A Latent Health Factor Model for Estimating Estuarine Ecosystem Health

by

Margaret A. C. Wu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Assessment of the "health" of an ecosystem is often of great interest to those interested in monitoring and conservation of ecosystems. Traditionally, scientists have quantified the health of an ecosystem using multimetric indices that are semi-qualitative. Recently, a statistical-based index called the Latent Health Factor Index (LHFI) was devised to address many inadequacies of the conventional indices. Relying on standard modelling procedures, unlike the conventional indices, accords the LHFI many advantages: the LHFI is less arbitrary, and it allows for straightforward model inference and for formal statistical prediction of health for a new site (using only supplementary environmental covariates). In contrast, with conventional indices, formal statistical prediction does not exist, meaning that proper estimation of health for a new site requires benthic data which are expensive and time-consuming to gather. As the LHFI modelling methodology is a relatively new concept, it has so far only been demonstrated (and validated) on freshwater ecosystems. The goal of this thesis is to apply the LHFI modelling methodology to estuarine ecosystems, particularly to the previously unassessed system in Richibucto, New Brunswick. Specifically, the aims of this thesis are threefold: firstly, to investigate whether the LHFI is even applicable to estuarine systems since estuarine and freshwater metrics, or indicators of health, are quite different; secondly, to determine the appropriate form that the LHFI model if the technique is applicable; and thirdly, to assess the health of the Richibucto system. Note that the second objective includes determining which covariates may have a significant impact on estuarine health. As scientists have previously used the AZTI Marine Biotic Index (AMBI) and the Infaunal Trophic Index (ITI) as measurements of estuarine ecosystem health, this thesis investigates LHFI models using metrics from these two indices simultaneously. Two sets of models were considered in a Bayesian framework and implemented using Markov chain Monte Carlo techniques, the first using only metrics from AMBI, and the second using metrics from both AMBI and ITI. Both sets of LHFI models were successful in that they were able to make

distinctions between health levels at different sites.

Acknowledgements

Firstly I would like to offer my sincere thanks to Dr. Grace S. Chiu, my thesis supervisor, for her guidance, enthusiasm and support. Many thanks also to Drs. Lin Lu and Jon Grant of the Department of Oceanography at Dalhousie University, particularly for Jon's NSERC grant that helped support this research. I am grateful to Professors David Matthews and Jeannette O'Hara-Hines for serving as readers, and to Ms. Mary Lou Dufton for addressing my numerous questions with patience.

I wish to thank Neelmoy C. Biswas, Bobby H. Katanchi and Carman H.K. Lee for their invaluable support. Lastly, and most importantly, I wish to express my gratitude towards my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Quantifying Marine and Estuarine Ecosystem Health

The assessment of benthic populations is an important part of many marine and estuarine environmental monitoring programmes. Generally, the goals of such programmes are to ensure that human health is not threatened, to ensure that unacceptable harm is not done to marine ecosystems or their resources, and to supply managers with information that allows them to make decisions on the continued use of such resources (Bilyard, 1987). Benthos, those organisms living on or in the sediment at the bottom of a body of water, are useful indicators of underlying health conditions as they are relatively sedentary (they cannot avoid deteriorating water/sediment quality conditions), have relatively long life spans (indicate and integrate water/sediment quality conditions), and consist of different species that exhibit different tolerances to stress (can be classified into functional groups) (Dauer, 1993; Bilyard, 1987). Thus, a study of the benthic population is often included in marine or estuarine monitoring programmes.

The overall condition, or "health", of an ecosystem is a complex concept, involving many diverse factors, and therefore can be difficult to evaluate in a quantitative

manner. Numerous methods exist for measuring estuarine health based mainly on benthic data. Many of these methods involve health indicators that are measures of species abundance, diversity, evenness, and many also examine specific groups of benthos, such as opportunistic species, species sensitive to stress, deep-dwelling species, etc.

1.2 Existing Methods

The conventional and most popular way of consolidating this assortment of data is via a multi-metric index, which produces a scalar number from a formula involving several of these indicator variables of health, called metrics. One such index is the AZTI Marine Biotic Index (AMBI), which was devised to establish the ecological quality of benthos within European estuarine and coastal environments. The calculation of AMBI is based upon the abundances of five species groups organized by their relative sensitivity to environmental stress (A. Borja and Erez, 2000). Another is the Infaunal Trophic Index (ITI), which aims to indicate changes in the amount of organic material present, and is based upon the abundances of four species groups organized by their methods of feeding and responses to sources of organic material (Word, 1978).

The remainder of this section is a summary of the main advantages and disadvantages of the broad group of conventional indices of the form described above, as brought up by Chiu et al. (2008).

The main advantage of these conventional indices is that their outputs are simple, and thus supposedly easily interpretable. This capability is particularly appealing to policy makers. They also contain a high amount of biological content from scientists being involved at all stages of index development.

On the other hand, conventional indices tend to be somewhat ad hoc since the process of building such indices is often highly arbitrary. For example, a formula is usually chosen for combining metrics to produce the index. It is hard to quantify

how much information is double-counted from metrics providing overlapping information, and thus the weighting scheme used in a typical formula can be viewed as somewhat arbitrarily chosen. Another disadvantage stems from the practice of comparing indices of test sites to so-called pristine sites as a control condition, where these control sites are assumed to be absolute and invariant. However, truly pristine sites virtually no longer exist, and the best available site is sometimes substituted, but these "best" sites themselves vary in quality. The conventional indices, however, do not take this variation into account. Thus, it is difficult to objectively compare index results across different locations and time periods. Results can be calibrated to account for the differences in local reference conditions with much effort, but there is no universally accepted calibration scheme. Therefore, calibration introduces additional subjectivity into the process. It is also often of interest to determine the relationships between health and environmental or impact-related covariates, such as water depth or urbanization, as these factors are known to affect the distribution of species and hence, health; however, conventional indices do not lend themselves easily to an analysis of these relationships except in a crude manner. Finally, properly assessing the variation in the indices as estimates of health is difficult, as is prediction of a certain site's health, under the conventional procedure. Typically, to carry out prediction of a new site's health, scientists use a regression of computed indices (i.e. health estimates) to environmental covariates after indices have already been computed. However, this method is problematic because computed indices are treated as observed data (Chiu et al., 2008). In fact, more statistically rigourous methods of prediction do not exist.

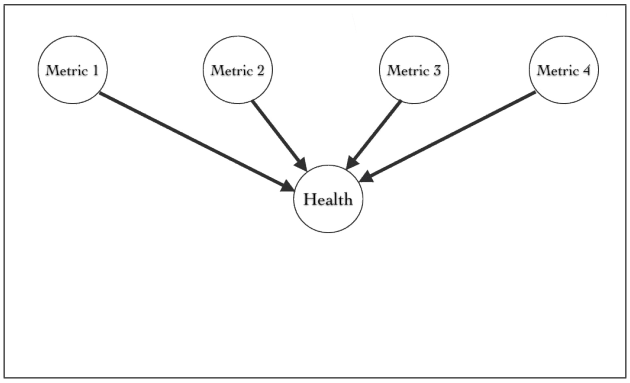
1.3 Latent Health Factor Index

Recently, a new statistical-model based index has been proposed by Chiu et al. (2008) called the Latent Health Factor Index (LHFI). This section describes the LHFI and provides a comparison to the conventional indices, as discussed originally in Chiu et al. (2008).

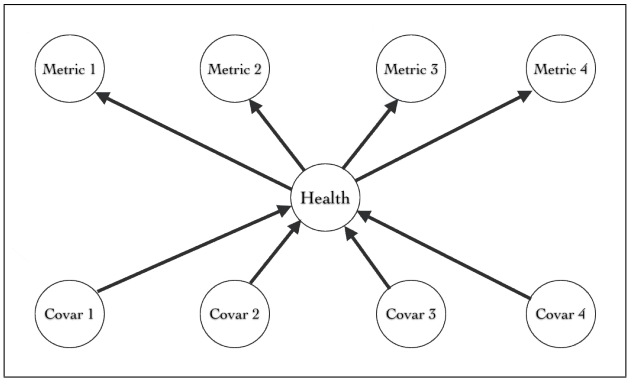
In the conventional indices, metrics are combined according to a formula to produce an estimate for health. This suggests that the metrics are the explanatory variables and health is a response variable. In reality, metrics provide indications of health and perhaps should be better considered as response variables. The LHFI adopts this reversal of roles: metrics are regressed upon health, and health can be further regressed upon other explanatory covariates, forming a multiple layered regression (see Fig. 1.1). Note that health is a latent variable since it is unobservable. With data on metrics and covariates, latent health and the effect of the explanatory covariates on health can be estimated. Thus, the LHFI is a multi-level, mixed-effects, Analysis of Covariance (ANCOVA), generalized linear regression model. Note that it is an ANCOVA since it can involve both categorical factors and continuous variables: the model involves factors, and in addition, explanatory covariates can be either continuous or factors, as we shall see in Section 3.1.

Chiu et al. (2008) designed the LHFI with the goals of retaining the advantages of the conventional indices while addressing their shortcomings. Specifically, the LHFI produces a single number, like the conventional indices, so it retains easy interpretability. Thus, sites within an ecosystem can still be ranked easily by their estimated health. The LHFI is less arbitrary than conventional indices since it is based upon standard modelling procedures. It is expected to be less sensitive to random variability since it uses field data directly instead of requiring any intermediate stages to adjust the data. As it involves not only metrics but also environmental and impact-related covariates, it allows for proper analysis of relationships between health and covariates, i.e. identifying those covariates that affect health and quantifying their impacts on health. This last capability could be particularly useful to policy makers. The LHFI also allows proper comparisons of different time periods and different locations through the addition of geographical and temporal blocking factors[1].

[1]Here we are interested in the difference between strata, or blocks, although blocks are not usually introduced for this purpose, but instead with the aim of controlling variability.



(a)



(b)

Figure 1.1: Relationships between metrics, health and covariates implied by (a) conventional indices, and (b) the latent health factor model

The LHFI appears to be more quantitatively sound than conventional indices, since it uses standard modelling procedures while conventional indices do not. Under the conventional scheme, model inference and prediction are quite difficult due to the scheme's ad hoc nature. Model inference and prediction can however be accomplished in an unambiguous manner with the LHFI model. Confidence intervals for estimates are straightforward to calculate, which means that the reliability of health estimates can be properly assessed. As well, once an LHFI model has been specified, prediction of a new site's health can be accomplished simply with information on covariates, thus bypassing the expensive benthic taxonomic laboratory procedures that are required to gather the metric data for conventional indices. This capability could prove an enormous asset to scientists, particularly if budgets or time are limited.

Thus, if the LHFI produces similar results about health compared to the conventional indices, the LHFI would generally be preferred. Note that the LHFI is potentially applicable to any context in which a latent "health" factor is desired to be estimated.

1.4 Thesis Objectives

In order to demonstrate the validity of the LHFI, Chiu et al. (2008) applied the LHFI modelling methodology to stream ecosystems, focussing on a 1997 data set for the Puget Sound Lowlands in Washington State.

This thesis aims to apply the LHFI modelling methodology to estuarine ecosystems. Since metrics, or indicators of health, are quite different for estuarine and freshwater ecosystems, two specific objectives of this thesis are to investigate if the LHFI is even applicable to estuarine systems, and if so, the form of the LHFI model. A third objective is to assess the health of the previously unassessed Richibucto, NB estuarine system using the LHFI. Note that the second objective includes determining which environmental covariates may have a significant impact on estuarine

health in order to facilitate the prediction of health in the absence of benthic taxonomic data. To address the objectives, this thesis considers building LHFI models using metrics from AMBI and ITI, since they have been previously used by scientists to measure estuarine ecosystem health.

In this chapter, we discussed existing methods for quantifying ecosystem health, the motivation behind the design of the latent health factor model as a new alternative method, and presented the objectives of this thesis. The second chapter provides a general latent health factor model for estuarine ecosystems and the accompanying methods for inference in a Bayesian framework. The third and fourth chapters discuss the results of fitting LHFI models for the Richibucto estuary using the AMBI data alone, and the AMBI and ITI data combined, respectively. The fifth and last chapter presents some general conclusions and considers future work.

Chapter 2

An LHFI Model for Estuarine Ecosystems

2.1 Building the Model

Recall Fig. 1.1(b) illustrating the relationships between metrics, health and covariates in the LHFI. We begin the model building process by developing mathematical equations to explain these relationships.

