 | 16 | (966, 1049) |

*Table 3.13:* Summary statistics of the negative binomial penalised MLE using penalty 3, calculated for 200 data sets of 1000 species simulated using a negative binomial distribution for each sampling depth/ negative binomial parameter combination.

| Depth | Mean | Median | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|
| 0.2 | 203 | 204 | 14 | 797 | 797 | (178, 231) |
| | 204 | 204 | 14 | 797 | 796 | (176, 230) |
| | 204 | 203 | 13 | 796 | 796 | (175, 231) |
| | 206 | 205 | 15 | 794 | 794 | (180, 231) |
| 0.3 | 304 | 306 | 15 | 696 | 696 | (274, 334) |
| | 303 | 302 | 15 | 698 | 697 | (271, 334) |
| | 308 | 308 | 17 | 692 | 692 | (273, 348) |
| | 324 | 317 | 32 | 676 | 676 | (276, 396) |
| 0.4 | 408 | 408 | 15 | 593 | 592 | (374, 438) |
| | 407 | 407 | 16 | 593 | 593 | (376, 441) |
| | 429 | 418 | 41 | 572 | 571 | (379, 542) |
| | 456 | 452 | 40 | 545 | 544 | (388, 532) |
| 0.5 | 511 | 508 | 23 | 490 | 489 | (476, 545) |
| | 520 | 515 | 32 | 481 | 480 | (481, 621) |
| | 558 | 546 | 47 | 444 | 442 | (491, 649) |
| | 587 | 586 | 45 | 415 | 413 | (508, 672) |
| 0.75 | 766 | 733 | 66 | 243 | 234 | (688, 900) |
| | 780 | 784 | 54 | 226 | 220 | (695, 882) |
| | 796 | 805 | 51 | 210 | 204 | (705, 888) |
| | 791 | 784 | 67 | 220 | 214 | (714, 898) |
| 0.9 | 936 | 925 | 34 | 72 | 65 | (891, 1007) |
| | 954 | 958 | 31 | 55 | 47 | (954, 1059) |
| | 956 | 957 | 26 | 51 | 46 | (948, 1055) |
| | 954 | 951 | 31 | 56 | 49 | (901, 1010) |

*Table 3.14:* Summary statistics of the Neyman Type A-gamma penalised MLE using penalty 3, calculated for 200 data sets of 1000 species simulated using a Neyman Type A-gamma distribution for each sampling depth/ Neyman Type A-gamma parameter combination.

the naive estimator of Chao (1984), $\hat{N}_{C1}$. Therefore I would strongly consider using an alternative.

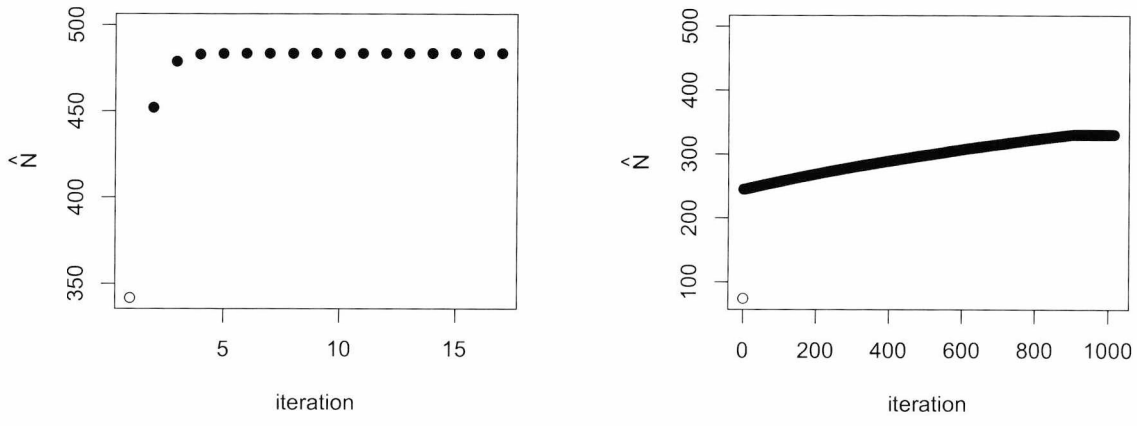
*Iterative approach to penalty 3*

I investigated an iterative approach, which started by calculating the penalised estimate using penalty 3, and then used this species richness estimate as an alternative to the $Chao_1$ species richness estimate in $\eta$. I looked at the convergence of an iterative penalty, because if it did not converge then it would not act to reduce the bias of penalty 3.

I found that the iterative penalised likelihood estimator converged towards a final estimate. When there was no boundary problem present, and the estimate of species richness without a penalty was not spuriously large, the iterative species richness estimate converged towards this non-penalised estimate, reaching it after two or three iterations even when the $Chao_1$ estimate was not close to the true species richness, as in the example in Figure 3.7a. In this example 252 species were observed, and the $Chao_1$ estimate was 342 while the non-penalised MLE was 483. Using the penalised MLE with iterative penalty 3, the estimate converged to the same value after three iterations.

When the boundary problem was present, as in the example shown in Figure 3.7b, the iterative penalised likelihood estimator converged towards a final estimate. For this example, I simulated abundance data for 500 species over 5 grabs from a negative binomial distribution with negative binomial parameters $\alpha = 0.1$ and $\beta = 0.1$.

Eighteen species were observed in the sample, and the non-penalised MLE was 7701 and the $Chao_1$ estimate was 74. Applying the penalised MLE with penalty 3, without iteration, gave a species richness estimate of 76, which is not a good estimate of the species richness. After one-step of the iteration, using this estimate in place of the $Chao_1$ estimate within penalty 3, the estimate improved greatly to 245. This estimate was still low, but a vast improvement on the estimate given using penalty




(a) $\alpha = 0.5$, $\beta = 0.5$ (b) $\alpha = 0.1$, $\beta = 0.1$

*Figure 3.7:* Plot of the convergence of the negative binomial MLE using iterative penalty 3 when applied to simulated abundance data for 500 species over 5 grabs from a negative binomial distribution, with (a) $\alpha = 0.5$, $\beta = 0.5$ (b) $\alpha = 0.1$, $\beta = 0.1$. The open circle shows the $Chao_1$ estimate.

3 without iteration. The fully iterative approach gave a species richness estimate of 330, converging after 901 iterations, which is again an improvement, but did not reach the true species richness of $N = 500$.

However, in some cases where the non-penalised estimate was relatively large, the iterative penalty still converged to that value. For example, $\hat{N} = 3441$, and the iterative penalty converged to 3367 after 76 iterations. The $Chao_1$ estimate was 744. Here the iterative penalty was not improving the non-penalised estimate. However it also did not degrade it, and if there was a possibility of the boundary problem occurring, it might be wise to utilise this penalty.

A problem with an iterative approach is the high computation time required when it applied to more complicated distributions, such as the Neyman Type A-gamma. To compute each iterative step for the Neyman Type A-gamma takes appropriately ten minutes, and so 901 iterations would take around 150 hours. Therefore I opted

for a one-step iteration, which as we can see in Figure 3.7 can already improve the species richness estimate greatly. If computation time could be improved, an area I am currently working on, then it might be possible to use more iterative steps. Also, using more iterations may be appropriate when fitting the negative binomial model, where computational costs are not high.

Another issue that may occur when applying this iterative penalty was that when the $Chao_1$ estimate was greater than the true species richness, then iterating the penalty could put more bias on the estimate.

I performed a simulation study to examine the performance of penalty 3 using one-step iteration, again simulating data sets from the negative binomial and Neyman Type A-gamma distributions for 1000 species. Comparing the results for the negative binomial model from this method to those of penalty 3 without iteration, we see that using iteration means that the true value of the number of species, $N = 1000$, always falls within the 95% central interval of the estimates over the 200 simulations (Tables 3.13 and 3.15).

When sampling depth was low, we see that this method gave smaller variance compared to the penalty without iteration. However, if I increased the proportion of species seen the variance also increased, and all the mean estimates were larger. At sampling depth greater than 0.5 the results were similar for both methods, and the results without a penalty. A penalty was no longer needed here as the boundary problem was not occurring.

We still see a significant negative bias in the estimates for a low sampling depth, which could be reduced if I introduced further iterative steps into the estimation. This would not be too computationally intensive for the negative binomial case, however, it may not be possible for more complex distributions such as the Neyman Type A-gamma.

| Depth | $(\alpha, \beta)$ | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|---|
| 0.2 | (4,0.06) | 982 | 940 | 243 | 244 | 197 | (630, 1530) |
| | (2,0.11) | 890 | 846 | 235 | 260 | 220 | (580, 1480) |
| | (1,0.25) | 779 | 748 | 200 | 298 | 255 | (473, 1201) |
| | (0.5,0.56) | 690 | 673 | 187 | 362 | 320 | (396, 1087) |
| 0.3 | (4, 0.09) | 1041 | 967 | 244 | 248 | 195 | (723, 1604) |
| | (2,0.20) | 982 | 909 | 266 | 267 | 219 | (623, 1624) |
| | (1,0.43) | 940 | 906 | 239 | 247 | 207 | (566, 1449) |
| | (0.5,1.04) | 891 | 877 | 202 | 229 | 193 | (555, 1258) |
| 0.4 | (4,0.13) | 1038 | 990 | 216 | 220 | 167 | (746, 1611) |
| | (2,0.29) | 1040 | 1002 | 231 | 234 | 177 | (713, 1584) |
| | (1,0.67) | 1053 | 1019 | 220 | 226 | 173 | (739, 1576) |
| | (0.5,1.79) | 1003 | 974 | 211 | 211 | 166 | (669, 1501) |
| 0.5 | (4,0.19) | 1003 | 988 | 140 | 140 | 103 | (800, 1336) |
| | (2,0.41) | 1000 | 983 | 136 | 136 | 106 | (803, 1392) |
| | (1,1) | 1004 | 971 | 158 | 158 | 121 | (778, 1407) |
| | (0.5,3) | 1030 | 1016 | 154 | 156 | 121 | (799, 1385) |
| 0.75 | (4,0.41) | 1002 | 1000 | 40 | 40 | 31 | (924, 1093) |
| | (2,1) | 997 | 990 | 43 | 43 | 35 | (929, 1109) |
| | (1,3) | 999 | 996 | 46 | 46 | 37 | (920, 1101) |
| | (0.5,15) | 1004 | 999 | 49 | 50 | 39 | (923, 1101) |
| 0.9 | (4,0.78) | 1001 | 1001 | 16 | 17 | 13 | (970, 1037) |
| | (2,2.17) | 1002 | 1001 | 16 | 16 | 12 | (972, 1032) |
| | (1,9) | 1000 | 999 | 18 | 18 | 14 | (968, 1041) |
| | (0.5,100) | 1001 | 1001 | 18 | 18 | 14 | (966, 1044) |

*Table 3.15:* Summary statistics of the negative binomial penalised MLE using penalty 3 with one-step iteration, calculated for 200 data sets of 1000 species simulated using a negative binomial distribution for each sampling depth/ negative binomial parameter combination.

| Depth | $\hat{N}$ | $\hat{M}$ | SD | RMSE | MAE | Central 95% |
|---|---|---|---|---|---|---|
| 0.2 | 205 | 204 | 14 | 795 | 795 | (180, 232) |
| | 204 | 203 | 12 | 796 | 796 | (182, 231) |
| | 203 | 203 | 12 | 797 | 797 | (180, 225) |
| | 210 | 205 | 32 | 791 | 790 | (179, 330) |
| 0.3 | 303 | 303 | 15 | 697 | 697 | (273, 332) |
| | 306 | 306 | 15 | 694 | 694 | (275, 337) |
| | 308 | 303 | 34 | 693 | 692 | (279, 333) |
| | 396 | 323 | 118 | 615 | 604 | (285, 637) |
| 0.4 | 409 | 409 | 21 | 591 | 591 | (378, 443) |
| | 409 | 405 | 39 | 592 | 591 | (379, 440) |
| | 479 | 415 | 132 | 538 | 521 | (380, 800) |
| | 639 | 665 | 154 | 392 | 361 | (393, 895) |
| 0.5 | 507 | 509 | 17 | 493 | 493 | (473, 538) |
| | 557 | 512 | 122 | 459 | 443 | (480, 932) |
| | 706 | 760 | 179 | 344 | 296 | (485, 1016) |
| | 840 | 839 | 109 | 194 | 167 | (536, 1067) |
| 0.75 | 832 | 727 | 149 | 224 | 191 | (691, 1129) |
| | 896 | 934 | 138 | 173 | 131 | (696, 1105) |
| | 981 | 979 | 90 | 92 | 70 | (739, 1164) |
| | 974 | 978 | 93 | 96 | 72 | (746, 1169) |
| 0.9 | 948 | 922 | 46 | 70 | 60 | (895, 1050) |
| | 988 | 997 | 41 | 43 | 31 | (903, 1053) |
| | 998 | 1000 | 31 | 31 | 23 | (918, 1057) |
| | 992 | 991 | 27 | 28 | 22 | (946, 1051) |

*Table 3.16:* Summary statistics of the Neyman Type A-gamma penalised MLE using penalty 3 with one-step iteration, calculated for 200 data sets of 1000 species simulated using a Neyman Type A-gamma distribution for each sampling depth/ Neyman Type A-gamma parameter combination.

For the Neyman Type A-gamma, introducing an iterative step to the species richness estimation did have the desired effect of reducing negative bias when the sampling depth was larger (Table 3.16). However, for sampling depth of 0.4 and lower, the iterative step had little effect on the species richness estimates.

Comparing these results to those using no penalty, for depths greater than 0.4 the non-penalised estimator performed better in terms of bias (Table 3.7). We therefore need to consider whether we are concerned primarily with variance or with bias in our species richness estimator.

*Confidence Intervals for penalised likelihood*

I found that the profile confidence intervals worked well for the MLE, so I used these intervals when applying the penalised likelihood. I ran a simulation for the negative binomial to check coverage when using penalties.

Table 3.17 shows the coverage of the profile confidence intervals for negative binomial simulated data using penalty 2 with $\gamma_2 = 0.5$, penalty 3 and penalty 3 using one-step iteration.

The coverage varied a lot between the penalties in some cases. At low sampling depth, when $\alpha$ was small the coverage of MLE with penalty 3 was very low. This illustrates the severity of the penalty, and suggests that using this penalised likelihood introduces too much bias to the estimate. Using the iterative step increased the coverage, but it can still be low. However, the coverage increased as the depth increased.

As expected, as depth increased, so did the coverage of the 95% profile confidence intervals. Penalty 2 had the best coverage, performing well in all cases, but when sampling depth was less than 0.75 the width of the confidence interval was significantly larger than those of the other penalties.

| | | Coverage | | | Width | | |
|---|---|---|---|---|---|---|---|
| Depth | $\alpha$ | P2 | P3 | P3 IT | P2 | P3 | P3 IT |
| 0.2 | 4 | 1.00 | 0.98 | 0.99 | 2542.76 | 969.72 | 1049.2 |
| | 2 | 0.99 | 0.93 | 0.96 | 2540.60 | 826.70 | 967.54 |
| | 1 | 0.97 | 0.77 | 0.79 | 2349.94 | 722.24 | 850.82 |
| | 0.5 | 0.92 | 0.38 | 0.68 | 2187.60 | 572.19 | 762.27 |
| 0.3 | 4 | 0.99 | 0.97 | 0.99 | 2304.90 | 875.84 | 1011.22 |
| | 2 | 0.97 | 0.97 | 0.94 | 2027.18 | 824.94 | 966.82 |
| | 1 | 0.96 | 0.90 | 0.93 | 2598.64 | 708.33 | 928.69 |
| | 0.5 | 0.91 | 0.73 | 0.88 | 2002.78 | 587.59 | 875.25 |
| 0.4 | 4 | 0.97 | 0.95 | 0.97 | 1341.97 | 726.32 | 849.36 |
| | 2 | 0.96 | 0.95 | 0.95 | 1575.17 | 693.21 | 896.58 |
| | 1 | 0.95 | 0.94 | 0.96 | 1608.88 | 621.60 | 862.33 |
| | 0.5 | 0.94 | 0.89 | 0.93 | 1610.04 | 536.46 | 798.71 |
| 0.5 | 4 | 0.94 | 0.97 | 0.96 | 637.65 | 553.97 | 567.6 |
| | 2 | 0.97 | 0.94 | 0.97 | 683.98 | 528.76 | 629.37 |
| | 1 | 0.93 | 0.95 | 0.93 | 1190.14 | 498.85 | 629.07 |
| | 0.5 | 0.94 | 0.94 | 0.96 | 1336.37 | 453.64 | 669.88 |
| 0.75 | 4 | 0.94 | 0.94 | 0.97 | 166.24 | 170.38 | 170.07 |
| | 2 | 0.98 | 0.95 | 0.95 | 172.60 | 175.79 | 176.26 |
| | 1 | 0.97 | 0.94 | 0.91 | 181.17 | 183.26 | 185.28 |
| | 0.5 | 0.97 | 0.94 | 0.96 | 195.41 | 191.95 | 198.58 |
| 0.9 | 4 | 0.94 | 0.93 | 0.96 | 63.93 | 64.52 | 64.85 |
| | 2 | 0.95 | 0.96 | 0.97 | 66.04 | 67.99 | 67.61 |
| | 1 | 0.96 | 0.97 | 0.94 | 69.45 | 69.37 | 69.71 |
| | 0.5 | 0.92 | 0.94 | 0.94 | 54.40 | 54.04 | 53.94 |

*Table 3.17:* Coverage and widths of 95% profile likelihood confidence intervals for $\hat{N}$ for the negative binomial MLE using penalty 2, P2, penalty 3, P3, and penalty 3 with one-step iteration, P3 IT, for various sampling depths and negative binomial parameter values, $\alpha$.

Since I have seen good coverage probabilities for the profile confidence intervals, I will use them for any further confidence interval estimation when applying the penalised maximum-likelihood estimator.

## 3.10 Analysis of real data

I applied the models to several real data sets: the well-known Lepidoptera data used in Fisher et al. (1943), Christmas Bird Count (CBC) data and benthic data sets from CEFAS, for which the species richness estimators including clustering were developed.

### 3.10.1 Lepidoptera data

The data contain 15,609 individuals, 240 species and maximum frequency 2,349. As estimation methods are affected by the long right tail of the distribution of frequencies, the data were truncated at $\tau = 112$, the same value used in Barger and Bunge (2010).

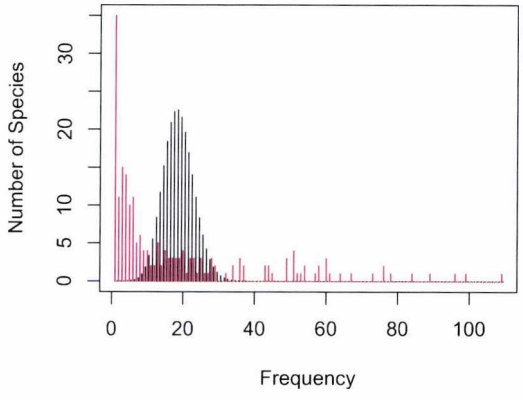
Table 3.18 shows the MLE of $N$ for the Lepidoptera data set based on different models, and the species richness estimates varied between the models. Both the negative binomial and Neyman Type A-gamma models were judged to provide an acceptable fit to the observed data by the Pearson chi-squared test. Using the AIC to compare models directly, the negative binomial model was selected as the best model for this data set. The Neyman Type A-gamma model had a $\Delta$AIC value below ten,

| Model | $\hat{N}$ | 95% Confidence | $\chi^2$ (df) | 5%$\chi^2$ point | AIC | $\Delta$AIC |
|---|---|---|---|---|---|---|
| PO | 241 | (241, 241) | 2282.00 (17) | 27.59 | 3733.0 | 3,964.20 |
| NB | 338 | (285, 487) | 21.17 (25) | 37.65 | -231.2 | 0.00 |
| NTAG | 475 | (330, 1172) | 25.75 (23) | 35.17 | -221.5 | 9.70 |

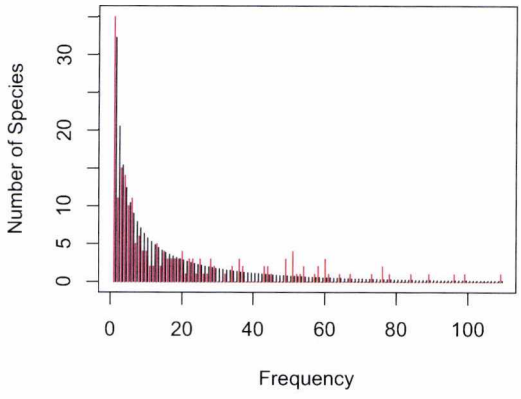
*Table 3.18:* Species richness estimates with 95% confidence interval for the Lepidoptera data set using the Poisson, PO, negative binomial, NB, and Neyman Type A-gamma, NTAG, maximum-likelihood estimator. The Pearson chi-squared value, degrees of freedom, *df*, 5%$\chi^2$ point, AIC and $\Delta$AIC are also reported.

so may be plausible for this data set.

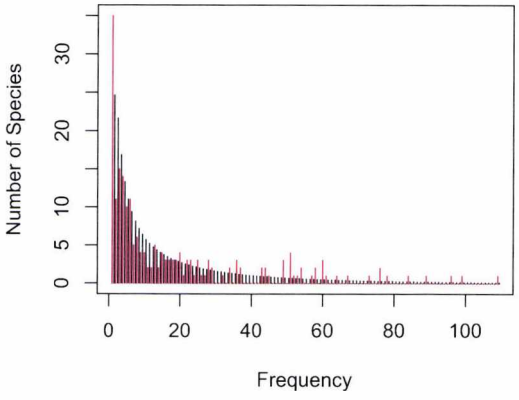
The AIC and chi-squared values of the Poisson model were very high, indicating a poor fit to the data and that this model was unsuitable (Figure 3.8a). The extremely narrow confidence interval for the Poisson model, which in fact was the same as the estimate of only one unseen species, showed that any parameter values away from the optimal ones were very different in likelihood value.




*(a)* Poisson




*(b)* Negative binomial




*(c)* Neyman Type A-gamma

*Figure 3.8:* Observed data (red) and fitted data (black) for the (a) Poisson, (b) negative binomial and (c) Neyman Type A-gamma MLE for the Lepidoptera data set.

The fit of the negative binomial model looks good for this data set (Figure 3.8c), and Figure 3.9 shows that the profile log-likelihood for the number of unseen species is asymmetric.

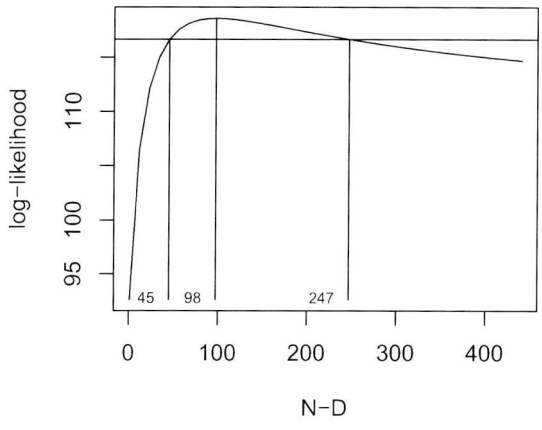



*Figure 3.9:* Profile log-likelihood for the number of unseen species in the Lepidoptera data set assuming the negative binomial model.

Since the estimates obtained from these models were not spuriously large, use of the penalised log-likelihood was not warranted. However, a suitable penalty would allow for this and only penalise the likelihood where necessary. When calculating the penalised MLE for the negative binomial and Neyman Type A-gamma models, the estimated number of species was not greatly affected by the use of penalties (Table 3.19).

For the negative binomial model, penalty 3 with one-step iteration gave the closest result to the non-penalised MLE, which I would expect as it was demonstrated in Section 3.9.3 that the iterative penalty converged to the non-penalised estimate. Penalties 2 and 3 gave estimates close to the non-penalised MLE, and the closeness of the AIC values shows that there was not much difference in the fit of the models to the data, with penalty 2 giving the best fit. The standard chi-square percentage points for these degrees of freedom led me to accept the fit of all of these models at the 5% level of significance for the Lepidoptera data.

| Model | $\hat{N}$ | 95% Confidence | $\chi^2$ (df) | $5\%\chi^2$point | AIC | $\Delta$AIC |
|---|---|---|---|---|---|---|
| NB | 338 | (285, 487) | 21.17 (25) | 37.65 | -241.2 | 0.8 |
| NB P2 | 330 | (282, 459) | 20.02 (25) | 37.65 | -242.0 | 0.0 |
| NB P3 | 333 | (285, 437) | 21.67 (25) | 37.65 | -241.1 | 0.9 |
| NB P3 IT | 337 | (285, 461) | 21.19 (25) | 37.65 | -241.2 | 0.8 |
| NTAG | 475 | (330, 1172) | 25.75 (23) | 35.17 | -235.5 | 10.5 |
| NTAG P2 | 452 | (322, 1084) | 25.84 (23) | 35.17 | -235.5 | 10.5 |
| NTAG P3 | 400 | (318, 547) | 31.39 (24) | 36.42 | -234.0 | 12.0 |
| NTAG P3 IT | 452 | (329, 723) | 25.85 (23) | 35.17 | -235.4 | 10.6 |

*Table 3.19:* Species richness estimates with 95% confidence interval for Lepidoptera data set using negative binomial and Neyman Type A-gamma model in the penalised MLE, with penalty 2, P2, penalty 3, P3 and penalty 3 with one-step iteration, P3 IT. The Pearson chi-squared value, degrees of freedom, *df*, $5\%\chi^2$ point, AIC and $\Delta$AIC are also reported.

The Neyman Type A-gamma model showed slightly greater differences in estimates when the penalties were used. Penalty 3 was the worst model according to AIC, and we can see that it was quite a harsh penalty, reducing the species richness estimate by 75 from the non-penalised estimate. It also had a much narrower confidence interval. Using one-step iteration gave the same estimate as using penalty 2, which is interesting, but the confidence interval was narrower.

Therefore, I concluded that for this data, where the boundary problem was not present, use of the penalised MLE did not adversely effect the species richness estimate when using the negative binomial model, but when using the Neyman Type A-gamma model one must be more wary.

### 3.10.2 CBC data

In the CBC data set 20,042 individuals from 126 species were seen. The data were truncated at 150, as a value above this cannot be calculated for the Neyman Type A-gamma model due to computational limitations. Barger and Bunge (2010) used a truncation point of 221 selected using the criteria described in Section 3.6. Again we see the extremely narrow confidence interval for the Poisson model (Table 3.20). However, the negative binomial model and Neyman Type A-gamma model fitted the data well according to the Pearson goodness-of-fit test.

Figures 3.10 and 3.11 show the observed and fitted values for these two models to the data, and we notice that the negative binomial fits better to the low frequency data. The AIC judged the negative binomial model as the most suitable model, but the Neyman Type A-gamma model was still plausible.

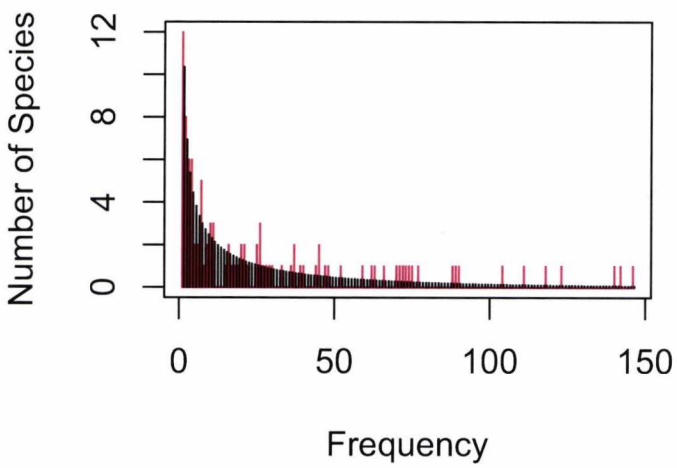



*Figure 3.10:* Observed data (red) and fitted data (black) for the negative binomial MLE for the CBC data set.

| Model | $\hat{N}$ | 95% Confidence | $\chi^2$ (df) | 5%$\chi^2$point | AIC | $\Delta$AIC |
|---|---|---|---|---|---|---|
| PO | 126 | (126, 127) | 7535 (3) | 7.82 | 24228.0 | 24097.1 |
| NB | 154 | (134 218) | 14.77 (17) | 27.59 | 130.9 | 0.0 |
| NTAG | 202 | (148 432) | 20.92 (16) | 26.30 | 138.1 | 7.2 |

*Table 3.20:* Species richness estimates with 95% confidence interval for the CBC data set using the Poisson, PO, negative binomial, NB, and Neyman Type A-gamma, NTAG, maximum-likelihood estimator. The Pearson chi-squared value, degrees of freedom, *df*, 5%$\chi^2$ point, AIC and $\Delta$AIC are also reported.

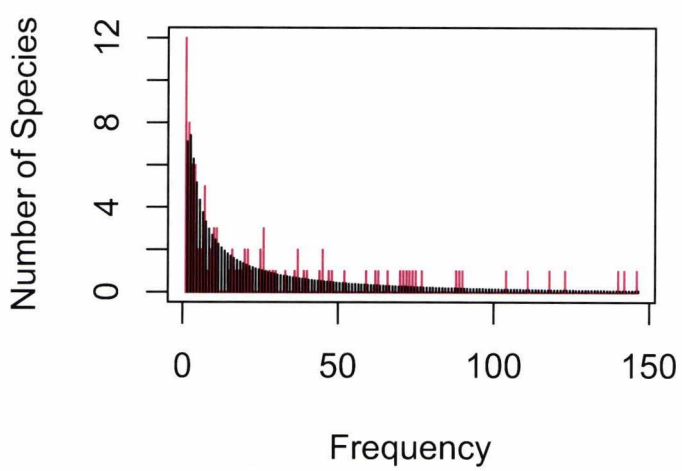
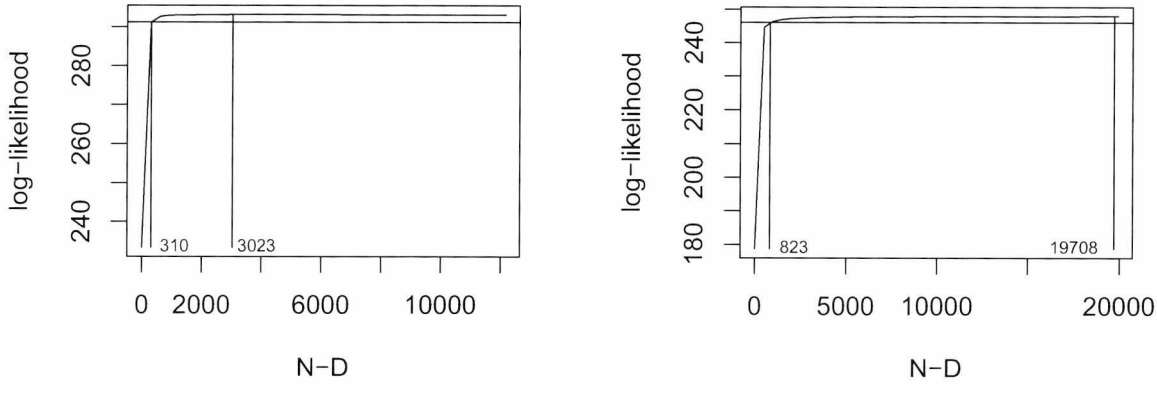



*Figure 3.11:* Observed data (red) and fitted data (black) for the Neyman Type A-gamma MLE for the CBC data set.

### 3.10.3 Isle of Wight benthic data sets

I applied the models to benthic data, and the first data set was from the Isle of Wight area, where a survey collected data using two grabs sizes, large $0.25m^2$ and small $0.1m^2$. Overall 273 species were recorded in the area, 240 using the large grabs, and 198 using the small grabs. Ten grabs were collected for each grab size. I hoped that estimators would be able to give approximately the same species richness estimate regardless of the grab size used to collect the data, by accounting for the distribution pattern of individuals on the sea bed.

Tables 3.21 and 3.22 clearly show the boundary problem was present when using the negative binomial model and the Neyman Type A-gamma model, which reinforces the requirement for a penalised MLE. The profile log-likelihood of the MLE using the negative binomial model was very flat, and showed no upper limit to the 95% confidence interval for this data set and model (Figure 3.12a). The same can be seen for this model applied to the Isle of Wight $0.1m^2$ data (Figure 3.12b).




*(a)* $0.25m^2$ grabs

*(b)* $0.1m^2$ grabs

*Figure 3.12:* Profile log-likelihood for the number of unseen species using the negative binomial MLE applied to the Isle of Wight data sets collected with (a) large grabs and (b) small grabs.

| Model | $\hat{N}$ | 95% Confidence | $\chi^2$ (df) | 5%$\chi^2$point | AIC | $\Delta$AIC |
|---|---|---|---|---|---|---|
| PO | 241 | (241, 241) | 3571.98 (17) | 27.59 | 5357.0 | 5937.5 |
| PO P2 | 241 | (241, 241) | 3571.98 (17) | 27.59 | 5340.0 | 5920.5 |
| PO P3 | 241 | (241, 241) | 3571.98 (17) | 27.59 | 5357.0 | 5937.5 |
| PO P3 IT | 241 | (241, 241) | 3571.98 (17) | 27.59 | 5357.0 | 5937.5 |
| NB | 3263 | (5500, $\infty$ ) | 20.46 (22) | 33.92 | -580.5 | 0.0 |
| NB P2 | 895 | (459, 2859) | 24.36 (23) | 35.17 | -578.8 | 1.7 |
| NB P3 | 496 | (388, 650) | 34.20 (24) | 36.42 | -571.2 | 9.3 |
| NB P3 IT | 720 | (469, 1116) | 29.78 (23) | 35.17 | -577.9 | 2.6 |
| NTAG | 13,816 | (1234, 54,535) | 46.40 (23) | 35.17 | -553.5 | 27.0 |
| NTAG P2 0.5 | 4206 | (904, 16,103) | 50.47 (23) | 35.17 | -550.1 | 30.4 |
| NTAG P2 0.25 | 6818 | (1026, 26,541) | 47.12 (23) | 35.17 | -551.7 | 28.8 |
| NTAG P3 | 561 | (441, 719) | 79.85 (25) | 37.65 | -533.1 | 47.4 |
| NTAG P3 IT | 936 | (616, 1380) | 61.04 (24) | 36.42 | -545.8 | 34.7 |

*Table 3.21:* Species richness estimates with 95% confidence interval for Isle of Wight benthic data set with $0.25m^2$ grabs, using the Poisson, PO, negative binomial, NB, and Neyman Type A-gamma, NTAG, maximum-likelihood estimator, with penalty 2, P2, penalty 3, P3 and penalty 3 with one-step iteration, P3 IT. The Pearson chi-squared value, degrees of freedom, df, 5%$\chi^2$ point, AIC and $\Delta$AIC are also reported.

| Model | $\hat{N}$ | 95% Confidence | $\chi^2$ (df) | 5%$\chi^2$point | AIC | $\Delta$AIC |
|---|---|---|---|---|---|---|
| PO | 198 | (198, 198) | 12351.26 (3) | 7.81 | 17789.0 | 18288.8 |
| PO P2 | 198 | (198, 198) | 12349.22 (3) | 7.81 | 17790.0 | 18289.8 |
| PO P3 | 198 | (198, 198) | 12320.67 (3) | 7.81 | 17839.0 | 18338.8 |
| PO P3 IT | 198 | (198, 198) | 12351.26 (3) | 7.81 | 17789.0 | 18288.8 |
| NB | 19,906 | (1021,$\infty$) | 27.11 (18) | 28.87 | -499.8 | 0.0 |
| NB P2 | 1862 | (595, 6854) | 26.05 (18) | 28.87 | -496.2 | 3.6 |
| NB P3 | 469 | (363, 611) | 34.71 (20) | 31.41 | -481.2 | 18.6 |
| NB P3 IT | 767 | (498, 1143) | 29.17 (19) | 30.14 | -492.3 | 7.5 |
| NTAG | 17,009 | (1704, 67,444) | 51.20 (19) | 30.14 | -469.8 | 30.0 |
| NTAG P2 0.5 | 6344 | (1198, 24,782) | 51.93 (19) | 30.14 | -465.9 | 33.9 |
| NTAG P2 0.25 | 9859 | (1393, 38,820) | 51.48 (19) | 30.14 | -467.7 | 32.1 |
| NTAG P3 | 515 | (404, 660) | 86.62 (20) | 31.41 | -441.3 | 58.5 |
| NTAG P3 IT | 920 | (615, 1326) | 67.52 (20) | 31.41 | -458.4 | 41.4 |

*Table 3.22:* Species richness estimates with 95% confidence interval for Isle of Wight benthic data set with $0.1m^2$ grabs, using the Poisson, PO, negative binomial, NB, and Neyman Type A-gamma, NTAG, maximum-likelihood estimator, with penalty 2, P2, penalty 3, P3 and penalty 3 with one-step iteration, P3 IT. The Pearson chi-squared value, degrees of freedom, df, 5%$\chi^2$ point, AIC and $\Delta$AIC are also reported.

The Poisson model was clearly inadequate for modelling these data, as shown by the high AIC and chi-squared values. We know that the estimates given for the Isle of Wight area by the Poisson model were too low, as overall between the two grab sizes, the total number of observed species in the area was 273. Three models were plausible for the data according to the AIC, the negative binomial, and the negative binomial with penalty 2 and one-step iterated penalty 3.

Using penalties decreased the estimates of species richness considerably, and penalty 3 was much harsher than the others. The iterated penalty performed well in terms of combatting the boundary problem and not penalising the estimate as much as penalty 3 in both cases, and gave a reasonably similar estimate of species richness from the two data sets using the negative binomial model and Neyman Type A-gamma model.

The estimates given by the penalised log-likelihood estimator using the negative binomial model and penalty 2 were not very similar for the two data sets. I hoped that an estimator would be able to provide a similar estimate of species richness using the two data sets. The choice of $\gamma_2 = 0.5$ played a part in these estimates, and I saw in Section 3.9.3 through simulations that a penalty parameter between 0.25 and 1 performed well, and the choice of $\gamma_2 = 0.5$ was fairly effective. However, the best choice of penalty parameter did vary with sampling depth.