First consider the top regression layer between metrics and health. We introduce some notation: let Y_{ijk} denote the value of the kth replicate of the jth metric for the ith site, where there are n sites in total, J metrics measured at each site, and r_i replicates of each of the J metrics at each site. Let H_i represent the latent health of site i and let β_j represent the effect of metric j. As, among all the variables under consideration, Y_{ijk} is governed by the latent health at site i and the effect on health of metric j, Chiu et al. (2008) naturally consider the following generalized linear mixed model for the regression between metrics and health:

$$\nu_{ij} = H_i + \beta_j \tag{2.1}$$

where $\nu_{ij} = g(E(Y_{ijk}))$ and g is a link function chosen based on the form of the metrics Y_{ijk}. The metric effects β_j's are block effects and since their estimates are

not of primary interest, they are essentially error terms. Thus, we consider them to be random effects and we set their means to be zero, since there is no reason to introduce a mean parameter to error terms. As metrics are not necessarily independent, an appropriate covariance structure for the β_j's should be chosen.

Note that one could argue that β_j's should be fixed effects since they correspond to the same metrics for a specific type of ecosystem, and thus are not chosen randomly from a wider population. However, having β_j's as fixed effects would, firstly, not allow for covariance structures, which form an important part of the model, and, secondly, introduce more parameters to estimate. Hence, we consider β_j's as random effects instead.

As well, the H_i's are random effects, and considered to be independent of each other and of the β_j's, since the H_i's correspond to sites which are typically chosen at random.

Next, recall Fig. 1.1(b) and consider the lower, optional regression layer between H_i and any explanatory covariates that are of interest. Let $\boldsymbol{x}_i$ denote a vector of the values of covariates at site i that may influence health. Chiu et al. (2008) consider for this layer the following latent regression:

$$H_i = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}_i) + \varepsilon_i \tag{2.2}$$

where α_0 is the overall health of the region from which the sample sites were taken, $f()$ is an appropriately chosen regression function of the covariates, $\boldsymbol{\alpha}$ is the corresponding vector of regression coefficients, and the ε_i's are independent and identically distributed errors with mean zero. If there are no covariates, (2.2) reduces to

$$H_i = \alpha_0 + \varepsilon_i \tag{2.3}$$

Together (2.1) and (2.2) constitute the basic LHFI model, which is a hierarchical mixed-effects, ANCOVA, generalized linear regression model.

2.1.1 Extension to Multiple Time Periods or Locations

Next, consider an extension of the basic LHFI model to allow for proper comparisons of different time periods and locations. We describe here an extension to incorporate multiple time periods, but the principles are the same for an extension to multiple locations. Note that here we discuss the situation where each site is measured once and not all sites are measured at the same point in time, but the LHFI can also be adapted to situations where each site is measured at multiple time points (not discussed in this thesis).

The extended model does not directly address complex temporal correlation patterns, but as Chiu et al. (2008) point out, ecological data used in indices are typically too variable and too sparse for such complex models to be feasible. Their proposed model is a compromise between statistical complexity and practicality. Let Y_{ijkl} denote the value of the kth replicate of the jth metric for the ith site in the lth temporal domain. Let λ_l denote a blocking factor representing the temporal effect on health:

$$\nu_{i(l)\times j} = H_{i(l)} + \beta_j \tag{2.4}$$

$$H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}_{i(l)}) + \lambda_l + \varepsilon_{i(l)} \tag{2.5}$$

Note that () in subscripts represents nesting and $\times$ represents crossing. The temporal effect λ_l could be a fixed or random effect depending on the context. The placement of λ_l in the equation for H_i and not $\nu_{i(l)\times j}$ acknowledges that health directly depends on time.

Note in the above model that health is now nested in temporal domain but the metric effect is unchanged. Depending on the data structure, having both $H_{i(l)}$ and $\beta_{j\times l}$ in the model could lead to inseparability issues, and we do not include both to safeguard against any weakness in identifiability in practice. Specifically we can employ the following reasoning.

Implicit in the term $H_{i(l)}$ is a breakdown into several factors, i.e.

$$H_{i(l)} = \text{intercept} + \text{effect due to } i \text{ nested in } l$$

Similarly, $\beta_{j\times l}$ can be regarded as

$$\beta_{j\times l} = \text{effect due to } l + \text{effect due to } j + \text{interaction due to } j \times l$$

Since the last term of each of the two preceding equations contains an l component, the LHFI approach could run into problems of not being able to estimate these two components well enough, especially since the type of ecological data that will be used with the LHFI is typically sparse, and there are many parameters to estimate. Thus, a choice had to be made between a model with $H_{i(l)}$ and β_j, and a model with H_i and $\beta_{j\times l}$. Since it is necessary for site to be nested within temporal domain, and additionally since it is unlikely that the effect of metrics on health (i.e. the concept of abundance and diversity being indicators of health in this context) would change over time, the first option was chosen.

2.2 Model Inference in a Bayesian Framework

Hierarchical models with latent factors are often best handled with a Bayesian framework. Bayesian inference is especially appropriate for dealing with ecological data since it does not rely on asymptotics which are often inappropriate for the small sample sizes and unbalanced designs that occur frequently with ecological studies (Chiu et al., 2008). This section provides background information on the Bayesian school of thought and discusses methods of Bayesian inference for the latent health factor model.

2.2.1 Bayesian Standpoint

In statistical inference, there are two main philosophies: the frequentist, and the Bayesian. The methods used by frequentists and Bayesians are not very different from each other; however, it is in the interpretation of the results where the two philosophies disagree. Frequentists consider parameters as deterministic, whose

true values one attempts to estimate; and data are realizations of random experiments and are thus variable. On the other hand, Bayesians consider parameters as random variables, in addition to the data. Any existing knowledge about a parameter θ can be expressed through its prior distribution $\pi(\theta)$. Data can be used to update the knowledge on θ, and the combined knowledge from the prior and the data can be expressed through the posterior distribution $\pi(\theta|\mathbf{x})$ which is conditional on the data $\mathbf{x}$. This fundamental difference in philosophy leads, for example, to very different interpretations of confidence intervals. To a frequentist, a confidence interval resulting from a specific set of data is not random nor is the parameter around which it is focussed. Thus, a specific confidence interval either contains or does not contain the parameter, and statements such as, "this confidence interval will contain the parameter 95% of the time" do not make sense. To a Bayesian, however, this statement makes perfect sense as the parameter itself is random. Please refer to Hogg et al. (2005), Chapter 11 for more details on Bayesian methods.

2.2.2 Inference

We return to the basic LHFI model, that is (2.1) and (2.2), in discussing methods for inference which were outlined originally in Chiu et al. (2008).

Let $\boldsymbol{H} = (H_1, \ldots, H_n)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$. Additionally let $\boldsymbol{\nu}$ denote the vector of ν_{ij}'s; $\boldsymbol{Y}$ denote the vector of Y_{ijk}'s which has length $J \times \sum_{i=1}^{n} r_i$; $\mathbb{X}$ denote the matrix whose ith rows is $\boldsymbol{x_i}$ and which has dimension $n \times r$ where r is the number of covariates; and $\boldsymbol{\Omega}$ denote the vector of the remaining parameters, i.e. α_0, $\boldsymbol{\alpha}$ and any parameters from the distributions of ε_i and $\boldsymbol{\beta}$.

Let $P()$ denote a distribution, so that $P(\boldsymbol{\Omega})$ is the prior distribution for $\boldsymbol{\Omega}$, $P(\boldsymbol{Y}|\boldsymbol{\nu}) = P(\boldsymbol{Y}|\boldsymbol{H}, \boldsymbol{\beta})$ is the likelihood function of the data, $P(\boldsymbol{H}|\boldsymbol{\Omega}, \mathbb{X})$ is the distribution of $\boldsymbol{H}$, and $P(\boldsymbol{\beta}|\boldsymbol{\Omega})$ is the distribution of $\boldsymbol{\beta}$. The forms of these distributions depend on the knowledge of the relationships between the data and the parameters.

We are mainly interested in estimating H_i and the regression coefficients $\boldsymbol{\alpha}$ which provide information on which covariates are useful in determining health. Bayesian inference on H_i and $\boldsymbol{\alpha}$ proceeds from the joint posterior distribution of $\boldsymbol{H}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$, i.e. $P(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega} | \boldsymbol{Y}, \mathbb{X})$. By rules of conditional probability,

$$\begin{aligned} P(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega} | \boldsymbol{Y}, \mathbb{X}) &\propto P(\boldsymbol{Y}, \mathbb{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega}) \\ &= P(\boldsymbol{Y} | \mathbb{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega}) P(\mathbb{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega}) \\ &= P(\boldsymbol{Y} | \mathbb{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega}) P(\boldsymbol{H}, \boldsymbol{\beta} | \boldsymbol{\Omega}, \mathbb{X}) P(\boldsymbol{\Omega}, \mathbb{X}) \end{aligned} \tag{2.6}$$

But $P(\boldsymbol{Y} | \mathbb{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega}) = P(\boldsymbol{Y} | \boldsymbol{H}, \boldsymbol{\beta})$ since $\boldsymbol{Y}$ depends on $\mathbb{X}$ and $\boldsymbol{\Omega}$ only through $\boldsymbol{H}$ and $\boldsymbol{\beta}$; $P(\boldsymbol{\Omega}, \mathbb{X}) = P(\boldsymbol{\Omega}) P(\mathbb{X})$ since $\boldsymbol{\Omega}$ and $\mathbb{X}$ are independent; and $P(\boldsymbol{H}, \boldsymbol{\beta} | \boldsymbol{\Omega}, \mathbb{X}) = P(\boldsymbol{H} | \boldsymbol{\Omega}, \mathbb{X}) P(\boldsymbol{\beta} | \boldsymbol{\Omega})$ since $\boldsymbol{H}$ and $\boldsymbol{\beta}$ are independent. As well, $P(\mathbb{X})$ is constant since $\mathbb{X}$ is considered fixed in a regression. Thus,

$$P(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega} | \boldsymbol{Y}, \mathbb{X}) \propto P(\boldsymbol{Y} | \boldsymbol{H}, \boldsymbol{\beta}) P(\boldsymbol{H} | \boldsymbol{\Omega}, \mathbb{X}) P(\boldsymbol{\beta} | \boldsymbol{\Omega}) P(\boldsymbol{\Omega}) \tag{2.7}$$

We take the posterior mean of H_i to be the latent health factor index:

$$\widehat{H}_i = E(H_i | \boldsymbol{Y}, \mathbb{X}) \tag{2.8}$$

$$= \int H_i \; P(H_i | \boldsymbol{Y}, \mathbb{X}) \; dH_i \tag{2.9}$$

$$= \int H_i \int \int \int P(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega} | \boldsymbol{Y}, \mathbb{X}) \; d\boldsymbol{\beta} \; d\boldsymbol{\Omega} \; d\boldsymbol{H}_{-i} \; dH_i \tag{2.10}$$

where $\boldsymbol{H}_{-i}$ is $\boldsymbol{H}$ without H_i and $P(H_i | \boldsymbol{Y}, \mathbb{X})$ is the marginal posterior distribution of H_i.

Measures of uncertainty for H_i are provided by posterior credible intervals, or more simply, credible intervals, which are Bayesian analogies of frequentist confidence intervals and are derived from the posterior distribution. Credible intervals serve for more informal tests instead of formal Bayesian hypothesis testing: they can be used to compare health at different sites; for example, one can determine if two sites have the same level of health by comparing their credible intervals and

examining the amount of overlap between them. Point and interval estimates of the elements of $\boldsymbol{\alpha}$ can be obtained similarly.

The posterior distribution also provides a simple method of predicting a new site's health once a model has been specified and given that measurements of covariates are available. We postpone the details for this until Chapter 5.

Chapter 3

Fitting an LHFI Model to Richibucto AMBI

Benthic data for the Richibucto estuary had already been gathered and the AMBI and ITI indices had accordingly been calculated and analyzed using this data before the start of this thesis project. Dr. Jonathan Grant and Dr. Lin Lu, Department of Oceanography at Dalhousie University, who had collected the benthic data for the indices, both believe that AMBI is a more appropriate measure than ITI for Richibucto since it is an estuary in which benthic fauna are not only affected by organic enrichment but also by freshwater input (salinity gradient), variability of sediment particle size, and topography (channel, water depth). Thus, we begin the modelling process by fitting a latent health factor model to the AMBI data alone and call these models LHFI-AMBI.

3.1 The AMBI data

The Richibucto benthic data were collected from 18 sites ranging through the entire estuarine system (see Fig. 3.1). Due to time constraints, the 18 sites could not all be sampled in one day, and thus were sampled during two days separated by approximately a month. Sites 1-3 and 9-18 were sampled during a single day in

September 2008, and sites 4-8 during a single day in October 2008. At each site, grab samples[1] of the benthic organisms were taken underwater and the organisms in each sample were identified in a laboratory.