For the Neyman Type A-gamma model, the estimates of species richness were higher than with the other models. The fit of the Neyman Type A-gamma models were rejected by the chi-squared test, and the $\Delta$AIC suggested that the models were not plausible to describe this data. Using the Neyman Type A-gamma model, the results of the MLE with penalty 2 showed very wide confidence intervals and large estimates, and suggested that not enough penalty was applied.

### 3.10.4 Eastern Channel benthic data set

In this data set the sample size is much larger, as 225 grabs were collected and 649 species observed, and therefore I expected the estimators to perform better for this data set. Table 3.23 shows a large estimate of species richness for the Eastern Channel data using the negative binomial model, with infinite upper confidence limit. However, the fit of the negative binomial model to the observed data without using a penalty was good (Figure 3.13 ) and this model was not rejected by a Pearson goodness-of-fit test and it was selected as the best model by the AIC. However, Figure 3.14 shows that the profile log-likelihood was flat and the species richness estimate was not well defined. Therefore, we must exercise caution in relying on chi-squared tests and AIC to determine the most suitable model for our data.

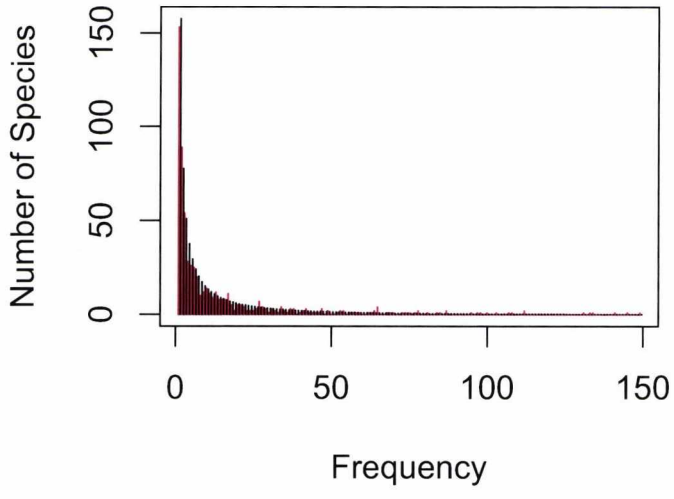



*Figure 3.13:* Observed data (red) and fitted data (black) for the negative binomial MLE fitted to the Eastern Channel data set.

Applying penalty 2 with $\gamma_2 = 0.5$ still resulted in a flat profile log-likelihood for $N$, as seen by the large confidence interval, and so a harsher penalty was required to avoid the spuriously large estimates associated with the boundary problem. The

| Model | $\hat{N}$ | 95% Confidence | $\chi^2$ (df) | 5%$\chi^2$point | AIC | $\Delta$AIC |
|---|---|---|---|---|---|---|
| PO | 650 | (650, 650) | 16009.81 (18) | 28.87 | 10278 | 13321 |
| PO P2 | 650 | (650, 650) | 16009.81 (18) | 28.87 | 10264 | 13307 |
| PO P3 | 650 | (650, 650) | 16009.81 (18) | 28.87 | 10278 | 13321 |
| PO P3 IT | 650 | (650, 650) | 16009.81 (18) | 28.87 | 10278 | 13321 |
| NB | 63,956 | (3691, $\infty$) | 31.53 (38) | 53.38 | -3043 | 0 |
| NB P2 | 11,143 | (2787, 42,607) | 32.45 (38) | 53.38 | -3040 | 3 |
| NB P3 | 1686 | (1382, 2075) | 51.56 (40) | 55.76 | -3017 | 26 |
| NB P3 IT | 2883 | (1981, 4117) | 37.64 (39) | 54.57 | -3035 | 8 |
| NTAG | 44,145 | (8845, 174,629) | 99.50 (37) | 52.19 | -2970 | 73 |
| NTAG P2 0.5 | 28,740 | (6958, 112,527) | 99.64 (37) | 52.19 | -2968 | 75 |
| NTAG P2 0.25 | 34,890 | (7781, 137,560) | 99.32 (37) | 52.19 | -2970 | 73 |
| NTAG P3 | 1961 | (1629, 2360) | 143.20 (39) | 54.57 | -2921 | 122 |
| NTAG P3 IT | 3900 | (2799, 5277) | 113.14 (38) | 53.38 | -2955 | 88 |

*Table 3.23:* Species richness estimates with 95% confidence interval for Eastern Channel benthic data set, using the Poisson, PO, negative binomial, NB, and Neyman Type A-gamma, NTAG, maximum-likelihood estimator, with penalty 2, P2, penalty 3, P3 and penalty 3 with one-step iteration, P3 IT. The Pearson chi-squared value, degrees of freedom, df, 5%$\chi^2$ point, AIC and $\Delta$AIC are also reported.

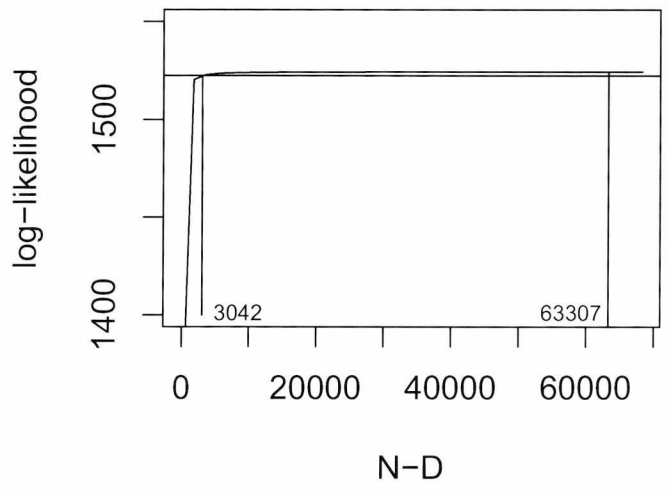



*Figure 3.14:* Profile log-likelihood for the number of unseen species for the negative binomial MLE fitted to the Eastern Channel data.

harsher penalty of penalty 3 decreased the estimate to a reasonable value. Using the one-step iterative penalty improved the fit of the model to the data according to the $\Delta$AIC. Applying the iterative step also widened the associated confidence interval, however it was still within the limit of number of known species in UK coastal waters of 10,000. As the Eastern Channel data was collected over a very large area, it was plausible that many of the species present around the UK coast could be present in this area.

Again the Poisson model performed poorly, with estimates not much larger than the observed number of species, and very high AIC in all cases.

The Neyman Type A-gamma model also performed poorly according to the chi-squared values and AIC, which was disappointing. We would hope that by modelling spatial clustering we would improve the fit of the model, although this was fixed across species, so perhaps by allowing the number of individuals per cluster to vary between species we can describe the data better.

## 3.11 Discussion

This chapter presented a maximum-likelihood approach to species richness estimation which incorporated the spatial pattern of the species' members using the Neyman Type A distribution to model the spatial pattern of individuals within a species. Chapter 2 described a range of species richness estimators that could be used to analyse sample data, however it was shown that these methods, which did not account for spatial clustering, were not adequate for analysing benthic data. Therefore I hoped that by using the Neyman Type A I would account for this clustering and improve the species richness estimates.

Three models were investigated: the Poisson, the negative binomial, and the Neyman Type A-gamma. A maximum-likelihood approach was taken to fit the models to the data, and obtain a species richness estimate. Fewster and Jupp (2009) showed that the difference between using the conditional and unconditional estimators is of order 1, and I used both methods during my investigation into the approach.

Simulation studies were carried out and it was necessary to consider approaches to estimate confidence intervals for the MLE, and several methods were investigated. As far as I was aware, a comparison of the coverage of these particular confidence interval methods had not been published before.

Although the Horvitz-Thompson confidence intervals performed well, profile confidence intervals performed equally well in terms of coverage. van der Heijden et al. (2003) suggested the use of confidence intervals which can be asymmetric, and hence the consideration of the confidence intervals for $\log(N)$ suggested by Fewster and Jupp (2009). However, these were found to be narrow and therefore coverage was lower than expected, especially when detection rates were low. This was in contrast to the results of Fewster and Jupp (2009), who found the method performed well in the examples they analysed. However, detection rates in the examples they looked at were higher than those considered in my simulations. This warrants further investigation,

but since the other methods performed well I did not pursue this further.

The bootstrap confidence intervals were very wide in comparison to the other methods, and this may have been attributed to the boundary problem when sampling depth was low. van der Heijden et al. (2003) also found the bootstrap confidence intervals performed poorly in comparison to their Horvitz-Thompson variance estimators.

When considering coverage of the confidence intervals for the negative binomial case, spuriously large species richness estimates arose due to the boundary problem. The profile confidence interval method had good coverage probabilities and these were used for confidence interval estimation when applying the maximum-likelihood estimator. I would recommend the use of profile confidence intervals, as they had the advantage that they could be asymmetric. When calculating profile confidence intervals for $N$ it was simple to use the full log-likelihood, and therefore this method was used rather than the conditional MLE.

Evidence of the boundary problem was found at various sampling depths, dependant on the model used and the parameters of that model. For the Poisson model fitted to simulated Poisson data there was no significant evidence of the boundary problem occurring when sampling depth was more than 10%. However, for the negative binomial case, when sampling depth fell below 50% there was evidence of the boundary problem. For the Neyman Type A-gamma model there were signs of the boundary problem at higher sampling depths, and when sampling depth was low some extreme species richness estimates occurred.

Several possible solutions have been considered to combat the boundary problem, and I opted to use penalties, and considered those suggested by Wang and Lindsay (2005). In simulations, penalty 2, based on the odds function, worked well for particular penalty parameter values, however it relied on the specification of a penalty

parameter. Simulations showed that a value of 0.5 worked well in the negative binomial case, however as the parameters changed the estimator became more sensitive to the penalty parameter chosen. When I considered the Neyman Type A-gamma model I noticed more sensitivity in the estimator to the value of the penalty parameter chosen.

The adaptive penalty 3 was also considered, which used the the $Chao_1$ species richness estimate as a naive estimate and penalised the estimate towards it. Chapter 2 showed that the $Chao_1$ estimator can severely underestimate the true species richness, and therefore introduce too much bias to the MLE when using penalty 3. Evidence of the over-penalisation was seen when this MLE estimator was applied to simulated data. Therefore I would exercise caution when using this penalty on the MLE applied to clustered data.

To combat the negative bias of penalty 3, I considered using an iterative penalty. Due to computational constraints, I opted for a one-step iteration and using this penalty the MLE performed much better than using penalty 3 in simulations. Using the iterative step meant that the true value of the number of species always fell within the 95% central interval when using the negative binomial model in simulations, whereas it had not using penalty 3 alone. However, for Neyman Type A-gamma model I still saw a significant negative bias in the estimates at a low sampling depth using this penalty, where data were simulated from a clustered population.

In reducing the bias of the MLE using penalties, I did increase the variance of the estimate and the width of the confidence interval. Ideally we would reduce both variance and bias, and it is important to consider which is preferred during application of the estimators. In our case, I would argue that a species richness estimate with less bias would be preferable when applied to benthic data sets. Alternatively, we could use the mean square error, which incorporates both the variance and bias of the estimator.

When applying the MLE to the real data sets, I found that the species richness estimates varied substantially between the models. According to the AIC, the negative binomial model gave the best fit for all the data sets. However, the Neyman Type A-gamma model was plausible for the Lepidoptera data and the CBC data and provided an acceptable fit according to the chi-squared values. This was encouraging, and suggested that modelling the spatial clustering of individuals within species could be beneficial when estimating species richness.

The chi-squared values and AIC for the benthic data sets suggested that the negative binomial model with no penalty fitted the data well, and was the best model, but the species richness estimates obtained in these cases were very large. This suggested that we can not always rely on these measures to select the most appropriate model. Chao (2004) highlights a general problem with parametric methods

> *'A model which gives a good fit to the data does not necessarily result in a satisfactory species richness estimate.'*

We have seen this illustrated in some of the species richness estimates we obtained, as there have only been approximately 10,000 benthic species recorded in UK coastal waters (personal communication, Keith Cooper) so we would not expect the true species richness to be as high as the estimate of 63,956 that we obtained for the Eastern Channel using the negative binomial model.

Using a penalised MLE on benthic data had an effect on the species richness estimates obtained, and decreased the estimates of species richness to something more realistic. Penalty 2 was not always harsh enough to combat the flatness of the log-likelihood with the penalty parameters of $\gamma_2 = 0.5$ and $\gamma_2 = 0.25$ I used. Wang and Lindsay (2005) stated that the optimal value for the penalty parameter of penalty 2 depended on the value of the odds function, and our estimates showed that for benthic data, where we have not observed a high proportion of the species, the penalty parameter we used was important. A method to select an appropriate value for this parameter

would be very useful, and should improve estimation, and this is an area for future investigation. One link considered by Wang and Lindsay (2005) was that penalties may be considered as priors in a Bayesian framework, and I consider this in Chapter 6.

In contrast, penalty 3 was much harsher. For the Isle of Wight data set both penalty 3 and the penalised estimator using one-step iteration gave reasonably similar estimates of species richness from the two grab sizes using both the negative binomial and Neyman Type A-gamma models, suggesting that this may be working well. The one-step iterative penalty performed well in both application to the benthic data sets and to the simulated data, and I would suggest the use of this penalty over penalty 2 while there is no method to select the best penalty parameter $\gamma_2$, as otherwise we may not avoid the spuriously large estimates associated with the boundary problem. If the sampling depth was known we could be more confident in the estimates gained using penalty 2, however in practice knowledge of the actual sampling depth is highly unlikely.

To check the suitability of using the penalised likelihood when the boundary problem was not present, I considered the estimates given by the penalised MLE for the Lepidoptera data. A suitable penalty would only penalise the likelihood where necessary, and the estimated number of species was not affected much by the use of penalties. Therefore I can conclude that where the boundary problem is not present, use of the penalised MLE will not greatly effect the species richness estimate.

The results showed the inadequacy of the Poisson model in describing benthic data, with estimates not much larger than the observed number of species, and very high AIC in all cases. When applying the MLE I allowed the abundance to vary between species, but not the spatial clustering. I suspect that this could be the cause of the lack of fit of the Neyman Type A-gamma given by the chi-squared values, because in reality species are likely to exhibit differing spatial clustering patterns. When I applied both the negative binomial and Neyman Type A-gamma models to data simulated

from Neyman Type A where both abundance and spatial clustering varied between species, the Neyman Type A-gamma model fitted the data better. This reassures me that the Neyman Type A-gamma model is more suitable for use in estimating species richness for spatially clustered data than the negative binomial.

To build the extra variation into the model involves computation of a double integral, and the computation time of fitting this model compared with the negative binomial and simple Neyman Type A-gamma is greatly magnified. Currently the Neyman Type A-gamma takes 8 hours to fit, when profile confidence intervals are to be calculated, whereas the negative binomial takes less than ten minutes. Adding in the extra parameter allowing clustering to vary between species greatly increased the number of integrations to be carried out within the optimisation, and takes sixteen days to complete. It is possible that the optimisation gets stuck in a flat area of the likelihood, and future work will look at ways to avoid this, and other possibilities to decrease computation time.

Another possible area of future research is to consider alternative distributions to describe the spatial clustering of individuals within benthic species. Neyman and Scott (1958) suggested three contagious distributions, so perhaps the Neyman Type B or C models could be more appropriate. I wished to consider a distribution that was a simplification of the Matérn process introduced in Chapter 2; however an alternative clustering distribution not based on parents and daughters may be more suitable.

## 3.12 Conclusions

In this chapter I extended the maximum-likelihood approach from the negative binomial model to the Neyman Type A-gamma distribution in an attempt to incorporate the spatial pattern of the species' members into species richness estimation. I have seen through simulations that this model does perform well for clustered data, and also when applied to real data sets.

I considered several approaches to estimate confidence intervals for the MLE, and found that confidence intervals based on profile log-likelihoods perform well in terms of coverage. Therefore I used this method in estimating confidence intervals for the MLE and penalised MLE.

To improve species richness estimates for benthic data, further work will include an extension to allow clustering intensity to vary between species, and I would expect that then the model fit may improve. Also, further work on investigating alternative clustering distributions to describe the spatial patterns of benthic organisms could be a valuable contribution to the field of species richness research.

I found that the boundary problem caused spuriously large species richness estimates, and attempted to combat this by penalising the log-likelihood. However, some of the penalties used were subjective, and an alternative to this is to use priors within a Bayesian framework to avoid spuriously large estimates. Using the Bayesian approach can also remove the need to compute the marginal likelihood during optimisation, and this should reduce the computation time of the estimate. The Bayesian approach to species richness estimation is considered in Chapter 4.

# 4. SPECIES RICHNESS ESTIMATION - A BAYESIAN APPROACH

## *4.1 Introduction*

In Chapters 2 and 3 I explored non-parametric and parametric frequentist approaches to estimating species richness. This chapter considers parametric Bayesian approaches and the aim is to explore these methods for species richness estimation, and extend them for the Neyman Type A-gamma model.

I first introduce both parametric and non-parametric approaches to Bayesian species richness estimation from the literature. I then proceed with a parametric Bayesian approach. First I describe the methodology used, introduce the priors investigated and the models used, and then provide species richness estimates for the Lepidoptera data set and the CBC data set. The chapter concludes with a discussion of the Bayesian approach.

One standard consideration in the use of the Bayesian approach is the requirement to specify an appropriate prior distribution for the parameters. In many applications, including ecological, the prior is a convenient way to incorporate expert opinion or information from previous or related studies.

There are two situations that could occur (King et al., 2010, p76):

- there is no prior information;
- there is prior information which needs to be expressed in the form of a suitable probability distribution.

In this chapter I investigate the use of non-informative (objective) priors. Objective priors are used in cases where prior information is unavailable or controversial, because an objective Bayesian procedure may be regarded as a default method which can be applied in cases where prior information is sparse or not well understood, or differs between the stakeholders (Sweeting, 2001).

Chapter 5 considers the Bayesian approach to estimating species richness using priors that have been informed by expert opinion.

## 4.2 Bayesian species richness estimation methods

There are several approaches that one could take to tackle the species richness problem from a Bayesian viewpoint. These can be categorised into parametric and non-parametric approaches, and a brief review of some of the methods developed in each approach is given below.

### 4.2.1 Parametric Bayesian approach

In the parametric Bayesian approach a prior is placed on the number of species, $N$, and on the parameters of the abundance distribution. Several models and priors have been suggested for the species richness estimation problem. Barger and Bunge (2008) and Barger and Bunge (2010) prefered objective priors, and presented Jeffrey's prior, $p(N) \propto N^{\frac{m-1}{2}}$ (where $m$ is the dimension of the nuisance parameter), and the reference prior, $p(N) \propto N^{-\frac{1}{2}}$, for several abundance models including the Poisson and negative binomial.

Others used informative priors, including Poisson and negative binomial priors for the number of species again applied with various abundance models (Madigan and York, 1997; Rodrigues et al., 2001; Wang et al., 2007). Wang et al. (2007) considered several priors of the form $p(N) \propto 1/N^c$, where $c$ is a non-negative constant, which include the improper uniform prior when $c = 0$. Rodrigues et al. (2001) made a comparison between the hierarchical Bayesian approach and the empirical Bayes approach, both

of which are discussed in more detail later in this chapter.

Advantages of the parametric Bayesian approach highlighted by Barger and Bunge (2010) are that smoothing the abundance data stabilises the estimate of the number of unobserved species, that the point and interval estimates will be greater than the number of species seen, and that credible intervals can be asymmetric.

However, problems with the approach do exist, and include the justification of the use of specific priors, and model selection as in the frequentist approach. In addition, there is the issue of truncating the data before performing the analysis, which some do (Barger and Bunge, 2010), while others do not (Quince et al., 2008).

### 4.2.2 *Non-parametric Bayesian approach*

Bayesian non-parametric inference is a relatively recent area of research (Lijoi et al., 2007), and the use of these methods has increased more recently, for example in statistical machine learning (Blei et al., 2010).

The non-parametric Bayesian approach requires the specification of a prior to assign probability distributions to function spaces (Barger and Bunge, 2010). Random discrete probability measures, such as the commonly used Dirichlet process, are used as priors within hierarchical mixture models for density estimation (Favaro et al., 2011). The class of Dirichlet process priors was first presented by Ferguson (1973) when interest in Bayesian non-parametric inference was in its infancy, and alternative ways to define a Dirichlet prior and to establish its properties were given in Blackwell and MacQueen (1973) and Blackwell (1973).

An alternative non-parametric approach to the species richness problem presented by Lijoi et al. (2007) looked at the probability of discovering new species during further sampling. Additional work in this area was detailed in Favaro et al. (2011), who studied the limiting behaviour of the number of new species to be observed in

a new sample. These approaches use a general class of discrete random probability measures, termed species sampling models, which were introduced by Pitman (1996) and discussed in Favaro et al. (2009). However, this addresses a slightly different question to the one I wish to answer.

## 4.3 Parametric Bayesian methods

### 4.3.1 Bayesian inference

In the Bayesian approach we specify a model for some observed data, $y = y_1, ..., y_i$, given a vector of unknown parameters, $\theta$, as $f(y|\theta)$, and suppose that $\theta$ is a random quantity with a prior distribution $p(\theta)$.

If we express the joint probability distribution for $y$ and $\theta$ as $p(y, \theta) = p(\theta)f(y|\theta)$, we base inference for $\theta$ on its posterior distribution:

$$\pi(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(\theta)f(y|\theta)}{p(y)}, \tag{4.1}$$

where

$$p(y) = \int p(y, \theta)\mathrm{d}\theta = \int p(\theta)f(y|\theta)\mathrm{d}\theta, \tag{4.2}$$

or the sum over all possible values of $\theta$ when $\theta$ is discrete. This may be expressed as

$$\pi(\theta|y) \propto f(y|\theta)p(\theta), \tag{4.3}$$

the likelihood times the prior if the factor $p(y)$, which does not depend on $\theta$ for fixed $y$, is considered as a constant (Gelman et al., 2004, p 8).

Figure 4.1 shows an example of a directed acyclic graph (DAG) for this approach, where the data $x_i \sim \text{Poisson}(\lambda_i)$ and $\lambda_i \sim \text{Gamma}(\alpha, \beta)$.

### 4.3.2 Hierarchical Bayes

In the hierarchical Bayes (HB) approach, the parameters of the priors are assumed to be unknown, and are themselves given 'hyperpriors' (George et al., 1993). This

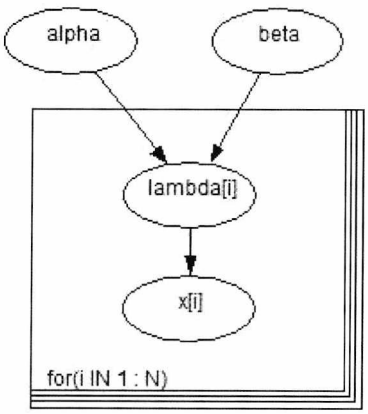



*Figure 4.1:* Directed acyclic graph of the Poisson-gamma model where the data $x_i \sim \text{Poisson}(\lambda_i)$ and $\lambda_i \sim \text{Gamma}(\alpha, \beta)$.

essentially incorporates random effects in the model parameters, and reduces the influence of prior assumptions on the posterior.

Therefore, we suppose that in our model, $f(y|\theta)$, $\theta$ has prior distribution $p(\theta|\nu)$ and the hyperparameter $\nu$ is unknown. Then $\nu$ has a prior distribution, p($\nu$). The joint prior distribution is

$$p(\nu, \theta) = p(\nu)p(\theta|\nu) \tag{4.4}$$

and the corresponding joint posterior is

$$\begin{aligned} \pi(\theta, \nu|y) &\propto f(y|\nu, \theta)p(\nu, \theta) \\ &= f(y|\theta)p(\nu, \theta). \end{aligned}$$

This holds as the hyperparameters $\nu$ only affect $y$ through $\theta$ (Gelman et al., 2004, p124).

Figure 4.2 shows an example of a DAG for this approach, where the data $x_i \sim \text{Poisson}(\lambda_i)$ and $\lambda_i \sim \text{Gamma}(\alpha, \beta)$ but $\alpha$ and $\beta$ also have priors.

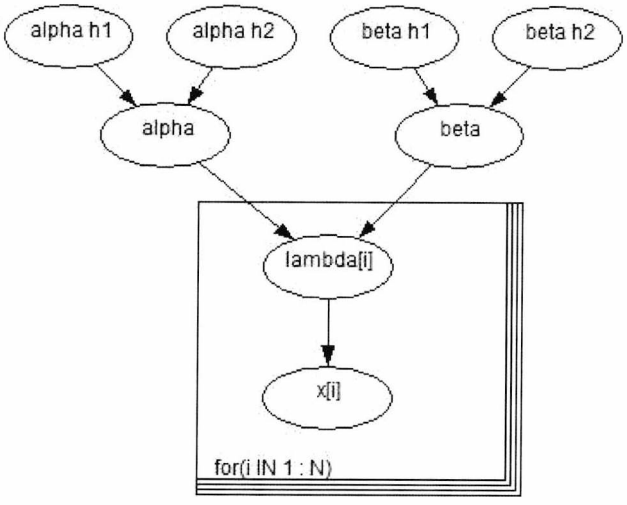



*Figure 4.2:* Directed acyclic graph of the Poisson-gamma model where the data $x_i \sim \text{Poisson}(\lambda_i)$ and $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, and $\alpha$ and $\beta$ have hyperpriors with parameters $\alpha_{h1}$ and $\alpha_{h2}$, and $\beta_{h1}$ and $\beta_{h2}$ respectively.

### 4.3.3 Markov chain Monte Carlo

To estimate the posterior distribution, I use Markov chain Monte Carlo (MCMC) methods. There are two components to MCMC; Monte Carlo integration, and Markov chains. Monte Carlo integration allows us to obtain an estimate of a given integral which is too complex to evaluate explicitly.

For example, given a sample of observations, $\theta^1, ..., \theta^n$ from the posterior distribution, we can estimate the expectation of a function $g(.)$ of parameter $\theta$ given observed data $\boldsymbol{x}$

$$\mathbb{E}_\pi[g(\theta)] = \int g(\theta)\pi(\theta|\boldsymbol{x})\mathrm{d}\theta \tag{4.5}$$

by the average

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(\theta^i) \tag{4.6}$$

(Morgan, 1984, p. 163), which, for independent samples, tends to $\mathbb{E}_\pi[g(\theta)]$ as $n$ tends to infinity, by the Law of Large Numbers.

When we are unable to sample the posterior directly, we are able to generate these samples via the use of a Markov chain; a discrete time stochastic process where each value in the sequence depends only on the last (Roberts, 1996, p. 45). The Markov chain is successful because at each step in the chain the distribution of the samples gets closer to the posterior. This approach is often the easiest way to get reliable results (Gelman et al., 2004, p. 287).

To simulate a Markov chain we generate the new state of the chain, say $\theta^{k+1}$, from some density dependent on $\theta^k$ such that

$$\theta^{k+1} \sim K(\theta^k, \theta),$$

where $K$ is the transition kernel for the chain (King et al., 2010, p. 101), which represents the probability distribution of moving from $\theta^k$ to a point in the target distribution.

The aim is to create a Markov process whose stationary distribution is the specified target distribution. For the distribution of the values of the Markov chain to converge to a stationary distribution, it must be irreducible, aperiodic and positive recurrent (Appendix B). Our chain will be positive recurrent if we can show it is irreducible, and a Markov chain is irreducible if we can get from any state $\theta^k$ to any other. For an aperiodic, positive recurrent Markov chain, according to the Ergodic Theorem (Appendix C), the target distribution is the limiting distribution of the chain

$$\bar{g}_n \to \mathbb{E}_\pi[g(\theta)] \text{ as } n \to \infty$$

(Roberts, 1996, p. 47).

As we update the parameters in the Markov chain, the probability distribution associated with the $k^{th}$ observation gets closer and closer to $\pi(\theta|\boldsymbol{x})$ as $k$ increases. Therefore, if the chain is run for long enough, the distribution of the chain will converge to the posterior distribution of interest. We must discard any observations

from before the chain converges, and we refer to this initial period of the chain as the burn-in.

### 4.3.4 Metropolis-Hastings algorithm

To construct a Markov chain we can use a form of rejection sampling, whereby we draw samples from a candidate distribution conditional on the last observation and accept the move with some probability. This forms the transition kernel for the chain. This method was developed by Metropolis et al. (1953) and generalised in Hastings (1970).

The proposal distribution, $q(\phi|\theta^k)$, from which we generate these candidates typically depends on the current state of the chain, plus some noise. We can chose any sensible distribution for $q$, such as $\phi|\theta \sim N(\theta, \sigma^2)$, where $\sigma^2$ is to be specified; which is easily sampled and symmetric.

We introduce an acceptance function, and accept the candidate observation and set $\theta^{k+1} = \phi$ with probability $\alpha(\theta^k, \phi)$; otherwise if the candidate is rejected the chain remains at $\theta^k$, and $\theta^{k+1} = \theta^k$.

The optimal form for the acceptance function, in terms of not rejecting candidate values too frequently, is given by

$$\alpha(\theta^k, \phi) = \min\left(1, \frac{\pi(\phi|\boldsymbol{x})q(\theta^k|\phi)}{\pi(\theta^k|\boldsymbol{x})q(\phi|\theta^k)}\right). \tag{4.7}$$

which is determined by requiring that the reversibility condition

$$\pi(\theta^k|\boldsymbol{x})q(\phi|\theta^k)\alpha(\theta^k, \phi) = \pi(\phi|\boldsymbol{x})q(\theta^k|\phi)\alpha(\phi, \theta^k) \tag{4.8}$$

is met (Peskun, 1973). This ensures that the Metropolis-Hastings kernel has $\pi(\theta|\boldsymbol{x})$ as its invariant density (Chib and Greenberg, 1995).

Thus the algorithm for the Metropolis-Hastings (MH) method is:

1. Draw $\phi \sim q(\phi|\theta^k)$

2. Set $\theta^{k+1} = \phi$ with probability $\alpha(\theta^k, \phi)$, otherwise set $\theta^{k+1} = \theta^k$.

3. Repeat steps 1 and 2.

If we specify a proposal distribution that is symmetric, such as a normal distribution, the proposal densities cancel in the acceptance function (Equation 4.7). We are left with

$$\alpha(\theta^k, \phi) = \min\left(1, \frac{\pi(\phi|\boldsymbol{x})}{\pi(\theta^k|\boldsymbol{x})}\right), \tag{4.9}$$

which is the Metropolis algorithm.

One construct of the MH algorithm is the single update MH algorithm; where MH is used in stages, updating each parameter one at a time. Gibbs sampling can be viewed as a special case of the single update MH algorithm, in which the proposal distribution for any parameter is set as the conditional posterior distribution of that parameter given the current value of the others (Gelman, 1996, p328). Therefore the acceptance probability will always equal one. Gibbs sampling is especially useful for conjugate models, that is where one can directly sample from the conditional posterior distribution as it will follow a known parametric form (Gelman, 1996, p 40).

### *4.3.5 Block updates*

If parameters are correlated, using block updates can be beneficial to reduce computation time and improving mixing (exploration of the parameter space). In this case correlated parameters can be updated in a single MH step. This requires a multivariate proposal distribution for these parameters.

To construct a covariance matrix for a multivariate normal proposal distribution, I can perform pilot tuning and obtain posterior standard deviations and correlations for the parameters of the model using the single-update MH algorithm; these are used

in the construction of a suitable covariance matrix.

We simulate a set of candidate values for the parameters, $\boldsymbol{\omega}$, from a multivariate normal proposal distribution given by

$$\boldsymbol{\omega}|\theta^k \sim \mathcal{N}_i \left( \begin{pmatrix} \theta_1^k \\ . \\ . \\ \theta_i^k \end{pmatrix} \begin{pmatrix} \sigma^2_{\theta_1} & \sigma_{\theta_1\theta_2} & \dots & \sigma_{\theta_1\theta_i} \\ . & . & . & . \\ . & . & . & . \\ \sigma_{\theta_1\theta_i} & \dots & \dots & \sigma^2_{\theta_i} \end{pmatrix} \right) \tag{4.10}$$

where $\theta^k$ are the parameter values at iteration $k$ of the Markov chain, $\sigma^2_\theta$ is the proposal variance of $\boldsymbol{\omega}$, and $\sigma_{\theta_i\theta_j}$ is the covariance of $\theta_i$ and $\theta_j$.

Following (King et al., 2010, p134), I allow a slightly larger covariance between parameters when performing the MCMC than given by the single-update pilot tuning, to explore the parameter space sufficiently.

The acceptance probability of the MH algorithm is given by

$$\alpha(\theta^k, \boldsymbol{\omega}) = \min(1, A), \tag{4.11}$$

where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\omega}|\boldsymbol{x})q(\theta^k|\boldsymbol{\omega})}{\pi(\theta^k|\boldsymbol{x})q(\boldsymbol{\omega}|\theta^k)} \\ &= \frac{f(\boldsymbol{x}|\boldsymbol{\omega})p(\boldsymbol{\omega})}{f(\boldsymbol{x}|\theta^k)p(\theta^k)} \end{aligned}$$

since the proposal distribution is symmetric and cancels in the acceptance probability.

### 4.3.6 *Performing MCMC within* BUGS *and* R

The BUGS (Bayesian inference Using Gibbs Sampling) software has been developed to perform Bayesian analysis of complex statistical models using MCMC methods (Lunn et al., 2000). It is a useful tool which allows one to perform MCMC simply by specifying a model and the data to be analysed.

WinBUGS (designed to run BUGS in Windows), is a user-friendly interface which not only performs MCMC, but also incorporates several of the diagnostic tools that I will mention in this chapter. As suggested by the name, BUGS uses Gibbs sampling, however alternative methods, such as the Metropolis-Hastings algorithm, are used for difficult full conditional distributions.

Several common distributions are built into WinBUGS, and those that are not can be implemented by expressing the negative log-likelihood, $-l$, and using the 'zeros trick' which operates using the following code:

```
for (i in 1:N) {
     zeros[i] <- 0
     phi[i] <- -log(l[i])
     zeros[i] ~ dpois(phi[i])
}
```

Here, the observed data `zeros` is a vector of zeros, and a Poisson($\phi$) observation of zero has likelihood $\exp(-\phi)$. As `phi[i]` is set to `-log(l[i])`, we will obtain the correct likelihood contribution for our specified distribution (Lunn et al., 2000).

However, for more flexibility and control one can use bespoke code in R to run MCMC. One can also run WinBUGS from within R, and use established R functions to analyse the MCMC output.

## 4.4 Summarising the posterior distribution

We use statistics to summarise the posterior distribution of $\theta$ and obtain a point estimate from the posterior by selecting one of these summary features, such as its mean, median or mode.

When the prior is flat the mode will be equal to the maximum-likelihood estimate, however the mean is commonly used as it minimises the posterior variance with

respect to the point estimate of the posterior (Carlin and Louis, 2000, p. 34). I report the mean and the median, which is preferred for asymmetric posteriors as it is intermediate to the mode and the mean.

To determine the accuracy of a point estimate, the standard deviation of the posterior with respect to that estimate is reported. However, this gives no information on the skewness of the distribution, and so we describe the spread of the distribution through credible intervals (King et al., 2010, p. 86).

The interval $(a, b)$ is defined as a $100(1-\alpha)\%$ credible interval for $\theta$ if

$$1 - \alpha \leq P(\theta \in [a, b]|\boldsymbol{x}) = \int_a^b p(\theta|\boldsymbol{x})\mathrm{d}\theta, \tag{4.12}$$

where integration is replaced by summation for discrete parameters (Carlin and Louis, 2000, p. 35).

A $100(1-\alpha)\%$ credible interval is an exact interval for which the probability that $\theta$ lies in the interval $(a, b)$ given the observed data is at least $(1-\alpha)$ (Carlin and Louis, 2000, p. 36). This is in contrast to the confidence interval of the frequentist approach, which is the confidence that if a trial were repeated several times, $\theta$ would be within the interval.

The $100(1-\alpha)\%$ credible interval is not unique. To calculate a $100(1-\alpha)\%$ credible interval, one can take the $\alpha/2$ and $(1-\alpha/2)$ quantiles. This gives a central credible interval with equal tails, which is invariant to one-to-one transformations and easy to compute. This is the highest posterior density interval (HPDI) if the posterior is symmetric, otherwise it is wider than the HPDI.

The HPDI is the shortest possible interval having a given credible level $(1-\alpha)$ (King et al., 2010, p. 86). Assuming there is a single mode, this interval is centered around the mode, and is always computable from the posterior density. A $100(1-\alpha)\%$

credible interval is a $100(1-\alpha)\%$ highest posterior density interval if for all $\theta' \in [a, b]$ and $\theta'' \notin [a, b]$, $p(\theta'|x) \geq p(\theta''|x)$ (King et al., 2010, p. 86).

I assume that the posterior distribution is uni-modal, and this can be confirmed by looking at posterior density plots.

## 4.5 Diagnostic tools

### 4.5.1 Convergence

As stated in Section 4.3.3, we must discard any observations previous to the convergence of the chains to estimate the posterior density. Therefore I must determine how long it takes the chain to converge and where to set this burn-in period. We can do this by looking at trace plots. However, these may be misleading and it is preferable to use multiple chains run from different initial values to observe if all chains converge to the same result.

A more formal analysis of convergence can be made using diagnostics, such as the Brooks-Gelman-Rubin (BGR) diagnostic (Brooks and Gelman, 1998). The BGR diagnostic uses a comparison similar to a classical analysis of variance to determine whether or not there are differences in estimates from different chains. When the BGR statistic is close to unity, one can assume that convergence has been achieved and the outputs from all chains are indistinguishable.

Another useful tool is the Heidelberger and Welch convergence diagnostic, which uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values in the Markov chain come from a stationary distribution (Brooks and Gelman, 1998). The test is applied successively, first to the whole chain, then after discarding the first $10\%, 20\%, \ldots$ of the chain. This continues until either the null hypothesis that the sampled values in the MCMC chain come from a stationary distribution is accepted, or 50% of the chain has been discarded.

If the stationarity test is passed, the function reports the number of iterations to keep and the number to discard. However, if 50% of the chain has been discarded the test has failed and indicates that a longer MCMC run is needed.

The half-width test calculates a 95% confidence interval for the mean of the posterior, using the portion of the chain that passed the stationarity test (Heidelberger and Welch, 1983). Half the width of this interval is compared with the posterior estimate of the mean, and if the ratio between the half-width and the mean is lower than a certain value $\epsilon$, the half-width test is passed. Otherwise the length of the sample is deemed not long enough to estimate the mean with sufficient accuracy.

Such diagnostics are useful tools, however they can only provide evidence for lack of convergence rather than proof of convergence. Therefore, it is sensible to be conservative when setting burn-in levels, and choose a longer burn-in than indicated by trace plots and diagnostics.

### *4.5.2 Pilot tuning*

To ensure adequate convergence of the Markov chains and reasonable acceptance probabilities within the MH algorithm, one can perform pilot tuning. The variances of the proposal distributions are adjusted to obtain MH acceptance probabilities in the range $35-40\%$. This level of acceptance should ensure adequate mixing (Gelman, 1996).

Excessive pilot tuning is often unnecessary for fast simulations, because we can tolerate a higher rejection rate if not computationally costly. However, for more complicated models pilot tuning becomes more important to reduce computational cost.

### 4.5.3 Autocorrelation

A useful tool to assess the performance of the Markov chain is the autocorrelation function (ACF), which examines the correlation between successive values in the Markov chain. The autocorrelation at lag $t$, that is the values of the chain separated by $t$ iterations, is defined as $\text{cor}(\theta^k, \theta^{k+t})$ (King et al., 2010, p136).

For good mixing we would like low levels of autocorrelation. The ideal is a fast decrease in the value of the autocorrelation as the lag increases, showing that values of the chain are not highly correlated as in the second plot of Figure 4.3.

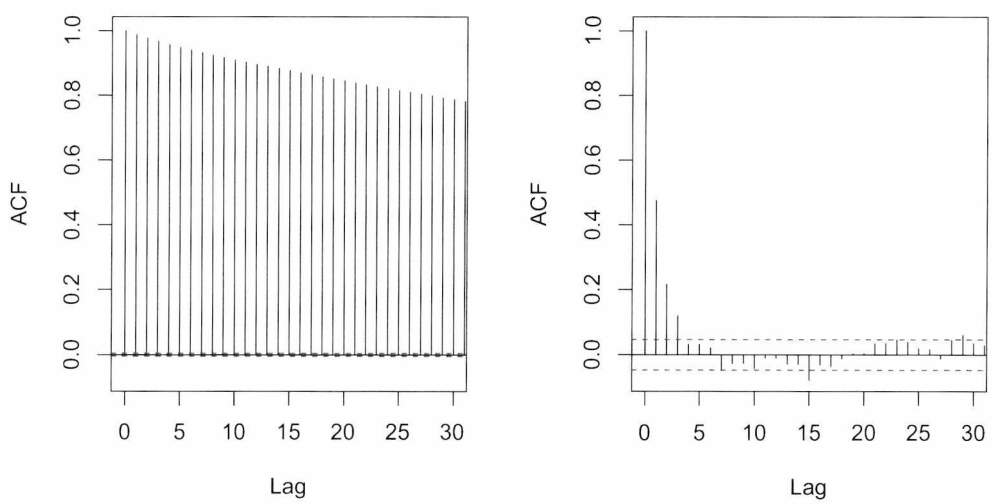



*Figure 4.3:* Example ACF plots showing (a) high and (b) low levels of autocorrelation. The chain in the second plot has been thinned to every $100^{th}$ iteration, and autocorrelation has greatly decreased.

One possibility to reduce autocorrelation is to make adjustments within the pilot tuning to increase the acceptance rate of the MH algorithm, which will allow the chain to move around more within the parameter space. To increase the acceptance rate, we decrease the variance of the proposal distribution and vice versa.

Thinning could also be used to reduce autocorrelation, that is where we keep only

every $i^{th}$ iteration of the sample. However, this will discard a very large number of sampled values, which is not favourable when computation is expensive. Thinning will not solve the problem of poor mixing so it is preferable to resolve the cause of the autocorrelation.

Another possibility to reduce autocorrelation is the use of block updates (Section 4.3.5). If parameters are correlated, their range of acceptable new values are constrained in a single-update MH by the values of the correlated parameters. If parameters are updated in blocks this should increase the movement around the parameter space.

### *4.5.4 Model checking and discrimination*

A commonly used Bayesian method to check goodness-of-fit is calculating Bayesian $p$-values, which match predicted or imputed data against observed data.

Given observed data $x$, at each MCMC iteration, $k$, a new data set, $\boldsymbol{x}^k$, is generated by simulating from the model with parameters $\boldsymbol{\theta}^k$, and a discrepancy statistic, $D(\boldsymbol{x}^k, \boldsymbol{e}_k)$, such as the negative of the log-likelihood function is used to measure the difference between this generated data and the expected values at that iteration, $\boldsymbol{e}_k$. This value is then compared to the discrepancy function evaluated at the observed data, $D(\boldsymbol{x}, \boldsymbol{e}_k)$ (King et al., 2010, p138).

The Bayesian $p$-value is the probability that the simulated data could be more extreme than the observed data, that is the proportion of times $D(\boldsymbol{x}, \boldsymbol{e}_k) < D(\boldsymbol{x}^k, \boldsymbol{e}_k)$. If the model is a good fit to the data, then we would expect the Bayesian $p$-value to be close to 0.5. Figure 4.4 shows an example of a Bayesian $p$-value scatter plot of the discrepancy function. The Bayesian $p$-value is the proportion of points lying above the line of unit slope in the plot of the discrepancy statistic of the observed data versus the discrepancy statistic of the simulated data.

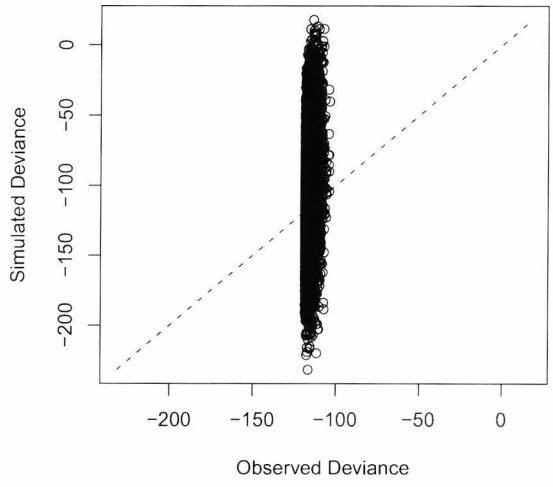



*Figure 4.4:* Example scatter plot of the negative log-likelihood of the observed data versus the negative log-likelihood of data simulated from the parameter values at each MCMC iteration. The proportion of points lying above the dashed line gives the Bayesian $p$-value, which is close to 0.5 when the model is a good fit to the data.

There are several suitable choices for the discrepancy statistic, including a measure of 'deviance', defined as -2 times the log-likelihood of the data $\boldsymbol{x}^k$ at $\boldsymbol{\theta}^k$. One could alternatively use the Freeman-Tukey statistic

$$\sum_i (\sqrt{\boldsymbol{x}_i} - \sqrt{\boldsymbol{e}_i})^2;$$

or the Pearson chi-squared statistic,

$$\sum_i (\boldsymbol{x}_i - \boldsymbol{e}_i)^2 / \boldsymbol{e}_i.$$

In some cases Bayesian $p$-values can differ substantially depending on the discrepancy function used (King et al., 2010, p140), reflecting different aspects of the model.

To make comparisons between competing models, we can again use discrepancy measures to compare the data to different models. We work with the 'deviance', defined

$$D(x, \theta) = -2\log f(x|\theta), \tag{4.13}$$

which can be averaged over the posterior distribution to estimate the expected deviance, $\hat{D}_{avg}(x)$. The model with lowest expected deviance will have the highest posterior probability (Gelman et al., 2004, p 181).

Using a point estimate for $\theta$, such as the mean of the posterior simulations, we define $D_{\hat{\theta}}(x) = D(x, \hat{\theta}(x))$ and we can use this within the Deviance Information Criterion (DIC):

$$DIC = 2\hat{D}_{avg}(x) - D_{\hat{\theta}}(x), \tag{4.14}$$

to estimate the expected predictive deviance, that is the error which would be expected when applying the model to future data (Gelman et al., 2004, p 182).

The DIC was developed to compare complex hierarchical models in which the number of parameters is not clearly defined (Spiegelhalter et al., 2002). The effective number of parameters for the model can be measured using the difference between the posterior mean deviance and the deviance at $\hat{\theta}$,

$$p_D = \hat{D}_{avg}(x) - D_{\hat{\theta}}(x). \tag{4.15}$$

DIC can be problematic as it is possible to obtain a negative effective number of parameters (King et al., 2010, p. 151), and so caution will be applied when using this method.

## 4.6 Estimating species richness using Bayesian methods

This section explores Bayesian methods for species richness estimation, and extends them for the Neyman Type A-gamma distribution.

Recall the species richness model likelihood (Equation 3.1 of Chapter 3)

$$L(N, \boldsymbol{\theta}) = \frac{N!}{(N-D)!} p_0(\boldsymbol{\theta})^{N-D} \{1 - p_0(\boldsymbol{\theta})\}^D \prod_{k \geq 1} q_k(\boldsymbol{\theta})^{f_k}. \tag{4.16}$$

where $N$ is the unknown number of species in the population, $D$ is the number of species in our sample, $f_k$ is the number of species observed $k$ times in the sample, $p_0(\boldsymbol{\theta})$

is the probability of not seeing a species, $q_k(\boldsymbol{\theta})$ are the zero truncated probabilities of seeing a species $k$ times, and $\boldsymbol{\theta}$ is the vector of parameters describing the abundance distribution .

Under a Bayesian framework, the posterior for our species richness model can be written as

$$\begin{aligned} \pi(N, \boldsymbol{\theta}|\boldsymbol{x}) &\propto p(N, \boldsymbol{\theta}) L(N, \boldsymbol{\theta}; \boldsymbol{x}) && (4.17) \\ &= p(N, \boldsymbol{\theta}) \frac{N!}{(N-D)!} p_0(\boldsymbol{\theta})^{N-D} \{1 - p_0(\boldsymbol{\theta})\}^D \prod_{k \geq 1} q_k(\boldsymbol{\theta})^{f_k}, && (4.18) \end{aligned}$$

where the parameters of the abundance distribution, $\boldsymbol{\theta}$, are treated as nuisance parameters as our goal is to estimate $N$.

### 4.6.1 Marginal probability calculation

The posterior distribution for the model can be evaluated by calculating the marginal probabilities as in the frequentist approach, by evaluating the integral

$$p_k(\boldsymbol{\theta}) = \int_0^\infty g(k|\lambda) f(\lambda; \boldsymbol{\theta}) \mathrm{d}\lambda, \tag{4.19}$$

and setting priors on the parameters of the mixing distribution, $\boldsymbol{\theta}$.

### 4.6.2 Hierarchical Bayes approach to species richness estimation

One possible advantage of using the Bayes approach is that we can eliminate the need to evaluate an integral to calculate the non-hierarchical likelihood. If we construct a hierarchical model we remove this integral, and model the data using a random effects approach. This approach gives us less informative priors, because we are adding an extra level to the model.

For each observation we estimate an abundance parameter, and the abundance parameters are distributed with some distribution. For example, observations $x_i \sim$ Poisson$(\lambda_i)$ and $\lambda_i \sim$ Gamma$(\alpha, \beta)$, for all $i = 1, ...., N$. This model is simple to implement, and can be defined in `WinBUGS` using the following short code:

```
model{
    for (i in 1 : N) {
       theta[i] ~ dgamma(alpha, beta)
       x[i] ~ dpois(theta[i])
    }
    alpha ~ dexp(1)
    beta ~ dgamma(0.1, 1.0)
}
```

where $\alpha$ and $\beta$ have been given particular exponential and gamma priors respectively.

However, this requires that $N$ is a fixed value, so to implement a hierarchical Bayes (HB) approach in `WinBUGS` when we have missing species we can utilise reversible jump Markov chain Monte Carlo (RJMCMC) or data augmentation. These methods enable us to model the data using this random effects approach, and give us an estimate of the total species richness.

### 4.6.3 *Varying clustering between species*

An advantage of the Bayesian approach is that we can allow the clustering parameter to vary between species within the Neyman Type A-gamma model.

Previously, in the frequentist approach of Chapter 3, I have given the abundance data a Neyman Type A distribution, and to estimate $\hat{N}$ I assumed that one parameter is constant across all species and assume that the mean abundance, $\lambda\phi = \mu$, follows a gamma distribution. The probability that a species is observed $k$ times in a sample is

$$p_k(\boldsymbol{\theta}) = \int_0^\infty \frac{e^{-\lambda}\phi^k}{k!} \sum_{j=0}^{\infty} \left\{ \frac{(\lambda e^{-\phi})^j j^k}{j!} \right\} \frac{\beta^\alpha (\lambda\phi)^{\alpha-1} e^{-\beta\lambda\phi}}{\Gamma(\alpha)} \mathrm{d}\lambda, \tag{4.20}$$

which allows the species density to vary, but I assumed that clustering intensity does not vary between species.

Considering this from a hierarchical Bayes viewpoint, I can extend the model so that not only the mean abundance but also the clustering parameter, $\phi$, can vary between species. The probability that a species is observed $k$ times in a sample becomes

$$p_k(\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty \frac{e^{-\lambda}\phi^k}{k!} \sum_{j=0}^{\infty} \left\{ \frac{(\lambda e^{-\phi})^j j^k}{j!} \right\} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \frac{b^a \phi^{a-1} e^{-b\phi}}{\Gamma(a)} \mathrm{d}\lambda \mathrm{d}\phi, \tag{4.21}$$

if I give $\phi$ a gamma prior with parameters $a$ and $b$. However, we do not need to evaluate this double integral when using the hierarchical Bayes approach, and therefore calculation of the species richness estimate of this model becomes simpler.

By allowing clustering to vary between species, I should be able to describe benthic data more accurately using this model, and I would hope that the fit of the Neyman Type A-gamma-gamma would improve over the Neyman Type A-gamma.

When applying this model within `WinBUGS`, multiple errors were produced. A possible source of error could be that there is insufficient data to inform so many parameters when each species is given a different clustering parameter and abundance parameter. This theory could be tested by running the model on larger data sets. To combat this problem, one approach might be to group species and allow the clustering parameter to vary between these groups. Using fewer parameter values should then allow the model to be fitted.

The question arising then is how to group the data or species, especially when dealing with the species that have not been seen in the sample. One option could be to use particular traits of the species, such as body length, or feeding type. Alternatively, the groupings could be made with respect to the abundance of a species. This area requires further research, as we would not have data on the missing species, and therefore we must consider how such species will be grouped within the model.

*Reversible jump Markov chain Monte Carlo*

Reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995) is used when we consider more than one model with different numbers of parameters in each, and allows us to choose between them. It allows a change in dimension of the Markov chain which is not possible using a standard MH algorithm.

RJMCMC produces a Markov chain with stationary distribution equal to the joint posterior of the models and parameters; that is:

$$\pi(\boldsymbol{\theta}_m, m|x) \propto f(x|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m)p(m), \tag{4.22}$$

where $m$ is the model, $\boldsymbol{\theta}_m$ denotes the corresponding parameters of that model and $p(m)$ is the prior probability for model $m$.

I can use this approach to tackle the missing data problem, by treating each value of $N$ as a separate model, with the corresponding number of random effects.

The RJ algorithm is a form of the MH algorithm. At each iteration there are two steps (King et al., 2010, p156);

1. update the parameters, $\theta_m$, conditional on the model using the MH algorithm,
2. update the model, $m$, conditional on the current parameter values using RJMCMC.

For each iteration the second step consists of two parts,

1. propose to move to a different model with some given parameter values,
2. accept this move with some probability.

Within this step, the shared parameters between the models are set to their previous values, and for any new ones a parameter value $u$ is simulated from a proposal distribution $q$ as in the MH algorithm.

The acceptance probability for the model move becomes

$$A = \frac{\pi(\theta', m|x)P(m|m')}{\pi(\theta, m|x)P(m'|m)q(u)} \left| \frac{d(\beta_0', \beta_1')}{d(\beta_0, u)} \right| \tag{4.23}$$

where $\left| \frac{d(\beta_0', \beta_1')}{d(\beta_0, u)} \right|$ is the Jacobian if $\beta_1'$ and $\beta_0'$ are the parameters of the proposed model $m'$, and model $m$ only has parameter $\beta_0$, and $P(m|m')$ is the probability of proposing to move to model $m$ from model $m'$ (King et al., 2010, p158). So the model move to state $(\theta', m')$ is accepted with probability min(1,$A$).

For the reverse move, from $(\theta', m')$ to $(\theta, m)$, we set $u = \beta_1'$ and $\beta_0 = \beta_0'$ and the move is accepted with probability min(1,$A^{-1}$). The posterior model probabilities are then estimated as the proportion of time the Markov chain is in any model.

The specification of the proposals, $q$, is important in achieving efficient random jump algorithms, and due to the form of the acceptance probability it is not possible to specify improper priors on parameters which are not common to all models (King et al., 2010, p159).

*Data augmentation*

An alternative to RJMCMC is the data augmentation (DA) approach. This has been used in capture-recapture and occupancy models, to estimate population size or species richness over one or more sites (Royle et al., 2007; Kéry and Royle, 2008; Royle and Young, 2008; Royle, 2009). The strategy, developed by Royle et al. (2007), is to augment the observed data set with a fixed number of all-zero observations, and model the augmented data set as a zero-inflated version of the complete data model using an unknown, but estimable zero inflation parameter (Figure 4.5).

This augmented data set can be thought of as a super-population, of size $N_S$, from which the sampled community was drawn, and we formulate the model for the data as if all $N$ species were observed, and include a zero-inflation parameter $\psi$ to allow for

| *Species* | *Count* | |
|---|---|---|
| 1 | 1 | Observed species from population |
| 2 | 2 | |
| . | . | |
| . | . | |
| . | . | |
| $D-2$ | 3 | |
| $D-1$ | 2 | |
| $D$ | 1 | |
| $D+1$ | 0 | Unobserved species from population |
| $D+2$ | 0 | |
| . | . | |
| . | . | |
| . | . | |
| $N-1$ | 0 | |
| $N$ | 0 | |
| $N+1$ | 0 | Species not from population |
| $N+2$ | 0 | |
| . | . | |
| . | . | |
| . | . | |
| $N_S-1$ | 0 | |
| $N_S$ | 0 | |

*Figure 4.5:* Concept of the super-population model, where species $1, ..., D$ are observed in the data set, species $D+1, ..., N$ are in the population but not in the data set, and species $N+1, ..., N_S$ are not within the population, but form a super-population of possible species in the population.

excess zeros (Kéry and Royle, 2008). This can be justified in a Bayesian framework as the prior on $N$ is induced in the super-population, the marginal distribution for $N$ arising from a binomial mixture, the mixing given by the prior on $\psi$ (Schofield and Barker, 2010).

A convenient way to model the zero-inflated outcomes is to specify $x_i$ conditional on a latent value $z_i$. More formally, if $\psi$ is the probability that a data point $x_i$, $i = 1, ..., N_S$, is from the population of interest, then we define an indicator variable $z_i$ such that

$$z_i = \begin{cases} 1 & \text{if the } i\text{th element of the data set is in the population, size } N \\ 0 & \text{otherwise (an excess zero).} \end{cases}$$

So we assume $z_i \sim \text{Bernoulli}(\psi)$, and $z_i$ are independent (Royle et al., 2007).

In this approach $N$ is a derived parameter, such that

$$N = \sum_{i=1}^{N_S} z_i.$$

Therefore we can assume that $N$ has a binomial prior distribution with index $N_S$ and success parameter $\psi$. If $\psi \sim \text{Uniform}(0, 1)$, removing $\psi$ from the joint prior $p(N|N_S, \psi)p(\psi)$ by integration

$$\int_0^1 \text{Binomial}(N|N_S, \psi)d\psi = \text{Discrete Uniform}(0, N_S) \tag{4.24}$$

gives a discrete uniform prior for $N$ on the integers between 0 to $N_S$ (Royle et al., 2007).

This approach is fast and easy to implement in `WinBUGS`. However since $N$ is a derived parameter we are restricted to implied priors for $N$ that are based on binomial mixture models (Schofield and Barker, 2010). $N_S$ must be chosen large enough such that the posterior of $\psi$ can be adequately explored, so there is no risk of underestimating the value of $N$ (Royle et al., 2007). However, if we assign too large a value to $N_S$ then this will increase computational costs, which could become very high when dealing

with complex distributions. Royle et al. (2007) advocate $N_S$ to be chosen by trial-and-error, such that the mass of the posterior of $N$ is not concentrated near $N_S$.

Schofield and Barker (2010) show that the RJMCMC approach is equivalent to the data augmentation approach, but RJMCMC is more general because it does not restrict the prior on $N$ to a family based on binomial mixtures. They also showed that RJMCMC can result in substantial gains in efficiency over the data augmentation method, because fewer parameters need be sampled at every step, especially when the posterior for $N$ is skewed. This is something to be considered during analysis.

### 4.6.4 *Incorporating information from multiple grabs*

A benefit of the hierarchical Bayes approach, highlighted in Kéry and Royle (2008), is that we are able to incorporate information from more than one sample. Previously I have pooled data across grabs for use in the estimators, but the hierarchical set-up allows us to use all of the data by specifying $x_{ij} \sim \text{Poisson}(\lambda_i)$, where $j = 1, ..., g$ and $g$ is the number of grab samples.

To implement this in `WinBUGS` requires little change to the code (See Appendix D.3). This should improve the species richness estimates, because we are obtaining as much information as possible from our samples. Typically in benthic surveys we have five or ten replicates from each site.

### 4.6.5 *Non-informative priors*

When there is an absence of prior information on the model parameters, we would like to reflect this in the priors used. That is we wish the prior to play a minimal role in the posterior distribution. Flat priors can be used, which assign equal probability to all possible parameter values. However these priors are improper distributions unless there are bounds on the parameter space, which may restrict the posterior values unrealistically (King et al., 2010, p77). A density is proper if it does not depend on data, and integrates to one (Gelman et al., 2004, p61).

Improper priors can lead to improper posteriors (although not always), and this could mean that the posterior mean does not exist (King et al., 2010, p77). Barger and Bunge (2010) show that the posterior in Equation 4.18 can be integrated with respect to $N$ and $\theta$, therefore the posterior is proper. The posterior will be finite as long as the prior for the nuisance parameters, $p(\theta)$, is proper (Barger and Bunge, 2010).

Since in this chapter I wish the priors to be non-informative, the choice among them should not matter as the likelihood should be dominant in the posterior distribution (Gelman et al., 2004, p65). The two non-informative priors to be considered are the reference and Jeffrey's priors, as used by Barger and Bunge (2010), which allows us to make a direct comparison with their results.

Alternatively one could chose vague priors which are proper and with large variance and use hyperparameters to dilute the influence of any prior assumptions on the posterior. This essentially creates random effects of the model parameters, as discussed in Section 4.3.2.

*Jeffrey's prior*

The Jeffrey's prior attempts to minimise the influence of the prior on the posterior and is based upon the expected Fisher information matrix. If $(N, \theta)$ are the model parameters for data $x$, Jeffrey's prior is given by

$$p(N, \theta) \propto [F(N, \theta|x)]^{1/2},$$

where $F(N, \theta|x)$ is the Fisher Information given by

$$F(N, \theta|x) = -E\left[\frac{\mathrm{d}^2 \log L(N, \theta|x)}{\mathrm{d}(N, \theta)^2}\right], \tag{4.25}$$

and $L(N, \theta|x)$ is the likelihood for $N$ and $\theta$. The Fisher information can only be found for likelihoods which are differentiable with respect to the parameters, using Equation 4.25. Since $N$ is a discrete parameter, the likelihood is not differentiable in $N$. However I can use the linear difference score to define the Fisher information.

The linear difference score is defined as

$$U(N) := \frac{L(N) - L(N-1)}{L(N)}. \tag{4.26}$$

If $U(N)$ is of the form

$$U(N) = (Y - \mu_N)/c_N$$

where $\mu_N$ and $c_N$ are functions of $N$, and $Y$ is random data, then var$(U(N))$ is the inverse information in $N$ (Lindsay and Roeder, 1987). This is termed the linear difference property, and the species likelihood can be shown to satisfy this property (Barger and Bunge, 2010).

Using this method, Barger and Bunge (2008) obtain the information for $N$ and $\theta$

$$F(N,\theta) = \begin{pmatrix} \frac{1}{N}\frac{1-p_\theta(0)}{p_\theta(0)} & \left(-\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(0)\right)^T \\ -\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(0) & N\varrho(\theta) \end{pmatrix}$$

where $\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(0)$ is the column vector of partial derivatives

$$\left[\left(\frac{\partial}{\partial\theta_1}\log p_\theta(0), \frac{\partial}{\partial\theta_2}\log p_\theta(0), ..., \frac{\partial}{\partial\theta_m}\log p_\theta(0)\right)\right]^T,$$

and

$$\varrho(\theta) = E_X\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log\, p_\theta(X)\right)^2\right].$$

Expectation is taken with respect to $p_\theta$ and $m$ is the dimension of the nuisance parameter $\theta$ (Barger and Bunge, 2010).

As the diagonal elements of this matrix contain terms that factor into a function of $N$ multiplied by a function of $\theta$, these can be treated independently. Taking the square root of the determinant of the Fisher Information, Jeffrey's prior is

$$p(N,\theta) \propto \det[F(N,\theta|x)]^{1/2} \tag{4.27}$$

$$= N^{\frac{m-1}{2}} p(\theta) \tag{4.28}$$

where $p(\theta)$ is some function of the nuisance parameters (Barger and Bunge, 2010),

$$p^2(\theta) = |\varrho(\theta)| \times \left|\frac{1-p_\theta(0)}{p_\theta(0)} - \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(0)\right)^T (\varrho(\theta))^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(0)\right)\right|. \tag{4.29}$$

Jeffrey's prior is often used as it is invariant to reparameterisations of the model (King et al., 2010, p77). However it is an improper prior, and therefore integrability of the posterior must be shown.

This multi-parameter form of Jeffrey's prior, used by Barger and Bunge (2010), is not of the form generally used. Instead, a product of the priors for each parameter is often used, which is less informative on the posterior and gives better results. However, I have chosen to use Jeffrey's prior in multi-parameter form so that my results are comparable to those of Barger and Bunge (2010), who also analysed the Lepidoptera and CBC data sets using their species richness estimation approach.

*Reference prior*

The reference prior is a non-informative prior based on maximising the expected entropy of an experiment, i.e. maximising the distance between the posterior and the prior.

The derivation of the reference prior for our model depends on the asymptotic results of Sanathanan (1972), and is defined in Theorem 1 of Barger and Bunge (2010). It is based on the general method for deriving a reference prior for continuous-valued parameters (Bernardo and Ramon, 1998) and the information for integer-valued parameters as described in the previous section (Lindsay and Roeder, 1987).

Barger and Bunge (2010) show that the conditional reference priors are

$$p(\theta_m|N,\theta_1,...,\theta_{m-1}) \propto \rho(\theta)_{mm}^{1/2} \tag{4.30}$$

and

$$p(\theta_k|N,\theta_1,...,\theta_{k-1}) \propto \exp\left[\int ... \int (\log h_{kk}^{1/2}) \times \left\{ \prod_{j=k+1}^{m} p(\theta_j|N,\theta_1,...,\theta_{j-1}) \right\} \mathrm{d}\boldsymbol{\theta_{k+1}}\right] \tag{4.31}$$

where $\mathrm{d}\boldsymbol{\theta_{k+1}} = \mathrm{d}\theta_{k+1} \times ... \times \mathrm{d}\theta_m$ if all of the $p(\theta_k|N,\theta_1,...,\theta_{k-1}), k = 1,...m$ are proper. If any are not, then a compact approximation is required for the corresponding

integrals (Barger and Bunge, 2010).

Again the joint prior, $p(N, \theta)$ factors into two independent priors for $N$ and $\theta$. The marginal reference prior for $N$ is

$$p(N) \propto N^{-1/2}. \tag{4.32}$$

Regardless of the abundance distribution used, the form of the reference prior for $N$ is the same (Barger and Bunge, 2010).

When we have a one parameter problem, that is assuming all nuisance parameters are known, it can be seen that the reference and Jeffrey's prior are equivalent.

## 4.7 Analysis of methods via simulation

I explored the performance of the Bayesian methods, using non-informative priors on simulated data. MCMC was run in `R` or `WinBUGS`, and the output analysed using using package `coda` in `R` . Manual tuning was carried out when running MH in `R`, and for the single-update MH I used a normal proposal distribution for sampling each of the parameters for simplicity. Although in reality $N$ is a discrete parameter, I set the model up this way for ease of computation.

I explored the effects of:

1. using single versus block updates,
2. the choice of uninformative prior on $N$,
3. using the data augmentation method versus RJMCMC within a hierarchical Bayes framework,
4. using multiple grabs versus pooled data within a hierarchical Bayes framework,
5. using the non-hierarchical likelihood versus the hierarchical Bayes method.

To do this, I simulated a data set $Y$ such that $y_i \sim \text{Poisson}(\lambda_i)$, and $\lambda_i \sim \text{Gamma}(2, 1)$. This gave a sampling depth of 75%, which meant a 0.75 probability of catching a randomly selected species over all grabs. $N$ was chosen as 200. The data were generated for 10 grabs, and the data when pooled are shown in Table 4.1. The number of species seen in the sample was $D = 150$.

In all of these simulations unless stated otherwise I used the negative binomial model for the abundances, a reference prior for $N$ and non-informative half Cauchy priors on the gamma parameters $\boldsymbol{\theta} = \{\alpha, \beta\}$. Half Cauchy priors were used so that the results would be comparable with those in Barger and Bunge (2010), although results are given there for this model only when the data are truncated to frequency counts less than 10 (The other results are given for an abundance distribution that is a mixture of two exponentials).

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_k$ | 54 | 33 | 34 | 10 | 10 | 3 | 2 | 1 | 2 | 1 |

*Table 4.1:* Data simulated from the negative binomial and pooled over grabs, $\alpha = 2, \beta = 1, N = 200$; $D = 150$.

### 4.7.1 Single versus block updates

It is likely that the nuisance parameters of the model will be highly correlated, and I know that the estimate of $N$ depends on these parameters, therefore block updates should prove useful in reducing computation time and improving mixing for the Markov chains. Therefore, I investigated the advantage of using block updates to allow better exploration of the parameter space, and therefore better estimates. This should decrease autocorrelation and speed up convergence.

I used the non-hierarchical likelihood method using MH MCMC, within R, with acceptance probabilities tuned to approximately $35-40\%$. The proposal distributions

for the single-update method were normal, with the previous value of the chain as the mean, and standard deviations of 27, 0.4, and 0.2 respectively, defined through pilot tuning of 1,000 iterations with the first 500 as a burn-in. The initial values for the pilot tuning were set at 150, 5.1 and 0.5 respectively. The initial value for $N$ was chosen as the number of species seen in the sample, and the initial values for the gamma parameters were estimated from the data by setting $\beta = \text{Var}(X)/\mathbb{E}(X)(1 - \mathbb{E}(X)/\text{Var}(X))$ and $\alpha = \mathbb{E}(X)/\beta$.

For the single updates I ran 150,000 iterations for three chains, after a burn-in of 80,000. I used the BGR, HW and half-width diagnostics to look for convergence of the chains, which was apparent after 40,000 iterations (Figure 4.6). Figure 4.7 shows that there was a high autocorrelation between successive values of the chains.

The posterior densities and summary statistics are shown in Figure 4.8 and Table 4.2, showing that the posterior estimates were close to the true parameter values of $N = 200, \alpha = 2$ and $\beta = 1$. The Bayesian $p$-value was close to 0.5 showing that this model was a good fit for the data, as one would expect.

Using the results from one of the single-update chains as a pilot for the block update

| Method | Mean | Median | SD | 95% Credible | Bayesian $p$-value |
|---|---|---|---|---|---|
| Single updates | 221 | 215 | 33.11 | (173, 282) | 0.49 |
| Block updates | 219 | 213 | 33.22 | (181, 290) | 0.51 |

*Table 4.2:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using MH single and block updates for the negative binomial model applied to negative binomial simulated data. The data were simulated with $N = 200$. The Bayesian $p$-value showing the fit of the model to the data is also reported.

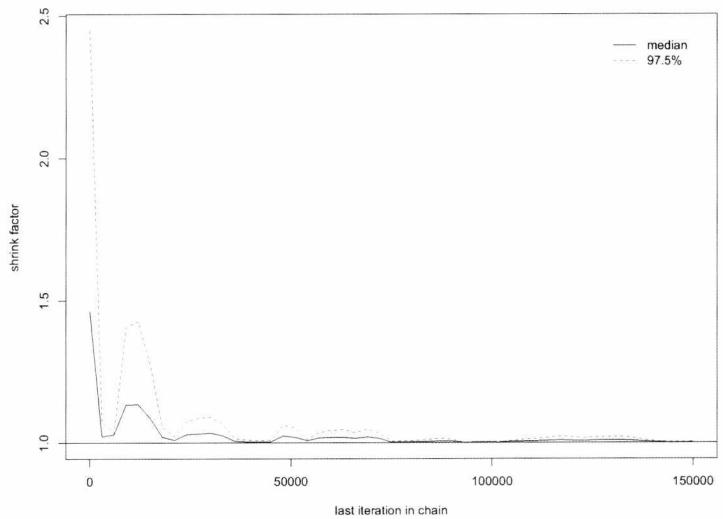



*Figure 4.6:* Convergence (black line) and confidence of the BGR statistic (dashed line) for the MCMC chain for $N$ for the negative binomial model applied to the negative binomial simulated data, using single update MH. Plots for $\alpha$ and $\beta$ show a very similar pattern. I used the reference prior on $N$ and half-Cauchy priors on $\alpha$ and $\beta$.

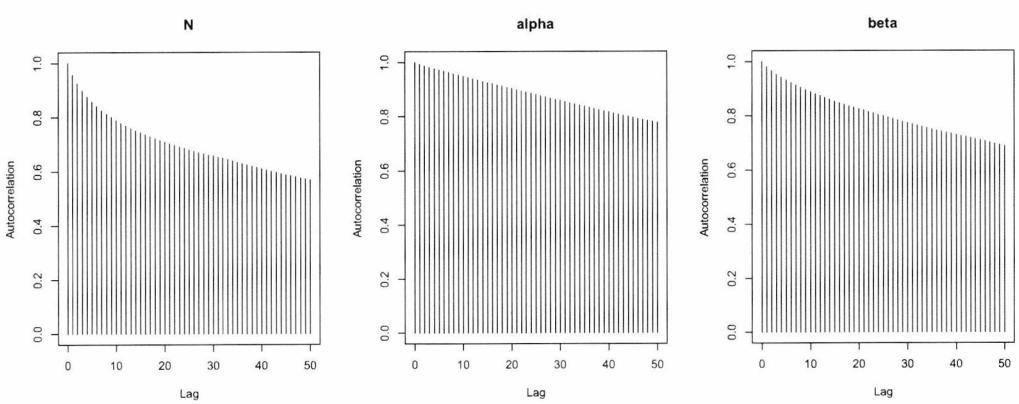



*Figure 4.7:* ACF plots for the MCMC chain for $N$, $\alpha$ and $\beta$ for the negative binomial model using single update MH applied to the negative binomial simulated data. I used the reference prior on $N$ and half-Cauchy priors on $\alpha$ and $\beta$.

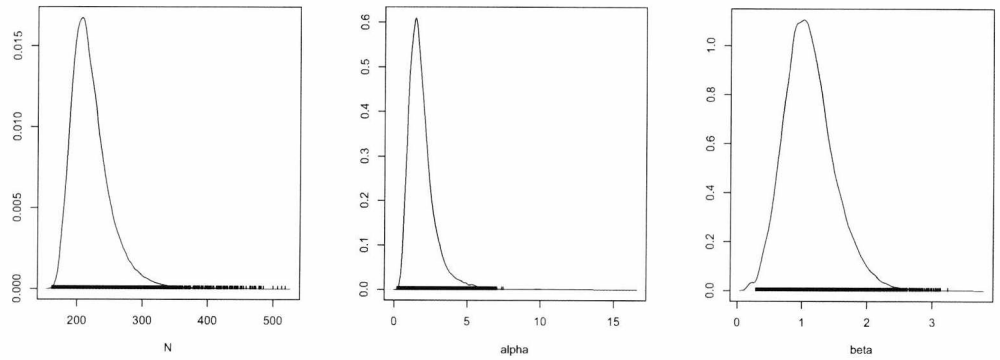



*Figure 4.8:* Posterior density plots for $N$, $\alpha$ and $\beta$ for the negative binomial model using single update MH applied to the negative binomial simulated data. I used the reference prior on $N$ and half-Cauchy priors on $\alpha$ and $\beta$.

method, I constructed a multivariate normal proposal distribution as

$$\boldsymbol{\omega}|\theta^k \sim \mathcal{N}_i\left(\begin{pmatrix}\theta_1^k\\ \theta_2^k\\ \theta_3^k\end{pmatrix}\begin{pmatrix}\sigma^2_{\theta_1} & \sigma_{\theta_1\theta_2} & \sigma_{\theta_1\theta_3}\\ \sigma_{\theta_1\theta_2} & \sigma^2_{\theta_2} & \sigma_{\theta_3\theta_2}\\ \sigma_{\theta_1\theta_3} & \sigma_{\theta_3\theta_2} & \sigma^2_{\theta_3}\end{pmatrix}\right) \tag{4.33}$$

using the posterior correlations from the pilot run. This gave a covariance matrix of

$$\Sigma = \begin{pmatrix} 1096.418 & -19.820 & 10.070\\ -19.820 & 0.676 & -0.279\\ 10.070 & -0.279 & 0.157\end{pmatrix}. \tag{4.34}$$

The posterior standard deviations from the single update MH MCMC of 33.11, 0.82 and 0.40 were reasonably close to those which were used in the proposal distribution of 27, 0.4, and 0.2. For direct comparison, I used the same standard deviations for the block updates. The correlation between parameters was high (Figure 4.9), over 0.7 between all pairs of parameters. Therefore I expected the block updates to perform better.