Two or three replicates of benthic data were taken at each site; the aim was to take three samples at each site, but at some sites only two were obtained due to physical difficulty. As well, 9 supplementary covariates were measured once at each site, i.e not replicated: *water depth* (m), which is the distance from water surface to estuary bed at the location of the station; *water temperature* (°C); *salinity* (parts per thousand or ppt); *silt-clay* (%), which is the fraction of sediment grains of size $< 63\mu m$; $md\phi$, which is the median grain size of the sediment; *sorting*, which is a measure of the variability of sediment grain size [2]; *organic content* (%) of the sediment; and *distance downstream* (km), which is the perpendicular distance from Site 1 along a straight line between Sites 1 and 18 as illustrated in Fig. 3.1. The data are included in Appendix A.

Each organism observed at Richibucto belonged to one of 88 species, and the number of animals observed per species was quite variable, ranging from 0 to several hundred. Grant and Lu believe, based on their ecological expertise, that they have observed all the main species and most of the species in Richibucto, although physically proving this would be difficult.

To calculate AMBI, the benthos were sorted into five disjoint groups organized by their relative sensitivity to environmental stress. A. Borja and Erez (2000), in their paper proposing AMBI, describe the five taxonomic groups as follows:

> "*Group I*: Species very sensitive to organic enrichment and present under unpolluted conditions (initial state). They include the specialist carnivores and some deposit-feeding tubicolous polychaetes.

[1]For more details on grab samples, please refer to the National Oceanic and Atmospheric Administration Coastal Services Center, `http://www.csc.noaa.gov/benthic/mapping/techniques/sensors/grab.htm`.

[2]The formula for sorting is: sorting $= \frac{\phi_{84}-\phi_{16}}{4} + \frac{\phi_{95}-\phi_5}{6.6}$ where ϕ_α denotes the αth percentile of the grain size distribution. Sorting is unit-less.

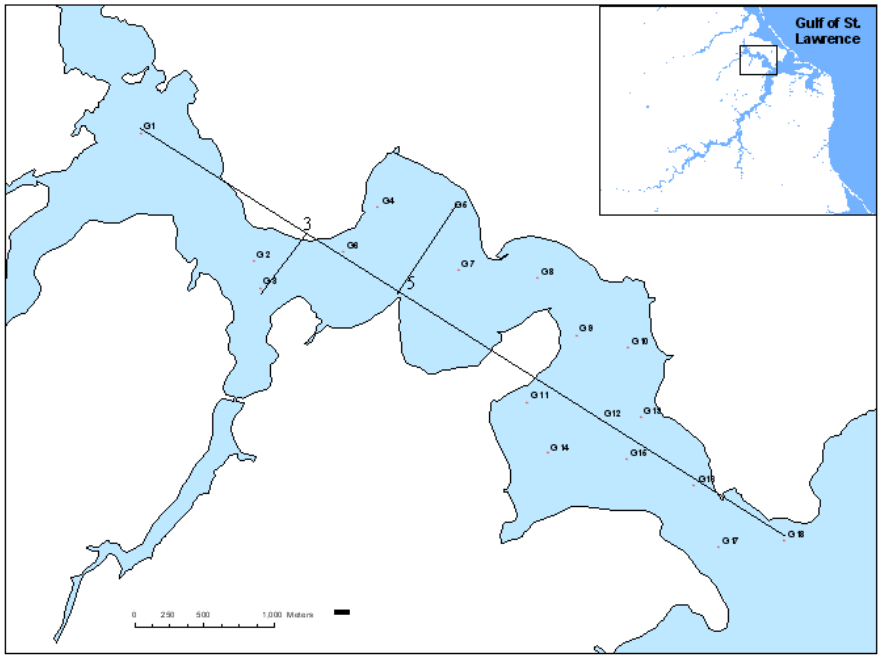


Figure 3.1: Map of Richibucto estuarine system with the 18 sample sites labelled; the straight black lines illustrate the calculation of *distance downstream* for sites 3 and 5; this figure was provided by Lin Lu

Group II: Species indifferent to enrichment, always present in low densities with non-significant variations with time (from initial state to slight unbalance). These include suspension feeders, less selective carnivores and scavengers.

Group III: Species tolerant to excess organic matter enrichment. These species may occur under normal conditions, but their populations are stimulated by organic enrichment (slight unbalance situations). They are surface deposit-feeding species, as tubicolous spionids.

Group IV: Second-order opportunistic species (slight to pronounced unbalanced situations). Mainly small sized polychaetes: subsurface deposit-feeders, such as cirratulids.

Group V: First-order opportunistic species (pronounced unbalanced situations). These are deposit- feeders, which proliferate in reduced sediments."

The abundances of the five taxonomic groups relative to the entire sample are combined via the following formula to produce a continuous scalar coefficient of health, which can also be mapped to a discrete index. AMBI is negatively correlated with health:

$$\text{AMBI} = [(0 \times \text{GI}) + (1.5 \times \text{GII}) + (3 \times \text{GIII}) + (4.5 \times \text{GIV}) + (6 \times \text{GV})] \quad (3.1)$$

where GI, GII, GIII, GIV and GV are the proportions of abundance of taxonomic Groups I-V relative to the entire sample.

A high abundance in taxonomic Group I is considered an indicator of "good health" while a high abundance in taxonomic Group V is considered an indicator of "poor health". Taxonomic Groups II-IV are on a gradually changing scale between the two extremes; high abundance in taxonomic Group IV is also an indicator of "poor health" but to a lesser extent than taxonomic Group V; and the same is true with taxonomic Group III to an even lesser extent; on the other hand, abundance in taxonomic Group II is neither an indicator of "good health" nor of "poor health". Thus, taxonomic Group I is positively related to health, taxonomic Groups III, IV and V are negatively related to health, and taxonomic Group II is indifferent to

health.

3.2 An Initial Model

The abundances of the five disjoint taxonomic groups in the calculation of AMBI are the metrics modelled in the LHFI framework. It was possible to exclude the second taxonomic group as it is not a strong indicator of health. However, it was included, as it may be statistically significant even though it is biologically insignificant: since the indifferent taxonomic group is of the same type of biological data in the same context referring to biodiversity and abundance, even if it is indifferent to health, it is not necessarily irrelevant to modelling of abundances and diversity.

3.2.1 Metric Groups

The metrics are considered in two groups depending on whether they are positively or negatively related to health. Thus, we denote metric j as nested in metric group s, where $s = 1$ for the metric group negatively related to health, and $s = 2$ for the metric group positively related to health. We place the indifferent second metric in metric group $s = 2$ as this group contains only one other metric, taxonomic Group I; providing more data may improve parameter estimation associated with this metric group.

As the five metrics are disjoint and exhaustive, an appropriate distribution for these data would be multinomial, specifically a quint-nomial, and the corresponding link function would be a generalized logit. However, since we want to consider the metrics in two groups, $s = 1$ and $s = 2$, we split them into a quadrinomial (III, IV, V, non-III-IV-V) and trinomial (I, II, non-I-II), respectively. Note that there are overlap and dependency between the two resulting multinomials, and we address these issues later in the section.

For metric group $s = 2$, the trinomial, the distribution and logit are

$$\left[Y_{i\times 1(2)\times k}, Y_{i\times 2(2)\times k}, N_{i\times k} - \sum_{j=1}^{2} Y_{i\times j(2)\times k} | N_{i\times k}, p_{i\times 1(2)}, p_{i\times 2(2)} \right]$$
$$\sim \text{multinomial}\left(N_{i\times k}, p_{i\times 1(2)}, p_{i\times 2(2)}, 1 - \sum_{j=1}^{2} p_{i\times j(2)} \right) \tag{3.2}$$

$$\nu_{i\times j(2)} = \ln \frac{p_{i\times j(2)}}{1 - \sum_{j=1}^{2} p_{i\times j(2)}}, \quad j = 1, 2 \tag{3.3}$$

where $Y_{i\times j(s)\times k}$ denotes the value of the kth replicate of the jth metric (nested within the sth metric group) for the ith site, $N_{i\times k}$ denotes the sample *cardinality* (i.e. total # organisms) for the kth replicate of the ith site, and $p_{j(s)}$ denotes the probability of an organism being in the $j(s)$th taxonomic group or metric.

Taken together with (2.1), link function (3.3) implies that each metric j is non-negatively related to health; so while (3.3) is appropriate for the positively related metric group $s = 2$, it is not so for the negatively related metric group $s = 1$. Thus, the link for the negatively related metric group is inverted, and the corresponding quadrinomial and logit for this metric group are

$$\left[Y_{i\times 3(1)\times k}, Y_{i\times 4(1)\times k}, Y_{i\times 5(1)\times k}, N_{i\times k} - \sum_{j=3}^{5} Y_{i\times j(1)\times k} | N_{i\times k}, p_{i\times 3(1)}, p_{i\times 4(1)}, p_{i\times 5(1)} \right]$$
$$\sim \text{multinomial}\left(N_{i\times k}, p_{i\times 3(1)}, p_{i\times 4(1)}, p_{i\times 5(1)}, 1 - \sum_{j=3}^{5} p_{i\times j(1)} \right) \tag{3.4}$$

$$\nu_{i\times j(1)} = \ln \frac{1 - \sum_{j=3}^{5} p_{i\times j(1)}}{p_{i\times j(1)}}, \quad j = 3, 4, 5 \tag{3.5}$$

Notice here that link functions (3.3) and (3.5) are designed such that higher values of LHFI, i.e. estimates of H_i, indicate better ecosystem health, as we now explain. Link function (3.3) is such that higher abundances of the positively related metrics ($j = 1$ and 2) imply higher values of corresponding $p_{i\times j(2)}$'s and $\nu_{i\times j(2)}$'s; and these in turn imply higher values of H_i due to the positive relationship between ν and H as shown in (2.1). Similarly, link function (3.5) is such that higher abundances of the negatively related metrics ($j = 3$ and 4) imply higher values of corresponding $p_{i\times j(1)}$'s, thus, lower values of $\nu_{i\times j(1)}$'s, and thus, lower values of

H_i. Recalling that a high abundance in metric 1 indicates "good health", high abundances in metrics 3-4 indicate "poor health", and a high abundance in metric 2 indicates neither, these two link functions therefore are such that high values of estimated H_i indicate better health.

To address overlap between the metric groups, we add a parameter, θ_s, to (2.1), to account for the effect on ν of different metric groups:

$$\nu_{i\times j(s)} = H_i + \beta_{j(s)} + \theta_s \tag{3.6}$$

Note that we choose not to model θ_s to explain H_i in (2.2) because it is a nuisance parameter. However, it would be acceptable to model θ_s as a component of $\beta_{j(s)}$, giving $\beta_{j(s)}$ a non-zero mean. Such a change would affect the interpretation of θ_s but not the estimation of H_i.

Moreover, we consider θ_s a fixed effect, as there is no reason to or gain from considering it as a random effect, it having only two levels. A constraint must accordingly be placed upon θ_s; we take $\theta_2 = 0$ since $s = 2$ contains the indifferent group and thus may be regarded as the baseline.

We did not explicitly address dependency between the metric groups; instead, metric effects $\beta_{j(s)}$ were considered independent for simplicity and as a starting point for our initial models. The multinomial distribution already accounts for some dependency in the metrics. As well, we assume that each metric has a different variance as it is possible that the distributions of metrics are quite different. Thus, we have

$$\left[\beta_{j(s)}|\sigma_{j(s)}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{j(s)}) \tag{3.7}$$

It may appear that (3.2)-(3.7) is unidentifiable in $\sigma_{j(s)}$ since for each $\sigma_{j(s)}$ we have a single $\beta_{j(s)}$. This situation could likely lead to trouble under the frequentist framework, but it is not an issue in the Bayesian framework as shown by Chiu (2008).

3.2.2 Temporal Blocks

Since the sites were measured in two days separated by a month, the model should incorporate blocking by time as well. As stated previously, sites 1-3 and 9-18 were sampled in September 2008, and sites 4-8 were sampled in October 2008. Thus, let blocks $l = 1$ and $l = 2$ correspond to September and October sites, respectively.

As discussed in Section 2.1, health is nested in block, while metric effects remain as they would have been without blocking. We consider $\varepsilon_{i(l)} \sim N(0, \sigma^2_{H(l)})$ which signifies that the variation in health may differ over time. A blocking factor λ_l is introduced into the model to account for an effect on health of the different time periods, which was also discussed in Section 2.1. We consider λ_l as a fixed effect for the same reasons as with θ_s and adopt the constraint $\lambda_1 = 0$.