Figures 4.10 to 4.12, and the summary statistics of Table 4.2, show some differences in the results when using block updates rather than single updates. The most notable difference when running the MCMC was that fewer iterations were required using

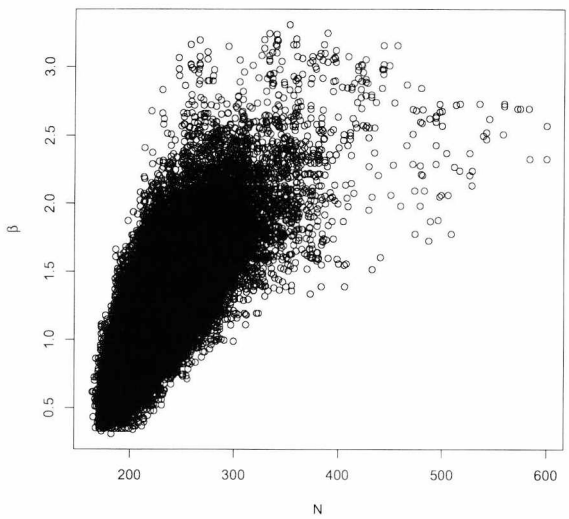



*Figure 4.9:* Correlation between values of the MCMC chain for $N$ and $\beta$ using single update MH for the negative binomial model applied to negative binomial simulated data.

block updates to see convergence in the chain. It was only necessary to run 100,000 iterations using block updates, and convergence is indicated after 20,000 iterations (Figure 4.10).

Figure 4.11 shows that the autocorrelation has decreased. Therefore the mixing of the chains has improved. This indicated that the chains converged quickly, which we have already seen in Figure 4.10.

The results produced by the two methods were very similar (Figure 4.12 and Table 4.2). There were slight differences in the credible intervals for $N$, but the posterior estimates for all parameters were close to the true parameter values from which I simulated the data, $N = 200$, $\alpha = 2$, $\beta = 1$.

Only considering the faster convergence rates of the block updating method, it would seem that this method is preferable. However, the autocorrelation is still high, and the results produced were similar for both methods.

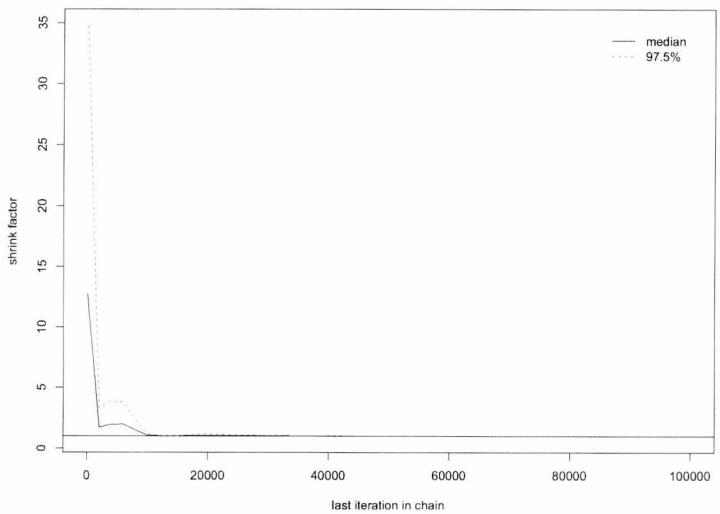



*Figure 4.10:* Convergence (black line) and confidence of the BGR statistic (dashed line) for the MCMC chain for $N$ for the negative binomial model applied to the negative binomial simulated data, using block updates. Plots for $\alpha$ and $\beta$ show a very similar pattern. I used the reference prior on $N$ and half-Cauchy priors on $\alpha$ and $\beta$.

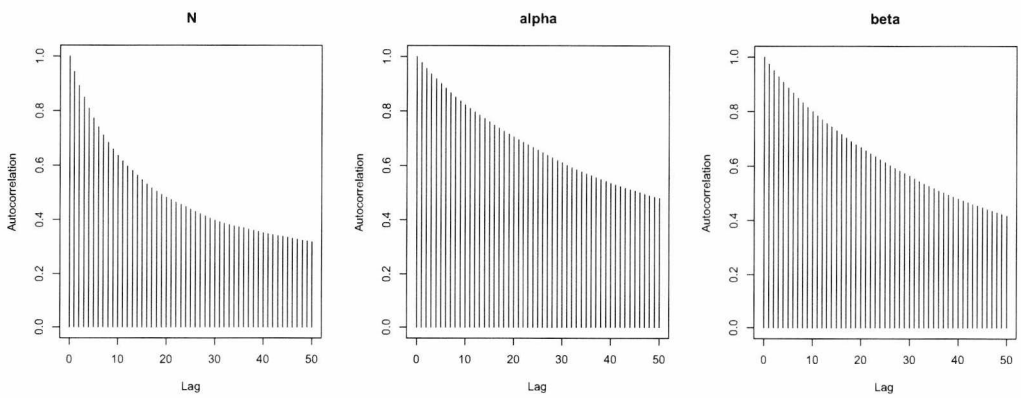



*Figure 4.11:* ACF plots for the MCMC chain for $N$, $\alpha$ and $\beta$ for the negative binomial model using block updates applied to the negative binomial simulated data. I used the reference prior on $N$ and half-Cauchy priors on $\alpha$ and $\beta$.

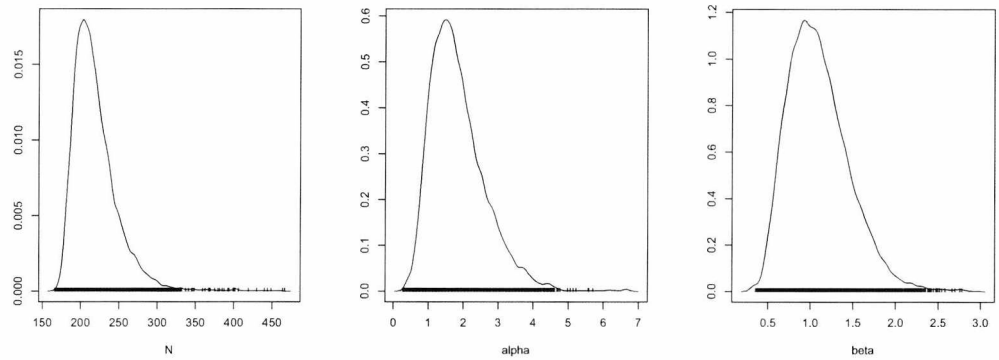



*Figure 4.12:* Posterior density plots for $N$, $\alpha$ and $\beta$ for the negative binomial model using block update MH applied to the negative binomial simulated data. I used the reference prior on $N$ and half-Cauchy priors on $\alpha$ and $\beta$.

As the block update method required a pilot study using the single-update MH to estimate posterior correlations and variances, it may not be beneficial to use it overall. In addition, if we were using a hierarchical approach in `WinBUGS`, it would not be possible to apply block updates for this type of model. Therefore I use single-updates as the preferred MH method.

### *4.7.2 Choice of prior for N*

Section 4.6.5 introduced two uninformative priors for $N$, Jeffrey's prior, $N^{\frac{m-1}{2}}$, and the reference prior, $N^{-\frac{1}{2}}$, where $m$ is the dimension of the nuisance parameters. I investigated the effect of these priors on the posterior. As they were objective priors, by design, they should have little effect on the posterior. I also compare them to using an uninformative flat, uniform prior, and highlight any advantages or disadvantages of each.

I again used the negative binomial simulated data set and the non-hierarchical likelihood method using MH MCMC, with tuned acceptance probabilities and normal proposal distributions for all parameters. I ran 150,000 iterations using each prior, using three chains each time; again using a burn-in of 80,000. This meant that I could

compare the results of the Jeffrey's prior and uniform prior to those already given for the reference prior in Section 4.7.1. Uninformative half-Cauchy priors were again used for the nuisance parameters $\boldsymbol{\theta} = \{\alpha, \beta\}$.

Since there were two nuisance parameters, Jeffrey's prior became $N^{\frac{1}{2}}$. When a uniform$(0, M)$ prior on $N$ was used, this cancelled in the acceptance probability. Therefore, the value of $M$ used was arbitrary.

Table 4.3 shows the summary statistics for the three posteriors, which are plotted in Figure 4.13. All priors gave very similar results, which was what I would have hoped. This implies that the likelihood was contributing all the information to the posterior. I would expect this as in the simulated data set there was a fairly high probability of seeing a species (0.75).

| Method | Mean | Median | SD | 95% Credible | Bayesian $p$-value |
|---|---|---|---|---|---|
| Reference | 221 | 215 | 33.11 | (173, 282) | 0.50 |
| Jeffrey's | 225 | 218 | 36.80 | (174, 290) | 0.49 |
| Uniform | 223 | 215 | 34.55 | (173, 289) | 0.50 |

*Table 4.3:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using reference, Jeffrey's and uniform priors on $N$ for the negative binomial model applied to negative binomial simulated data. The data were simulated with $N = 200$. The Bayesian $p$-value measuring the fit of the model to the data is also reported.

When using data with high probability of seeing a species, the choice of uninformative prior that I used on $N$ appeared to be arbitrary. I could therefore select the prior which was most convenient for the application. The uniform prior has the disadvantage of being bounded on the parameter space, and as Jeffrey's prior is improper, so I preferred to use the reference prior. Barger and Bunge (2010) also preferred the reference prior, especially as it does not depend on the dimension of the

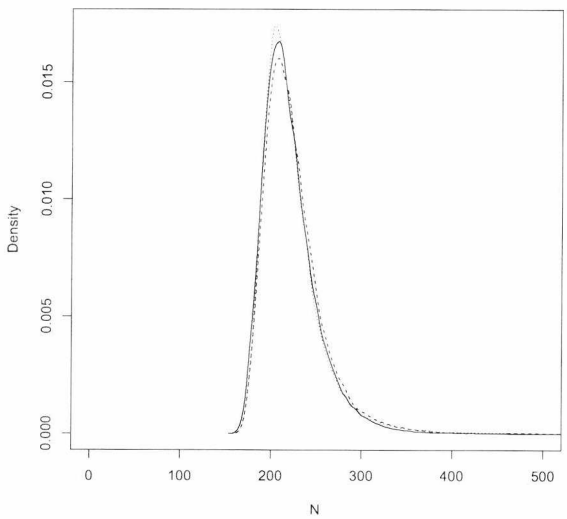



*Figure 4.13:* Posterior density plots for the simulated data for the negative binomial model with reference (solid), Jeffrey's (dashed) and uniform (dotted) priors on $N$. I used half-Cauchy priors on $\alpha$ and $\beta$, and the data were simulated with $N = 200$.

nuisance parameter. When $m = 1$ Jeffrey's prior is flat, but as $m > 1$, Jeffrey's prior is increasing, such that larger values of $N$ are more likely a priori (Barger and Bunge, 2010).

It was expected that the prior would have a larger effect on the posterior if the probability of seeing a species decreased. Therefore, I repeated the analysis for a simulated data set where only 40% of species were observed.

It was necessary to increase the number of iterations used appreciably, because the chains moved wildly around the parameter space and did not converge. Even after increasing the number of iterations to over three million per chain, the chains still did not converge. Therefore, I cannot give a posterior distribution for this data set. This lack of convergence was an indication of the boundary problem which we have seen in Chapter 3.

### 4.7.3 Data augmentation method versus Reversible Jump MCMC

The equivalence of the RJMCMC approach to the data augmentation (DA) approach is interesting. These HB approaches could reduce computational costs and reduce the likelihood of underestimating $N$.

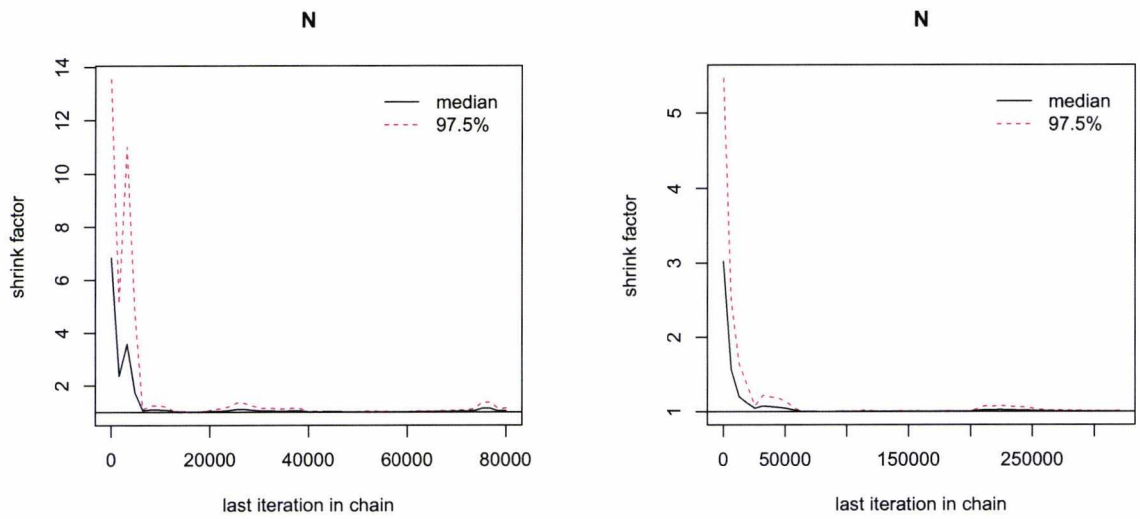
I ran the the MCMC in `WinBUGS` for the negative binomial model using the simulated data set, and initial values used in the previous MCMCs (For code see Appendix D). Uninformative half-Cauchy priors were placed on the gamma parameters, as in the previous examples. I gave the zero-inflation parameter $\psi$ a uniform prior over (0,1), which was equivalent to a Discrete uniform$(0, M)$ prior on $N$ (Equation 4.24).

The DA method was sensitive to the initial values used, in that unless they were well specified it could get stuck in an area of unrealistic parameter values, i.e. $N = 20$. When I specified initial values for all parameters this problem disappeared. I used a super-population size of $M = 750$ to ensure that the whole posterior was captured.

The DA method seemed to converge after 40,000 iterations (Figures 4.14a, 4.15), so I used a burn-in of 40,000, and ran another 40,000 for the posterior sample. The RJMCMC method required 200,000 iterations with a burn-in of 80,000 (Figure 4.14b). Data augmentation was a great deal faster to run in `WinBUGS` than RJMCMC, and required fewer iterations.

The posterior for the DA method was very similar to that of the RJMCMC method (Figure 4.16). The summary statistics in Table 4.4 confirmed the similarity of the posteriors, and indicated that the two methods gave essentially indistinguishable results for this particular example.

I showed that the two methods of reversible jump MCMC and data augmentation were essentially equivalent in this case. It was important to consider alternative specifications of distributions to avoid compiling errors within `WinBUGS`, especially




*(a)* data augmentation *(b)* RJMCMC

*Figure 4.14:* Convergence (black line) and confidence of the BGR statistic (dashed line) for the MCMC chain for $N$ for the negative binomial model applied to the negative binomial simulated data using data augmentation and RJMCMC. Plots for $\alpha$ and $\beta$ show a very similar pattern.

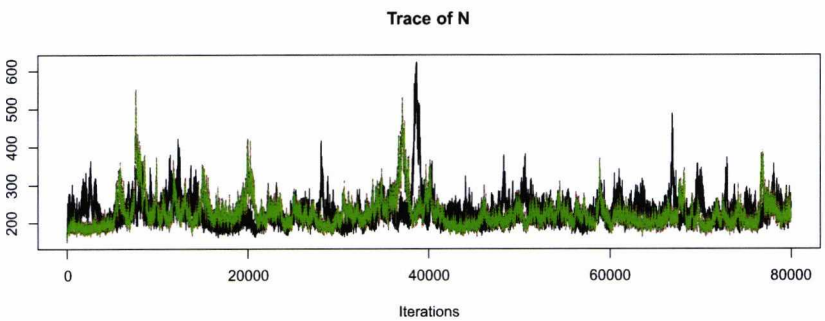



*Figure 4.15:* Trace plot of the MCMC chain for $N$ for the negative binomial model applied to the negative binomial simulated data using data augmentation.

| Method | Mean | Median | SD | 95% Credible interval |
|---|---|---|---|---|
| DA | 221 | 216 | 27.92 | (175, 273) |
| RJMCMC | 218 | 202 | 29.52 | (172, 275) |

*Table 4.4:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using DA and RJMCMC for the negative binomial model applied to negative binomial simulated data. The data were simulated with $N = 200$.

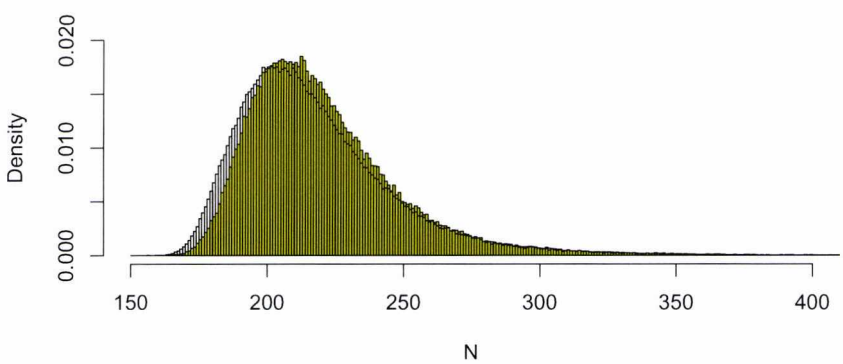



*Figure 4.16:* Posterior density plots for $N$ for the negative binomial model fitted to the negative binomial simulated data set in `WinBUGS` using data augmentation (yellow) and RJMCMC. The data were simulated with $N = 200$.

when using the zeros trick, and it was also important to input initial values to get realistic posterior values.

RJMCMC was more computationally intensive than DA for my models, and more iterations were required to ensure adequate mixing. This was in contradiction to Schofield and Barker (2010), who found that RJMCMC resulted in substantial gains in efficiency over the DA method, because fewer parameters needed to be sampled at every step. Schofield and Barker (2010) suggested that implementation of MCMC in software `JAGS` (Just Another Gibbs Sampler)(Plummer, 2003) might mix better than `BUGS`, and there might also be other computational implications. To specify the Neyman Type A distribution using the zeros trick was not as straightforward in `JAGS`,

and therefore I used WinBUGS for all computations using HB models.

### 4.7.4 Multiple grabs versus pooled data within a hierarchical Bayes framework

I ran MCMC in WinBUGS for the negative binomial model using the simulated data set over 10 grabs to investigate whether there was an advantage to use sample data from individual grabs rather than to pool. I used the DA method, because this showed computational advantages over RJMCMC in Section 4.7.3. I used the same priors as previously, so that the results from the multiple grab MCMC could be compared directly to those already calculated.

To implement this in WinBUGS required little change to the code (see Appendix D.3). For the multiple grabs data I ran 100,000 iterations using a 50,000 burn-in period to allow for convergence. Figure 4.17 shows the posterior for the multi-grab method and the pooled data method, and the distributions are fairly similar, with that of the pooled data slightly positively skewed. The summary statistics are shown in Table 4.5.

There was not much difference in the posterior using pooled versus multiple grabs. The species richness estimate for the pooled data was slightly higher, with a wider confidence interval. This indicated a less precise estimate in this instance, with more bias. However the data set was simulated from negative binomial, so I considered a data set with more variation, simulated from the Matérn distribution, to see if this

| Method | Mean | Median | SD | 95% Credible interval |
|---|---|---|---|---|
| Pooled data | 221 | 216 | 27.92 | (175, 273) |
| Multiple grabs | 238 | 228 | 39.62 | (177, 319) |

*Table 4.5:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ for the negative binomial model applied to negative binomial simulated data using pooled data and multiple grabs. The data were simulated with $N = 200$.

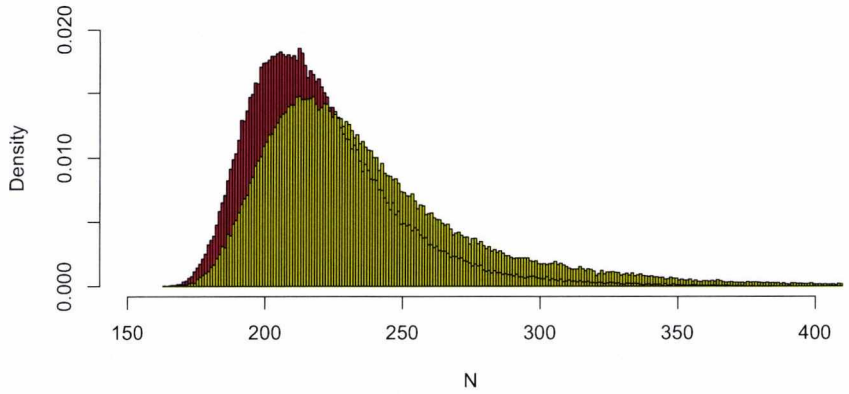



*Figure 4.17:* Posterior density for $N$ for the negative binomial model fitted to the negative binomial simulated data set in `WinBUGS` using DA. Pooled data in red, multiple grabs in yellow. The data were simulated with $N = 200$.

showed a greater difference between the two methods.

*Matérn simulated data*

I again simulated a population of 200 species, with 75% sampling depth. I used the Matérn process and allowed the mean abundance to vary between species. I used parameters of $\alpha = 1, \beta = 1, \phi = 5$, and $R = 0.1$ and set the number of grabs at ten. So the mean number of individuals per unit area, $\lambda\phi$, for each species, was simulated from a Gamma$(\alpha, \beta)$, and the number of parents, $\lambda$, was equal to the mean over the number of children per parent, where the number of children per parent is distributed Poisson$(\phi)$. $R$ is the radius of the circle around the parent in which the children are located. So in this simulation all the species had an equal number of children per parent.

Using a 50,000 burn-in, and 50,000 iterations I obtained the posteriors summarised in Table 4.6 (Figure 4.18) for the pooled data and multiple grabs using the negative binomial model. Again, there was little difference in the results and I can see no advantage to using multiple grabs rather than pooling data.

| Method | Mean | Median | SD | 95% Credible interval |
|---|---|---|---|---|
| Pooled data | 154 | 154 | 2.12 | (150, 159) |
| Multiple grabs | 154 | 153 | 1.99 | (150, 158) |

*Table 4.6:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ for the negative binomial model applied to Matérn simulated data using pooled data and multiple grabs and DA. The data were simulated with $N = 200$.

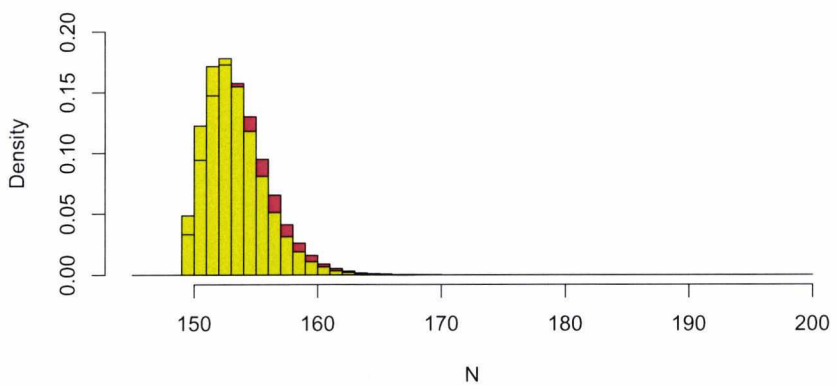



*Figure 4.18:* Posterior density for $N$ for the negative binomial model fitted to Matérn simulated data set fitted in `WinBUGS` using DA. Pooled data in red, multiple grabs in yellow.

I extended the model to account for species distribution as well as abundance, replacing the Poisson by the Neyman Type A distribution. More priors were specified. I already had a gamma prior for the mean abundance of the Neyman Type A, $\mu = \lambda\phi$, for which I have uninformative half-Cauchy hyper-priors. So I specified a prior on one of the Neyman Type A parameters, $\phi$, the clustering parameter, as a gamma prior, again with uninformative half-Cauchy hyper-priors. I fixed $\phi$ across species as this was how I simulated the data. Appendix D.4 shows the code used to fit the model.

| Method | Mean | Median | SD | 95% Credible interval |
|---|---|---|---|---|
| Pooled data | 187 | 178 | 29.48 | (156, 265) |
| Multiple grabs | 187 | 184 | 14.99 | (162, 216) |

*Table 4.7:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ for the Neyman Type A-gamma model applied to Matérn simulated data using pooled data and multiple grabs and DA. The data were simulated with $N = 200$.

Table 4.7 shows that the Neyman Type A-gamma model gave a better estimate for the true species richness of 200. The 95% Credible interval for the negative binomial model was very narrow, and did not include the true species richness, however the Neyman Type A-gamma model did. The main difference in the results was that the credible interval for the pooled data was wider, suggesting that using multiple grab data was more accurate.

### *4.7.5 The non-hierarchical likelihood versus the hierarchical Bayes method*

Comparing the results between the non-hierarchical likelihood and the hierarchical Bayes method for the negative binomial model, where a uniform prior was used for $N$ and data were pooled over grabs, we see that the two methods produce very similar results as we would expect (Table 4.8). Speed and number of iterations required were roughly equal, therefore it is our choice to use whichever method is easiest to implement for each data set when using the negative binomial model.

## *4.8 Analysis of data*

I investigated the performance of the Bayesian approach to species richness estimation using simulations, but now present the results of applying the models to real data examples of the Lepidoptera data and Christmas Bird Count data.

| Method | Mean | Median | SD | 95% Credible interval |
|---|---|---|---|---|
| Non-hierarchical | 223 | 215 | 34.55 | (173, 289) |
| Hierarchical Bayes | 221 | 216 | 27.92 | (175, 273) |

*Table 4.8:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ for the negative binomial model applied to negative binomial simulated data using non-hierarchical and hierarchical Bayes (DA) methods. The data were simulated with $N = 200$.

### *4.8.1 Lepidoptera data*

The data set used was the Lepidoptera data, containing 15,609 individuals, 240 species and maximum frequency 2,349. The data were truncated at a point $\tau = 112$, as in Barger and Bunge (2010), and I did the same so that the results were comparable.

Results are presented from the Poisson, Poisson-exponential, negative binomial and Neyman Type A-gamma models. The priors that were used were uniform on $N$, with uninformative half-Cauchy priors on the nuisance parameters.

Results are shown in Table 4.9. The negative binomial estimates of species richness were slightly higher than those found by Barger and Bunge (2010) when using the Jeffrey's prior of 342 with SE = 666.9, and 95% HPDI (272, 910), although our model had a much narrower credible interval. I used the uniform prior on $N$ for ease of computation, but I expected to obtain a similar estimate as it was shown in Section 4.7.2 that the choice of uninformative prior had little effect on the estimate for the negative binomial model.

However, Barger and Bunge (2010) posterior estimate using the reference prior was less than 300 using a mixture of two exponentials for the abundance distribution instead of a gamma, so I did not expect the results to be exactly the same. This estimate lies between the estimates obtained by the Poisson-exponential and negative

| Method | Mean | Median | SD | 95% Credible | B. $p$-value | DIC | $\Delta$DIC |
|---|---|---|---|---|---|---|---|
| PO | 240 | 240 | 0.09 | (240, 240.23) | 0.00 | 2656.29 | 2901.30 |
| PE | 252 | 252 | 3.52 | (245, 259) | 0.52 | -237.49 | 7.63 |
| NB | 357 | 344 | 60.03 | (275, 469) | 0.48 | -245.12 | 0.00 |
| NTAG | 450 | 445 | 223.95 | (292, 1130) | 0.25 | -236.78 | 8.34 |

*Table 4.9:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ for the Poisson, PO, Poisson-exponential, PE, negative binomial, NB and Neyman Type A-gamma, NTAG, models applied to the Lepidoptera data using the non-hierarchical approach. As a measure of fit, the Bayesian $p$-value is also reported.

binomial models in Table 4.9.

The Bayesian $p$-values showed the best fitting distributions were the Poisson-exponential and the negative binomial. The DIC judged the negative binomial as the best model, although the Poisson-exponential was plausible. The negative binomial was also found to be the best model in terms of AIC in the frequentist approach. The Neyman Type A-gamma was a poorer fit to the data set, and the confidence interval was wide compared to those of the other models.

### *4.8.2 CBC data*

For the analysis of the CBC data, where 126 species were seen, results are presented from the Poisson, Poisson-exponential, negative binomial and Neyman Type A-gamma models. The priors that were used were again uniform on $N$ with uninformative half-Cauchy priors on the nuisance parameters. Results are shown in Table 4.10. Truncation of the data was at 221, except when fitting the Neyman Type A-gamma model, which cannot cope with abundance values of over 150 due to limitations within the software when calculating the probabilities of the Neyman Type A distribution. Therefore, data above 150 were truncated and added to the results after MCMC.

| Method | Mean | Median | SD | 95% Credible | B. $p$-value | DIC | $\Delta$DIC |
|---|---|---|---|---|---|---|---|
| PO | 126 | 126 | 0.05 | (126, 126.1) | 0.00 | 3021.29 | 2898.02 |
| PE | 129 | 129 | 1.27 | (126, 132) | 0.50 | 123.27 | 0.00 |
| NB | 166 | 160 | 31.24 | (131, 227) | 0.45 | 131.63 | 8.36 |
| NTAG | 241 | 238 | 88.74 | (133, 450) | 0.23 | 139.50 | 16.23 |

*Table 4.10:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ for the Poisson, PO, Poisson-exponential, PE, negative binomial, NB and Neyman Type A-gamma, NTAG, models applied to the CBC data using the non-hierarchical approach. As a measure of fit, the Bayesian $p$-value is also reported.

According to the Bayesian $p$-values, the best fitting distribution was the Poisson-exponential. The Poisson was not a good fit at all, with a Bayesian $p$-value of 0. The Neyman Type A-gamma was also not a good fit to the data, agreeing with the conclusions of the frequentist analysis in Chapter 3.

The DIC judged the Poisson-exponential to be the best model, and the negative binomial also to be plausible. The Poisson model did not describe the data well at all.

Barger and Bunge (2010) found a species richness estimate of approximately 140 (130,160) for the CBC data using a finite mixture of two exponentials, which is higher than the estimate using the Poisson-exponential, but lower than the estimate using the negative binomial.

### *4.8.3 Benthic data*

I attempted to analyse the benthic data using non-informative priors on $N$ and the nuisance parameters. However, as seen in Section 4.7.2, when the probability of seeing a species falls below a particular level, the MCMC algorithms will not converge using non-informative priors.

An example of this is shown for the Hastings data set (Figure 4.19). When using the reference prior the MCMC algorithm had not converged even after ten million iterations using the negative binomial model. The chains moved wildly around the parameter space. The same pattern was apparent when using the uniform

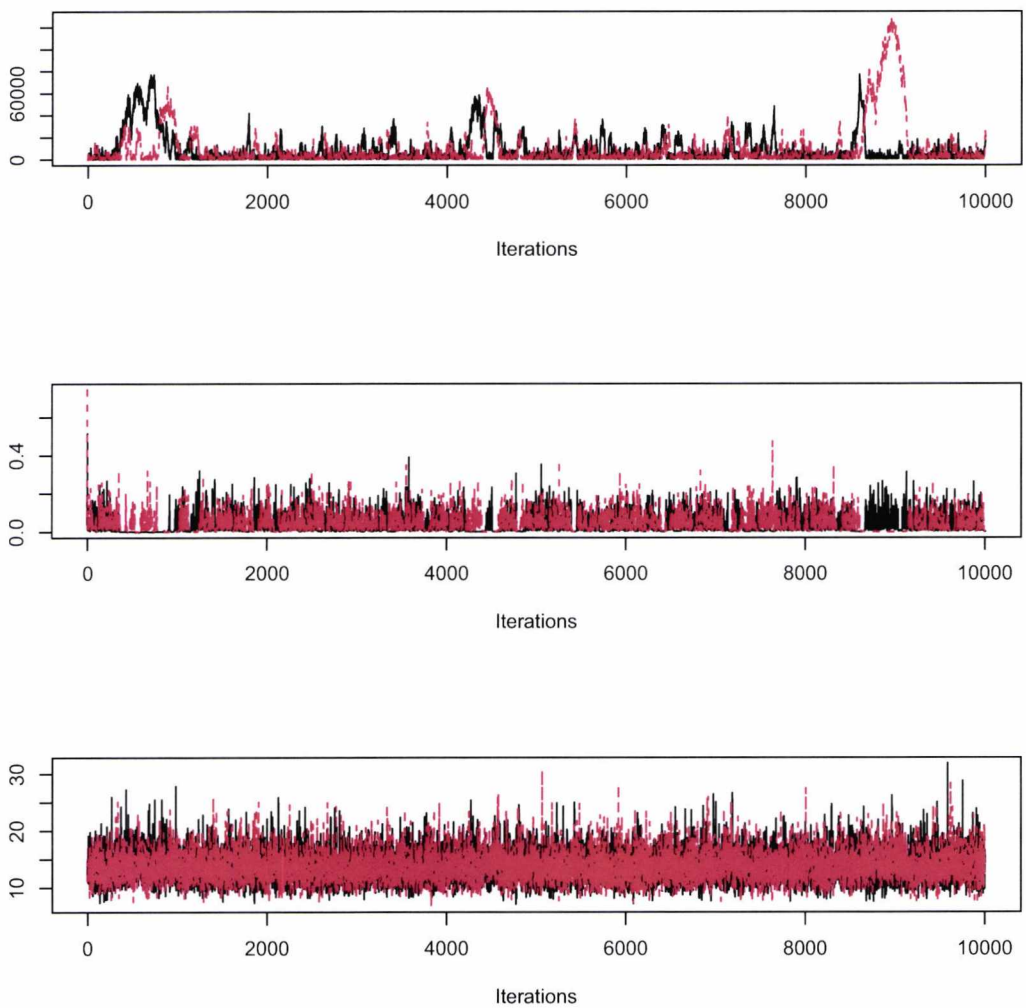



*Figure 4.19:* Trace plots for $N$, $\alpha$ and $\beta$ for the negative binomial model fitted to the Hastings data using reference prior for $N$ and uninformative priors on the nuisance parameters (iterations shown in thousands).

and Jeffrey's priors for $N$, and also for alternative benthic data sets. Using the Neyman Type A-gamma did not solve the problem, as we might expect. I believe this lack of convergence was attributed to the boundary problem.

## 4.9 Discussion

This chapter aimed to explore parametric Bayesian approaches for species richness estimation using noninformative priors, and extend them for the Neyman Type A model. I briefly summarised both parametric and non-parametric Bayesian approaches to the problem from the literature, and proceeded with a parametric Bayesian approach.

The Bayesian methodology was described, and the reference and Jeffrey's prior were introduced, along with the models to be used. I demonstrated some of the properties of the alternative approaches to model data with unobserved species using simulated data. I then analysed the Lepidoptera data and CBC data, using various models, and demonstrated the boundary problem appearing as a lack of convergence in MCMC chains for the benthic data.

After exploring Bayesian species richness methods, I found the Bayesian approach to species richness estimation to be flexible. There were a number of approaches to model data with unobserved species that were outlined in the literature, including data augmentation and RJMCMC, and I have applied these using the negative binomial model and extended them to the Neyman Type A-gamma model.

I found some interesting results from the investigation using simulated data. Single-update MH was the preferred MH method, because although fewer iterations were required using block updates, the block update method required a pilot study using the single-update MH to estimate posterior correlations and variances, so block-updates did not seem beneficial.

I showed that when using data with high probability of seeing a species, the choice of uninformative prior applied to $N$ was arbitrary. However, when I decreased the probability of seeing a species, the MCMC algorithm failed to converge. Therefore, if we wish to estimate species richness for data sets where a high proportion of the

species have not been recorded, an uninformative prior may not be sufficient to ensure convergence of the MCMC chains.

The non-hierarchical likelihood and the hierarchical Bayes method for the negative binomial model produced very similar results in terms of the posterior distribution, but also in speed of calculation and number of iterations required. Therefore, I would suggest personal choice of whichever method is easiest to implement for each case. However, to incorporate varying clustering between species, it would be advantageous to use the hierarchical approach and view the variation between species as random effects. This allows for more straightforward set up within `WinBUGS`.

Within the hierarchical Bayes framework, the data augmentation method and RJMCMC gave essentially indistinguishable results, although in contrast to Schofield and Barker (2010), DA was faster and required fewer iterations than RJMCMC. When using these methods it was important to consider alternative specifications of distributions to avoid errors within `WinBUGS`, and to set initial values to get realistic posterior values.

We might expect that when we use data from multiple grabs, rather than pooling data across grabs, the species richness estimate would improve. However there was not much difference in the posterior distributions of the two methods.

When I investigated the performance of the Bayesian approach to species richness estimation applied to real data examples, the estimate of species richness using the negative binomial model appeared high compared to the results of Barger and Bunge (2010) for the Lepidoptera data. However, this model was judged the best of the models investigated. The best estimate was also higher than the estimate found by Barger and Bunge (2010) for the CBC data, and the best fitting model I fitted here was the Poisson-exponential. According to the Bayesian $p$-values the Neyman Type A-gamma model did not fit the data well, and this could be due to the spatial distribution

of birds and butterflies, which may not be found in clusters. In addition, alternative sampling methods such as transects may not fit a clustering model.

When applying the species richness models to benthic data I found that when using an uninformative prior, the MCMC chains did not converge even after millions of iterations. I supposed that this was due to the boundary problem that arises when the probability of seeing a species falls below a particular level. A standard consideration in the use of the Bayesian approach is the requirement to specify a prior of an appropriate form. In ecological applications the prior is a convenient way to incorporate expert opinion or information from previous or related studies. Chapter 5 considers the Bayesian approach to estimating species richness using priors that have been informed using expert opinion. Using an informative prior on $N$ should combat the non-convergence of the MCMC chains that we have seen when analysing the benthic data with uninformative priors.

Software has been developed to perform Bayesian analysis of complex statistical models using MCMC methods including `WinBUGS`, which not only performs MCMC, but also incorporates several diagnostic tools. However, not all distributions, including the Neyman Type A, are incorporated in `WinBUGS`, and for more flexibility and control one might prefer to use bespoke code in `R` to run MCMC.

Additional hurdles to application of the parametric Bayesian approach include the justification of the use of specific priors, and, as in the frequentist approach, model selection. To mitigate the effect of choosing priors, a hierarchical framework can be adopted, which allows more uncertainty in the model.

## 4.10 Conclusions

I explored parametric Bayesian approaches for species richness estimation using noninformative priors, and extended them for the Neyman Type A-gamma model using a hierarchical and non-hierarchical approach, which has not previously appeared

in the literature for this model. I analysed data sets using this model, however uninformative priors were not sufficient to model benthic data, and the boundary problem arose again through lack of convergence of chains within MCMC.

The difficulty in analysing benthic data has been highlighted, and it is expected that the use of informative priors will combat the boundary problem and that this will allow us to obtain realistic species richness estimates for the study areas. Chapter 5 considers the elicitation of informative priors for benthic data and how these can be incorporated into a Bayesian framework.

# 5. ELICITATION OF INFORMATIVE BAYESIAN PRIORS

## *5.1 Introduction*

In Chapter 4, a Bayesian approach using uninformative priors was unable to estimate species richness for benthic data. This chapter considers the Bayesian approach to estimating species richness using informative priors.

Recall that there are two situations that could occur. Firstly that there is no prior information, and then I proceed with the method outlined in Chapter 4 using non-informative priors. Secondly, there is prior information which needs to be expressed in the form of a suitable probability distribution. The use of an informative prior will have an effect on the posterior, which is proportional to the product of the likelihood and the prior. The objective was to elicit priors for benthic data with the help of CEFAS scientists. Using an informative prior on $N$ would hopefully contribute more information to the posterior for large proposed values of $N$ than the objective priors used previously, and stop the spurious estimates associated with the boundary problem. I could also elicit information on suitable informative priors for the nuisance parameters.

This chapter summarises the current literature and techniques used in elicitation, and then presents an outline of the process used to elicit information for benthic data. I present a summary of the pilot study that was undertaken, and describe the main elicitation process. The results of the process, and converting the elicited information to priors will be discussed, and I present the results of using the elicited priors in a Bayesian analysis of several benthic data sets.

## 5.2 Eliciting expert knowledge

Expert knowledge is used to solve problems every day, from large scale management decisions, to deciding what to wear. There is a growing trend to utilise expert judgements in a more structured and formal way to inform decisions and processes, in areas such as conservation science and landscape ecology (Martin et al., 2012; McBride and Burgman, 2012).

Although not a new topic, there has been a recent surge of interest in eliciting expert knowledge, with many papers published and software developed for the purpose of elicitation over the last year or so (Martin et al., 2012; McBride and Burgman, 2012; Fisher et al., 2012; Burgman et al., 2011; Kuhnert, 2011). However, there is little literature on the use of expert knowledge to inform models involving benthic organisms (Allan et al., 2011), and none is apparent for marine benthos.

Expert knowledge can be the result of not only personal experience, but also training, research and skills; and what counts as expertise often depends on the context (Burgman et al., 2011). Expert judgements can be used directly to inform management decisions or indirectly to provide information about model parameters when data are scarce. However, there is some controversy surrounding the use of expert knowledge because of its subjective and potentially biased nature (Kuhnert, 2011).

If little information is available regarding a system, then expert opinion can be indispensable, and as long as the elicitation process is undertaken rigorously and is carefully structured into a model, it has a vital role in ecological analysis. Use of expert knowledge could be criticised because of the difficulty of testing or evaluating the accuracy of the informed models. However, in cases when appropriate data are difficult to obtain it can be the only reasonable option (Allan et al., 2011).

The elicitation process generally includes several steps: deciding how information will

be used, determining what to elicit, designing the process of elicitation, performing the elicitation and translating the information to be used into a model (Martin et al., 2012). How the information will be used will determine what variables to elicit, and the best process to obtain the information. Discussion of the questions with experts, and training in the elicitation format and procedure should help to manage bias and subjectivity in the results. This can be checked by performing a pilot elicitation.

Additional steps can be taken to reduce bias and uncertainty in elicited judgements. In a situation where participants estimate an unknown value, by starting from some initial value which is then adjusted to yield a final answer, different starting points can typically yield different estimates, which are biased towards the initial value. This is the phenomenon of anchoring (Tversky, 1974). Subjective probability distributions can be obtained by using one of two formally equivalent procedures: by asking the subject to select values that correspond to specified percentiles of the probability distribution, or by asking the subject to assess the probability that the true value of the quantity will exceed some specified values (Tversky, 1974).

Other sources of bias may be motivational, if the expert has a personal stake in a decision, or accessible, when judgement is influenced by the information that comes more easily to mind (Martin et al., 2012). Some of this bias can be overcome by asking experts to evaluate rather than produce intervals, and also by ensuring regular systematic feedback (Martin et al., 2012). Bias could also arise from experts' overconfidence in their judgements, and a possible mitigation technique for this source of bias would be to ask the same question more than once with alternative wording, to ascertain uncertainty (Martin et al., 2012).

Multiple experts can be involved in the elicitation procedure, in order to obtain a group consensus. Stakeholders from a variety of backgrounds may have differing opinions on a topic, so this could be a complicated process. Methods have been developed to ease this process, including the Delphi method (Linstone and Turoff,

1975), which elicits individual estimates from experts and allows each one to adjust their estimate in light of the response of others by means of successive iterations of given questions (Grupp and Linstone, 1999).

An alternative approach to incorporate the opinions of multiple experts is Cooke's method, in which the opinion of each expert is weighted on the basis of its accuracy (Cooke and Goossens, 2004). This involves the expert answering a set of test questions, the accuracy of answers to which is used as a basis to weight their judgement. However, there is a trade-off between the number of variables that can be elicited accurately and the need to retain the experts' focus (Martin et al., 2012).

## 5.3 Elicitation of priors within a Bayesian framework

I wanted to build expert knowledge into the models used for estimating species richness in Chapter 4. To choose a suitable family of prior distributions for each model parameter, and select parameters for those priors, I first obtained an idea of plausible values from experts. Once the values were elicited they could be used to inform priors in a Bayesian framework.

An approach to elicit information on the distribution of parameters, following the methodology of O'Hagan (1998), is firstly to get estimates of upper and lower bounds, $U$ and $L$, and the most likely value, the mode $M$, for a parameter. Then experts are asked to give probabilities for the quantity lying in the following intervals, denoted $p_1, ..., p_5$ respectively:

1. (L,M),
2. (L,(L+M)/2),
3. ((M+U)/2,U),
4. (L,(L+3M)/4),
5. ((3M+U)/4,U).

The questions are chosen and asked in this order to avoid asking experts to assess very small probabilities and to avoid problems of anchoring (O'Hagan, 1998).

These elicited probabilities are then used to calculate six probabilities that can be used to fit a distribution using the least squares method.

$$\begin{aligned}
q_1 &= P\left(L \leq X \leq \frac{L+M}{2}\right) = p_2, \\
q_2 &= P\left(\frac{L+M}{2} \leq X \leq \frac{L+3M}{4}\right) = p_4 - p_2, \\
q_3 &= P\left(\frac{L+3M}{4} \leq X \leq M\right) = p_1 - p_4, \\
q_4 &= P\left(M \leq X \leq \frac{3M+U}{4}\right) = 1 - p_1 - p_5, \\
q_5 &= P\left(\frac{3M+U}{4} \leq X \leq \frac{M+U}{2}\right) = p_5 - p_3, \\
q_6 &= P\left(\frac{M+U}{2} \leq X \leq U\right) = p_3.
\end{aligned}$$

Least squares minimises the sum of squared differences between the observed values and the fitted values. If $\mathbf{Y}$ is a vector random variable of dimension $d$ with expectation $\mu(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters of the distribution to be fitted, then we can estimate $\boldsymbol{\theta}$ by minimising

$$S(\boldsymbol{\theta}) = \{\mathbf{y} - \mu(\boldsymbol{\theta})\}'\{\mathbf{y} - \mu(\boldsymbol{\theta})\} = \sum_{i=1}^{d} \{y_i - \mu_i(\boldsymbol{\theta})\}^2 \tag{5.1}$$

(Morgan, 2009, p150). This can be minimised using function `optim` in `R`, or the least squares function `nls`.

An alternative approach to elicit the same underlying distribution, as mentioned in Section 5.2, is to ask the subject to select values that correspond to specified percentiles of the probability distribution, such as the quartiles of the distribution. In this approach we would elicit estimates of upper and lower bounds, and the median of the parameters of interest.

The elicited distributions can then be used within a Bayesian framework as priors.

## 5.4 Tools to aid elicitation

The focus on the subject of eliciting expert knowledge in recent years, along with the increase in use of technology, has led to the development of various tools to aid the elicitation process.

The Sheffield Elicitation Framework (SHELF) (Oakley and O'Hagan, 2010) is a formal procedure for elicitation. Various computer packages have been developed to aid the elicitation process including within SHELF (Oakley and O'Hagan, 2010), `Elicitator` (James et al., 2010), and a new package `ElictN` in `R` for eliciting expert knowledge about species richness (Fisher et al., 2012).

The Elicitator (http://elicitator.uncertweb.org/) is a website where one can upload problems and get experts to log on and give their opinions. It is necessary to specify the experts before uploading and defining the problem. There are benefits to specifying the experts before considering the approach, however this does not allow the elicitor to get a feel for the process and there are no clear instructions available. Thus this is only any good if the experts are inaccessible; otherwise a face-to-face process would be preferable.

`Elicitor` (Kynn, 2005) implements an indirect approach to elicitation for Generalised Linear Models, and is implemented as a module within `WinBUGS`. The elicitation tool `Elicitator` (James et al., 2010) assists in quantifying expert knowledge for use as a prior model in Bayesian regression, using an alternative to estimating probabilities. Instead, experts estimate the response in a regression for a set of cases given covariates corresponding to each case. This elicitation approach suits experts who are less comfortable estimating probabilities, but is only applicable to regression problems.

An exciting recent development is the `ElictN` software developed in `R` by Fisher

et al. (2012). This is a template developed for eliciting expert knowledge of species richness, and uses coral reefs as a case study. This software is designed to support elicitation of knowledge from a single expert, and has an interface withinR that allows data to be entered directly by the elicitor to generate graphical output for feedback to the expert during elicitation.

The approach of Fisher et al. (2012) is based on taxonomy and estimates the total species richness of particular taxon, by breaking it down into subcomponents based on the number of species discovered and named, the number of species discovered and unnamed and the number of species yet to be discovered. `ElictN` is designed to estimate species richness within this framework. However the benefit of the software is that a simplified version is available, which can be modified to suit any application where elicitation is desired for frequency or count data (Fisher et al., 2012). The simplified version of the software is designed to be modified for the elicitation of species richness in any ecosystem, at any spatial scale by someone with only limited knowledge of programming in `R` and the `tcl` programming language. Therefore, it could prove a useful tool in future elicitation of information on benthic species.

## 5.5 Pilot study

### 5.5.1 Method

Considering the Bayesian framework for which I wished to elicit priors, elicitation was carried out for questions in three areas: number of species, abundance of species and clustering.

I produced an elicitation spreadsheet in `Excel` on which the experts carried out the exercise, since I was unable to be present at the time. The pilot was carried out with Jon Barry of CEFAS as facilitator. Three scientists were involved as experts, and each considered a different area when completing the elicitation.

I used `Excel` as most ecologists are familiar with the software. The spreadsheet contained a brief introduction to myself and the aim of the exercise. Then there were cells to fill in for each elicitation, to give probabilities for the quantity lying in particular intervals, following the framework of O'Hagan (1998) mentioned in Section 5.3. An example of this was shown on a second sheet of the `Excel` workbook.

For the number of species, which is the key variable for which I hoped to elicit a prior, an estimated frequency graph was produced in `Excel` during the process to aid understanding, and offered the experts the chance to modify their estimates. Due to the limitations of using `Excel` this was not an estimate of a fitted distribution, but a representation of the probabilities given.

After the spreadsheet was completed, I used the elicited probabilities to fit a smooth density function, and plots and summaries were produced and shown to the experts via email. The experts were then asked to provide feedback on these distributions, indicating whether they were consistent with their beliefs.

### 5.5.2 Distribution fitting to the elicited information

Chapter 4 showed that the hierarchical Bayes data augmentation (HBDA) approach was useful for estimating species richness. However, it was restricted to implied priors for $N$ that were based on a binomial distribution. In the HBDA approach, $N$ was a derived parameter, such that $N = \sum_{i=1}^{N_S} z_i$, where the inclusion parameters $z_i \sim \text{Bernoulli}(\psi)$ and $N_S$ was the size of the super-population. This results in $N \sim \text{Binomial}(N_S, \psi)$ and $\psi \sim \text{Uniform}(0, 1)$. Within this approach $N$ had a discrete uniform prior.

To use a more informative prior for $N$ I extend this approach by letting $\psi$, be a random variable with a beta distribution. Then removing $\psi$ from the joint prior $p(N|N_S, \psi)p(\psi)$ by integration:

$$\int \text{Binomial}(N|N_S, \psi)\text{Beta}(\psi|\alpha, \beta)\text{d}\psi \tag{5.2}$$

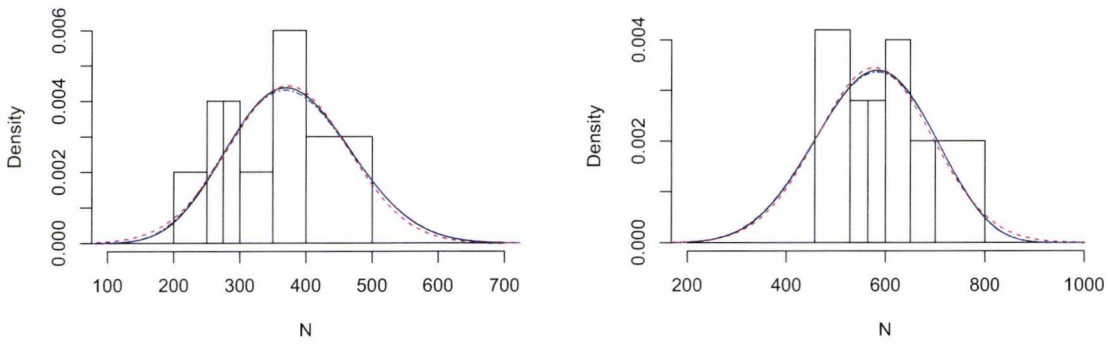
gives a beta-binomial prior for $N$.

Therefore, I fitted a beta-binomial distribution to the elicited probabilities for the number of species present in the population for the pilot study. This prior was flexible as it can also be used in the non-hierarchical Bayesian method. I fitted the distribution in R by minimising the sum of the squared differences between $q_1, ..., q_6$ and the corresponding probabilities $q_1^*, ..., q_6^*$ (See Section 5.3) implied by the beta-binomial distribution for each of the areas.

I also fitted a scaled beta distribution and a normal distribution to the elicited information, because it may be preferable to use an alternative prior. When using the non-hierarchical Bayesian method it is possible to select from a range of priors. I can then select the best-fitting prior to use on $N$ in the species richness estimation.

Figure 5.1a shows the elicited distributions for the Eastern Channel, and Figure 5.1b for the Hastings Shingle Bank. For the Eastern Channel, the elicited estimate of the mode of 300, upper bound of 500 and lower bound of 200 did not correspond well with the fitted distributions. However, for the Hastings Shingle Bank, although the mode and upper bound of 800 fitted the elicited distribution well, the lower bound of 457 was too high compared with the distributions shown in Figure 5.1b. As this figure was exact, I assumed that it was the number of species that had been observed in the area. So without prior knowledge of this value, perhaps the fitted distribution would have reflected what one would expect before undertaking a survey.

Adjusting the distribution to see if different bounds could fit the experts beliefs more accurately, could improve the fit of the distribution. This is described as 'stepping back' by (O'Hagan, 1998), and involves adjusting the distribution and consulting the experts as to if the adjusted distribution represents their beliefs better than the original elicited distribution. A 'stepping back' step would likely be useful here to gauge what the experts' beliefs really were, especially as there was little feedback




*(a)* Eastern Channel *(b)* Hastings Shingle Bank

*Figure 5.1:* Prior distributions for $N$, elicited from benthic experts during the pilot study for (a) the Eastern Channel and (b) the Hastings Shingle Bank. The elicited information is shown in the form of a histogram, and three distributions have been fitted: the beta-binomial (solid black line), normal (dashed red line) and scaled beta (dashed blue line).

during the actual elicitation process.

Although the fits of the elicited distributions to the elicited information were poor, it would be possible to use them as priors for the number of species within MCMC. However, the elicited priors did not always correspond to the species abundance data that I hoped to use. However, at least we can see that the priors used for the Bayesian analysis should be very different to the non-informative ones.

In the Hastings Shingle Bank sampling programme, there were 353 species found in the survey area between 2001 and 2004. The lower bound as given by the expert was 457, which despite being larger than the observed number of species in the survey of interest, does not seem unreasonable if the expert had knowledge of the particular area. However, in the Eastern Channel data set, 649 species were observed over 225 grabs, but in the elicitation the mode for the number of species was given as 300, with lower bound 200 and upper bound 500. When questioned, the expert explained that there are many data sets available within each area, and within a single sampling

program it is usual that around 300 species would be found over multiple grabs.

One of the elicitation results was not useable, because the upper and lower bounds were mixed up, and the probability estimates did not make sense, in the way that, for example, the probability $N > 50$ was less than the probability $N > 75$. I was unable to deduce what the expert meant, and unfortunately did not have the opportunity to clarify it further. I put the issues down to unfamiliar language of bounds used on the spreadsheet, and also a lack of feedback to the expert during the process. There was no facilitator present during this pilot elicitation.

### *5.5.3 Issues arising from the pilot study*

There were several issues that arose from the pilot study. O'Hagan (1998) highlighted the importance of making the elicitation process simple and familiar for experts in terms of language and quantities used. I attempted to achieve this, but issues arose during the process that made it clear that I had not succeeded entirely. The first was that some estimates by the experts of number of species were made per sample not for the population as a whole.

In some cases, the results produced did not tie in with the data sets that I was thinking of when I asked for the elicited information. More clarity was required to avoid a misunderstanding. More information given to experts prior to the elicitation and a discussion of what I hoped to achieve could have gone some way to improving clarity. Being there in person to facilitate elicitation and assist with any queries would have improved understanding of the task. These issues highlighted the importance of the 'facilitator'.

One interesting issue that arose was the definition of species richness. One scientist said that the answer to the number of species present in the survey was highly dependent on the sampling method, sample processing, and what fraction of the fauna you are targeting (Megafauna, Macrofauna, Meiofauna, or bacteria). It was

clear that I did not define the problem well enough and I should have been present to clarify that these estimators are being developed for data on Macrofauna greater in size than $1mm$.

The elicitation process could also have been improved through use of interactive computer software. This enables experts to verify their ideas, by gaining feedback on their opinions and values adjusted if required. However, I did not have the resources to achieve this during the pilot study. This point reaffirms the need for further work in elicitation, highlighted by O'Hagan (1998), in developing general-purpose software to encourage serious elicitation.

Using interactive computer software would also have avoided confusion filling in the elicitation form, and the definition of terms such as upper and lower bounds, and clustering. This would be reaffirmed with a facilitator present.

For each of the benthic data sets considered, a different expert was consulted. It might have been worthwhile to use more than one expert per data set and gain a consensus to improve the elicitation process, but this was not possible within the scope of this pilot study.

In this pilot study I fitted the beta-binomial distribution along with the scaled beta and the normal distribution to the elicited information about the number of species. Using a beta-binomial prior I am able to apply the HBDA method. However, the non-hierarchical Bayesian method produced valid estimates and had a similar computation time to the HBDA method, and using this approach I could use a normal prior on $N$ to simplify computations in terms of the acceptance probability within the MCMC.

## 5.6 Elicitation using SHELF

Many important issues were highlighted during the pilot study. One of these was the importance of feedback during the elicitation process. In addition, the procedure must

be rigorous to ensure validity of the results. I therefore decided to complete the formal elicitation process following the Sheffield Elicitation Framework (SHELF) (Oakley and O'Hagan, 2010), and adapting it to my needs. SHELF is a formal procedure for elicitation, including briefing documents, and R software to utilise during the process.

### 5.6.1 Method

In the pilot study, following O'Hagan (1998), I elicited information on probabilities, which some of the scientists had difficulty with. Therefore where necessary in the main elicitation procedure I elicited information using the quartile (Q) method in SHELF. After defining the bounds of the distribution, the experts were asked for their estimates of the median, and upper and lower quartiles. This method gives adequate distributions, while keeping the process simple. The percentile (P) method was used when experts were more comfortable working with probabilities; in which case a probability was elicited for the intervals $(L, (2M + L)/3)$ and $((2M + U)/3, U)$ after defining the bounds and the median.

Each participant was given a pre-elicitation briefing, to make clear what I hoped to achieve and clarify any notation. This also included a brief presentation of my work so far and how the information elicited would contribute, including an example on how the present methods used for species richness estimation are not adequate for benthic communities.

We then decided on an area to elicit information for, and recorded the scientists' expertise in this area. I then went through an example in which I hoped to elicit the distance between Lowestoft and London, to demonstrate the process and clarify any misunderstandings. This also enabled me to determine which method (Q or P) the expert felt most comfortable with. A mixture of percentile and quartile methods was used, because some questions lent themselves better to a particular method, and it was important to keep the process user-friendly for the experts involved.

The process was carried out using the `shelf2` function in `R`. Using the SHELF software in `R` enabled me to fit distributions to the elicited probabilities while the experts were present. This had the advantage of immediate feedback, and therefore included the 'stepping back' step. Figure 5.2 gives an example of the output produced during the SHELF elicitation procedure using the quartile method. The values for the median and quartiles were adjusted during the elicitation procedure, producing the fitted distribution shown. The choice of distribution fitted could be easily adjusted during the process, and the best fitting of the available distributions could be selected, as calculated by least squares.

During the SHELF process one was able to compare the fitted distributions, and the software calculated which distribution best fitted the data. However, it was useful to consider several of these distributions with the experts and select the one which best represented their view, because often the shape of the distributions could be quite different, and the best-fitting one might not always have been the one the expert agreed with. A drawback of using the `shelf2` software was that only a limited selection of distributions were built-in, namely the normal, the Student-t, scaled beta, log normal, log Student-t and gamma distributions, which is a good range but does not include the beta-binomial which would enable use of the HBDA method.

The sampling areas considered by the experts were Hastings, Isle Of Wight (IOW), Eastern Channel, East Coast Regional Environmental Characterisation Survey (REC) and Norfolk. During the process I made clear that I wanted the experts to consider their answers over the whole area and not per sample.

I elicited information on four parameters: $N$, the number of species in the area, $G$ the number of grabs required to find a certain percentage of the observed species, $C$ the average number of individuals found in a cluster, and $U$ the total species richness of UK coastal waters.

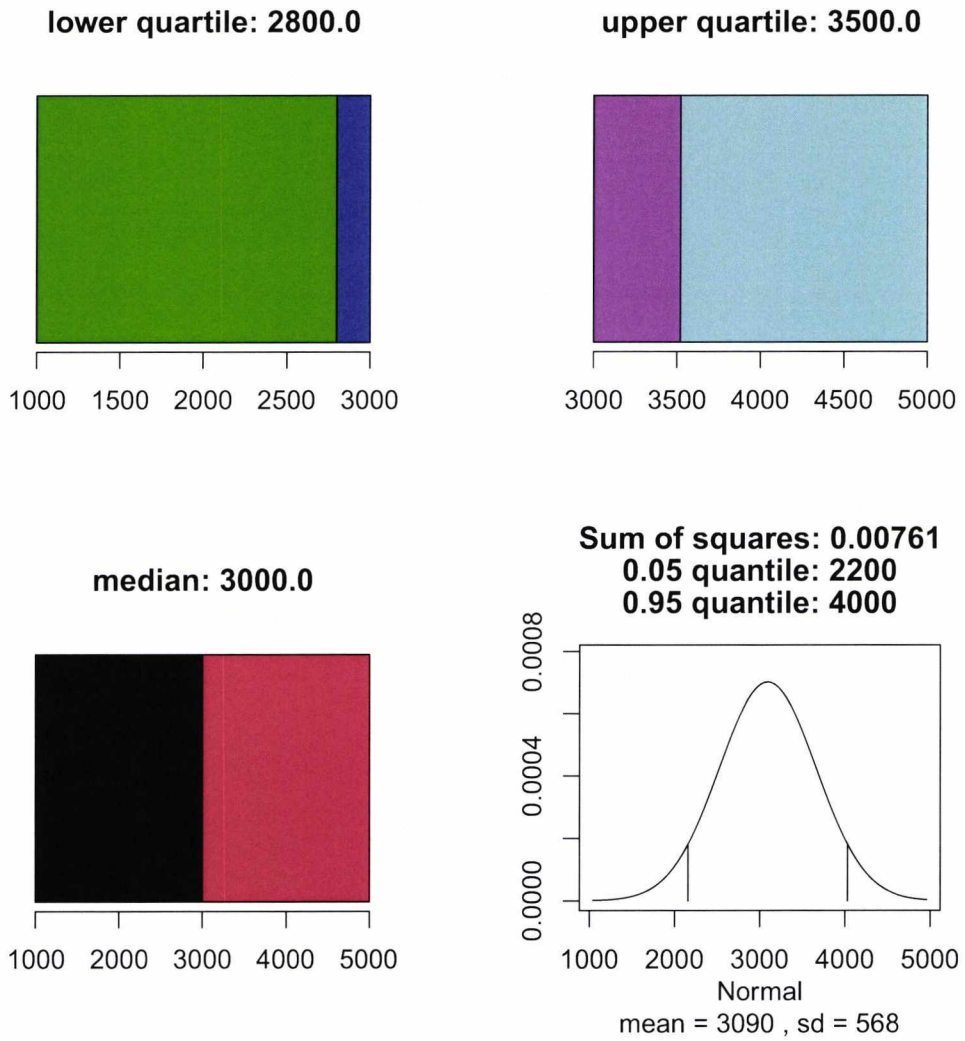



*Figure 5.2:* SHELF graphical output produced using the quartile method. First the bottom left graph is displayed and the median is elicited. Then the lower quartile and the upper quartile graphs are displayed and can be adjusted. The fitted distribution is then displayed in the bottom right.

After each distribution had been elicited, the expert was able to adjust his or her answer to make sure it reflected their beliefs. I then noted down any additional comments and finally gained feedback on the process and thanked the experts for taking part. Each elicitation meeting took approximately 40 minutes.

## 5.7 Distribution fitting to elicited information

Although the elicitation process was carried out using the `shelf2` function in `R`, the limited number of distributions available within the software meant that some fitting of additional distributions to the elicited values was necessary subsequently. Some alternative distributions could be useful if the distribution was severely skewed, and in some cases the `shelf2` function was unable to fit a distribution that reflected the expert's views, especially when the range of possible values was high, as with the clustering parameter.

The elicited information is shown in Table 5.1. Due to the two methods of collecting the information, using the quartile method or the percentile method, some cells in the table are blank. In addition, no information was elicited for $C$ for the East Coast REC.

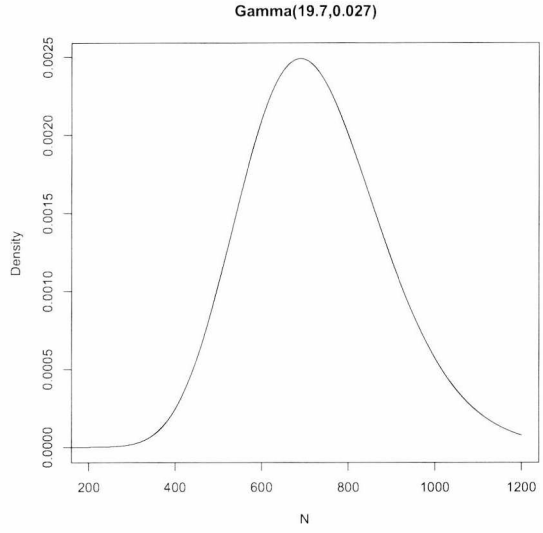
### 5.7.1 Number of species

Fitting the elicited information for $N$ gave the priors shown in Figures 5.3 and 5.4, which were the best fitting distributions during the elicitation process. Table 5.2 shows the observed number of species in each of the data sets, and we can see if the values given for the number of species during elicitation were realistic by comparing the values in Table 5.1 to those of Table 5.2.

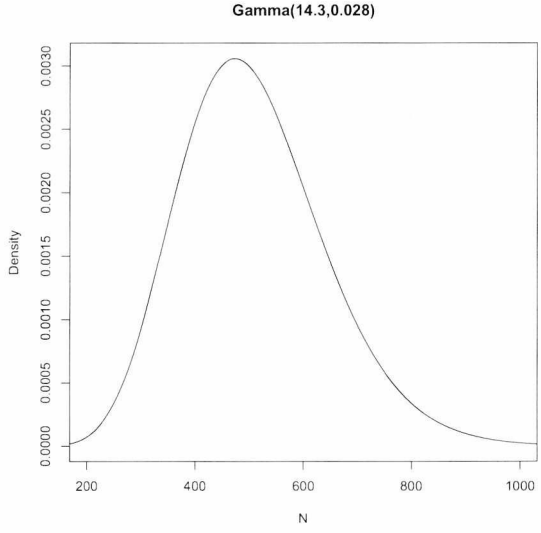
From these data, it would seem that the estimates given for the East Coast REC were below the observed number of species, and therefore were not very useful. For the Eastern Channel, the observed number of species was at the top end of the elicited distribution. The other estimates were close to the observed values, except for the

| Area | $\theta$ | M | Q1 | Q3 | A | B | prob(A) | prob(B) |
|---|---|---|---|---|---|---|---|---|
| IOW | $N$ | 750 | - | - | 590 | 800 | 0.28 | 0.36 |
| | $C$ | 0.70 | - | - | 0.47 | 0.80 | 0.15 | 0.11 |
| | $G$ | 2 | 1 | 3 | - | - | - | - |
| | $U$ | 1,500 | - | - | 1,400 | 1,600 | 0.10 | 0.34 |
| Eastern | $N$ | 500 | 398 | 626 | - | - | - | - |
| | $C$ | 10-15 | 1 | 200 | - | - | - | - |
| | $G$ | 158 | 115 | 180 | - | - | - | - |
| | $U$ | 3,702 | 3,295 | 4,364 | - | - | - | - |
| East Coast | $N$ | 220 | - | - | 210 | 250 | 0.2 | 0.1 |
| REC | $C$ | - | - | - | - | - | - | - |
| | $G$ | 3 | 1 | 5 | - | - | - | - |
| | $U$ | 2,100 | 1,000 | 2,500 | - | - | - | - |
| Norfolk | $N$ (pre) | 248 | 199 | 277 | - | - | - | - |
| | $N$ (post) | 20 | 7 | 22 | - | - | - | - |
| | $C$ | 400 | 1 | 100s/1,000s | - | - | - | - |
| | $G$ | 2 | 1 | 3 | - | - | - | - |
| | $U$ | 2,038 | 1,519 | 3,519 | - | - | - | - |
| Hastings | $N$ | 340 | 220 | 430 | - | - | - | - |
| | $C$ | 5 | 1 | 7.5 | - | - | - | - |
| | $G$ | 3 | 2 | 8 | - | - | - | - |
| | $U$ | 10,000 | 9,840 | 11,400 | - | - | - | - |

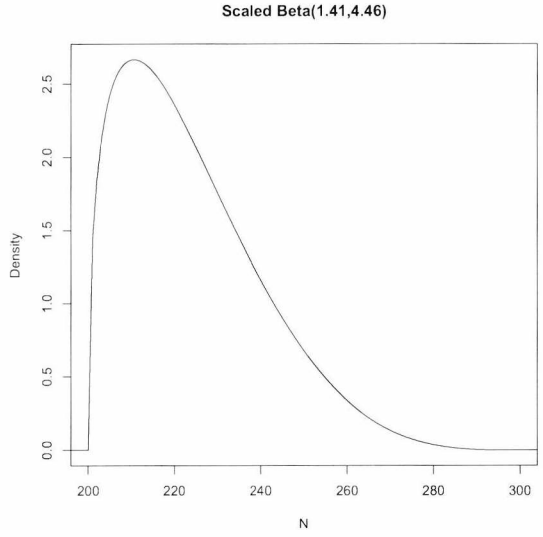
*Table 5.1:* Elicited information for the parameters: $N$ the number of species in the area, $G$ the number of grabs required to find a certain percentage of the observed species, $C$ the average number of individuals found in a cluster, and $U$ the total species richness of UK coastal waters. M: median of $\theta$, Q1 and Q3: elicited lower and upper quartiles of $\theta$ such that $P(L < \theta < Q1) = P(Q1 < \theta < M) = 0.25 = P(M < \theta < Q3) = P(Q3 < \theta < U)$, A: (2M+L)/3, B: (2M+U)/3, prob(A), prob(B): the elicited probabilities of A and B respectively.



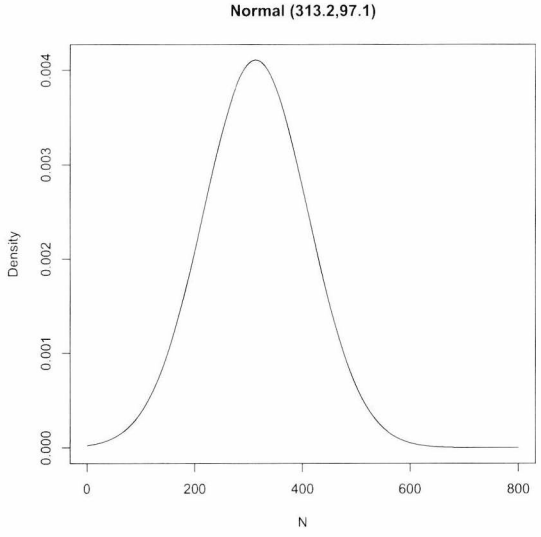

(a) Isle of Wight




(b) Eastern Channel




(c) East Coast REC

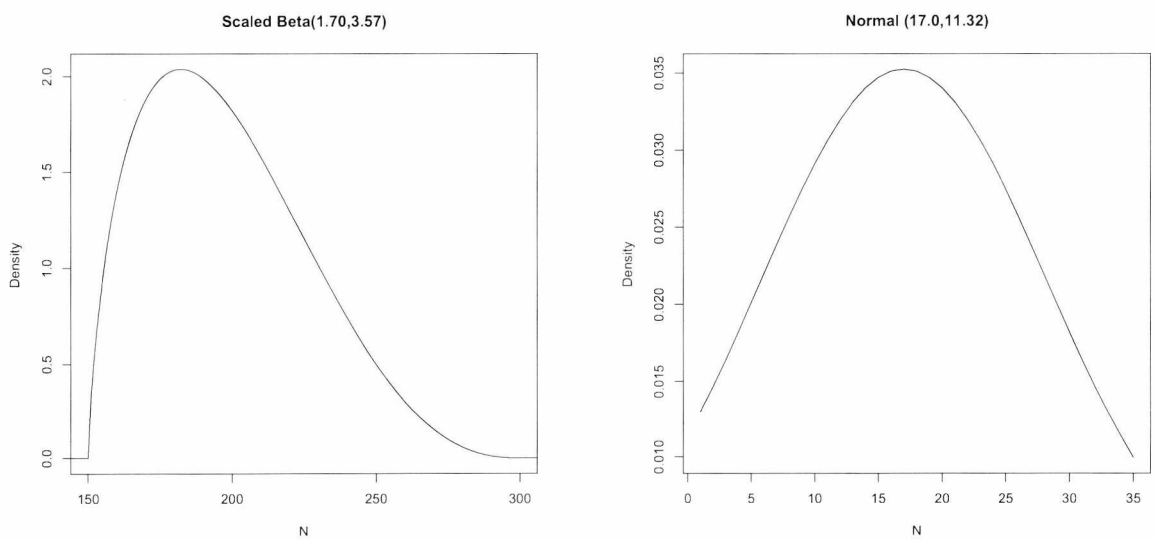



(d) Hastings

*Figure 5.3:* Elicited prior distributions for $N$ for (a) Isle of Wight, (b) Eastern Channel, (c) East Coast REC and (d) Hastings.

| Area | Number of species observed |
|---|---|
| Isle of Wight $0.1m^2$ | 198 |
| Isle of Wight $0.25m^2$ | 240 |
| Isle of Wight overall | 273 |
| Eastern | 649 |
| East Coast REC | 391 |
| Norfolk pre-dredge | 64 |
| Norfolk post-dredge | 26 |
| Hastings | 141 |

*Table 5.2:* Number of species observed in each benthic data set.




*(a)* Pre Dredging

*(b)* Post Dredging

*Figure 5.4:* Elicited prior distributions for $N$ for Norfolk, (a) pre- and (b) post-dredging.

Isle of Wight, where the estimate was considerably more than the observed number of species. However, this was not a problem as I expected that there were many species that were not observed in the samples due to their clustering nature.

### 5.7.2 Number of species in UK coastal waters

The results for the parameter $U$, the total species richness of UK coastal waters were interesting, because the estimates spanned a significant range. The lowest estimate given was 1,200 and the highest was 12,000. The number of recorded species on the UNICORN©(copyright©1995-2004 Unicomarine) database of known benthic species is approximately 10,000 (personal communication, Keith Cooper), so we can see that the lowest estimate of 1,200 was far too small.

I also elicited some information informally from a researcher in marine biodiversity and macroecology, who gave a range of 2,000-10,000 benthic species, with a best guess of around 4,500. However, he also had a list of $\sim$ 825 benthic invertebrate species which are more regularly found around Britain (Tyler et al., 2012). These species were defined as those that occurred in more than 1% of all samples, or in more than ten individual samples, whichever was greater, from five spatially extensive surveys of benthic habitats which sampled 2,641 unique taxa, fully identified to species level.