3.2.3 The Model

An initial LHFI model for the Richibucto AMBI data incorporating metric group divisions and blocking is as follows.

$$
\begin{gathered}
[Y_{i3kl1}, Y_{i4kl1}, Y_{i5kl1}, N_{ikl} - Y_{i3kl1} - Y_{i4kl1} - Y_{i5kl1} | N_{ikl}, p_{i3l1}, p_{i4l1}, p_{i5l1}] \\
\sim \text{multinomial}(N_{ikl}, p_{i3l1}, p_{i4l1}, p_{i5l1}, 1 - p_{i3l1} - p_{i4l1} - p_{i5l1}) \\
\nu_{ijl1} = \ln \frac{1 - p_{i3ls} - p_{i4ls} - p_{i5ls}}{p_{ijls}}, \quad j = 3, 4, 5 \\
[Y_{i1kl2}, Y_{i2kl2}, N_{ikl} - Y_{i1kl2} - Y_{i2kl2} | N_{ikl}, p_{i1l2}, p_{i2l2}] \\
\sim \text{multinomial}(N_{ikl}, p_{i1l2}, p_{i2l2}, 1 - p_{i1l2} - p_{i2l2}) \\
\nu_{ijl2} = \ln \frac{p_{ijl2}}{1 - p_{i1l2} - p_{i2l2}}, \quad j = 1, 2 \\
\nu_{ijls} = H_{i(l)} + \beta_{j(s)} + \theta_s, \qquad H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}_{i(l)}) + \varepsilon_{i(l)} + \lambda_l \\
\left[\beta_{j(s)} | \sigma_{j(s)}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{j(s)}), \qquad \left[\varepsilon_{i(l)} | \sigma_{H(l)}\right] \overset{\text{iid}}{\sim} N(0, \sigma^2_{H(l)}) \\
\theta_{s=2} = 0, \qquad \lambda_{l=1} = 0
\end{gathered}
\tag{3.8}
$$

where

$$\text{temporal block } l = \begin{cases} 1 & \text{if data collected in September} \\ 2 & \text{if data collected in October} \end{cases}$$

$$\text{site } i = \begin{cases} 1, \ldots, 3, 9, \ldots, 18 & \text{if } l = 1 \\ 4, \ldots, 8 & \text{if } l = 2 \end{cases}$$

$$\text{metric group } s = \begin{cases} 1 & \text{negatively related to health} \\ 2 & \text{positively related to health} \end{cases}$$

$$\text{metric } j = \begin{cases} 3, 4, 5 & \text{in metric group } s = 1 \\ 1, 2 & \text{in metric group } s = 2 \end{cases}$$

$$\text{replicate } k = 1, \ldots, r_i \quad \text{where } r_i \text{ is the number of replicates at site } i$$

The notation has become rather complicated: the value of the kth replicate of the jth metric (nested within the sth metric group) for the ith site (nested within the lth temporal block) is now denoted by $Y_{i(l)\times j(s)\times k}$, and the sample *cardinality* for the kth replicate of the ith site (nested within l) is now denoted by $N_{i(l)\times k}$. These, however, have been written above as Y_{ijks} and N_{ik} to reduce clutter. Other parameters have been similarly simplified: $p_{i\times j(s)}$ as p_{ijs}, and $v_{i\times j(s)}$ as v_{ijls}.

We adopt diffuse priors for the elements of $\boldsymbol{\Omega} = (\alpha_0, \boldsymbol{\alpha}, \lambda_2, \theta_1, \sigma_{H(1)}, \sigma_{H(2)}, \sigma_{1(2)}, \sigma_{2(2)}, \sigma_{3(1)}, \sigma_{4(1)}, \sigma_{5(1)})^T$ as we do not have any specific prior information about these. We apply normal and inverse-gamma distributions as priors, which are commonly used for unbounded parameters and parameters left-bounded at 0, respectively, due to their conjugate properties under some conditions (but not necessarily here):

$$\begin{aligned} \alpha_0, \lambda_2, \theta_1, \text{elements of } \boldsymbol{\alpha} &\overset{\text{iid}}{\sim} N(0, 100) \\ \sigma_{H(l)}, \sigma_{j(s)} &\overset{\text{iid}}{\sim} \text{Inv-Gamma}(1, 1) \end{aligned} \tag{3.9}$$

It may appear to be a concern that we are attempting to fit a highly complex model with many parameters to only 18 sites. However, recall that we have a somewhat large set of benthic data including replicates: 5 metrics $\times \sum_{i=1}^{18} r_i =$ 245 benthic data-points in total. Even still, this could have been an issue in the

frequentist framework since there are not many replicates. However, under the Bayesian framework, this issue is somewhat less acute since it is countered partially by introducing priors to parameters.

3.2.4 A Note on Replicates

In this and several previous subsections, we have discussed in detail the various components of the initial LHFI model for Richibucto. The remainder of this chapter is mainly devoted to describing methods and results of implementing this model (and variations of it) using the Richibucto metrics. However, before continuing to implementation, we revisit the topic of replicates per site, an important side-note to the model. Recall that there are either 2 or 3 replicates at each site. Here, we view the number of replicates measured per site as fixed, although alternatively one could consider each site as having three replicates with some randomly missing. One could make an argument for both viewpoints. However, adopting the alternative view would require imputation of the missing values; this would increase model complexity and likely be computationally intensive, which is not desirable as we do not wish to begin our modelling exercise with a very complex model. As well, imputation could increase the variance of the estimates. As it was not crucial to view these as missing data in the context of an unbalanced design, we chose not to adopt this alternative view.

3.2.5 Markov Chain Monte Carlo Methods

The LHFI models in this thesis were all implemented using Markov chain Monte Carlo methods. The goal of Markov chain Monte Carlo methods is to generate a random sample via a Markov chain from a distribution that approximates the target distribution, which in this case is the posterior distribution of $(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega})^T$. An often large number of steps must be generated before a stationary distribution is reached; these initial steps before convergence are known as the "burn-in" and

must be discarded from the analysis.

One of the major difficulties with Markov chain Monte Carlo methods is determining if and when a Markov chain has reached convergence. This is a current topic of research and many "convergence diagnostic" methods have already been developed. It is important to note that these methods cannot determine if and when convergence has been reached; they can only ascertain when convergence has *not* been reached. Please refer to Gentle (2002) for more details on Markov chain Monte Carlo methods.

3.2.6 Markov Chain Monte Carlo Software

The latent health factor models of this thesis were run using OpenBUGS, an open-source computer software which specializes in the Bayesian analysis of statistical models using Markov chain Monte Carlo methods. OpenBUGS is available online at `http://mathstat.helsinki.fi/openbugs/`. Methods of checking convergence used for the running of the LHFI models in this thesis, many of which are built into OpenBUGS since they are popular tools, are described in Appendix B.

3.2.7 Implementing a Baseline Model

We began the model-fitting process by implementing a baseline model (i.e. without covariates $\boldsymbol{x}$), and refining that model by removing superfluous parameters. Once a final baseline model was determined, covariates were incorporated and the form of $f()$ was determined e.g. linear, quadratic etc. To improve *mixing*, which roughly refers to behaviour of the Markov chains with respect to the number of iterations required to reach the target distribution, the model specified by equations (3.8)-(3.9), denoted Model 3.8, was partially *hierarchically centred*, as were all subsequent models (details in Appendix C). For this and all subsequent models, two chains from the posterior distribution of $(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega})^T$ were generated from different starting values: one chain's starting values were numbers randomly generated from normal

distributions centred at 0 with variances around 20 to 30, and the other chain's were 0 for means and 1 for variances. Estimates of $(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega})^T$ were obtained from these two chains.

Health estimates $\widehat{H}_i$, denoted LHFI(3.8), along with their 95% centred posterior credible intervals (i.e. constructed using the 2.5% and 97.% percentiles)[3] are plotted in Fig. 3.2 in black. We also are interested in the precision of health estimates. Thus estimates and 95% credible intervals for $\sigma_{H(l)}$ are plotted in Fig. 3.3 along with similar estimates for all later AMBI models for easy comparison. In this figure, estimates for Model 3.8 are the furthest left. As well, estimates and 95% credible intervals for $\sigma_{j(s)}$ are plotted in Fig. 3.4. In addition, summary statistics for $\boldsymbol{\Omega}$ are in Table 3.1[4]. In Table 3.1, the column "MC error" represents the Monte Carlo standard error of the mean, i.e. an estimate of $s/N^{1/2}$; more details on the MC error are included in Appendix B, third bullet point. Note here that some values of MC error in Table 3.1 are listed as 0.00; this does not indicate that the MC error is actually 0, but is due to it being rounded to two decimal points.

It is clear from the plot of the 95% credible intervals for health (Fig. 3.2) that the LHFI has somewhat succeeded in its purpose of distinguishing the health levels of different sites. Some of the estimates are distinct from each other: although many of the 95% intervals overlap a great deal, the LHFI makes reasonable quantitative distinction between several sites (e.g. sites 1, 2, 6, 7).

The 95% credible interval for λ_2 contains 0 (Table 3.1), which suggests that the sampling time period does not significantly affect average health; in fact, the confidence level of λ_2 was 50-60% i.e. the widest centred credible level that does not

[3]Note that all of the posterior credible intervals mentioned in this thesis are centred intervals. Alternatively, we could have used highest posterior density intervals, but opted not to for convenience. A highest posterior density interval at a specified confidence level is the interval with the smallest width of all possible intervals at the same confidence level (for more details please see Hogg et al. (2005)).

[4]Focus should be on medians instead of means for σ's since their distributions are left bounded, somewhat asymmetrical and therefore produce some extreme values.

contain 0 has a confidence level between 50 and 60%. However, posterior credible intervals for $\sigma_{H(1)}$ and $\sigma_{H(2)}$ are distinctly different although they do overlap a large amount (Fig. 3.3)[5]. These two together signify that temporal blocking is important although λ_l is a superfluous parameter and can be removed. As well, the autocorrelation for Monte Carlo draws of λ_2 was low but long-living, which could be an artifact of λ_l being superfluous.

As well, there is significant overlap in the credible intervals for $\sigma_{j(s)}$'s within metric group, which could mean that distinct variances are not required for each metric, but perhaps only for each metric group (Fig. 3.4). Thus, the model could be reduced by using $\sigma_{\beta(s)}$ in place of $\sigma_{j(s)}$.

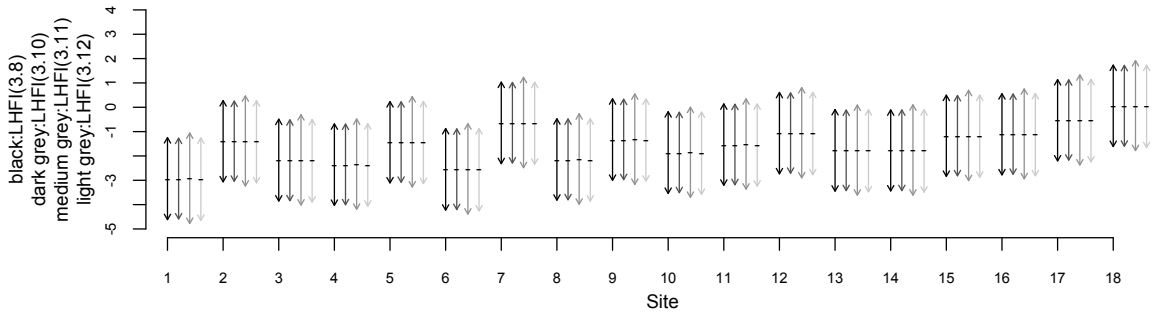


Figure 3.2: Estimates and 95% credible intervals for LHFI-AMBI baseline models; '-' denotes $\widehat{H}_i$

In addition to analyzing point and interval estimates of parameters as described above, we also examined the Deviance Information Criterion (DIC) for this model (and for all subsequent models). The DIC is a useful tool in model comparison and selection and we use it for this purpose: it is a measure of model fit penalized for the complexity of the model, and is similar to the Akaike Information Criterion

[5]It is quite possible that the variance in $\sigma_{H(l)}$ (hence, the length of the corresponding credible interval) in the second temporal block is much higher than in the first block simply because the second block contains only 5 sites whereas the first block contains 13 sites. However, we cannot confidently assume that this is the case, and so we do not attempt to collapse $\sigma_{H(1)}$ and $\sigma_{H(2)}$ into a single term.