As these estimates referred to the total number of species in UK waters, they may not be the most insightful for informing the species richness model for a particular area, because the species present will be dependent on variables such as substrate or activity level. This may account for the low estimate given by one of the experts, if they had a particular substrate or habitat in mind when answering this question.

These results could provide an upper limit on the number of species that we would expect to see on a particular surveying programme. The estimates obtained using the maximum-likelihood method in Chapter 3 were considerably larger than these values, and this helps justify the requirement for using a Bayesian approach with informative

priors, and confirms the reality of the boundary problem that I have discussed.

### 5.7.3 Clustering of individuals within a species

The concept of clustering is difficult to build into the model. For example, in some cases there could be isolated individuals, and in other cases there could potentially be hundreds of individuals within $0.1m^2$. The variation in scale of the species could cause difficulties in expressing these values as a distribution. The extent of clustering also depends on the substrate.

Table 5.1 highlights some of these issues, as we can see that some results for $C$ are left blank, and in another the estimate is given only as in the hundreds or thousands. The expert was unable to be more precise. Therefore for some surveys I was not able to use this information to generate an informative prior for the model.

For the Isle of Wight area, the expert did not find the question of the number of species that cluster realistic, so the question was changed to 'What proportion of species in the area can be found in clusters?'. I am unsure how this can be incorporated into the model, however it was interesting to get some kind of information on the clustering from the expert even if it wasn't exactly what I wanted. Also, this would assist in developing a better question for future elicitation procedures.

For the Hastings data I was able to fit distributions to the clustering parameter during the elicitation process. Figure 5.5 shows the normal, gamma and scaled-beta distributions fitted to the elicited information. The scaled-beta gave the best fit by least squares, and the normal distribution was not suitable as a non-trivial portion of it is negative, whereas we could not have a negative number of species within an area.

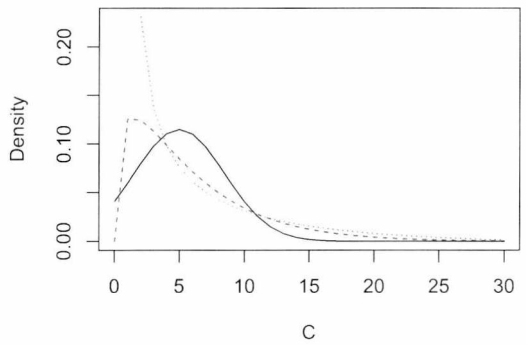



*Figure 5.5:* Normal (solid), gamma (dashed) and scaled-beta (dotted) distributions fitted to elicited information for $C$, the number of species that cluster, for Hastings data.

### 5.7.4 Number of grabs

Elicitation results on the number of grabs required to find a given percentage of observed species were interesting. This reaffirmed my belief that there were many rare species. Experts stated that a high proportion of species would be found in the first few grabs, with a few more found in each additional grab, but not many. However, experts were of the opinion that however many grabs you took you would not find all the species in the area due to the extreme rarity of some species.

An interesting aspect to consider here is that if a species is so rare that you would be highly unlikely to find it during a survey, is it playing a significant role in the ecosystem or is it negligible? If the latter, then does it matter if it is included in the species richness estimate or not? We must bear in mind the application of these estimates when considering the importance of estimating the species richness of an area.

In many cases, benthic surveys consist of only a few grabs samples, such as five or ten, of a relatively small size, $0.1m^2$ or $0.25m^2$, over a large area. These small sample sizes cause concern when estimating species richness for benthic organisms.

If organisms were distributed randomly across the seabed, we could define the probability that a member of a particular species is found by at least one grab as:

$$P = 1 - (1 - a/A)^{yg}, \tag{5.3}$$

where $a$ is the area of the grab, $y$ is the number of individuals of that species in the region, $g$ is the number of samples and $A$ is the area of the region.

This can be rearranged to calculate the number of individuals that you would need to have in the region in order to give a probability, $P$, that the species is found:

$$y = \frac{\log(1 - P)}{g \log(1 - a/A)}. \tag{5.4}$$

| Number of grabs | Abundance (Density) needed for | | |
|---|---|---|---|
| | $P = 50\%$ | $P = 75\%$ | $P = 90\%$ |
| 5 | 74,860 (0.55) | 149,720 (1.11) | 248,679 (1.84) |
| 10 | 37,430 (0.28) | 74,860 (0.55) | 124,340 (0.92) |
| 25 | 14,972 (0.11) | 29,944 (0.22) | 49,736 (0.37) |
| 50 | 7,486 (0.05) | 14,972 (0.11) | 24,868 (0.18) |
| 100 | 3,743 (0.03) | 7,486 (0.05) | 12,434 (0.09) |

*Table 5.3:* Number (and density per $m^2$) of individuals required to give probability levels of 0.5, 0.75 and 0.9 to see a species for 5, 10, 25, 50 and 100 grabs of $0.25m^2$ across a region of $135,000m^2$ if individuals were randomly distributed.

Table 5.3 shows how many individuals would need to be present in the study area to give a 50%, 75% and 90% chance of a finding that species, when taking 5, 10, 25, 50, 100 grabs of $0.25m^2$. I set the area of the region as $135,000m^2$, which is the size of the Norfolk coast study area (Barry et al., 2010).

We can see that we require a large number of individuals within a region in order

for the species to be found. However, if we look at the density per $m^2$ required, it is always less than two individuals per square metre. This does not seem much for organisms that are as small as $1mm$, however for a larger species such as a starfish this density is less likely. If we introduced clustering within species, the number of individuals required would increase.

## 5.8 Quantifying the influence of elicited priors

When estimating species richness for the data sets, I wanted to quantify the influence of the elicited priors on the posteriors and evaluate whether the data supply information about certain parameters. Aside from displaying the prior-posterior pair plots as a visual aide, one could evaluate the overlap for each non-hierarchical prior-posterior pair numerically (Gimenez et al., 2009, p1057).

The overlap between the two distributions can be computed as:

$$\tau_\theta = \int \min(p(\theta), \pi(\theta|X))\mathrm{d}\theta, \tag{5.5}$$

where $\pi(\theta|X)$ is a marginal posterior distribution for data $X$, parameter $\theta$ and prior distribution $p(\theta)$ (Schmid and Schmidt, 2006). When $\pi(\theta|X) \approx p(\theta)$, the data are supplying little information to the posterior for parameter $\theta$.

This can be regarded as the sum of two error probabilities (Schmid and Schmidt, 2006):

$$\begin{aligned}
\tau_\theta &= \int \min(f(x), g(x))\mathrm{d}x \\
&= \int 1_{\{f(X)<g(X)\}} f(x)\mathrm{d}x + \int 1_{\{g(X)\leq f(X)\}} g(x)\mathrm{d}x \\
&= \mathbb{E}(1_{\{f(X)<g(X)\}}) + \mathbb{E}(1_{\{g(X)\leq f(X)\}}) \\
&= P(f(X) < g(X)) + P(g(X) \leq f(X))
\end{aligned}$$

To estimate $\tau_\theta$, firstly the posterior distribution $\pi(\theta|X)$ is estimated using a kernel

density estimator:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right), \tag{5.6}$$

where $K$ is a kernel function centred at the MCMC generated values $x_i, i = 1, ..., n$, and $h$ is the bandwidth.

A standard Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \tag{5.7}$$

is used, along with its associated optimal bandwidth

$$h^* = 1.06\hat{\sigma}n^{-0.2}$$

where $\sigma = \min(\text{standard deviation, interquartile range}/1.34)$ (Silverman, 1986, p48).

We obtain a sample of the posterior from the MCMC generated values $x_i, i = 1, ..., n$, and generate a sample from the prior, $y_j, j = 1, ..., m$. We can then use these to estimate the error probabilities (Schmid and Schmidt, 2006):

$$P(f(X) < g(X)) \quad \text{by the relative frequency} \quad \frac{1}{n}\sum_{i=1}^{n} 1_{\{\hat{f}_n(x_i)<\hat{g}_m(x_i)\}}$$

$$P(g(X) \leq f(X)) \quad \text{by the relative frequency} \quad \frac{1}{m}\sum_{j=1}^{m} 1_{\{\hat{g}_m(y_i)\leq\hat{f}_n(y_i)\}},$$

such that the estimator of overlap becomes

$$\hat{\tau}_\theta = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\hat{f}_n(x_i)<\hat{g}_m(x_i)\}} + \frac{1}{m}\sum_{j=1}^{m} 1_{\{\hat{g}_m(y_i)\leq\hat{f}_n(y_i)\}}. \tag{5.8}$$

Values of $\tau_\theta$ lie in the interval (0,1). When $\tau_\theta$ is above some pre-determined threshold then the parameters are declared weakly identifiable (Gimenez et al., 2009, p1057). A threshold of 0.35 has been suggested (Garrett and Zeger, 2000).

## 5.9 Use of priors applied to benthic data

I continued by using the Norfolk, Hastings and Isle of Wight data sets, because the elicitation process provided sensible looking priors for the number of species for

these areas. These priors could be utilised within the MCMC framework to aid the estimation of species richness.

I used the negative binomial model under the non-hierarchical likelihood method, because the priors elicited were not in the form of a binomial mixture. However, the pilot study showed that the fitted distributions were very close for the normal, scaled beta and beta-binomial so to use the hierarchical approach it would be possible to use the elicited information to generate a prior of this form for $N$.

### 5.9.1 *Bayesian analysis using an elicited prior on N and negative binomial model*

Table 5.4 shows the results of MCMC to produce a posterior using the elicited priors on $N$ for the benthic data sets, and non-informative half-Cauchy priors for the nuisance priors. The overlap of prior and posterior varies between the areas (Table 5.5).

| Area | Mean | Median | SD | 95% Credible | B. $p$-value | DIC |
|---|---|---|---|---|---|---|
| Hastings | 410 | 405 | 98.92 | (279, 546) | 0.71 | -426.97 |
| Isle of Wight $0.1m^2$ | 818 | 805 | 216.93 | (535, 1,117) | 0.80 | -421.36 |
| Isle of Wight $0.25m^2$ | 776 | 758 | 215.96 | (494, 1,084) | 0.63 | -407.96 |
| Norfolk pre-dredging | 171 | 168 | 37.79 | (104, 245) | 0.79 | -28.05 |
| Norfolk post-dredging | 38 | 38 | 8.09 | (28, 50) | 0.75 | 24.56 |

*Table 5.4:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using elicited priors on $N$ and the negative binomial model. The Bayesian $p$-value and DIC are also reported.

Figure 5.6 shows prior and posterior distributions for $N$ for the Hastings data. In the samples, 141 species were seen, and the estimate of the species richness was 410 (279, 546). The prior had a relatively large effect on the posterior, as shown by the $\tau_\theta$ value in Table 5.5, which was well above the 0.35 suggested by Garrett and Zeger

| Area | $\tau_\theta$ |
|---|---|
| Hastings | 0.58 |
| Isle of Wight $0.1m^2$ | 0.79 |
| Isle of Wight $0.25m^2$ | 0.87 |
| Norfolk pre-dredging | 0.40 |
| Norfolk post-dredging | 0.17 |

*Table 5.5:* Coefficient of overlap, $\tau_\theta$, values for the negative binomial model fitted to the benthic data.

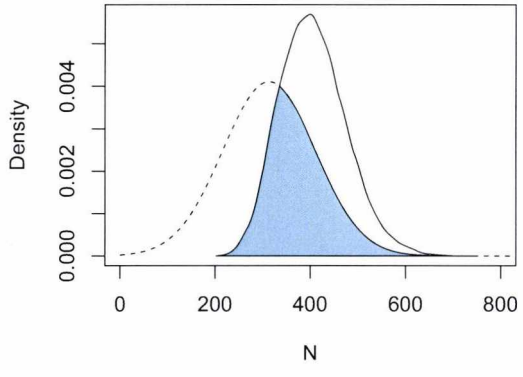
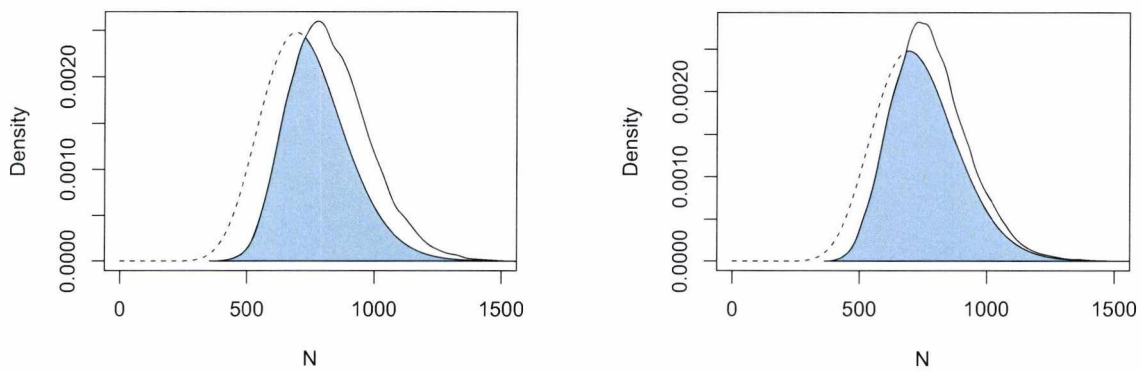



*Figure 5.6:* Prior (dashed line) and posterior (solid line) for $N$ for Hastings data using elicited prior and negative binomial model.

(2000).



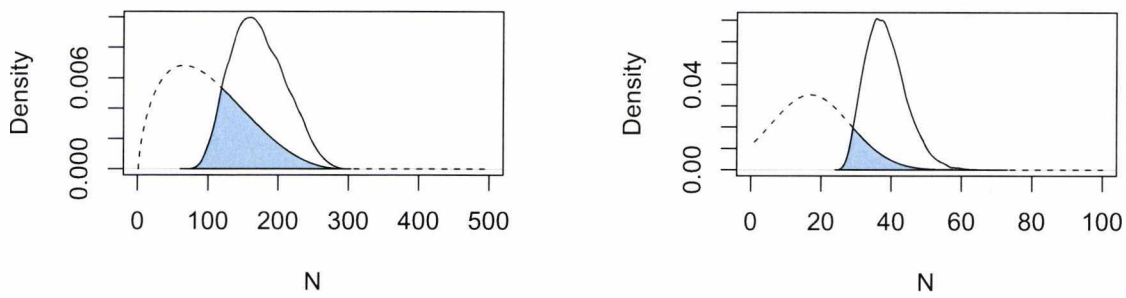

*(a)* Small grabs *(b)* Large grabs

*Figure 5.7:* Prior (dashed line) and posterior (solid line) for $N$ for Isle of Wight data using elicited prior.

Figure 5.7a shows the prior and posterior for Isle of Wight with small grabs. The prior had quite an influence on the species richness estimate as the posterior distribution was very close to the prior distribution. The prior contributed most of the information to the posterior, rather than the likelihood. The same applied for the data collected with large grabs (Figure 5.7b).

Figure 5.8 shows the posterior estimates pre- and post-dredging for the Norfolk data. Post-dredging, the prior had much less influence over the posterior distribution than pre-dredging. Dredging took place over three days in April 1992, resulting in the removal of 50,000 tonnes of marine aggregate. Surveys of the marine benthos were carried out pre-dredging in March 1992 and post-dredging in May 1992 (Barry et al., 2010).

There is a shift to the right in all of these graphs from the prior to the posterior. In all cases apart from Norfolk, the prior was having a large effect on the posterior, as shown by the $\tau_\theta$ values in Table 5.5. Therefore I concluded that either the model




*(a)* Pre-dredging *(b)* Post-dredging

*Figure 5.8:* Prior (dashed line) and posterior (solid line) for $N$ for Norfolk data using elicited prior.

was very well informed by this prior, and the data matched well to the expectation of the experts, or that the sample size was not large enough for the data to contribute much to the posterior. The effect of increasing sample size is investigated in Section 5.10.

The Bayesian $p$-values for the fit of the negative binomial model to the data were all above 0.64, and we would expect a value close to 0.5 for a good-fitting distribution. Therefore in the next section I fit the Neyman Type A-gamma model to the data which will account for clustering and hopefully improve the fit and the species richness estimates.

### *5.9.2 MCMC using elicited prior on N and Neyman Type A - gamma model*

Table 5.6 shows the results of the MCMC using elicited priors on $N$ for the benthic data sets and the Neyman Type A-gamma model. The priors for the nuisance parameters of the abundance distribution remained uninformative half-Cauchy priors.

The estimates of species richness increased slightly, but the fit to the data, as given by the Bayesian $p$-values, were not as good as the negative binomial model apart

| Area | Mean | Median | SD | 95% Credible | B. $p$-value | DIC |
|---|---|---|---|---|---|---|
| Hastings | 429 | 426 | 70.33 | (297, 567) | 0.74 | -408.58 |
| Isle of Wight $0.1m^2$ | 903 | 889 | 159.64 | (605, 1,220) | 0.87 | -388.95 |
| Isle of Wight $0.25m^2$ | 890 | 875 | 154.73 | (600, 1,194) | 0.73 | -483.85 |
| Norfolk pre-dredging | 165 | 161 | 39.60 | (94, 243) | 0.76 | -252.94 |
| Norfolk post-dredging | 40 | 40 | 6.08 | (29, 52) | 0.75 | 27.78 |

*Table 5.6:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using elicited priors on $N$ and the Neyman Type A-gamma model. The Bayesian $p$-value and DIC are also reported.

from for the Norfolk data. The pre-dredging data had a Bayesian $p$-value closer to 0.5 after fitting the Neyman Type A-gamma model, and for the post-dredging data the Bayesian $p$-values were the same for each model. A comparison of DIC values also suggested that the Neyman Type A-gamma outperformed the negative binomial model for the Norfolk pre-dredging survey, and that both models were plausible.

One reason that the Neyman Type A-gamma model might not have fitted as well to the data, was the use of uninformative prior on the clustering parameter. Table 5.1 showed a lot of variation in the distribution of the clustering parameter between sampling programs, and therefore it is likely that an informative prior for the clustering parameter would play a significant role in the posterior. However, I was unable to use an elicited prior for $C$ for all of these data sets.

### *5.9.3 MCMC using elicited prior on N and informative prior for C*

Through the elicitation process I was able to elicit information on clustering that could be used to inform a prior for the Hastings data set. Therefore I compared the results using an informative prior for the clustering aspect of the Neyman Type A distribution to those obtained above. I expected that by including the prior on $C$ the fit of the model to the data would improve, as quantified by the Bayesian $p$-value.

| Area | Mean | Median | SD | 95% Credible | B.$p$-value | DIC |
|---|---|---|---|---|---|---|
| NB | | | | | | |
| $N$ | 410 | 405 | 69.94 | (279, 546) | 0.71 | -426.97 |
| $\alpha$ | 0.1772 | 0.17 | 0.03 | (0.1061, 0.2945) | | |
| $\beta$ | 10.9348 | 10.71 | 1.40 | (7.7098, 15.3958) | | |
| NTAG $N$ | | | | | | |
| $N$ | 429 | 426 | 70.33 | (297, 567) | 0.74 | -408.58 |
| $\alpha$ | 0.1979 | 0.19 | 0.05 | (0.1216, 0.3173) | | |
| $\beta$ | 9.2847 | 9.10 | 1.67 | (6.5535, 13.0905) | | |
| $\phi$ | 0.6496 | 0.64 | 0.11 | (0.4815, 0.8960) | | |
| NTAG $N$ and $C$ | | | | | | |
| $N$ | 436 | 433 | 67.59 | (309, 567) | 0.76 | -1262.85 |
| $\alpha$ | 0.215 | 0.21 | 0.05 | (0.1342, 0.3396) | | |
| $\beta$ | 8.501 | 8.36 | 1.53 | (5.9413, 11.9285) | | |
| $\phi$ | 1.027 | 1.01 | 0.06 | (1.0000, 1.1606) | | |

*Table 5.7:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ assuming the negative binomial model, NB, using elicited prior on $N$ and the Neyman Type A-gamma using elicited prior on $N$, NTAG $N$, and Neyman Type A-gamma using elicited prior on $N$ and $\phi$, NTAG $N$ and $C$.

Table 5.7 shows that when I included an informative prior on $C$, the fit of the model to the data got slightly worse according to the Bayesian $p$-value. The species richness estimate increased, however the 95% credible interval narrowed slightly. However, the DIC showed this model as by far the best of the three, with a much lower DIC value. However, caution must be taken when interpreting this value, as in some cases the DIC can have undesirable properties such as obtaining a negative number of effective parameters. In fact the Neyman Type A-gamma model with informative prior on $C$ does have a negative number of effective parameters here, and so the DIC cannot reliably be used to select the best model.

Previously I used a half-Cauchy reference prior for the clustering parameter in the Neyman Type A-gamma model. The posterior values for $\phi$ in Table 5.7 show that the credible interval for this parameter was (0.4815, 0.8960). This parameter corresponded to the number of individuals per cluster, so it is estimating rather low when compared to the elicited information. To quantify the influence of the prior on $\phi$ I considered the coefficient of overlap, $\tau_\theta$ (Table 5.8).

Table 5.8 shows that as I included an informative prior on $\phi$, the coefficient of

| Parameter | NB model | NTAG model $N$ | NTAG model $N$ and $C$ |
|---|---|---|---|
| $N$ | 0.55 | 0.49 | 0.43 |
| $\alpha$ | 0.16 | 0.18 | 0.18 |
| $\beta$ | 0.08 | 0.08 | 0.08 |
| $\phi$ | - | 0.23 | 0.54 |

*Table 5.8:* Coefficient of overlap, $\tau_\theta$, values for the models fitted to the Hastings data, assuming the negative binomial model, NB, using elicited prior on $N$ and the Neyman Type A-gamma using elicited prior on $N$, NTAG $N$, and Neyman Type A-gamma using elicited prior on $N$ and $\phi$, NTAG $N$ and $C$.

overlap for $\phi$ increased. However, the coefficient of overlap for $N$ decreased slightly. We might expect this as the prior on $\phi$ will have contributed more to the joint posterior than before, and so we would expect the prior on $N$ to have less influence on the joint posterior.

The coefficient of overlap for $\phi$ is below the 0.35 threshold when an uninformative prior is used on the clustering parameter of the Neyman Type-A model. This is to be expected, because an uninformative prior should not influence the posterior greatly. When I included an informative prior for the clustering parameter $\phi$, the overlap between prior and posterior increased and is above this threshold. This suggested that the prior is having a significant influence on the posterior.

The overlap for $\alpha$ and $\beta$ respectively was fairly similar for the three models, and well below the suggested identifiability threshold of 0.35. This suggested that the priors for these parameters were not significantly influencing the posterior, as we would expect when using uninformative priors.

## 5.10 Identifiability and sample size

A model may be subject to weak identifiability when the sample size is not large enough to estimate parameters accurately. Therefore I increased the sample size artificially to investigate the relationship between prior and posterior. I expected that as I increased the sample size, the influence of the prior on the posterior distribution should decrease.

I used the Isle of Wight data to investigate the link between sample size and identifiability as this showed the greatest influence of prior on posterior for the negative binomial model (Table 5.9). In addition, I combined the two Isle of Wight datasets to increase the sample size, because both were taken at the same locations (although we must remain aware that different grab sizes were used for each survey).

| Area | Mean | Median | SD | 95% Credible | Bayesian $p$-value |
|---|---|---|---|---|---|
| Isle of Wight $0.1m^2$ | 818 | 805 | 216.93 | (535, 1117) | 0.80 |
| Isle of Wight $0.25m^2$ | 776 | 758 | 215.96 | (494, 1084) | 0.63 |
| Isle of Wight combined | 838 | 825 | 202.40 | (577, 1127) | 0.71 |

*Table 5.9:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using elicited priors on $N$ and the negative binomial model fitted to the Isle of Wight data sets. The Bayesian $p$-value is also reported.

| Area | $\tau_\theta$ |
|---|---|
| Isle of Wight $0.1m^2$ | 0.79 |
| Isle of Wight $0.25m^2$ | 0.87 |
| Isle of Wight combined | 0.68 |

*Table 5.10:* Coefficient of overlap, $\tau_\theta$, values for $N$ for the negative binomial model using elicited prior on $N$ fitted to the Isle of Wight data, for each data set, and combined.

Table 5.10 shows that combining the data from the large and small grabs did decrease the overlap between prior and posterior for the Isle of Wight data. The posterior distribution moved away from the prior, which implies that the data were having a greater influence in the posterior than before, however the overlap value was still very high (Figure 5.9).

However, the observed number of species also increased, (198 for small, 240 for large, 273 overall), so this may have caused the higher mean and confidence intervals of the posterior for the combined data set, therefore decreasing the overlap between prior and posterior.

To clarify the cause of the extent of the overlap, I used the data set for the larger grabs and sampled with replacement over the ten grabs. I took a sample of 20, 50

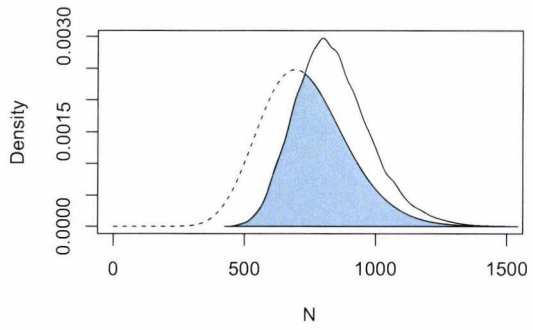



*Figure 5.9:* Prior (dashed line) and posterior (solid line) for $N$ for combined Isle of Wight data using elicited prior and negative binomial model.

and 100 to create new data sets of 20 grabs, 50 grabs, and 100 grabs. I expected that as the sample size increased, the overlap between prior and posterior would decrease if the data were being swamped by the prior in the posterior.

Tables 5.11 and 5.12 show that as the sample size increased, there was not much impact on the posterior distribution or the overlap between prior and posterior. This indicated that it was not the prior swamping the data which was causing the overlap between prior and posterior in this case.

| Sample size | Mean | Median | SD | 95% Credible | B. $p$-value |
|---|---|---|---|---|---|
| Isle of Wight $0.25m^2$ | 776 | 758 | 152.71 | (494, 1,084) | 0.63 |
| 20 grabs | 772 | 760 | 150.10 | (491, 1,064) | 0.64 |
| 50 grabs | 781 | 766 | 153.03 | (501, 1,083) | 0.63 |
| 100 grabs | 783 | 768 | 149.17 | (518, 1,076) | 0.64 |
| 1000 grabs | 764 | 752 | 145.78 | (497, 1045) | 0.64 |
| 10,000 grabs | 771 | 761 | 145.78 | (513, 1068) | 0.63 |

*Table 5.11:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using elicited priors on $N$ and the negative binomial model fitted to the sampled Isle of Wight data sets.

| Sample size | $\tau_\theta$ |
|---|---|
| Isle of Wight $0.25m^2$ | 0.87 |
| 20 grabs | 0.87 |
| 50 grabs | 0.85 |
| 100 grabs | 0.85 |
| 1000 grabs | 0.89 |
| 10,000 grabs | 0.89 |

*Table 5.12:* Coefficient of overlap, $\tau_\theta$, values for $N$ for the negative binomial model using elicited prior on $N$ fitted to the sampled Isle of Wight data, after sampling the $0.25m^2$ data with replacement for 20, 50, 100, 1000 and 10,000 grabs.

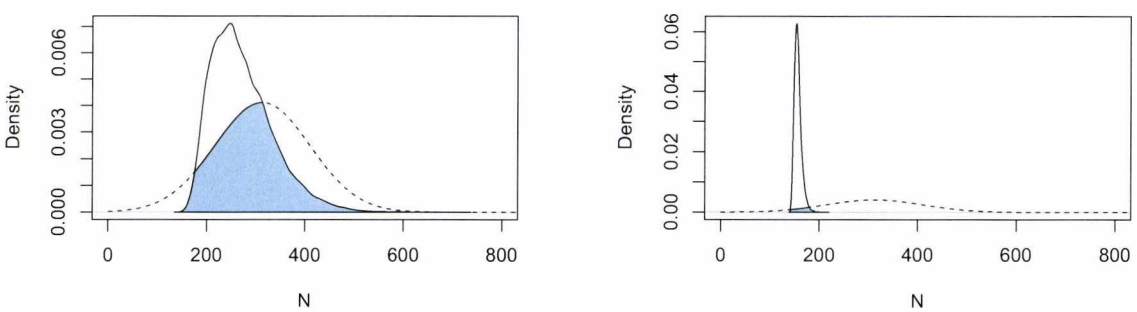
I repeated the above sampling method to investigate the effect of increasing sample size applied to the Hastings data as a comparison. Table 5.13 shows that as the sample size increased, the overlap between prior and posterior decreased rapidly. Table 5.14 shows that the posterior estimates decreased, and the credible 95% interval was narrowing. The point estimates approached the number of observed species in the sample, 141. We can see that as the sample size increased, the prior had less influence on the posterior (Figure 5.10). This is in contrast to the results from the Isle of Wight.

| Sample size | $\tau_\theta$ |
|---|---|
| Hastings | 0.58 |
| 20 grabs | 0.67 |
| 50 grabs | 0.07 |
| 100 grabs | 0.01 |

*Table 5.13:* Coefficient of overlap, $\tau_\theta$, values for $N$ for the negative binomial model using elicited prior on $N$ fitted to the sampled Hastings data, after sampling with replacement for 20, 50 and 100 grabs.

| Sample size | Mean | Median | SD | 95% Credible | Bayesian $p$-value |
|---|---|---|---|---|---|
| Hastings | 410 | 405 | 69.94 | (279, 546) | 0.71 |
| 20 grabs | 272 | 262 | 62.58 | (173, 397) | 0.45 |
| 50 grabs | 157 | 155 | 7.48 | (144, 172) | 0.45 |
| 100 grabs | 144 | 143 | 1.98 | (141, 148) | 0.46 |

*Table 5.14:* Mean, median, standard deviation, SD, and 95% credible interval of the posterior of $N$ using elicited priors on $N$ and the negative binomial model fitted to the sampled Hastings data sets, for 20, 50 and 100 grabs sampled with replacement.




*(a)* 20 grabs

*(b)* 50 grabs

*Figure 5.10:* Prior (dashed line) and posterior (solid line) for $N$ for 20 and 50 grab sampled Hastings data using elicited prior on $N$ and negative binomial model.

In some cases the data was swamped by the prior, however a higher overlap value might not always indicate this. It could just be that the prior could be a very good fit to the actual population, and this was most likely the case for the Isle of Wight data.

## 5.11 Discussion

The aim of this chapter was to explore the elicitation process, and produce informative priors which could be used within a Bayesian framework to estimate species richness and avoid spuriously large estimates. I did this for certain data sets, and the elicitation procedure could be repeated for additional data sets. A review of the relevant literature showed that the elicitation process consists of several steps and the entire process requires structure and rigor. This chapter highlights the importance of a pilot study, the role of the facilitator and the use of appropriate software.

From feedback given by the experts, during and after the elicitation process, on the whole they were happy with the elicitation procedure. The more structured framework and formal procedure of the main elicitation task was important in improving the process. The supply of more background information about why the elicitation was needed was valuable, because the experts realised that there was a need for the research being undertaken in this area.

Running through the example was useful in demonstrating the process and understanding the best way to approach the questions with the experts. One expert commented that the process was a 'different way of thinking to the scientists', in the way the questions were posed, so I made it explicit exactly what I wanted them to do. I also adjusted the procedure to be familiar in terms of language and quantities used.

The pilot study and elicitation procedure highlighted difficulties regarding the scale of benthic organisms. Data on benthic organisms are collected using Hamon grabs, and the small organisms are recorded to different levels depending on the study, so it was crucial to ensure that both myself and the experts were considering the same scale of identification. Throughout this research I have concentrated on Macrofauna greater than $1mm$ in size, but it was important to clarify this with the scientists who sometimes work to different scales.

In addition, the range of sizes of organisms and scales of habitat regarding clusters of species was raised. For a small species of $2mm$, a grab of $0.1m^2$ could be its whole world, whereas for a large species such as a starfish, an aggregation might extend well beyond the grab area sampled.

The pilot study highlighted the importance of the facilitator in the elicitation procedure, and this was confirmed in the main elicitation process. It was also interesting to note that additional unsolicited information arose while completing the elicitation, giving useful background information on benthic organisms and their relationship with the substrate and other species.

For each of the data sets, a different expert was consulted. As mentioned previously, it would be worthwhile to use more than one expert per data set and gain a consensus to improve the elicitation process. I did elicit the number of species throughout UK waters from all experts, so this could be useful as an upper bound to any species richness estimates, and also could be used as an indicator of the accuracy of the estimates of the other parameters. This is not something that I consider within this thesis, but is an interesting area for further research.

Priors were elicited for both $N$ and nuisance parameters, but not all were incorporated into the model. In some cases the elicited distributions of the number of species did not correspond well to the data sets available. This was possibly because there were several sampling programmes carried out in each area over time and space and therefore as more species were found the expert's opinion might change.

Information on clustering proved difficult to incorporate into priors, due to the extreme variation in cluster size between species, and within species. Also, the scale of the clusters in proportion to the grabs sizes was a barrier to elicitation of this parameter for some experts.

The area sampled in proportion to the study region in most benthic surveys, and the clustering of individuals means that there will be many reasonably common species that will be missed in a sampling program. This has been demonstrated for randomly distributed individuals, but will be accentuated when we incorporate clustering into the calculation. The importance of the inclusion of all species in a species richness estimate needs to be carefully considered when designing a sampling program and analysing data.

When applied to benthic data sets, the elicited priors influenced the posterior and avoided the spuriously large species richness estimates seen previously. This reaffirmed the fact that elicitation can play a pivotal role in informing models and decisions (Martin et al., 2012).

As the species demonstrate a clustering behaviour, we expected the Neyman Type A-gamma model to fit the data better than the negative binomial. However, according to the Bayesian $p$-values and DIC this was only the case for the Norfolk data. This could be occurring if the prior on the clustering parameter was not suitable to model the data. Including an informative clustering parameter had some influence on the species richness estimate. However, I would suggest caution in using such a prior as it was difficult to define during elicitation and therefore may not be very reliable. The number of effective parameters of the DIC for this model was negative, and this highlighted an undesirable property of this information criterion.

The influence of the prior on the posterior was quantified by calculating the coefficient of overlap, and in some cases the sample size influenced this overlap. When sampling from the Hastings data set, I found that the overlap between prior and posterior decreased rapidly as I increased the sample size. However, for the Isle of Wight data there were no significant changes in the posterior as sample size increased, suggesting that the prior fitted the data well.

Although benthic ecologists were consulted in order to construct priors for the number of species in an area, the limitations of expert knowledge can lead to mis-specification of these priors. Experts were asked to specify upper and lower bounds for the distribution of $N$, and a prior was fitted to the information given, in the form of gamma, normal and scaled-beta distributions. These priors could have a strong influence over the posterior. This was illustrated by the Hastings data, for which a normal prior was used. The prior was shown to have a strong influence over the posterior, in terms of prior-posterior overlap, which was greatly reduced by increasing the sample size.

A possible solution to reduce the influence of the prior would be to fit an alternative prior to the elicited information, such as the t-distribution, which has heavier tails than the normal distribution, meaning that it is more prone to producing values that fall further from the mean. This could decrease the influence of the prior over the posterior. An alternative approach might be to fit the elicited information firstly ignoring the upper and lower bounds, and then cut off the distribution above and below the bounds. The use of a fat-tailed distribution may not only reduce the influence of the prior, but also be less sensitive to mis-specification of the prior, as found by Chen et al. (2000) in the case of using elicited information to inform priors in fisheries-stock assessment.

This chapter highlighted a key barrier to elicitation, in that statisticians and biologists approach a problem from very different vantage points, and to perform a successful elicitation, one must consider both of these when designing the procedure, but also ensure that each party is benefitting from the process. Once I was able to highlight the reasoning behind my models and how they would benefit benthic ecologists, the scientists were much more receptive to what I was asking them, and very forthcoming with additional information that could assist in the models. This would be applicable to any application.

The process has highlighted several benefits of using software in the elicitation procedure. It ensured fast feedback to experts so that they could adjust their views, ensuring that elicited information adequately reflected the experts' opinions. This is especially important when dealing with non-statisticians who may not be comfortable with the concept of probability. Automation of model-fitting calculations and streamlining of the elicitation process, reducing the length of elicitation sessions, also keeps the expert engaged as much as possible (Fisher et al., 2012).

I found that it was much better to carry out elicitation face-to-face, using a facilitator. The facilitator is able to phrase the elicitation questions to suit the individual experts' understanding and knowledge of probability. Also it is important to use appropriate language, and questions should refer to terms and concepts familiar to the expert, including the way they measure things.

Elicitation is a mechanism for capturing not only an expert's best estimate of a value, but also the uncertainty of that estimate. Eliciting uncertainty is particularly important when only a single expert is available (Kuhnert et al., 2010), as in this case. For benthic organisms, as with other species, it is likely that there are few experts with suitable expertise for situations where elicitation of expert knowledge is useful.

As highlighted by Fisher et al. (2012), the quality and value of elicited knowledge is highly dependent on the specific experience and expertise of the experts involved. I assumed that the experts had the necessary knowledge and experience to provide reasonable answers to the questions posed. However, there was no guarantee that their answers were not biased (Kuhnert et al., 2010), and I found that estimates did not always correspond with reality. This could be seen especially when I elicited the number of species in UK waters, the results of which varied immensely.

## 5.12 Conclusions

Through elicitation I was able to obtain priors that gave much more realistic species richness estimates when applied to benthic data sets. Although the fit using the Neyman Type A-gamma was not as good as the negative binomial, which seems anomalous, this could perhaps be explained by the fact that clustering was not allowed to vary between species within the model. Further work will look at incorporating this variation into the model.

Elicitation of information about marine benthic data has not previously been discussed in the literature as far as I know, but this research has contributed to developing a structured procedure that can be used for this purpose. I have shown that the elicitation process is complex and requires careful planning and implementation. However, the results can be very beneficial. The elicitation process must be applied with rigor to ensure the validity of the priors and there are several ways that I could potentially improve the elicitation that I have performed, to get more accurate and useful information to include in a model. These include improving the use of software, appropriate language, and understanding of benthic communities.