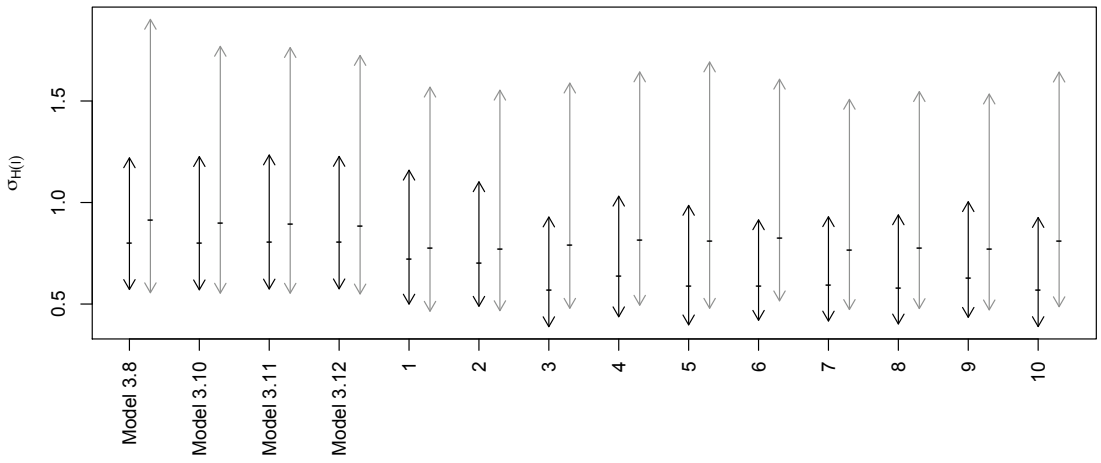


Figure 3.3: Estimates and 95% credible intervals for $\sigma_{H(1)}$ in black and $\sigma_{H(2)}$ in grey for all LHFI-AMBI models considered in this thesis; numbers 1 to 10 denote the 10 models run with covariates in Table 3.3 in that order; '-' denotes posterior median

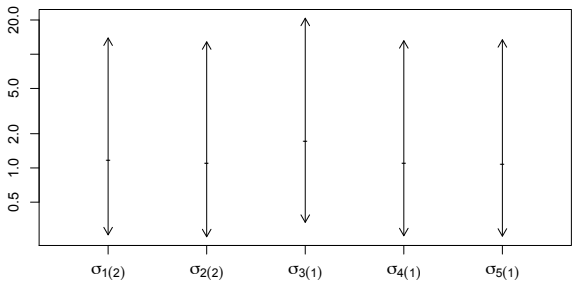


Figure 3.4: Model 3.8 estimates and 95% credible intervals for $\sigma_{j(s)}$ on a log-scale; '-' denotes posterior median; note that this and all following plots of credible intervals for metric effect standard deviations, which are represented by $\sigma_{j(s)}$'s here, use a log scale on the y-axis since the intervals are highly skewed. This, however, was not required for plots of $\sigma_{H(l)}$ which were less skewed.

Table 3.1: Summary statistics of posterior draws for LHFI-AMBI baseline models

		Mean	Median	2.5th %-ile	97.5th %-ile	MC Error	# Draws
Model 3.8	α_0	-1.46	-1.47	-3.14	0.29	0.01	40000
	λ_2	-0.40	-0.40	-1.43	0.64	0.01	
	θ_1	2.00	1.99	-0.19	4.24	0.01	
	$\sigma_{1(2)}$	3.13	1.16	0.26	13.89	0.33	
	$\sigma_{2(2)}$	2.67	1.10	0.25	12.85	0.07	
	$\sigma_{3(1)}$	4.36	1.70	0.33	20.72	0.13	
	$\sigma_{4(1)}$	2.86	1.10	0.25	13.14	0.21	
	$\sigma_{5(1)}$	2.79	1.09	0.25	13.41	0.10	
	$\sigma_{H(1)}$	0.82	0.80	0.57	1.22	0.00	
	$\sigma_{H(2)}$	1.00	0.92	0.56	1.90	0.00	
Model 3.10	α_0	-1.54	-1.55	-3.20	0.18	0.01	48000
	θ_1	1.99	1.98	-0.20	4.23	0.01	
	$\sigma_{1(2)}$	2.93	1.17	0.26	14.38	0.07	
	$\sigma_{2(2)}$	2.57	1.09	0.25	13.22	0.04	
	$\sigma_{3(1)}$	4.82	1.71	0.33	20.94	0.50	
	$\sigma_{4(1)}$	2.98	1.09	0.25	13.43	0.26	
	$\sigma_{5(1)}$	2.66	1.09	0.25	13.28	0.06	
	$\sigma_{H(1)}$	0.83	0.80	0.57	1.23	0.00	
	$\sigma_{H(2)}$	0.97	0.90	0.55	1.77	0.00	
Model 3.11	α_0	-1.55	-1.56	-3.40	0.38	0.01	54000
	θ_1	2.09	2.09	-0.41	4.55	0.01	
	$\sigma_{\beta(1)}$	2.12	1.28	0.39	8.68	0.02	
	$\sigma_{\beta(2)}$	1.94	0.91	0.23	9.54	0.02	
	$\sigma_{H(1)}$	0.83	0.81	0.57	1.23	0.00	
	$\sigma_{H(2)}$	0.96	0.89	0.55	1.76	0.00	
Model 3.12	α_0	-1.61	-1.61	-3.32	0.13	0.01	50000
	θ_1	2.09	2.10	-0.11	4.26	0.00	
	σ_β	1.46	1.02	0.34	5.19	0.01	
	$\sigma_{H(1)}$	0.83	0.81	0.58	1.23	0.00	
	$\sigma_{H(2)}$	0.95	0.88	0.55	1.72	0.00	

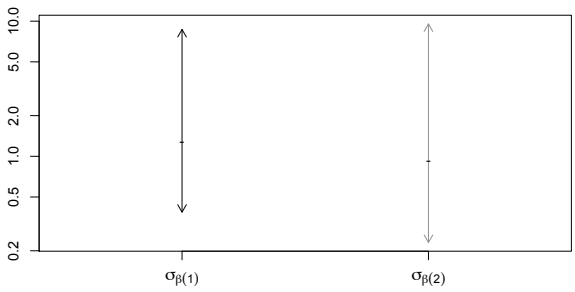


Figure 3.5: Model 3.11 estimates and 95% credible intervals for $\sigma_{\beta(s)}$ on a log-scale; '-' denotes posterior median

but more appropriate for Bayesian models implemented with MCMC (Spiegelhalter et al., 2002). For Model 3.8, the DIC was 4380 as evaluated by OpenBUGS.

3.2.8 Refining the Baseline Model

In the previous subsection, it was found that λ_2 was not statistically significant although there was evidence that temporal blocking might not be discounted. As well, it was found that distinct variances are perhaps only required for each metric group. It would have been equally acceptable to first reduce the model by addressing either of these two issues, but we chose to remove λ_l first. Thus, the equation for health in this second baseline model changed to:

$$H_{i(l)} = \alpha_0 + \varepsilon_{i(l)} \tag{3.10}$$

Posterior estimates for this Model 3.10, were quite similar to those of Model 3.8. Health estimates, denoted LHFI(3.10) and their 95% credible intervals were practically unchanged (Fig. 3.2 in dark grey), as were estimates and credible intervals for $\sigma_{j(s)}$ (not plotted). Comparing Models 3.8 and 3.10, estimates for $\sigma_{H(l)}$ were virtually unchanged, although 95% credible intervals were slightly narrower (Fig. 3.3). As well, the DIC remained the same and summary statistics for the remainder of $\mathbf{\Omega}$ were quite similar to those for Model 3.8 (Table 3.1). Since the overlap in credible intervals for $\sigma_{j(s)}$ was again quite substantial, the next step taken in refining the baseline model was to reduce $\sigma_{j(s)}$ to $\sigma_{\beta(s)}$.

Note that in this subsection we are mainly concerned with baseline models i.e. without covariates since our focus is on refining the baseline, but versions of Model 3.10 (and subsequent baseline models) with covariates also exist. The version of Model 3.10 with covariates would differ only in its equation for health, which would read $H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x_{i(l)}}) + \varepsilon_{i(l)}$, where covariates $\boldsymbol{x}$ are also nested in l. Refining models with covariates appears in Section 3.3.

Returning to the discussion of refining the baseline models, the next baseline model, denoted Model 3.11, takes the following distribution for metric effects $\beta_{j(s)}$:

$$\left[\beta_{j(s)} | \sigma_{\beta(s)}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{\beta(s)}) \tag{3.11}$$

The 95% credible intervals for $\sigma_{\beta(s)}$ (Fig. 3.5) are much narrower than those for $\sigma_{j(s)}$ in the previous two baseline models (Table 3.1), which is a good sign as a more appropriate model is frequently revealed through increased precision of estimates. For the most part, other estimates remained the same. Health estimates LHFI(3.11) virtually did not change, though their 95% credible intervals widened slightly (Fig. 3.2 in medium grey). Metric group effect θ_1 was not as precisely estimated as with the two previous models, since its 95% credible interval widened slightly (Table 3.1). The remainder of the summary statistics for $\boldsymbol{\Omega}$ were largely unchanged, as were the estimates and credible intervals for $\sigma_{H(l)}$, and the DIC was again 4380. There was substantial overlap in 95% credible intervals for $\sigma_{\beta(1)}$ and $\sigma_{\beta(2)}$, which meant that the baseline model could be further refined by reducing $\sigma_{\beta(s)}$ to σ_β.

The distribution of metric effects for the next baseline, Model 3.12, is thus:

$$\left[\beta_{j(s)} | \sigma_{\beta}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{\beta}) \tag{3.12}$$

Again, the 95% credible interval for σ_β is much narrower than those for $\sigma_{\beta(s)}$ (Table 3.1), thus validating the decision to use σ_β. Estimates for this model are again very similar to the previous baseline models: 95% credible intervals for health estimates LHFI(3.12) are slightly narrower than for LHFI(3.11) (Fig. 3.2), though the estimates themselves virtually did not change. Statistics for $\boldsymbol{\Omega}$ including $\sigma_{H(l)}$ virtually did not change (Table 3.1, Fig. 3.3), nor did the DIC.

3.2.9 Summary of Baseline Models

Overall, the four baseline models were very similar to each other. Markov chains for these models all mixed very well, with the slight exception (high autocorrelation) for several of the p's in all of the models and λ_2 in the first model. Slight changes in widths of credible intervals as the model was reduced step-by-step were likely due to the struggle between parsimony and goodness-of-fit. Model 3.12 was employed from this point onwards as a base upon which the environmental covariates described in Section 3.1 were added.

3.3 Incorporating Covariates

As mentioned previously, Jonathan Grant and Lin Lu, the scientists who collected the data, had conducted their own examination of the data before the start of this thesis (Lu et al., 2008). Based on their examination, they believed that salinity, silt-clay % (SC), organic content % (OC) and sorting (SI) should be important for determining health in the Richibucto estuary. They also note that SC, OC and SI are highly correlated; SC and OC are always correlated by definition, and SC/OC and SI happened to be correlated for the Richibucto data. Regarding depth, temperature and mdϕ, Grant and Lu believed that these three covariates were probably not important for the Richibucto sites, since the ranges of values for these covariates are all narrow, although these covariates may be useful in determining health in other estuaries. Finally, Grant and Lu note that distance downstream should be correlated with salinity since proximity to the estuary mouth implies a higher site label as well as higher salinity.

In practice, it is crucial to consider the scientists' viewpoints in incorporating the covariates; using only the results of exploratory analysis on data and nothing else to design the model, and then fitting the model to that same data would have been a circular and somewhat flawed approach. In this case, the biological meaning of the resulting LHFI models could be suspect. Thus, we consider the scientists'

viewpoints in conjunction with our own analysis.

3.3.1 Preliminary Analysis

With multiple covariates, a common difficulty is the presence of too many possible combinations of covariates to be modelled. Therefore, before running any models, we conducted a preliminary analysis on health estimates LHFI(3.12) and the covariates; the results of this analysis were considered in conjunction with the scientists' viewpoints in determining which models to run. We used plots and linear regressions to find covariates that potentially had relationships with health and to determine appropriate transformations or powers. This analysis focussed on the first block of data, as the second block did not contain enough data points (five) on its own. This preliminary analysis served as a guide, complementary to the scientists' expert knowledge, to what combinations of covariates would be worth modelling. Although this analysis implicitly assumed a single-level model where health is incorrectly viewed as an observed variable, it still provided some useful information towards arriving at a final model for constructing a useful LHFI.