It would be worthwhile to use more than one expert per data set and gain a consensus to improve the elicitation process. The results from the elicitation of information on the number of species throughout UK waters could be used as an indicator of the accuracy of the estimates of the parameters from each expert. Although this is not something that I considered within this thesis, it is an interesting area for further research.

This work could be extended by an investigation of the validity of elicitation when only one expert is available, along with the development of more specialised software for the elicitation of information particular to benthic organisms. This could be incorporated into a larger software package to be used by benthic ecologists which is able to estimate diversity of benthic organisms for a particular area. In addition, this

`R` package could be used for other taxonomic groups.

Using elicited priors avoided the spuriously large estimates associated with the boundary problem, and a link between Bayesian priors and penalties in a frequentist framework is explored in Chapter 6.

# 6. COMPARISON OF SPECIES RICHNESS ESTIMATION APPROACHES

## *6.1 Introduction*

This chapter considers the link between the various methods of species richness estimation. Firstly, I consider how elicited priors can be converted into penalties within a frequentist approach. I then compare the results of the frequentist approach to those of the Bayesian approach, and discuss my preferred method. I also consider the results of the non-parametric estimators, and how they compare.

Finally I make a recommendation of the best method to use to estimate species richness for benthic data, and use this method to investigate the impacts of dredging on the Norfolk coast. These results will be compared to those found by Barry et al. (2010), who utilised a clustering model to describe the spatial pattern of each species.

## *6.2 Penalties as priors*

In Chapter 3 I considered penalising the maximum-likelihood estimator to avoid spuriously large species richness estimates. Recall that if $l(N, \boldsymbol{\theta})$ is a log-likelihood function where $(N, \boldsymbol{\theta})$ represent the unknown parameters, then the penalised log-likelihood corresponding to penalty parameter $\gamma$ and penalty function $h(\boldsymbol{\theta})$ is defined as

$$l^{\gamma}(N, \boldsymbol{\theta}) = l(N, \boldsymbol{\theta}) - \gamma h(N, \boldsymbol{\theta}). \tag{6.1}$$

In the Bayesian approach we base inference for $\theta$ on its posterior distribution which

may be expressed as

$$\pi(\theta|x) \propto f(x|\theta)p(\theta), \tag{6.2}$$

and the log of the posterior is

$$\log(\pi(\theta|x)) = \log(f(x|\theta)) + \log(p(\theta)) + \text{constant}, \tag{6.3}$$

the log-likelihood plus the log of the prior. So interpreting the log-likelihood function in a Bayesian context, it is the log of the probability distribution, combined with a penalty, which corresponds to a prior contribution on the distribution of $\theta$.

For large samples and a uniform prior, the mode of the posterior and MLE will be equivalent. If we use an alternative prior this may be considered equivalent to using a penalty.

We have seen in Chapters 3 and 4 that if the true odds parameter of the population is low, that is if many of the species were sampled, I would want the prior to have less influence on the posterior to allow the species richness estimate to reach the appropriate value. This corresponds to using a less harsh penalty.

Conversely, if the true odds parameter is large, then I want to have an informative prior that will influence the posterior and avoid the spuriously large species richness estimates arising from the boundary problem.

In Chapter 3 I considered two penalties suggested by Wang and Lindsay (2005). The first was a penalty based on the odds function, $\psi = p_0/(1 - p_0)$, which was found to combat the boundary problem. As $\hat{N}_c = D/\{1 - p_0(\hat{\theta}_c)\}$ is equivalent to $\hat{N} = D(1 + \psi)$, imposing a penalty on $\psi$ reduces the magnitude of $\hat{N}$.

The penalised log-likelihood for this penalty is

$$l_2(N, \theta) = l(N, \theta) - \gamma_2\psi(\theta), \quad \gamma_2 > 0.$$

However, the optimal choice of $\gamma_2$ depended strongly on the value of $\psi$ (Wang and Lindsay, 2005). No method has been devised that can choose the best penalty for a particular problem, but using simulations in Chapter 3 I found that a value of $\gamma_2 = 0.5$ worked reasonably well in all cases. When I extended the estimator to the Neyman Type A-gamma model we saw that this was more sensitive to the value of $\gamma_2$ used in the penalty, and in some cases a value of $\gamma_2 = 0.25$ was more appropriate.

If we interpret this particular penalty function as a prior, then this penalty corresponds to an exponential prior for the odds function with mean $1/\gamma_2$ (Wang and Lindsay, 2005).

I also considered the penalised log-likelihood,

$$l_3(N, \theta) = l(N, \theta) - \gamma_3(\psi - \eta)^2 \boldsymbol{I}(\psi > \eta), \quad \gamma, \eta > 0,$$

where $\eta = \hat{N}_{C1}/D - 1$, $\gamma_3 = 1/2\eta$, and $\hat{N}_{C1} = D + f_1^2/(2f_2)$ is the lower bound estimator of Chao (1984).

This penalty corresponds to a uniform prior for the odds function on $(0, \eta)$ and a normal prior with mean $\eta$ and variance $\sigma^2 = \frac{1}{2\gamma}$ on $(\eta, \infty)$ (Wang and Lindsay, 2005). The shape of the posterior of the odds parameter is determined by the choice of $\gamma_3$ and $\eta$.

Using this penalty with one-step iteration I would run the MCMC algorithm with the prior described above, and then repeat MCMC using the posterior mean as a naive estimator in place of $\hat{N}_{C1}$.

### 6.2.1 Converting elicited priors to penalties

If I elicited information on the proportion of the species in the area that were caught during the sampling program, this would relate to the odds ratio. I did not elicit information on this, however, if I asked experts to give the proportion of species in the population that they estimate would be caught, then this information can be used

directly to form a Bayesian prior for the odds parameter.

A normal Bayesian prior on the odds parameter would correspond to a penalised log-likelihood of

$$l(N, \theta) = l(N, \theta) - \gamma_3(\psi - \eta)^2, \quad \gamma_3, \eta > 0,$$

where $\eta$ is the mean of the elicited distribution, and $\gamma_3 = 1/(2\sigma^2)$. Larger values of $\eta$ give a flatter prior, which implies more unsampled species and more uncertainty in the estimate (Wang and Lindsay, 2005).

Instead I elicited information on $N$, the total number of species that are in the population, which as previously stated is a function of the odds parameter. Since the family of normal distributions is closed under linear transformation, that is if $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$ then $aX + b$ is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$, I can use the elicited normal prior for $N$ to specify the values of $\eta$ and $\gamma_3$ to use as a penalty in the MLE.

The elicited prior for $N$ for the Hastings data was Normal$(313.2, 97.1^2)$. Under the conditional likelihood approach, $\hat{N} = D + D\psi$, and I can use the linear transform property of the normal family of distributions to convert the prior for $N$ into a penalty. For the Hastings data set $D = 141$, so $a = b = D$ and $\psi \sim$ Normal$(1.22, 0.474)$.

This prior is shown in Figure 6.1. The distribution extends below 0, which is not possible for this parameter. This reflects the inaccuracy of using elicited information on $N$, and the prior could be improved for the odds parameter by eliciting information on it directly. However, this is a less intuitive parameter, and therefore may cause confusion during the elicitation process.

I can then use this to construct the penalised likelihood

$$l(N, \theta) = l(N, \theta) - 1/0.948(\psi - 1.22)^2, \tag{6.4}$$

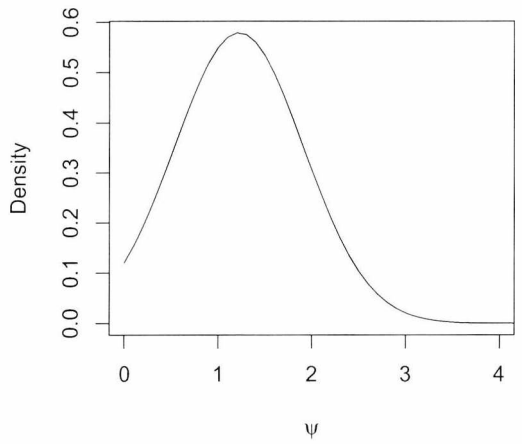



*Figure 6.1:* Elicited prior for the odds parameter, $\psi \sim \text{Normal}(1.22, 0.474)$, for the Hastings data.

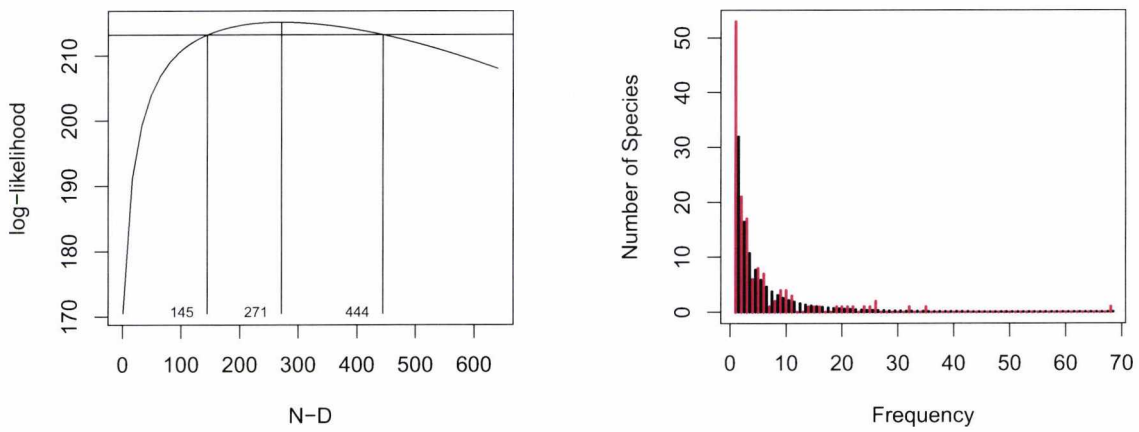
which should lead to the same estimate for $N$ as the point estimate given by the Bayesian approach using the elicited prior for $N$ and uniform priors for the nuisance parameters. I have disregarded the $\log(\sqrt{2\pi\sigma^2})$ term in the log likelihood, because this does not depend on the parameters and will not affect the maximisation.

This MLE approach gives a species richness estimate of 411 (285, 584) using the negative binomial model. The profile log-likelihood for the number of unseen species is shown in Figure 6.2a. This corresponds very well to the estimate of the Bayesian approach given in Chapter 5 which was 410 (279, 546) when using half-Cauchy priors on the nuisance parameters, and 415 (278, 554) using uniform priors on the nuisance parameters, which are non-informative.

Alternatively, I could use the prior for $N$ directly converted into a penalty. For the Hastings data that gives us

$$l(N, \theta) = l(N, \theta) - \frac{1}{2(97.1^2)}(N - 313.2)^2, \tag{6.5}$$

When calculating the MLE including profile confidence intervals I use the full likelihood. This approach gives us a result very similar to that above, namely




*(a)* Profile log-likelihood for $f_0$ *(b)* Goodness of fit

*Figure 6.2:* (a) Profile log-likelihood for the number of unseen species, $f_0 = N - D$, and (b) goodness of fit for the penalised MLE of Eqn 6.4 to the Hastings data assuming the negative binomial model.

411 (288, 568), where the confidence interval is skewed slightly to the right. The fit of this distribution (excluding the zero class) can be seen in Figure 6.2b.

Using this approach allows us easily to create a penalty using the information elicited for the other sampling areas, which were fitted with gamma and scaled beta distributions. For gamma distributed priors, $N \sim \text{Gamma}(\alpha, \beta)$ we have the penalised log-likelihood

$$l(N, \theta) = l(N, \theta) + (\alpha - 1)\log(N) - N/\beta, \tag{6.6}$$

which is just the log-likelihood plus the log of the gamma density, excluding terms which do not involve $N$. Similarly for the scaled beta we add the log density of the scaled beta to the log-likelihood to form the penalty term corresponding to the scaled beta prior.

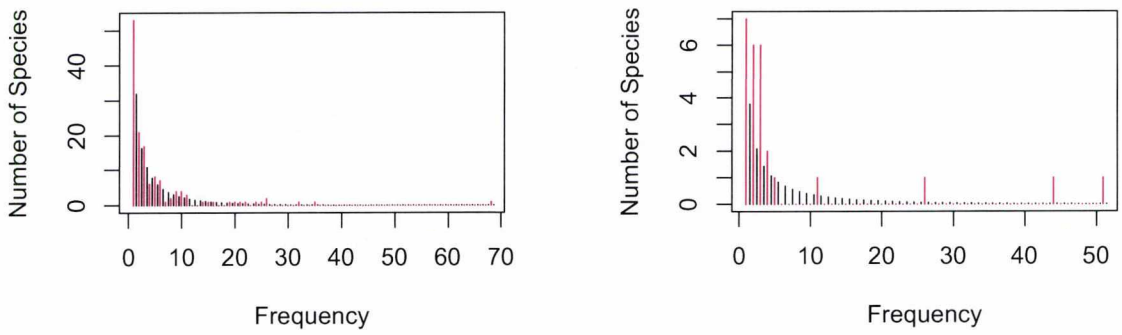
MLE using these penalties gave approximately equivalent results to the Bayesian approach using elicited priors (Tables 6.1 and 6.2).

| | Penalised MLE | | Bayes with elicited prior | |
|---|---|---|---|---|
| Area | $\hat{N}$ | 95% Confidence | Mean | 95% Credible |
| Hastings | 411 | (288, 568) | 410 | (279, 546) |
| IOW $0.1m^2$ | 822 | (567, 1,178) | 818 | (535, 1,117) |
| IOW $0.25m^2$ | 773 | ( 533, 1,120) | 776 | (494, 1,084) |
| Norfolk Pre | 174 | (105, 262) | 171 | (104, 245) |
| Norfolk Post | 38 | (29, 52) | 38 | (28, 50) |

*Table 6.1:* Maximum-likelihood estimates for $N$ for Hastings, Isle of Wight and Norfolk data sets, using elicited penalties on $N$ and fitting the negative binomial model.

| | Penalised MLE | | Bayes with elicited prior | |
|---|---|---|---|---|
| Area | $\hat{N}$ | 95% Confidence | Mean | 95% Credible |
| Hastings | 432 | (306, 586) | 429 | (297, 567) |
| IOW $0.1m^2$ | 897 | (634, 1,255) | 903 | (605, 1,220) |
| IOW $0.25m^2$ | 877 | (623, 1,231) | 890 | (600, 1,194) |
| Norfolk Pre | 189 | (116, 268) | 165 | (94, 243) |
| Norfolk Post | 39 | (29, 54) | 40 | (29, 52) |

*Table 6.2:* Maximum-likelihood estimates for $N$ for Hastings, Isle of Wight and Norfolk data sets, using elicited penalties on $N$ and fitting the Neyman Type A-gamma model.




*(a)* Hastings data

*(b)* Norfolk post-dredging data

*Figure 6.3:* Observed data (red) and fitted data (black) for the penalised MLE using elicited penalty and the negative binomial model fitted to (a) the Hastings data and (b) Norfolk post-dredging data.

For the Hastings data the estimates were very close, but the 95% confidence and 95% credible intervals did not quite correspond. This may have been due to the difference in the way the confidence interval was calculated.

The same pattern occurs for the Isle of Wight data sets and the Norfolk pre- and post-dredging survey estimates. For the latter data set the fit of the negative binomial distribution was not extremely good (Figure 6.3b). This was reflected in the Bayesian $p$-value of 0.75 that I found in Chapter 5.

The results using the Neyman Type A-gamma model were slightly more different between the two methods for the Isle of Wight $0.25m^2$ and Norfolk pre-dredging survey. In the former, the penalised MLE gave a lower estimate of species richness, and in the latter the penalised MLE was higher than using the Bayesian approach with elicited prior, although the differences were not that large.

Our results show that the two methods were approximately equivalent. There were

slightly more differences between the two approaches when using the Neyman Type A-gamma model, however we might expect this as there were more parameters to be estimated within the model. The differences in confidence intervals was more interesting, and due to the differing methods of calculation employed. The HPDI appeared more symmetrical than the corresponding profile confidence intervals, which were more positively skewed.

## 6.3 Comparison of species richness estimation methods

I compared the species richness estimates obtained using the various methods described throughout the thesis. Table 6.3 shows results for the two Isle of Wight data sets.

The non-parametric estimates of species richness were substantially lower than those given by the parametric approaches. I already showed in Chapter 2 that these estimators were insufficient to model benthic data, because the estimates did not reach the total observed species for this area of 273 in most cases. However, I included them here as these estimators are currently used by ecologists to estimate species richness of benthic data. The parametric species richness estimators did produce plausible estimates when using penalties or informative priors.

I have not shown the results of the non-penalised MLEs, as the spuriously large estimates given by these estimators clearly demonstrated they were not suitable. The best of these models, according to AIC, for each of the data sets was the negative binomial using penalty 2. The negative binomial model using penalty 3 with one-step iteration was also a plausible model.

Despite the negative binomial showing a better fit to the data than the Neyman Type A-gamma, the Neyman Type A-gamma model performed well in simulations when applied to clustered data, and should not be dismissed. The negative binomial model has shown negative bias when applied to clustered data,

| Grab size | Estimator | $\hat{N}$ | SD | 95% Confidence | AIC | ΔAIC |
|---|---|---|---|---|---|---|
| $0.25m^2$ | $D$ | 240 | - | - | - | |
| | $\hat{N}_{C1}$ | 286 | 16.71 | (263, 332) | - | |
| | $\hat{N}_{J1}$ | 293 | 10.26 | (273, 313) | - | |
| | NB P2 $\gamma_2 = 0.5$ | 895 | - | (459, 2859) | -578.8 | 0.0 |
| | NB P3 | 496 | - | (388, 650) | -571.2 | 7.6 |
| | NB P3 IT | 720 | - | (469, 1116) | -577.9 | 0.9 |
| | NB EP | 773 | - | (533, 1120) | -566.9 | 11.9 |
| | NTAG P2 $\gamma_2 = 0.5$ | 4206 | - | (904, 16,103) | -550.1 | 28.7 |
| | NTAG P2 $\gamma_2 = 0.25$ | 6818 | - | (1026, 26,541) | -551.7 | 27.1 |
| | NTAG P3 | 561 | - | (441, 719) | -533.1 | 45.7 |
| | NTAG P3 IT | 936 | - | (616, 1380) | -545.8 | 33.0 |
| | NTAG EP | 877 | - | (623, 1231) | -533.9 | 44.9 |
| $0.1m^2$ | $D$ | 198 | - | - | - | |
| | $\hat{N}_{C1}$ | 241 | 15.60 | (219, 284) | - | |
| | $\hat{N}_{J1}$ | 258 | 6.85 | (245, 271) | - | |
| | NB P2 $\gamma_2 = 0.5$ | 1862 | - | (595, 6854) | -496.2 | 0.0 |
| | NB P3 | 469 | - | (363, 611) | -481.2 | 15.0 |
| | NB P3 IT | 767 | - | (498, 1143) | -492.3 | 3.9 |
| | NB EP | 822 | - | (567, 1178) | -472.0 | 24.2 |
| | NTAG P2 $\gamma_2 = 0.5$ | 6344 | - | (1198, 24,782) | -465.9 | 30.3 |
| | NTAG P2 $\gamma_2 = 0.25$ | 9859 | - | (1393, 38,820) | -467.7 | 28.5 |
| | NTAG P3 | 515 | - | (404, 660) | -441.3 | 54.9 |
| | NTAG P3 IT | 920 | - | (615, 1326) | -458.4 | 37.8 |
| | NTAG EP | 897 | - | (634, 1255) | -433.5 | 62.7 |

*Table 6.3:* Comparison of species richness estimates for Isle of Wight $0.25m^2$ and $0.1m^2$ benthic data sets. Estimators shown are $D$: observed number of species, $\hat{N}_{C1}$: the $Chao_1$ estimator, $\hat{N}_{J1}$: the first order jackknife, NB: the negative binomial and NTAG: the Neyman Type A-gamma model, with penalty 2, penalty 3 and penalty 3 with one-step iteration, P3 IT, and the elicited penalty on $N$, EP.

and therefore we need to be wary not to underestimate species richness by using this model.

The approximate agreement in species richness estimates using the parametric models between the two Isle of Wight data sets improves our confidence in the results. The precision of the estimates for the two grabs sizes were very similar, and we might expect that the smaller grabs should give a less precise estimate. However, in patchy environments a more precise estimate can be gained using small grabs than large grabs, per overall unit grab area (Boyd et al., 2006). Since we analysed ten samples for each of the grab sizes, we sample a larger area using the $0.25m^2$ grabs, and the precision should increase. An additional analysis incorporating the grab size and area into the model would be beneficial to tease out these aspects.

The Neyman Type A-gamma model was constrained by having clustering fixed across species. However, by increasing the flexibility of the model to allow clustering to vary between species the fit of the model should improve, and hopefully also the species richness estimate. However, such a model would have many more parameters and there would be more possible sources of variation within the model. It is also possible that the model may be difficult to fit due to a flat likelihood, which is already a possible source of the slow computation time of fitting the Neyman Type A-gamma model.

Although the models incorporating expert opinion did not come out top, if possible I would recommend the use of elicitation to inform the model. Using penalty 2 is subjective, and the penalty parameter chosen can have a large effect on the resulting species richness estimate. Simulations in Chapter 3 showed that the choice of $\gamma_2 = 0.5$ was reasonable for the data analysed. However the estimates given by the Neyman Type A-gamma model using this penalty were high, and the upper confidence limit rises above the 10,000 known species to inhabit UK coastal waters. Although it is possible that some species may be missing from this list, an extensive number of

benthic surveys have been carried out and so we can be reasonably confident in this figure as an upper limit.

Several approaches could be used for species richness estimation in practice, and this could give some form of measure of confidence in the results.

## 6.4 Estimating the impact of dredging

The motivation behind the development of species richness estimators for benthic data, was to be able to monitor the impact on benthic organisms from activities such as marine dredging.

Approximately 16 million tonnes of sand and gravel were removed from around 70 offshore licensed extraction areas located around the coast of England and Wales in 2010 (Crown Estate, 2010). License decisions take into account the amount to be extracted, the rate and duration of extraction, the size of the area to be effected, and the proximity of sensitive areas such as fish feeding and breeding areas (Crown Estate, 2002), and predictions of the consequences of marine aggregate extraction are needed before a licence is awarded. A licence is only issued when predicted environmental impacts are deemed acceptable.

Currently there is a lack of scientific knowledge regarding cause and effect relationships of marine biota, and the assessment relies on a site-specific impact assessment, which can be subjective as there are no standardised criteria (Barry et al., 2010). By developing tools to aid the decision-making process we are able to reduce this subjectivity.

Both the initial impact and the predicted rates of recovery of marine benthos are important. However I concentrated on modelling one of the initial impacts of dredging, the reduction in species diversity. Monitoring programs are carried out before and after aggregate extraction, and these data can be used to assess impacts

on biodiversity.

Barry et al. (2010) used simulation of Matérn processes to estimate the number of species eliminated by dredging. This incorporated individual impact, individuals eliminated during dredging, and species impact, the elimination of an entire species for example due to removal of suitable habitat. The species richness estimator can be used to estimate the impacts of dredging on benthic organisms by comparing pre-and post dredging estimates.

### 6.4.1 Estimating the impact of dredging using the Matérn process

Barry et al. (2010) used a Matérn process to model the abundance and spatial clustering of individual benthic species. Recall that the Matérn process has three components:

1. 'Parent' events form a Poisson process with intensity $\lambda$,
2. Each 'parent' produces a Poisson number of 'daughters' with intensity $\phi$,
3. The positions of the 'daughters' relative to their 'parents' are randomly distributed within a circle of radius $R$.

Barry et al. (2010) estimated the Matérn parameters $\lambda$, $\phi$ and $R$ for each species using a pseudo-maximum-likelihood approach, assuming that species were located independently of each other, and $\lambda$ and $\phi$ were constrained such that the mean of the Matérn process was equal to the mean observed count.

Barry et al. (2010) considered two types of impact corresponding to a loss at the individual level and the species level respectively. The individual impact of dredging was estimated for species found in both the pre- and post-dredging surveys, conditioning on whether an individual was in a dredged or clear area, and assuming that an individual of species $j$ was killed with probability $p_{d_j}$ or $p_{c_j}$ respectively. It was assumed that species densities would be greater in the non-dredged areas.

The mortalities of each species, $p_{d_j}$ and $p_{c_j}$, were estimated for each species by simulation, by repeating the following steps over a grid of potential $p_{d_j}$ and $p_{c_j}$, and then finding the maximum likelihood of these potential values.

First a spatial realisation from the Matérn process was simulated on a 10x10 square and $2.5m$ wide dredge strips were randomly placed onto the spatial realisation to cover a certain percentage of the area to simulate dredging. The points of the spatial realisation were thinned with probability $p_{d_j}$ if in a dredged area, or $p_{c_j}$ in a non-dredged area, and a grab was randomly placed onto the square, and the number of individuals recorded.

This was repeated one thousand times to give a probability distribution for the number of individuals in a grab. The observed counts from the post-dredging survey and this probability distribution were used to calculate the log-likelihood of the Matérn parameters, and estimates of $\boldsymbol{\theta}_j = (R_j, \lambda_j, \phi_j)$ were found by maximising the log-likelihood over a 5x5 grid of potential values for $R_j$ and $\lambda_j$.

By using the estimated mortalities and the estimated Matérn parameters, the frequency distribution of the number of individuals of each species in a post-dredging grab was simulated, assuming the species was not eliminated. For those species observed in the pre-dredge survey, but not the post-dredge survey, a pair of mortality probabilities were randomly allocated from those of the species observed in both surveys.

This gave an estimate of the probability that species $j$ was absent from a post-dredging grab given individual impacts, $\hat{q}_j$. The log-likelihood for the species-level probability that a species was eliminated by dredging, $p_e$ was then estimated , assuming that the presence of each species in the post-dredging survey was independent across species Barry et al. (2010).

When maximised, this likelihood gave an estimate of the probability a species was eliminated, and Barry et al. (2010) obtained an estimate of the number of species eliminated by multiplying the change in the number of species across the two surveys, multiplied by the species elimination probability,

$$n_e = (n_p - n_b)\hat{p_e},$$

where $n_p$ was the number of species seen in the pre-dredging survey, and $n_b$ was the number of species seen in both surveys. A 95% likelihood interval was estimated by taking the values of $p_e$ satisfying $2(l(\hat{p_e}) - l(p_e)) \leq \chi^2_{1:5}$ (Barry et al., 2010).

### 6.4.2 Alternative approach to estimating impact of dredging

I improved the approach of (Barry et al., 2010), using a finer grid of possible parameters of the Matérn process and smoothing the likelihood function across this grid using cubic splines. The same was done when estimating the mortality probabilities of each species. Although not detailed here, this improved the accuracy of the estimates of the Matérn parameters.

However, there are several disadvantages of this approach. For those species observed in the pre-dredge survey, but not the post-dredge survey, a pair of mortality probabilities were randomly allocated from those of the species observed in both surveys, which may not represent the true mortality probabilities of those species. Another disadvantage is that the estimate of the number of eliminated species does not take into account species that were missed in the pre-dredging survey.

Therefore, I used the species richness estimators developed in Chapters 3, 4 and 5 of this thesis. I estimate species richness pre- and post-dredging, and use the difference between the two as an estimate of the number of species eliminated by dredging. This simplifies the calculations of the impact of dredging, and takes into account species that were missed in the pre-dredging survey. This assumes that no new species have entered the population since the pre-dredging survey.

No explicit estimates of the mortality probabilities are required in the species richness estimators, and therefore there is no need to assign arbitrary probabilities to these values for species unobserved in the post-dredging survey. In addition, the probability that a particular species is eliminated by dredging can vary across species.

### 6.4.3 Estimating the impact of dredging in Norfolk - an example

Kenny and Rees (1996) investigated the impacts of marine gravel extraction off the Norfolk coast in 1992, over an area of 135,000$m^2$. Dredging of 70% of this area was carried out over a period of three days, and surveys of marine benthos were carried out during the month before and the month after dredging. Survey data consisted of species counts from five randomly-placed 0.25$m^2$ Hamon grabs. The number of individuals per grab was recorded for each species, and Kenny and Rees (1996) found a significant reduction in the variety, abundance and biomass of benthic organisms after dredging.

A total of 64 species were found in the pre-dredging survey, and 26 in the post-dredging survey, of which three were not seen in the pre-dredging survey. These observed numbers of species are smaller than in the other data sets I have considered, and this is likely to be due to the more gravelly substrate in the Norfolk area, making it suitable for dredging. Also, since a dredging licence had been granted for the area only after an environmental impact assessment was carried out, we might expect that biodiversity was low before dredging took place.

Barry et al. (2010) estimated the number of species that were eliminated by dredging, of those seen in the pre-dredging survey, by modelling the spatial patterns of the 64 species seen in the pre-dredging survey, and the individual impact for the 23 species seen in both surveys. This gave an estimate of four species of those observed pre-dredging eliminated by dredging, with a 95% likelihood interval of 0-14 species.

Table 6.4 shows the estimates of species richness for the Norfolk area pre- and post-

| Data | Estimator | $\hat{N}$ | SD | 95% Confidence | AIC | ΔAIC |
|---|---|---|---|---|---|---|
| Pre-dredging | $D$ | 64 | - | - | - | - |
| | $\hat{N}_{C1}$ | 87 | 28.0 | (72, 129) | - | - |
| | $\hat{N}_{J1}$ | 84 | 4.2 | (76, 92) | - | - |
| | NB P2 | 463 | - | (123, 1658) | -43.5 | 0.0 |
| | NB P3 | 145 | - | (99, 218) | -39.5 | 4.0 |
| | NB P3 IT | 201 | - | (116, 336) | -42.7 | 1.2 |
| | NB EP | 174 | - | (105, 262) | 7.8 | 35.7 |
| | NTAG P2 | 780 | - | (156, 2919) | -25.7 | 17.8 |
| | NTAG P3 | 154 | - | (106, 228) | -19.5 | 23.9 |
| | NTAG P3 IT | 225 | - | (131, 367) | -24.0 | 19.5 |
| | NTAG EP | 189 | - | (116, 268) | 38.8 | 82.3 |
| Post-dredging | $D$ | 26 | - | - | - | - |
| | $\hat{N}_{C1}$ | 29 | 15.6 | (27, 42) | - | - |
| | $\hat{N}_{J1}$ | 34 | 4.6 | (25, 43) | - | - |
| | NB P2 | 89 | - | (34, 279) | 23.0 | 0.0 |
| | NB P3 | 42 | - | (30, 64) | 26.6 | 3.6 |
| | NB P3 IT | 57 | - | (34, 100) | 23.7 | 0.7 |
| | NB EP | 38 | - | (29, 52) | 35.9 | 12.9 |
| | NTAG P2 | 136 | - | (38, 466) | 25.7 | 2.8 |
| | NTAG P3 | 44 | - | (31, 66) | 30.7 | 7.8 |
| | NTAG P3 IT | 61 | - | (36, 107) | 27.0 | 4.0 |
| | NTAG EP | 39 | - | (29, 54) | 40.1 | 17.1 |

*Table 6.4:* Comparison of species richness estimates for Norfolk pre- and post-dredging benthic data sets. Estimators shown are $D$: observed number of species, $\hat{N}_{C1}$: the $Chao_1$ estimator, $\hat{N}_{J1}$: the first order jackknife, NB: the negative binomial and NTAG: the Neyman Type A-gamma model, with penalty 2 with $\gamma_2 = 0.5$, penalty 3 and penalty 3 with one-step iteration, P3 IT, and the elicited penalty on $N$, EP.

| Estimator | Estimated species mortality | 95% Confidence |
|---|---|---|
| Matérn modelling | 4 | (0, 14) |
| $D$ | 38 | - |
| $\hat{N}_{C1}$ | 58 | (30, 102) |
| $\hat{N}_{J1}$ | 50 | (33, 67) |
| NB P2 | 373 | (-157, 1624) |
| NB P3 | 103 | (35, 188) |
| NB P3 IT | 144 | (16, 302) |
| NB EP | 136 | (53, 233) |
| NTAG P2 | 644 | (-310, 2881) |
| NTAG P3 | 110 | (40, 197) |
| NTAG P3 IT | 164 | (24, 331) |
| NTAG EP | 150 | (62, 239) |

*Table 6.5:* Estimates of number of species eliminated by dredging in 1992 for the Norfolk area, for estimators $D$: observed number of species, $\hat{N}_{C1}$: the $Chao_1$ estimator, $\hat{N}_{J1}$: the first order jackknife, NB: the negative binomial and NTAG: the Neyman Type A-gamma model, with penalty 2 penalty 3 and penalty 3 with one-step iteration, P3 IT, and the elicited penalty on $N$, EP.

dredging, and the estimated numbers of eliminated species are given in Table 6.5, along with confidence intervals formed by taking the greatest and smallest differences between the pre- and post dredging confidence limits. Table 6.5 shows that the estimated impact of dredging using species richness estimators was very different from the estimated number of species eliminated of 4 (0, 14) estimated by Barry et al. (2010). However, this estimate was concerned with estimating only how many were eliminated of those species observed in the pre-dredging survey.

If we consider the number of observed species for each survey, $D$, we would estimate that 38 species were eliminated. However, this did not account for those which were unobserved un the pre-dredging survey, so this was not a good estimator for the number of eliminated species.

The rest of the estimators attempt to account for missed species by estimating species richness before and after impact, and the estimated species mortality given by these estimators was much higher than the estimate of Barry et al. (2010).

The estimates of mortality given by the negative binomial and Neyman Type A-gamma models using the elicited priors are fairly similar, suggesting that the prior is having a strong influence over the posterior or that the priors fit the data well. The coefficients of overlap for the Norfolk data, calculated in Chapter 5 for the negative binomial model, were fairly low values of $\tau_\theta = 0.4$ and $\tau_\theta = 0.17$ for the pre- and post-dredging data respectively. Therefore this suggested that the prior was not having total influence over the results and these estimates of species richness are reasonable. Therefore, the estimate of dredging impact from these models should also be reasonable.

In Section 5.9.2 the DIC judged the Neyman Type A-gamma to be the best fitting model for the Norfolk pre-dredging data, but the AIC values in Table 6.4 judged the negative binomial MLE using penalty 2 to be the best fitting. This highlighted that

caution should be used when considering these information criteria to choose between models, especially the DIC.

The estimated species mortality using both the negative binomial and Neyman Type A-gamma model and the penalised MLE with penalty 2 had negative mortality as a lower bound, because of the way the upper and lower limits for the estimate were calculated. The highest estimate within the 95% confidence interval of species richness post-dredging was greater than the lowest estimate of the 95% confidence interval of species richness pre-dredging. This suggested that there could be more species present post-dredging than there were pre-dredging.

This is not impossible, as some species could colonise the substrate very quickly after dredging, and also some species may prefer a more sandy habitat which is left post-dredging. However, this goes against our assumption of no new species entering the population, so this negative mortality estimate was due to the very wide confidence intervals associated with penalty 2 when the penalty parameter has not been tuned. A value of $\gamma_2 = 0.5$ was used here. Although this model was judged best by the AIC, I would advise against using this estimate until some way to choose the appropriate penalty parameter had been found, and we can be more confident in the results given by this estimator.

Excluding the estimators using penalty 2, the estimate of dredging impact given by the parametric models are fairly similar, in the order of 100-170 species. This is encouraging, and by considering several estimators we can gain a rough idea of the impact dredging has had on the biodiversity in the Norfolk area.

## 6.5 Discussion

The aim of this chapter was to consider the link between the frequentist and Bayesian approaches to species richness estimation, and to compare these methods with established methods of species richness estimation when applied to the assessment

of dredging impact on benthic organisms. I outlined the link between priors in the Bayesian approach, and penalties in a frequentist approach, and illustrated this using the penalties described in Chapter 3. For large samples and a uniform prior, the Bayes estimate and MLE will be equivalent, so using an informative prior may be considered equivalent to using a penalty in the frequentist approach.

I showed that since $N$ is a function of the odds parameter, I was able to transform a normal prior on $N$ to a penalty on the odds parameter. However, it was also possible to use the prior on $N$ directly as a penalty in the log-likelihood of the MLE, and this allowed the conversion of a prior on $N$ such as the gamma. The two methods were approximately equivalent, as we would expect.

This allowed us to incorporate expert opinion easily into the frequentist approach, which will overcome the boundary problem and allow use of the MLE for cases where this phenomenon is present. The frequentist approach may be preferred in some cases, especially when using the more complicated Neyman Type A-gamma to model the abundance distribution of benthic organisms. An original motivation behind using a Bayesian approach was to eliminate the calculation of the marginal likelihood for the Neyman Type A-gamma model, however using the DA method constrained us to using particular priors for $N$, and RJMCMC within `WinBUGS` also proved difficult when trying to incorporate the elicited priors.

I compared several species richness estimators for the Isle of Wight area, and found that the estimates given by the non-parametric estimators were appreciably lower than those given by the parametric approach. These estimators were insufficient to model benthic data. However the parametric species richness estimators did produce plausible estimates when using penalties or informative priors.

I would recommend, if possible, the use of elicitation to inform the prior or penalty to be used to combat the boundary problem. Simulations showed that the best penalised

method, MLE with penalty 2, was not always reliable, and using expert knowledge could enhance the accuracy of the species richness estimates. The negative binomial showed a better fit to the benthic data analysed than the Neyman Type A-gamma, but the Neyman Type A-gamma performed well in simulations, and should not be dismissed.

The precision of the species estimates using the MLE on the two Isle of Wight data sets were very similar, and an additional analysis incorporating the grab size and area into the model, as touched upon in Section 3.3, would be beneficial to tease out these aspects.

The motivation for the development of species richness estimators for benthic data was to be able to quantify impacts on benthic organisms caused by activities such as dredging, or climate change. As licences for dredging are only awarded after environmental impacts have been deemed acceptable, the development of good performing species richness estimators can contribute to improving the decision-making process, by reducing subjectivity and ensuring impacts are assessed accurately.