A matrix plot of the covariates and LHFI(3.12) indicated that depth should be log-transformed, as points were clumped towards the left side of the plot of health and depth. SC and OC were also log-transformed in order that they may not be restricted to the 0-100 range; extra constraints such as restricted ranges are undesirable since they may lead to computational issues and they can also affect dependent structures, i.e. regression errors may no longer be independent. Note that temperature was excluded from the model, and hence the preliminary analysis since the ranges of temperature during the two sampling times are very distinct from each other and thus temperature would be confounded with blocking. A matrix plot of the transformed covariates and LHFI(3.12) and a table of correlation values (Fig. 3.6(a), Table 3.2) reveal potential linear relationships between health and salinity, health and distance, and (possibly) health and log(SC). A quadratic relationship between health and log(depth) also appears somewhat plausible; such

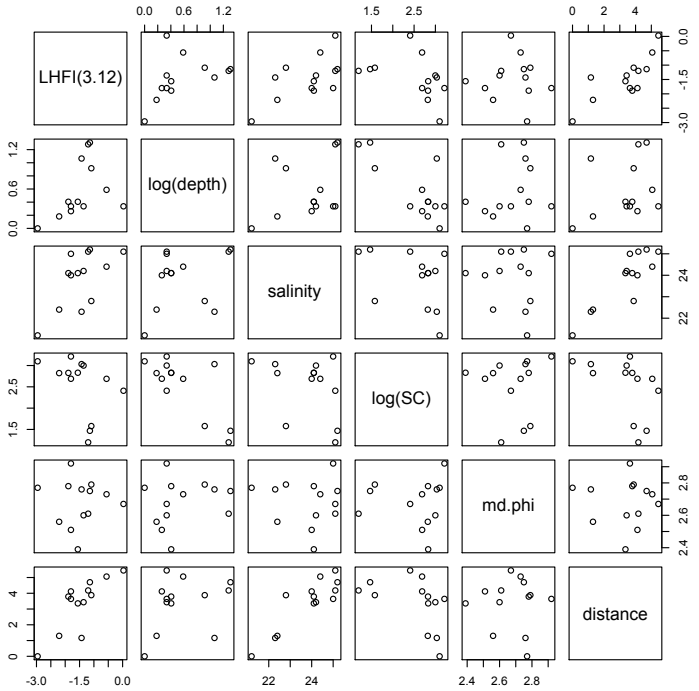
a relationship, however, is likely too convoluted for sensible interpretation in the physical context of the ecosystem. As well, since log(SC), log(OC) and SI were highly correlated as expected (Fig. 3.6(b), Table 3.2)[6], they were treated as one covariate from this point onwards, called the "trivariate", and only log(SC) was included in Fig. 3.6(a) for simplicity. Of the three, SI is the least valuable to the LHFI because it is a derived variable and not field data like the others; any derived variable is less valuable since it is in a sense further from the raw field data. Log(SC) was chosen to represent the trivariate as its data points are less clumped than those of log(OC), though either might be acceptable. As expected, distance was found to be highly correlated with salinity ($r = 0.88$). As well, it turns out that log(depth) is somewhat highly correlated with the trivariate ($r = -0.76$ with log(SC), -0.73 with SI, and -0.64 with log(OC)).

Table 3.2: Correlations between LHFI(3.12) and transformed covariates

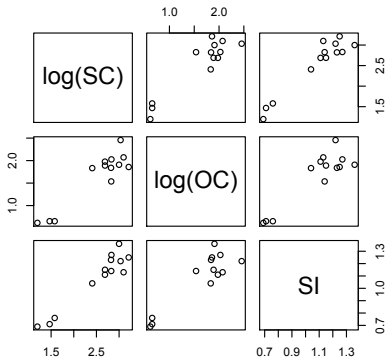
	LHFI(3.12)	log(depth)	salinity	log(SC)	mdϕ	distance	log(OC)	SI
LHFI(3.12)	1	0.43	0.64	-0.42	-0.01	0.8	-0.23	-0.33
log(depth)	0.43	1	0.32	-0.76	0.17	0.31	-0.64	-0.73
salinity	0.64	0.32	1	-0.37	-0.05	0.88	-0.37	-0.26
log(SC)	-0.42	-0.76	-0.37	1	0.02	-0.47	0.92	0.95
mdϕ	-0.01	0.17	-0.05	0.02	1	-0.04	-0.14	-0.2
distance	0.8	0.31	0.88	-0.47	-0.04	1	-0.44	-0.39
log(OC)	-0.23	-0.64	-0.37	0.92	-0.14	-0.44	1	0.9
SI	-0.33	-0.73	-0.26	0.95	-0.2	-0.39	0.9	1

Fitting simple and multiple linear regressions of LHFI(3.12) on various combinations of the five usable covariates (log(depth), salinity, log(SC), mdϕ and distance) provided further indications as to which covariates could be useful in an LHFI model. The adjusted R^2, t-test results and F-test results were particularly useful in comparing regressions. The regressions suggested that salinity, distance and (possibly) log(SC) were the most useful covariates of those available in determining health, and additionally that the three sets of interactions, log(depth)$\times$ log(SC),

[6]The correlation between these three appears to be mainly due to the three points in the bottom left corners of plots in Fig. 3.6(b). These points correspond to sites 12, 15, and 16. Speaking to Grant and Lu revealed no reason to consider these as outliers; thus, excluding these sites from analysis was not advisable as this could lead to nonsensical estimates.



(a)



(b)

Figure 3.6: Matrix plot using the first block of data of (a) LHFI(3.12) and transformed covariates and (b) the transformed trivariate

log(depth)× salinity, and log(SC)× distance, could be of use. Note that since distance and salinity were very highly correlated, we did not consider models that included both of these; however, we did consider including log(depth) and log(SC) in the same model as they were less highly correlated and we did not want to rule out anything yet in a preliminary model, and by the same reasoning we also considered the interaction log(depth)× log(SC).

3.3.2 Implementing the Models

Ten models with covariates of the form of Model 3.12 were implemented with MCMC. Determining the most appropriate form of $f()$ was an iterative process. Roughly speaking, if a specific covariate was found to have no significant impact on health, it was removed from the model, and the modified model was run. The results from the modified model would then be examined, and the process would be repeated. Or as another example, an interaction might be introduced, if the data exploration described above indicated it could be useful in determining health.

The motivation behind each choice of model together with the confidence levels for each regression coefficient in the model are described in Table 3.3 (recall that the confidence level of a coefficient is the level of the credible interval at which the coefficient becomes significant). The models are listed in roughly the order in which they were run, so as to illustrate the evolution of the final LHFI-AMBI model with covariates. For all of the models in Table 3.3, the estimates and 95% credible intervals for health, the estimates and 95% intervals for σ_β and θ_1, and the DIC values were very similar; and these values were also very similar to those of the baseline model (Model 3.12) upon which they were based. However, the estimated precision of health estimates $\sigma_{H(l)}$ varied somewhat with the combination of covariates; estimates and intervals for each model are plotted in Fig. 3.3 and labelled 1 to 10 according to their order in Table 3.3.

MCMC chains of the models mixed fairly well for the most part. Minor mixing problems were observed: autocorrelation was consistently low but long-living for

coefficients involving log(SC) (for both main effects and interactions), and autocorrelation was high for some of the p's. As well, the model with log(depth), salinity, log(SC), and log(depth)×log(SC) required an unusually long burn in (300,000 iterations). This could have been due to the high level of correlation between log(depth) and log(SC), or alternatively due to chance, if the randomly chosen initial values happened to be far from the convergence values.

The form of $f()$ for these models is a linear regression using centred covariates that have been transformed (if necessary). As an illustration of models with only main effects, the form of $f()$ for the first model in Table 3.3 with log(depth), salinity, log(SC), and mdϕ is:

$$f(\boldsymbol{\alpha}, \boldsymbol{x}_i) = \alpha_1(x_{1i} - \overline{x}_{1+}) + \alpha_2(x_{2i} - \overline{x}_{2+}) + \alpha_3(x_{3i} - \overline{x}_{3+}) + \alpha_4(x_{4i} - \overline{x}_{4+}) \quad (3.13)$$

where for each i,

$$x_{1i} = \log(\text{depth}),\ x_{2i} = \text{salinity},\ x_{3i} = \log(\text{SC}), \quad (3.14)$$
$$x_{4i} = \text{md}\phi,\ x_{5i} = \text{distance downstream}.$$

As an illustration of models with interactions and main effects, the form of $f()$ for the model with log(depth), salinity, log(SC), and log(depth)×log(SC) is as follows:

$$f(\boldsymbol{\alpha}, \boldsymbol{x}_i) = \alpha_1(x_{1i} - \overline{x}_{1+}) + \alpha_2(x_{2i} - \overline{x}_{2+}) + \alpha_3(x_{3i} - \overline{x}_{3+}) + \alpha^*_{13}(x_{1i} - \overline{x}_{1+})(x_{3i} - \overline{x}_{3+}) \quad (3.15)$$

Note that covariates are centred to reduce correlation (and thus perhaps improve mixing) between the intercept α_0 and coefficients α_r where r indicates the covariate. The motivation for centring covariates lies within the theory of linear regression: if there is one covariate, centring it in this way completely removes any correlation between the maximum likelihood estimators (assuming normality) of α_0 and α_r; if there are multiple covariates, centring reduces correlation in the same way, although not to 0. Since the likelihood is a major component of the Bayesian framework, centring covariates could also reduce correlation within a Bayesian hierarchical model, though it will not remove correlation entirely.

Table 3.3: LHFI-AMBI models implemented with covaraites

	Covariates	Motivation	Confidence levels of slopes associated with covariates
1	log(depth), salinity, log(SC), mdϕ	With data for only 18 sites, we were loath to begin with a complex model, e.g. one with interactions. Thus, the first model we implemented included the four covariates available at the time: log(depth), salinity, log(SC) and mdϕ[7]. As stated earlier, we did not avoid including both log(depth) and log(SC) although they are correlated ($r = -0.76$) since we do not want to exclude any possibilities in a preliminary model. As well, although mdϕ did not appear significant in the preliminary analysis, it was included to begin with, as it could have been easily removed if found unimportant.	log(depth): 60-70%, salinity: 95-99%, log(SC): <20%, mdϕ: 30-40%
2	log(depth), salinity, log(SC)	An adaptation of the model above: mdϕ was removed since it was the second least significant in the previous model, and it appeared to be least significant in the preliminary analysis. Although least significant in the previous model, log(SC) was not removed since it appeared significant throughout the preliminary analysis and according to the scientists' viewpoints.	log(depth): 70-80%, salinity: 95-99%, log(SC): <20%
3	log(depth), salinity, log(SC), log(depth)× log(SC)	An adaptation of the model above: log(depth)×log(SC) was added since it appeared promising in preliminary analysis. Again, note that we do not avoid including this interaction although its elements are correlated since this is a preliminary model.	log(depth): 30-40%, salinity: >99%, log(SC): 90-95%, log(depth)× log(SC): 95-99%

Continued on Next Page...

[7]Distance downstream data were obtained part-way through the modelling process.

Table 3.3 – Continued

	Covariates	Motivation	Confidence levels of slopes associated with covariates
4	log(depth), salinity, log(SC), log(depth)× salinity	Another adaptation of the 2nd model: log(depth)×salinity was added since it also appeared promising in the preliminary analysis.	log(depth): 40-50%, salinity: 85-90%, log(SC): 60-70%, log(depth)× salinity: 85-90%
5	log(depth), salinity, log(SC), log(depth)× salinity, log(depth)× log(SC)	An attempt at combining the previous two models since both log(depth)× log(SC) and log(depth)× salinity appeared significant.	log(depth): 30-40%, salinity: 95-99%, log(SC): 90-95%, log(depth)× salinity: 20-30%, log(depth)× log(SC): 90-95%
6	distance	At this point in time, the distance downstream data were obtained. Since the preliminary analysis revealed that distance was very highly correlated with salinity ($r = 0.88$), we avoided including salinity in the same model with distance. Other than distance, that left three covariates log(depth), log(SC) and mdϕ, but since mdϕ seemed by all indications to be not useful, that actually left only log(depth) and log(SC). Since the preliminary analysis also revealed a strong linear relationship between health and distance, we began with a model with only distance (here), and later implemented models with log(depth) and log(SC) as well.	distance: >99%
7	log(depth), distance	See previous note.	log(depth): 80-85%, distance: >99%
8	log(depth), distance, log(depth)× distance	See previous note. As well, the interaction log(depth)× distance also appeared promising in the preliminary analysis.	log(depth): 70-80%, distance: 95-99%, log(depth)× distance: 70-80%

Continued on Next Page...

Table 3.3 – Continued

	Covariates	Motivation	Confidence levels of slopes associated with covariates
9	log(SC), distance, log(SC)× distance	See previous note. Since log(depth) did not appear to be significant in the previous two models, we did not include it in this model. As well, the interaction log(SC)× distance is included since it appeared promising in preliminary analysis.	log(SC): 60-70%, distance: >99%, log(SC)× distance: 60-70%
10	log(depth), log(SC), distance, log(depth)× log(SC)	An adaptation of the third model, which was also one of the two most successful models at this point: substitute distance for salinity since they are strongly correlated.	log(depth): 70-80%, log(SC): 30-40%, distance: >99%, log(depth)× log(SC): 70-80%

Of the ten models with covariates implemented with MCMC, that with log(depth), salinity, log(SC), log(depth)×log(SC), and that with distance are the two that provided the "best" fits to the data, in that the confidence levels for the covariates in these models were among the highest and there were no extraneous non-significant covariates at the 90% confidence level. As well, the estimated precisions of health $\sigma_{H(l)}$ were among the lowest for these two models (numbers 3 and 6 in Fig. 3.3). Denote these models respectively as Model 3.12a and Model 3.12b. Note that it is hard to quantitatively determine if one of the two models is "better" than the other, since their posterior means and credible intervals for health and their DIC were empirically identical.