Barry et al. (2010) focused on estimating the initial impact of dredging, and the decline in species number, by modelling the impact on each species directly using a Matérn process to model the spatial distribution of individuals within a species. However, although clustering is accounted for, there was no accounting for species that were unobserved in the surveys. It is likely that several species were missed during surveying and by estimating species richness in the pre- and post-dredging surveys we can see the whole picture, rather than a snapshot of species observed. I compared several methods of estimating these impacts, and found that it was important to consider unobserved species in the estimates, otherwise the impact of dredging could be significantly underestimated. I found an impact of around 100-170 species eliminated by dredging, and the estimates of dredging impact given by the

parametric models were fairly similar, and I would recommend considering several estimators.

The use of species richness estimates also has advantages over the approach taken by Barry et al. (2010). For those species observed in the pre-dredge survey, but not the post-dredge survey, a pair of mortality probabilities were randomly allocated from those of the species observed in both surveys. These assigned values may not be very representative of the true mortality probabilities of such species, and if such unrepresentative values are used within the estimate of impact, the estimate given may also be unrepresentative. We did not make this assumption using the species richness estimators.

Some species are more sensitive to disturbance of the seabed than others, and by using the species richness estimates we are not constraining the probability that a species is eliminated due to dredging to be fixed across species. Rates of biodiversity recovery in dredged areas will depend on whether time between dredging activity is sufficient for organisms to reproduce and for new recruits to settle. Many species start out in an initial phase of flotation, which may allow the rapid colonisation of previously dredged areas if the substrate is in a suitable condition (Kenny, personal communication).

The approach of modelling species impacts using a Matérn process assumed that dredging was carried out in even strips across the study area. However in practice dredging tends to be targeted to particular deposits (Barry et al., 2010). This would also correspond to sediment type, so that if species were clustering in areas of preferred large sediment habitat, they were much more likely to be eliminated by targeted dredging of gravel. This could be incorporated into the Barry et al. (2010) approach, however to do so it would be necessary to have more information about the aggregate being targeted, the proposed tonnage to be extracted and the rate and duration of extraction, all things taken into account before a dredging licence is granted (Barry

et al., 2010). Therefore it should be possible to acquire the information needed to be able incorporate these aspects into the model without too much difficulty.

The application of the estimators was illustrated through an example looking at initial dredging impact, but these estimators can also be used to monitor the recovery of the seabed after dredging, or track changes in species richness over time in general. In addition, species interact with each other, and with their habitat, and it would be interesting to incorporated these aspects into the estimators.

## 6.6 Conclusions

In this chapter I highlighted the equivalence of the frequentist and Bayesian approaches incorporating penalties and priors, and by comparing several estimators I concluded that the best approach is to use several estimators in practice, and this would give some kind of measure of confidence in the results.

The negative binomial model proved to be the most plausible model, but I believe the Neyman Type A-gamma model could be improved by allowing clustering to vary between species. Theoretically this could be incorporated easily into the model. However, in practice this has proved more complicated to implement, and this is an area of further work.

Using the species richness estimators allows us to explore the effects of impacts such as dredging, incorporating the spatial distributions of benthic organisms. The species richness approach to estimating dredging impact is one that could be used in decision making if software was made freely available. If site-specific elements were also incorporated, then this would be a useful tool in understanding the implications of local changes due to dredging, and reduce subjectivity of decisions on granting of dredging licences.

# 7. CONCLUSIONS AND FURTHER WORK

The study and analysis of benthic organisms presents a number of statistical challenges, and this thesis has addressed some of these issues. The main focus of the thesis research has been on modelling the spatial distribution of benthic organisms, specifically by describing the clustering of individuals within a species using a Neyman Type A distribution.

The objectives of the research were to review the current state of the art in species richness estimation, develop methodology to incorporate the spatial clustering of individuals within a species into multinomial species richness models, and to apply these models to benthic data in order to assess impacts on marine biodiversity.

Chapter 2 showed that there is scope for the development of a new species richness estimator for use with benthic data. Current methods, including non-parametric estimators, do not deal well with spatially clustered data, and this was confirmed by a simulation study. In some instances the number of observed species is used as an estimate of species richness, but merely using the number of observed species is inadequate for species richness estimation for benthic organisms, as it will clearly be an underestimate.

I also introduced a clustering model that could be used to describe the spatial distribution of benthic organisms. Chapter 3 considered how this model could be built into a parametric species richness estimator via a multinomial model.

Multinomial models have been used for estimating species richness within a maximum-

likelihood framework for some time, and much work has been carried out in this area. However, the distributions used to describe the abundance of species within these models do not encompass the spatial clustering of individuals. I incorporated this into the model in the form of the Neyman Type A distribution. The mean abundance of species was allowed to vary according to a gamma distribution, and the mean number of individuals per cluster was fixed across species. Further work will include an extension to this model to allow clustering intensity to vary between species, and I would expect that this would improve the fit of the model to benthic data. Alternative clustering distributions, to describe the spatial patterns of benthic organisms, could also be investigated.

Problems associated with parametric species richness estimation included the choice of method for constructing confidence intervals, and how to deal with spuriously large estimates.

I considered several approaches to estimate confidence intervals, and found that confidence intervals based on profile log-likelihoods performed well in terms of coverage. Therefore, I used this method in estimating confidence intervals for the MLE and penalised MLE. However, further investigation into the theory behind using these confidence intervals for penalised likelihoods is required. The profile log-likelihood intervals were quite wide in some cases, and the question arises as to whether these are appropriate. In addition, the calculation of profile log-likelihood confidence intervals for the Neyman Type A-gamma model is computationally intensive, increasing the computation time considerably, so alternative methods may be preferred.

Wang and Lindsay (2005) used bootstrap intervals for their non-parametric MLE, and so some form of bootstrap interval could be appropriate. Wang and Lindsay (2005) suggested the use of multinomial-based bootstrap confidence intervals, and further investigation into these could prove fruitful. During simulations, the confidence

intervals for $\log(N)$, suggested by Fewster and Jupp (2009), were very narrow, and did not perform well in terms of coverage. However, these intervals did perform well for data sets described in Fewster and Jupp (2009) and it would be beneficial to investigate this method further.

I found that spuriously large species richness estimates obtained during maximum-likelihood estimation were caused by the boundary problem. I combatted this problem by penalising the log-likelihood, following the method of Wang and Lindsay (2005). One of these penalties used the $Chao_1$ species richness estimate as a naive estimate, and penalised the log-likelihood towards that value. However, simulations showed that this penalty could be too harsh for clustered data, as the $Chao_1$ estimator showed large negative bias. Therefore, I proposed the use of an iterative approach to decrease the harshness of this penalty, focussing on one-step iteration due to the computational cost of applying this method for the complex Neyman Type A-gamma model.

Another of the penalties considered was subjective, requiring the specification of a penalty parameter, the optimal value of which depended on the sampling depth of the data. Since there was no easy way to choose the optimal penalty parameter, as sampling depth cannot readily be determined for real data sets, an alternative approach to avoiding the 'boundary problem' was considered. This was the use of priors within a Bayesian framework. A Bayesian approach could also eliminate the need to evaluate the marginal likelihood, which would speed up the computation time of fitting the Neyman Type A-gamma model.

The Neyman Type A-gamma model is a reasonably complex model to use in species richness estimation, as the probability density function includes a summation, which complicates the calculation of the integrated or marginal likelihood. In addition, the calculation of confidence intervals is also time consuming. The profile log-likelihood confidence intervals take several hours to calculate, and it is likely that the optimisation is getting stuck. Further investigation into this problem could reduce

computation time. I used the built-in `optimise` and `optim` functions within `R` both based on the Nelder-Mead algorithm, but perhaps using the EM algorithm could increase computational speed.

Chapter 4 investigated the Bayesian approach to species richness estimation, using uninformative priors on the number of species and the parameters of the abundance distribution. To be able to model the species that were not seen in the sample, reversible jump MCMC and data augmentation approaches were considered in a hierarchical Bayesian framework. I analysed the Lepidoptera and CBC data sets using this approach with the Poisson, Poisson-exponential, negative binomial and Neyman Type A-gamma models. Unfortunately, uninformative priors were not sufficient to model benthic data. When fitting the models to these data, the boundary problem manifested itself through a lack of convergence of the Markov chains within MCMC.

Chapter 5 considered the elicitation of informative priors for benthic data, and how these could be incorporated into a Bayesian framework. A process of elicitation of information from experts was described, and the resulting priors for the total number of species were incorporated into the species richness estimates. This is a particularly interesting aspect of the research, which as far as I am aware has not been considered previously for benthic data. This chapter highlighted difficulties within the elicitation process and applied the elicited priors in estimating species richness for a number of benthic data sets.

Through elicitation I was able to obtain priors that gave realistic species richness estimates, and although the fit using the Neyman Type A-gamma using these priors was not as good as the negative binomial, the species richness estimates seemed realistic. The best-fitting model to the data may not always give better real-world answers. Gelman et al. (2004) illustrated this when estimating populations of municipalities in New York. They showed for this example that even when a model

appears to fit well to observed data, it can yield inaccurate inferences. Therefore, there is a need to include realistic prior assumptions into the model, such as the clustering of organisms in the benthic case. Evidence exists of the clustering nature of benthic organisms (Heip, 1975), therefore a suitable model to analyse benthic data should be flexible enough to incorporate this spatial clustering.

The elicitation process is complex and requires careful planning and implementation. However, the results can be very beneficial. As we have a database of 10,000 known species in UK waters, the results from the elicitation of the number of species found throughout UK waters could be used as an indicator of the accuracy of the other estimates given by the experts. Although this is not something that I considered within this thesis, it is an interesting area for further research. The number of species found throughout UK waters could also be incorporated into the model as an upper estimate of the species richness.

It has been shown that the use of appropriate software can greatly improve the elicitation process, and the development of more specialised software for the elicitation of information particular to benthic organisms could encourage this practice to be more widely used. This software could be incorporated into a larger package to be used to estimate species richness and diversity of benthic organisms. In addition, the software could be made flexible enough to be used for other taxonomic groups, or alternative applications.

The link between Bayesian priors, and penalties in a frequentist framework for species richness estimation was explored in Chapter 6. The parametric estimates were then compared to those of some of the non-parametric estimators that are currently used for species richness estimation of benthic organisms. If we are confident in the expert opinion available, we should try and incorporate this into species richness estimates. However, if this is not possible then the one-step iterative penalty 3 has proved reasonable. Several approaches could be used in practice, and this would give an

additional measure of confidence in the results.

It is sensible to start with a simple model, and expand it, which is what I have done, considering the Poisson, negative binomial and Neyman Type A-gamma models. I believe the model could be improved further by allowing clustering to vary between species. Theoretically, this could be incorporated easily into the model using the hierarchical Bayes approach and giving the clustering parameter a prior. However, in practice this has proved complicated to implement within `WinBUGS` using uninformative priors, but might be improved by using an informative prior on this clustering parameter or by considering grouping for the clustering parameter. Often a more complex model can make more sense and fit the data better, but will be more difficult to understand and compute (Gelman et al., 2004, p 180). Therefore, we should consider the application of the models when adding complexity to the model.

A result that has been shown throughout the thesis, is that the estimate of species richness given by the Neyman Type A-gamma model is always higher than that given by the negative binomial model for a particular data set. This could be an aspect of the increased variance in the Neyman Type A distribution. By allowing clustering of individuals within the model, we would expect to miss more species during sampling and therefore gain a larger species richness estimate. It would be interesting to investigate this further, and prove if this would always be the case.

The species-richness methodology allows us to explore the effects of impacts such as dredging, incorporating the spatial distributions of benthic organisms. The species-richness approach to estimating dredging impact is one that could be implemented by ecologists and other end users if software was made freely available. If site-specific elements, such as substrate, were also incorporated into the estimation of species richness, this would be a useful tool in understanding the implications of local changes due to dredging, and better predictions would be possible to aid the decision on the granting of dredging licences.

For some of the benthic data sets, the estimates arising from applying the different models and penalties were quite different. I would suggest a precautionary approach to estimating species richness. For example, if we are interested in determining whether a dredging licence should be granted for a particular area, we would not want to grant a licence if there was any possibility that the area could be rich in biodiversity. Therefore, adopting a precautionary approach, overestimation of species richness would be preferred to underestimation.

Several additional statistical questions arise from the analysis of benthic data, such as the design of benthic surveys to assess dredging impact. This area requires some consideration, as in many cases only five or ten samples are taken per survey. This limits the possibility to validate models, and only allows for estimates of low precision. More grabs are needed to make robust inferences about species, particularly those that will be rare after dredging. However, increasing the sample size greatly increases sampling effort and therefore may not be economically viable.

In this research I have focussed on estimating species richness, but there are several other biodiversity measures which could be used to analyse impacts. Numerous indices have been developed to calculate measures of biodiversity, and many of these measures incorporate not only the number of species, but also a measure of variation in numbers of individuals of each species. These indices assume that individuals are randomly sampled from an indefinitely large population and that all species are represented in the sample, and are therefore not suitable for use on benthic data in their current form. Despite this, they are commonly used by ecologists, so an adjustment to account for unobserved species and also clustering of organisms would be beneficial.

Other alternative measures of biodiversity consider traits of species within a community, believing that it is not the specific species but the functions they perform

within the ecosystem that are important. This is an area of growing interest (Petchey and Gaston, 2006). Under this approach, the loss of a particular species may not be as important as the loss of a particular function within an ecosystem.

It is also suggested that biodiversity measures should not only measure species richness and abundance, but also include a measurement of the relatedness of those species. Such measures will be higher in an area where species are evenly spread over many genera, rather than all in the same genus (Clarke and Warwick, 1998).

A key concern is to make sure that any measure of biodiversity, whether species richness or taxonomic distinctiveness, is easily understandable and computable. This will ensure that these approaches are taken up by scientists and decision makers. It is all very well developing complicated models and estimators, but without ease of application, it is time wasted.

There are already many species richness estimation methods available for analysis of data, and often a key concern for ecologists is which is the most appropriate statistical method to use. Although I have shown that several of the non-parametric species richness estimators are not suitable for clustered data, they did all show an improvement on using the observed number of species alone. Realistically, this is the measure of species richness that is used in many analyses, and so the use of any species richness estimator is better than this.

I would recommend the use of a number of estimators, instead of choosing the 'best' method. This will allow the user to make an informed decision, and also give some form of measure of uncertainty. This kind of thinking should be encouraged, and can be achieved through continued collaboration between statisticians and ecologists.

# BIBLIOGRAPHY


Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symnposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest.

Allan, J. D., Yuan, L. L., Black, P., Stockton, T., Davies, P. E., Magierowski, R. H., and Read, S. M. (2011). Investigating the relationships between environmental stressors and stream condition using Bayesian belief networks. *Freshwater Biology* doi: 10.1111/j.1365-2427.2011.02683.x.

Baltanas, A. (1992). On the use of some methods for the estimation of species richness. *Oikos* **65,** 484–492.

Barger, K. and Bunge, J. (2008). Bayesian estimation of the number of species using noninformative priors. *Biometrical Journal* **50,** 1064–1076.

Barger, K. and Bunge, J. (2010). Objective Bayesian estimation for the number of species. *Bayesian Analysis* **5,** 765–786.

Barry, J. (2009). A model-based framework for the estimation of species richness from grab or quadrat samples. unpublished.

Barry, J., Boyd, S., and Fryer, R. (2010). Modelling the effects of marine aggregate extraction on benthic assemblages. *Journal of the Marine Biological Association of the UK* **90,** 105–114.

Behnke, A., Bunge, J., Barger, K., Breiner, H.-W., Alla, V., and Stoeck, T. (2006). Microeukaryote community patterns along an $O_2/H_2S$ gradient in a supersulfidic anoxic fjord (Framvaren, Norway). *Applied and Environmental Microbiology* **72,** 3626–3636.

Bernardo, J. M. and Ramon, J. M. (1998). An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *Journal of the Royal Statistical Society. Series D (The Statistician)* **47,** 101–135.

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics* **1,** 356–358.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics* **1,** 353–355.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* **57,** 1–30.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* **9,** 176–200.

Böhning, D. (2009). Population size estimation, nonparametric mixtures, and empirical Bayesian smoothing. Symposium on Biostatistics and Statistical Genetics.

Bolam, S. (2011). Burial survival of benthic macrofauna following deposition of simulated dredged material. *Environmental Monitoring and Assessment* **181,** 13–27.

Boyd, S. E., Barry, J., and Nicholson, M. (2006). A comparative study of a $0.1m^2$ and $0.25m^2$ Hamon grab for sampling macrobenthic fauna from offshore marine gravels. *Journal of the Marine Biological Association of the UK* **86,** 1315–1328.

Britannica, E. (2012). Benthos. `http://www.britannica.com/EBchecked/topic/61141/benthos`. Encyclopaedia Britannica Online, Accessed: 11 June 2012.

Brooks, S. P. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative samples. *Journal of Computational and Graphical Statistics* **7,** 434–455.

Brose, U., Martinez, N. D., and Williams, R. J. (2003). Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology* **84,** 2364–2377.

Buckland, S. T., Magurran, A. E., Green, R. E., and Fewster, R. M. (2005). Monitoring change in biodiveristy through composite indices. *Philosophical Transactions of the Royal Society B* **360,** 243–254.

Burgman, M., Carr, A., Godden, L., Gregory, R., McBride, M., Flander, L., and Maguire, L. (2011). Redefining expertise and improving ecological judgment. *Conservation Letters* **4,** 81–87.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multi-model inference.* Springer-Verlag.

Burnham, K. P., White, G. C., and Anderson, D. R. (1995). Model selection strategy in the analysis of capture-recapture data. *Biometrics* **51,** 888–898.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes methods for data analysis.* Texts in Statistical Science Series. Chapman and Hall/CRC.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11,** 265–270.

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

Chao, A. (2004). Species richness estimation. In Balakrishnan, N., Read, C., and Vidakovic, B., editors, *Encyclopedia of Statistical Sciences*, pages 7906–7916. Wiley Press, New York, USA.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58,** 531–539.

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

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

Chao, A. and Lin, C.-W. (2012). Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics* doi: 10.1111/j.1541-0420.2011.01739.x.

Chen, Y., Breen, P. A., and Andrew, N. L. (2000). Impacts of outliers and mis-specification of priors on Bayesian fisheries-stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences* **57,** 2293–2305.

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

Clarke, K. R. and Warwick, R. M. (1998). A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology* **35,** 523–531.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* **345,** 101–118.

Condit, R., Hubbell, S. P., Lafrankie, J. V., Sukumar, R., Manokaran, N., Foster, R. B., and Ashton, P. S. (1996). Species-area and species-individual relationships for tropical trees: A comparison of three 50-ha plots. *Journal of Ecology* **84,** 549–562.

Cooke, R. M. and Goossens, L. H. J. (2004). Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research* **7,** 643 – 656.

Cooper, K., Boyd, S., Aldridge, J., and Rees, H. (2007). Cumulative impacts of aggregate extraction on seabed macro-invertebrate communities in an area off the east coast of the united kingdom. *Journal of Sea Research* **57,** 288–302.

Craig, C. C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika* **40,** 170–176.

Crown Estate, T. (2002). *Marine Mineral Guidance 1: Extraction by Dredging from the English Seabed.* The Crown Estate, London.

Crown Estate, T. (2010). *Marine Aggregates: The Crown Estate Licences Summary Statistics 2010.* The Crown Estate, London.

Cruyff, M. J. L. F. and van der Heijden, P. G. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal* **50,** 1035– 1050.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68,** 589–599.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82,** 171–185.

Evans, D. A. (1953). Experimental evidence concerning contagious distributions in ecology. *Biometrika* **40,** 186–211.

Fager, E. (1972). Diversity: a sampling study. *American Naturalist* **106,** 293–310.

Favaro, S., Lijoi, A., Mena, R. H., and Prnster, I. (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71,** 993–1008.

Favaro, S., Lijoi, A., Mena, R. H., and Prnster, I. (2011). On some issues related to species sampling problems. 7th Conference on Statistical Computation and Complex Systems.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1,** 209–230.

Fewster, R. M. and Jupp, P. E. (2009). Inference on population size in binomial detectability models. *Biometrika* **96,** 805–820.

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K., and Caley, M. J. (2012). A software tool for elicitation of expert knowledge about species richness or similar counts. *Environmental Modelling and Software* **30,** 1 – 14.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (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.

Foggo, A., Attrill, M. J., Frost, M. T., and Rowden, A. A. (2003). Estimating marine species richness: an evaluation of six extrapolative techniques. *Marine Ecology Progress Series* **248,** 15–26.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics* **56,** 1055–1067.

Gelman, A. (1996). Inference and monitoring convergence. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice*, pages 131–143. Chapman and Hall, London.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis.* Chapman and Hall/CRC.

George, E. I., Makov, U. E., and Smith, A. F. M. (1993). Conjugate likelihood distributions. *Scandinavian Journal of Statistics* **20,** 147–156.

Gimenez, O., Morgan, B. J., and Brooks, S. P. (2009). Weak identifiability in models for mark-recapture-recovery data. In Thomson, D. L., Cooch, E. G., Conroy, M. J., and Patil, G. P., editors, *Modeling Demographic Processes In Marked Populations*, volume 3 of *Environmental and Ecological Statistics*, pages 1055–1067. Springer US.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40,** 237–264.

Gotelli, N. and Colwell, R. (2010). Estimating species richness. In Magurran, A. and McGill, B., editors, *Biological Diversity: Frontiers In Measurement And Assessment*, pages 39–54. Oxford University Press, Oxford.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Grupp, H. and Linstone, H. A. (1999). National technology foresight activities around the globe: Resurrection and new paradigms. *Technological Forecasting and Social Change* **60,** 85 – 94.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109.

Heck, Kenneth L., J., Belle, G. v., and Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* **56,** 1459–1461.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research* **31,** 1109 – 1144.

Heip, C. (1975). On the significance of aggregation in some benthic marine invertebrates. In Barnes, H., editor, *Ninth European Marine Biology Symposium*, pages 527–538.

Hellmann, J. J. and Fowler, G. W. (1999). Bias, precision, and accuracy of four measures of species richness. *Ecological Applications* **9,** 824–834.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns.* Statistics in Practice. John Wiley and Sons Ltd, Chichester.

James, A., Choy, S. L., and Mengersen, K. (2010). Elicitator: An expert elicitation tool for regression in ecology. *Environmental Modelling and Software* **25,** 129 – 145.

Johnson, N., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, Inc.

Karakassis, I. (1995). $S_\infty$: A new method for calculating macrobenthic species richness. *Marine Ecology Progress Series* **120,** 299–303.

Keating, K. A. and Quinn, J. F. (1998). Estimating species richness: the Michaelis-Menten model revisited. *Oikos* **81,** 411–416.

Kenny, A. J. and Rees, H. L. (1996). The effects of marine gravel extraction on the macrobenthos: Results 2 years post-dredging. *Marine Pollution Bulletin* **32,** 615–622.

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

King, R., Morgan, B. J., Gimenez, O., and Brooks, S. P. (2010). *Bayesian Analysis for Population Ecology.* Interdisciplinary Statistics Series. Chapman and Hall/CRC.

Kuhnert, P. M. (2011). Four case studies in using expert opinion to inform priors. *Environmetrics* **22,** 662–674.

Kuhnert, P. M., Martin, T. G., and Griffiths, S. P. (2010). A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecology Letters* **13,** 900–914.

Kuhnert, R., Del Rio Vilas, V. J., Gallagher, J., and Böhning, D. (2008). A bagging-based correction for the mixture model estimator of population size. *Biometrical Journal* **50,** 993– 1005.

Kynn, M. (2005). *Eliciting Expert Knowledge for Bayesian Logistic Regression in Species Habitat Modelling.* PhD thesis, Queensland University of Technology.

Li, L., Huang, Z., Ye, W., Cao, H., Wei, S., Wang, Z., Lian, J., Sun, I. F., Ma, K., and He, F. (2009). Spatial distributions of tree species in a subtropical forest of china. *Oikos* **118,** 495–502.

Lijoi, A., Mena, R. H., and Pranster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94,** 769–786.

Lindsay, B. G. and Roeder, K. (1987). A unified treatment of integer parameter models. *Journal of the American Statistical Association* **82,** 758–764.

Linstone, H. and Turoff, M. (1975). *The Delphi Method: Techniques and Applications.* Addison-Wesley, Reading.

Ludwig, J. A. and Reynolds, J. F. (1988). *Statistical Ecology: A Primer on Methods and Computing*. Wiley-Blackwell.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* pages 325–337.

Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84,** 19–31.

Magurran, A. E. (1991). *Ecological diversity and its measurement*. Chapman and Hall, London.

Magurran, A. E. (2004). *Measuring biological diversity*. Wiley-Blackwell.

Mao, C. X. (2007). Estimating the number of species with multiple incidence-based subsamples. *Statistica Sinica* **17,** 1591–1600.

Martin, D. C. and Katti, S. K. (1962). Approximations to the Neyman Type A distribution for practical problems. *Biometrics* **18,** 354–364.

Martin, T. G., Burgman, M. A., Fidler, F., Kuhnert, P. M., Low-Choy, S., Mcbride, M., and Mengersen, K. (2012). Eliciting expert knowledge in conservation science. *Conservation Biology* **26,** 29–38.

Matérn, B. (1986). *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer, New York, 2nd edition.

McBride, M. F. and Burgman, M. A. (2012). What is expert knowledge, how is such knowledge gathered, and how do we use it to address questions in landscape ecology? In Perera, A. H., Drew, C. A., and Johnson, C. J., editors, *Expert Knowledge and Its Application in Landscape Ecology*, pages 11–38. Springer New York.

McIntosh, R. P. (1967). An index of diversity and the relation of certain concepts to diversity. *Ecology* **48,** 392–404.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21,** 1087–1092.

Moreno, M. and Lele, S. R. (2010). Improved estimation of site occupancy using penalized likelihood. *Ecology* **91,** 341–346.

Morgan, B. J. T. (1984). *Elements of simulation.* Chapman and Hall, London.

Morgan, B. J. T. (2009). *Applied Stochastic Modelling*, volume 79 of *Texts in Statistical Science Series.* Chapman and Hall/CRC.

Morgan, B. J. T. and Ridout, M. S. (2009). Estimating N: A robust approach to capture heterogeneity. In Thomson, D. L., Cooch, E. G., and Conroy, M. J., editors, *Modeling Demographic Processes In Marked Populations*, volume 3 of *Environmental and Ecological Statistics*, pages 1069–1080. Springer US.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)* **20,** 1–43.

Oakley, J. E. and O'Hagan, A. (2010). SHELF: the Sheffield Elicitation Framework (version 2.0). http://tonyohagan.co.uk/shelf.

O'Dea, N., Whittaker, R. J., and Ugland, K. I. (2006). Using spatial heterogeneity to extrapolate species richness: a new method tested on Ecuadorian cloud forest birds. *Journal of Applied Ecology* **43,** 189–198.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *Journal of the Royal Statistical Society. Series D (The Statistician)* **47,** 21–35.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60,** 607–612.

Petchey, O. L. and Gaston, K. J. (2006). Functional diversity: back to basics and looking forward. *Ecology Letters* **9,** 741–758.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In T. S.Ferguson, L. S. and MacQueen, J., editors, *Statistics, Probability and Game Theory*, pages 245–267. Hayward: Institute of Mathematical Statistics.

Pledger, S. and Phillpot, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal* **50,** 1022–1034.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. Technische Universit Wien, Vienna. 3rd International Workshop on Distributed Statistical Computing.

Quince, C., Curtis, T. P., and Sloan, W. T. (2008). The rational exploration of microbial diversity. *The ISME Journal* **2,** 997–1006.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice*, pages 45–58. Chapman and Hall, London.

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

Royle, J. A. (2009). Analysis of capture-recapture models with individual covariates using data augmentation. *Biometrics* **65,** 267–274.

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

Royle, J. A. and Young, K. V. (2008). A hierarchical model for spatial capture recapture data. *Ecology* **89,** 2281–2289.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics* **43,** 142–152.

Sanathanan, L. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association* **72,** 669–672.

Schmid, F. and Schmidt, A. (2006). Nonparametric estimation of the coefficient of overlappingtheory and empirical application. *Computational Statistics and Data Analysis* **50,** 1583 – 1596.

Schofield, M. R. and Barker, R. J. (2010). Data augmentation and reversible jump MCMC for multinomial index problems. arXiv:1009.3507v1 [stat.AP].

Shen, T.-J., Chao, A., and Lin, C.-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84,** 798–804.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, NewYork, London.

Skellam, J. G. (1958). On the derivation and applicability of Neyman's type A distribution. *Biometrika* **45,** 32–36.

Smith, E. P. and van Belle, G. (1984). Nonparametric estimation of species richness. *Biometrics* **40,** 119–129.

Snelgrove, P. V. R. (1997). The importance of marine sediment biodiversity in ecosystem processes. *Ambio* **26,** 578–583.

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

Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88,** 657–675.

Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)* **36,** 148–159.

Tyler, E. H. M., Somerfield, P. J., Berghe, E. V., Bremner, J., Jackson, E., Langmead, O., Palomares, M. L. D., and Webb, T. J. (2012). Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology. *Global Ecology and Biogeography* .

Ugland, K. I. and Gray, J. S. (2004). Estimation of species richness: Analysis of the methods developed by Chao and Karakassis. *Marine Ecology Progress Series* **284,** 1–8.

Ugland, K. I., Gray, J. S., and Ellingsen, K. E. (2003). The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology* **72,** 888–897.

van der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., and van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated poisson regression model. *Statistical Modelling* **3,** 305–322.

Walther, B. and Morand, S. (1998). Comparitive performance of species richness estimation methods. *Parasitology* **116,** 395–405.

Walther, B. A. and Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **28,** 815–829.

Wang, J.-P. and Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100,** 942–959.

Wang, J.-P. and Lindsay, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5,** 30–45.

Wang, X., He, C. Z., and Sun, D. (2007). Bayesian population estimation for small sample capture-recapture data using noninformative priors. *Journal of Statistical Planning and Inference* **137,** 1099–1118.

# APPENDIX

# A. DERIVATIVES OF THE CONDITIONAL LOG-LIKELIHOOD FUNCTION OF THE NEGATIVE BINOMIAL MODEL

For the truncated negative binomial conditional log-likelihood function,

$$\begin{aligned} l_c(\boldsymbol{\theta}) = & \sum_{i=1}^{D} \log(\Gamma(\alpha + x_i)) - \log(x_i!) - \log(\Gamma(\alpha)) \\ & + \alpha \log(\tfrac{1}{1+\beta}) + x_i \log(\tfrac{\beta}{1+\beta}) - \log(1 - \left(\tfrac{1}{1+\beta}\right)^{\alpha}), \end{aligned}$$

the first and second order partial derivatives of the function with respect to $\alpha$ and $p = 1/(1+\beta)$ required to calculate the Fisher information are given below:

$$\frac{\partial l}{\partial p} = -\sum_{ix=1}^{D} \frac{x_i}{1-p} + \frac{\alpha}{p} + \frac{p^{\alpha}\alpha}{p(1-p^{\alpha})}$$

$$\frac{\partial^2 l}{\partial p^2} = -\sum_{i=1}^{D} \frac{x_i}{(1-p)^2} + \frac{\alpha(p^{\alpha} + p^{\alpha}\alpha - 1)}{p^2(p^{\alpha}-1)^2}$$

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{D} \Psi(\alpha + x_i) - \Psi(\alpha) + \log(p) + \frac{p^{\alpha}\log(p)}{1-p^{r}}$$

$$\frac{\partial^2 l}{\partial \alpha^2} = \sum_{i=1}^{D} \Psi'(\alpha + x_i) - \Psi'(\alpha) + \frac{\log(p)^2 p^{\alpha}}{(p^{\alpha}-1)^2}$$

$$\frac{\partial^2 l}{\partial \alpha \partial p} = \frac{1 - p^{\alpha} + p^{\alpha}\alpha\log(p)}{p(p^{\alpha}-1)^2}.$$

# B. MARKOV CHAIN DEFINITIONS

**Definition 1** *(Roberts, 1996, p. 46)*

*Let $X$ be a Markov chain such that*

$$P[X_t \in A | X_0, X_1, ..., X_{t-1}] = P[X_t \in A | X_{t-1}]$$

*for any set $A$, where $P[\cdot|\cdot]$ denotes a conditional probability. We consider transition probabilities of the form $P_{ij}(t) = P[X_t = j | X_0 = i]$. Let $\pi(\cdot)$ be a stationary distribution, such that if the initial value $X_0$ is sampled from $\pi(\cdot)$ then all subsequent iterates will also be distributed according to $\pi(\cdot)$. Let $\tau_{ii}$ be the time of the first return to state $i$,($\tau_{ii} = \min\{t > 0 : X_t = i | X_0 = i\}$).*

*(i) $X$ is called irreducible if for all $i, j$ there exists a $t > 0$ such that $P_{ij}(t) > 0$.*

*(ii) An irreducible chain $X$ is recurrent if $P[\tau_{ii} < \infty] = 1$ for some (and hence for all) $i$. Otherwise, $X$ is transient. Another equivalent condition for recurrence is*

$$\sum_t P_{ij}(t) = \infty$$

*for all $i, j$.*

*(iii) An irreducible recurrent chain $X$ is called positive recurrent if $E\ [\ \tau_{ii}\ ] < \infty$ for some (and hence for all) $i$. Otherwise, it is called null-recurrent. Another equivalent condition for positive recurrence is the existence of a stationary probability distribution for $X$, that is there exists $\pi(\cdot)$ such that*

$$\sum_i \pi(i) P_{ij}(t) = \pi(j)$$

*for all $j$ and $t \geq 0$.*

*(iv) An irreducible chain $X$ is called aperiodic if for some (and hence for all) $i$,*

$$\textit{greatest common divider } \{t > 0 : P_{ii}(t) > 0\} = 1.$$

# C. ERGODIC THEOREM

**Theorem 1** *If $X$ is positive recurrent and aperiodic then its stationary distribution $\pi(\cdot)$ is the unique probability distribution satisfying $\sum_t \pi(i)P_{ij}(t) = \pi(j)$ for all $j$ and $t > 0$. We then say that $X$ is ergodic and the following hold:*

*(i) $P_{ij}(t) \to \pi(j)$ as $t \to \infty$ for all $i, j$.*

*(ii) (Ergodic theorem) If $E_\pi[|f(X)|] < \infty$, then*

$$P[\bar{f}_N \to E_\pi[f(X)]] = 1,$$

*where $E_\pi[f(X)]] = \sum_i f(i)\pi(i)$, the expectation of $f(X)$ with respect to $\pi(\cdot)$.*

# D. WINBUGS CODE

## *D.1 Negative binomial data augmentation model*

```
model{
     for (i in 1:2){          #   These lines set
        U[i]~dnorm(0,1)       #   up the half cauchy
        V[i]~dnorm(0,1)       #   priors for the
    }                         #   gamma parameters
    a<-abs(U[1]/V[1])         #
    b<-abs(U[2]/V[2])         #

    psi~dunif(0,1)
    for (i in 1:M){           #   These steps are repeated over the
        p[i]~dgamma(a,b)      #   super-population size M
        z[i]~dbin(psi,1)      #   This is the indicator variable
        pi[i]<-p[i]*z[i]
        x[i]~dpois(pi[i])
    }
    N<-sum(z[])               #   N is the sum of the indicator variables
}
```

## D.2 Negative binomial RJMCMC model

```
model{
    for (i in 1:2){          #   These lines set
        U[i]~dnorm(0,1)      #   up the half cauchy
        V[i]~dnorm(0,1)      #   priors for the
    }                        #   gamma parameters
    a<-abs(U[1]/V[1])        #
    b<-abs(U[2]/V[2])        #

    for (i in 1:M){
        p[i]~dgamma(a,b)     #   Gamma prior for Poisson means
        pi[i]<-p[i]*w[i]
        w[i]<-step(N-i)      #   w[i] is set to zero if i > N
        x[i]~dpois(pi[i])
    }
    N~dcat(pee[1:M])         #   N is a discrete value with some
    n~dbin(0.00001,N)        #   probability distribution
}
```

## D.3 Negative binomial data augmentation model for multiple grabs

```
model{
    for (i in 1:2){          #   These lines set
        U[i]~dnorm(0,1)      #   up the half cauchy
        V[i]~dnorm(0,1)      #   priors for the
    }                        #   gamma parameters
    a<-abs(U[1]/V[1])        #
    b<-abs(U[2]/V[2])        #

    psi~dunif(0,1)
    for (i in 1:M){          #   These steps are repeated over the
        p[i]~dgamma(a,b)     #   super-population size M
        z[i]~dbin(psi,1)     #   This is the indicator variable
        pi[i]<-p[i]*z[i]

        for (j in 1:g){      #   This sets up the model for g grabs
           x[i,j]~dpois(pi[i])
        }
    }

    N<-sum(z[])              #   N is the sum of the indicator variables
}
```

## D.4 Neyman Type A-gamma data augmentation model

```
model{
    for (i in 1:4){          #   These lines set
        U[i]~dnorm(0,1)      #   up the half cauchy
        V[i]~dnorm(0,1)      #   priors for the
    }                        #   gamma parameters
    a<-abs(U[1]/V[1])        #
    b<-abs(U[2]/V[2])        #
    c<-abs(U[3]/V[3])        #
    d<-abs(U[4]/V[4])        #
    psi~dunif(0,1)
    ph~dgamma(c,d)           #   Prior on the clustering parameter

    for (i in 1:M){
        z[i]~dbin(psi,1)
        mu[i]~dgamma(a,b)
        li[i]<-(mu[i]/ph[i])*z[i]
        lam[i]<- li[i] - x[i]*log(ph[i]) + logfact(x[i]) - log(sums[i])
    }
                             #   The above lines specify the
                             #   prior on the mean of the Neyman Type A
                             #   and the negative log-likelihood

    for (i in 1:M) {
        for (j in 1:50){
            ss[i,j]<-  exp((j-1)*(log(li[i])-ph[i])+x[i]*log(j-1)
                       -logfact(j-1))
        }
        sums[i]<-sum(ss[i,]) #   This section calculates the summation
    }                        #   in the Neyman Type A pdf
```

```
    zero<-0                 #    These lines are the 'zeros' trick
    zero~dpois(lik)         #    which allows us to specify any
    lik<-sum(lam[])         #    log-likelihood in the model

    N<-sum(z[])             #    N is a derived by summing the indicator
}
```