Some summary statistics for α_0 and $\boldsymbol{\alpha}$ for Model 3.12a and Model 3.12b are shown in Table 3.4. The credible intervals indicate confidence levels at which the covariates have statistically significant effects on health. For example, the 99% credible interval for α_2 in Model 3.12a reveals that salinity has a statistically significant impact on health at a 99% confidence level after adjusting for the other covariates. Furthermore, the sign of an estimated coefficient indicates the direction of its corresponding covariate's effect on health. For example, in Model 3.12a,

$\widehat{\alpha}_1 = 0.18$ and $\widehat{\alpha}_2 = 0.39$ signify that within the observed range, increasing depth (and log(depth)) and salinity are both associated with better health. As stated in the first chapter, this aspect of the LHFI could be particularly useful to policy makers: scientists may believe that a certain covariate has an effect to health, and the LHFI can provide quantitative evidence to support (or oppose) such beliefs. As well, as Chiu et al. (2008) point out, policy makers who are presented with several factors that have potential impact on ecosystem health and who must devise conservation policies in response to selected factors with limited resources would likely find this capability of the LHFI approach useful.

Table 3.4: Summary Statistics for α_0 and $\boldsymbol{\alpha}$ for Model 3.12a and Model 3.12b

	Covariate	Mean	99% Credible Interval	95% Credible Interval	90% Credible Interval	MC Error	# Draws
				Model 3.12a			
α_0	intercept	-1.33	(-3.92, 1.37)	(-3.04, 0.41)	(-2.69, 0.03)	0.009	80000
α_1	log(depth)	0.18	(-1.01, 1.37)	(-0.70, 1.05)	(-0.55, 0.90)	0.002	
α_2	salinity	0.39	(0.05, 0.77)	(0.13, 0.67)	(0.18, 0.62)	0.001	
α_3	log(SC)	-0.87	(-2.14, 0.47)	(-1.80, 0.09)	(-1.64, -0.08)	0.008	
α^*_{13}	log(depth)× log(SC)	1.96	(-0.36, 4.17)	(0.30, 3.58)	(0.60, 3.29)	0.019	
				Model 3.12b			
α_0	intercept	-1.57	(-4.21, 1.07)	(-3.27, 0.15)	(-2.91, -0.23)	0.008	50000
α_5	distance	0.37	(0.09, 0.66)	(0.17, 0.58)	(0.20, 0.54)	0.001	

We attempted to combine the covariates from these two models into one integrated model (see models 7-9 in Table 3.3) but were ultimately unsuccessful. When other covariates were added to the model with distance, they were not significant at any reasonable level (i.e. at the least with a confidence level of 90%), even if they had been significant in other models. One possible explanation for this phenomenon is that distance likely has much less measurement error than the other covariates, since it is easier to measure accurately in the field than other variables that required lab work. An effect on health from distance would therefore come across more clearly than an effect from the other covariates, i.e. distance might "smother" effects from the noisier covariates.

In another attempt to combine covariates from these two models, we experimented with building a new covariate from salinity and distance since these are highly correlated. Specifically, we attempted to build a density measure, that is, either salinity/distance or distance/salinity. These would be interpreted as a measure of saltiness per unit of distance, and a measure of how far sites are located relative to salinity level. We chose to build a density measure as it seemed to be more interpretable given the physical context than a traditional simple linear combination of the two covariates, salinity and distance having two different scales. We fit four separate simple linear regressions: LHFI(3.12) on salinity, LHFI(3.12) on distance, LHFI(3.12) on salinity/distance and LHFI(3.12) on distance/salinity, and compared the results of the four regressions. As the two fitted regressions on the density measures were each found to be quite similar to either the regression on salinity or the regression on distance, it appeared that this approach was not helpful and was thus abandoned.

In light of the scientists' expertise and our quantitative analyses, we would endorse both Model 3.12a and Model 3.12b equally as latent health factor models for the Richibucto estuary based on AMBI.

3.4 Comparing LHFI-AMBI, AMBI and ITI

A matrix plot of AMBI, ITI and LHFI-AMBI (i.e. LHFI(3.12), LHFI(3.12a), LHFI(3.12b) which are all equivalent empirically) is provided to illustrate the relationships between LHFI-AMBI and AMBI and ITI (Fig. 3.7). Note that we include ITI in this comparison although we have not yet discussed models involving ITI so as to compare AMBI and ITI as well.

LHFI-AMBI is strongly negatively correlated with AMBI itself ($r = -0.81$). The negative relationship is not unexpected since a high AMBI value indicates poor health and a low AMBI value indicates good health (A. Borja and Erez, 2000), while the LHFI is designed such that a high LHFI value will always indicate good health.

The strong linear relationship between LHFI-AMBI and AMBI testifies that the LHFI model is no less effective in estimating health as the index upon whose data it is based.

On the other hand, there is no strong relationship between LHFI-AMBI and ITI ($r = 0.01$), nor between AMBI and ITI ($r = -0.33$). Again, this is not unexpected since it is known that AMBI and ITI measure different aspects of health. The next chapter discusses a more comprehensive latent health model that incorporates metrics from both AMBI and ITI.

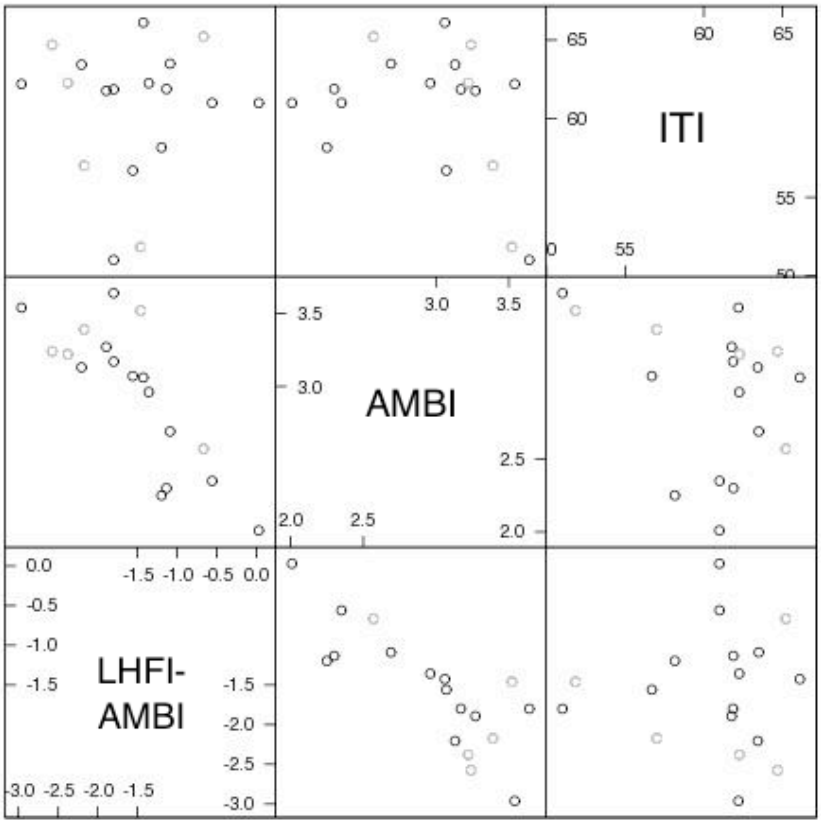


Figure 3.7: Matrix plot of LHFI(3.12) (empirically equivalent to LHFI(3.12a) and LHFI(3.12b)), AMBI and ITI; a black '∘' denotes a point in the first temporal block and a grey '∘' denotes a point in the second block

Chapter 4

Fitting an LHFI Model to Richibucto AMBI and ITI

Since AMBI and ITI focus on different aspects of health, it was our aim to create a more comprehensive index for Richibucto by incorporating metrics from both indices into a single model. We chose to fit a combined model instead of building a model for ITI on its own, since as mentioned in Chapter 3, Grant and Lu believe that the ITI data are not as important in indicating health as AMBI for Richibucto. We call the models in this chapter LHFI-AMBI+ITI models.

4.1 The ITI data

In many ways, the ITI data for Richibucto are similar to the AMBI data. The same taxonomic data set can be used to calculate both AMBI and ITI. Note, however, that while both indices rely on the same taxonomic data set, ITI still provides additional information beyond that from AMBI. For ITI, the 88 species observed at Richibucto were organized into four disjoint taxonomic groups according to species feeding habits (listed below as described by Cromey et al. (2002)), instead of five disjoint taxonomic groups by sensitivity for AMBI. As the same data set was used for both indices, many of the issues discussed in Chapter 3 regarding the AMBI data

apply to the ITI data as well, e.g. temporal blocking, diffuse priors, considering the number of replicates per site as fixed, and considering simple parameters as fixed.

> "*Group I (suspension feeders)*: these animals feed on detritus from the water column and usually lack sediment grains in their stomach contents, e.g. Spio, Spiophanes, Sabella, Ampelisca, Corophium, Phaxas pellucidus, Mya arenaria, Ophiothrix fragilis and Amphiura filiformis.
>
> *Group II (interface/surface detrital feeders)*: these animals obtain the same types of food as suspension feeders but usually from the upper 0.5 cm of the sediment, e.g. Nephtys incisa, Levinsenia gracilis, Polydora, Cirratulidae, Scalibregmatidae, Photis, Mysella and Ophiura.
>
> *Group III (deposit feeders)*: these invertebrates generally feed from the top few centimetres of the sediment and feed on encrusted mineral aggregates, deposit particles or biological remains. While carnivores have been classified as Group 2 by Word (1980), they are included here in Group 3 as in Codling and Ashley (1992), e.g. Anaitides, Goniada maculata, Nephtys hombergii, Scoloplos armiger, Nucula and Thyasira.
>
> *Group IV (specialised environment feeders)*: mobile burrowers that feed on deposited organic material. While exhibiting variable feeding behaviour, they are all adapted to live in highly anaerobic sediment, e.g. Ophryotrocha, Schistomeringos, Capitella capitata, Notomastus latericeus, Oligochaeta and Bittium."

The relative abundances of these four taxonomic groups are combined via the following formula to produce a continuous scalar index, with values ranging from 0 to 100, and is positively correlated with health:

$$\mathrm{ITI} = 100 - 33.3\left(\frac{0 \times \mathrm{GI} + 1 \times \mathrm{GII} + 2 \times \mathrm{GIII} + 3 \times \mathrm{GIV}}{\mathrm{GI} + \mathrm{GII} + \mathrm{GIII} + \mathrm{GIV}}\right) \tag{4.1}$$

where GI, GII, GIII and GIV are the abundances of taxonomic Groups I-IV.

High abundances in taxonomic Groups I and II indicate "good health" (taxonomic Group II to a lesser degree), while a high abundance in taxonomic Group

IV indicates "poor health". Abundance in taxonomic Group III is neither an indicator of "good health" nor of "poor health". Thus, taxonomic Groups I and II are positively related to health, taxonomic Group IV is negatively related to health, and taxonomic Group III is indifferent to health.

4.2 The Model

The LHFI-AMBI+ITI is essentially an extension of the LHFI-AMBI model; an extra component for ITI was constructed in the same fashion and with the same reasoning as for AMBI, and this was added to the existing LHFI-AMBI model to form the LHFI-AMBI+ITI model. Let m denote the data set, where $m = 1$ represents the AMBI data set and $m = 2$ represents the ITI data set.

Similar to the LHFI-AMBI model, the ITI metrics are split into two groups: $s = 1$, containing the third and fourth metrics which are indifferent and negatively related to health; and $s = 2$ containing the first and second metrics which are positively related to health. For $s = 1$, the link function is inverted as in (3.5). Since $s = 1$ and $s = 2$ have the same interpretation regardless of the value of m, we crossed s with m. Thus, metric effects are nested in $s \times m$; that is, $\beta_{j(m\times s)}$ denotes the effect on ν of the jth metric in the sth metric group and the mth data set. For AMBI, we take the variance of β to be as it was in the LHFI-AMBI model i.e. $\sigma_{\beta(m=1)}$. For ITI, however, we cannot make any prior assumptions about the metric effect variances, and so we take $\sigma_{j((m=2)\times s)}$.

Note that crossing s and m as outlined above may not be entirely precise since $s = 1$ contains the indifferent metric for ITI, but $s = 2$ contains the indifferent metric for AMBI. This choice of allocating the indifferent metric is so that each value of s is associated with at least two levels of j so as to avoid weak identifiability of θ_s (see Section 3.2). At this point, the implications of changing the current crossing scheme to address the issue of somewhat inconsistent denitions of s are unclear, and this issue can be investigated in future work. For now we return to

the model so far outlined above.

A term is needed in the model to explain the difference among the data set-metric combinations (i.e $m \times s$). Since we already chose to use $\beta_{j(m\times s)}$, it was most logical to use other terms associated with $m \times s$ to explain $\beta_{j(m\times s)}$. Since any crossed term can in principle be broken down into main effects and an interaction, we defined

$$\beta_{j(m\times s)} = \gamma_m + \theta_s + \xi_{m\times s} + \omega_{j(m\times s)} \tag{4.2}$$

where θ_s, γ_m, and $\xi_{m\times s}$ represent the effects of metric group, data set and their interaction on metric effect β, and they are all fixed effects. Under this formulation, β is no longer an error term with mean 0 as for LHFI-AMBI, but this change is acceptable as the β's are essentially nuisance parameters[1].

The LHFI-AMBI+ITI model includes temporal blocks, for the same reason as the LHFI-AMBI model. As with the latter, health is nested in block with different variances per block, but metric effect is not nested in or crossed with block. From the results of the LHFI-AMBI model, we saw that λ for AMBI is unnecessary, but a λ term was initially included for the ITI data since we had no prior quantitative information on ITI. However, the model does not allow for a $\lambda_{m\times l}$ term since health should not be affected by data set. If the model allowed it, the equation for health would read

$$H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}_{i(l)}) + \lambda_{m\times l} + \varepsilon_{i(l)} \tag{4.3}$$

Even if the model did allow it, potential idenitifiability issues with γ_m could arise in practice, similar to those described in Section 2.1, so it would not be recommended to include a $\lambda_{m\times l}$ term. A model with λ_l instead of $\lambda_{m\times l}$ is mathematically viable, but would be perhaps less realistic since the effect λ would be the same for ITI and AMBI. However, since λ_l was already found to be insignificant for the LHFI-AMBI models, it was quite possible that it a corresponding term would also be insignificant for the LHFI-AMBI+ITI models. Thus, we chose to fit a model with

[1]Alternatively, if a 0-mean β was desired, θ_s, γ_m, and $\xi_{m\times s}$ could be placed in the equation to explain ν instead of β.

λ_l and would have returned to consider this problem only if λ_l was found to be significant.

An initial LHFI model for the Richibucto AMBI and ITI data incorporating metric groups, temporal blocking, and multiple data sets is as follows. Let $Y_{i(l)\times j(m\times s)\times k}$ denote the value of the kth replicate of the jth metric (nested within the sth metric group of the mth data set) for the ith site (nested within the lth temporal block). For simplicity, this is written as Y_{ijklms}. Other parameters are similarly simplified: $p_{i(l)\times j(m\times s)}$ as p_{ijlms}, $v_{i(l)\times j(m\times s)}$ as v_{ijlms}, $\beta_{j(m\times s)}$ as $\beta_{j(ms)}$, $\sigma_{j((m=2)\times s))}$ as $\sigma_{j(2s)}$, and $\sigma_{\beta(m=1)}$ as $\sigma_{\beta(1)}$.

For $m = 1$ (AMBI):

$$[Y_{i3kl11}, Y_{i4kl11}, Y_{i5kl11}, N_{ikl} - Y_{i3kl11} - Y_{i4kl11} - Y_{i5kl11} | N_{ikl}, p_{i3l11}, p_{i4l11}, p_{i5l11}] \quad (4.4)$$
$$\sim \text{multinomial}(N_{ikl}, p_{i3l11}, p_{i4l11}, p_{i5l11}, 1 - p_{i3l11} - p_{i4l11} - p_{i5l11})$$

$$\nu_{ijl11} = \ln \frac{1 - p_{i3l11} - p_{i4l11} - p_{i5l11}}{p_{ijl11}}, \qquad j = 3, 4, 5 \quad (4.5)$$

$$[Y_{i1kl12}, Y_{i2kl12}, N_{ikl} - Y_{i1kl12} - Y_{i2kl12} | N_{ikl}, p_{i1l12}, p_{i2l12}] \quad (4.6)$$
$$\sim \text{multinomial}(N_{ikl}, p_{i1l12}, p_{i2l12}, 1 - p_{i1l12} - p_{i2l12})$$

$$\nu_{ijl12} = \ln \frac{p_{ijl12}}{1 - p_{i1l12} - p_{i2l12}}, \qquad j = 1, 2 \quad (4.7)$$

For $m = 2$ (ITI):

$$[Y_{i3kl21}, Y_{i4kl21}, N_{ikl} - Y_{i3kl21} - Y_{i4kl21} | N_{ikl}, p_{i3l21}, p_{i4l21}] \quad (4.8)$$
$$\sim \text{multinomial}(N_{ikl}, p_{i3l21}, p_{i4l21}, 1 - p_{i3l21} - p_{i4l21})$$

$$\nu_{ijl21} = \ln \frac{1 - p_{i3l21} - p_{i4l21}}{p_{ijl21}}, \qquad j = 3, 4 \quad (4.9)$$

$$[Y_{i1kl22}, Y_{i2kl22}, N_{ikl} - Y_{i1kl22} - Y_{i2kl22} | N_{ikl}, p_{i1l22}, p_{i2l22}] \quad (4.10)$$
$$\sim \text{multinomial}(N_{ikl}, p_{i1l22}, p_{i2l22}, 1 - p_{i1l22} - p_{i2l22})$$

$$\nu_{ijl22} = \ln \frac{p_{ijl22}}{1 - p_{i1l22} - p_{i2l22}}, \qquad j = 1, 2 \tag{4.11}$$

For both ITI and AMBI:

$$\begin{gathered}
\nu_{ijlms} = H_{i(l)} + \beta_{j(ms)}, \qquad H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}_{i(l)}) + \lambda_l + \varepsilon_{i(l)} \\
\left[\varepsilon_{i(l)} | \sigma_{H(l)}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{H(l)}), \qquad \beta_{j(ms)} = \gamma_m + \theta_s + \xi_{ms} + \omega_{j(ms)} \\
\left[\omega_{j(1s)} | \sigma_{\beta(1)}\right] \overset{\text{iid}}{\sim} N(0, \sigma^2_{\beta(1)}), \qquad \left[\omega_{j(2s)} | \sigma_{j(2s)}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{j(2s)})
\end{gathered} \tag{4.12}$$

And linear constraints:

$$\lambda_{l=1} = 0, \quad \gamma_{m=1} = 0, \quad \theta_{s=1} = 0, \quad \xi_{11} = \xi_{12} = \xi_{21} = 0 \tag{4.13}$$

where

$$\text{data set } m = \begin{cases} 1 & \text{if AMBI} \\ 2 & \text{if ITI} \end{cases}$$

$$\text{temporal block } l = \begin{cases} 1 & \text{if data collected in September} \\ 2 & \text{if data collected in October} \end{cases}$$

$$\text{site } i = \begin{cases} 1, \ldots, 3, 9, \ldots, 18 & \text{if } l = 1 \\ 4, \ldots, 8 & \text{if } l = 2 \end{cases}$$

$$\text{metric group } s = \begin{cases} 1 & \text{negatively related to health} \\ 2 & \text{positively related to health} \end{cases}$$

$$\text{AMBI metric } j = \begin{cases} 3, 4, 5 & \text{in metric group } s = 1 \\ 1, 2 & \text{in metric group } s = 2 \end{cases}$$

$$\text{ITI metric } j = \begin{cases} 3, 4 & \text{in metric group } s = 1 \\ 1, 2 & \text{in metric group } s = 2 \end{cases}$$

$$\text{replicate } k = 1, \ldots, r_i \quad \text{where } r_i \text{ is the number of replicates at site } i$$

This model, which encompasses equations (4.4)-(4.12) and is denoted Model 4.4, considers metric effects $\beta_{j(ms)}$ to be independent for simplicity as a preliminary step, like the LHFI-AMBI models. Again, we wanted to avoid complications in modelling and implementation to begin with. Diffuse priors are adopted for parameters, similar to those in 3.9.

Note as well that the linear constraint $\theta_{s=1} = 0$ here is different from the corresponding constraint on the LHFI-AMBI Model (3.8), that is, $\theta_{s=2} = 0$. This discrepancy was due to an oversight, but should not have any serious repercussions:

changing the constraint to be consistent with that of LHFI-AMBI could possibly lead to a slight offset in the health estimates, which is inconsequential since health estimates are only considered relative to each other and not in absolute terms.

Again, we are attempting to fit a highly complex model with multiple parameters to only 18 sites, which may be of concern to the reader. But, as stated before in Chapter 3, we do have a somewhat large benthic data set, which is even larger than that in Chapter 3 with (5 AMBI metrics + 4 ITI metrics) $\times \sum_{i=1}^{18} r_i = 441$ data points, and additionally, the use of priors in the Bayesian framework helps to mitigate concerns.

4.2.1 The Baseline Model

The model building process for the combined AMBI and ITI data was very similar to the process for LHFI-AMBI. We began by fitting the baseline model above (Model 4.4), refined that model, and then proceeded to add various combinations of the covariates to the model to see which covariates are most useful in determining health.

Health estimates for this baseline, denoted LHFI(4.4), and corresponding 95% credible intervals are plotted together with those for all later AMBI+ITI models and with all AMBI models for comparison in Fig. 4.1 (those for LHFI(4.4) are in the bottom panel, in black). Estimates and 95% credible intervals for $\sigma_{H(l)}$ are plotted in Fig. 4.3. Summary statistics for $\mathbf{\Omega}$ are in Table 4.1, and estimates and 95% credible intervals for $\sigma_{j(s)}$ are also plotted in Fig. 4.2.

Rankings of health by site are somewhat similar to rankings using the AMBI data alone (Fig. 4.1). However, health is more precisely estimated for AMBI+ITI than for AMBI, since the credible intervals using AMBI+ITI data are substantially narrower. As well, the estimated precision of health $\sigma_{H(l)}$ for Model 4.4 is somewhat improved in terms of posterior credible intervals compared to that of the LHFI-AMBI models (compare Fig. 4.3 to Fig. 3.3). On the other hand, it appears that this increased precision of estimates has not greatly changed the ability of the LHFI

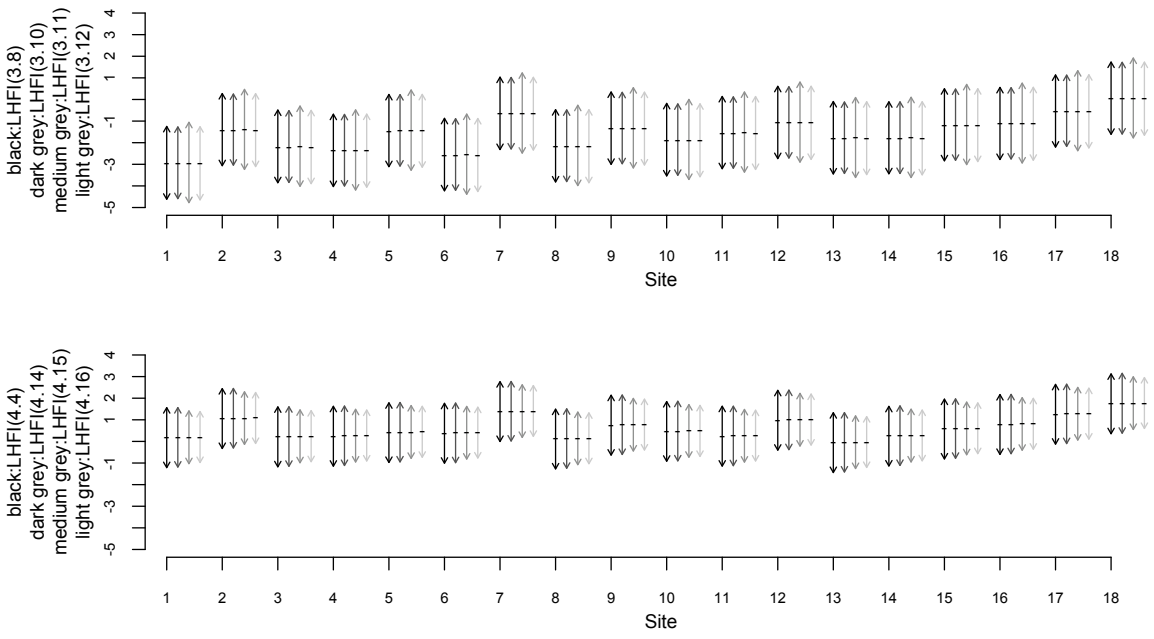


Figure 4.1: Estimates and 95% credible intervals for LHFI-AMBI baseline models (top panel) and LHFI-AMBI+ITI baseline models (bottom panel); '-' denotes $\widehat{H}_i$

to distinguish between different sites, as credible intervals for health for LHFI-AMBI and LHFI-AMBI+ITI appear to overlap by roughly the same amount. It is difficult to interpret this phenomenon, and it is unclear whether one set of models is superior to the other.

The temporal block effect λ_2 is insignificant in this model, which is not surprising given the results of LHFI-AMBI; its 95% credible interval contains 0, and in fact its confidence level was very low at 20-30%. Minor mixing problems, i.e. high autocorrelation, were observed for λ_2 as well. Also unsurprisingly, credible intervals for $\sigma_{H(1)}$ and $\sigma_{H(2)}$ are distinctly different. Thus, temporal blocking is important although λ_l is inessential. The interaction ξ_{22} was also found to be insignificant considering its 95% credible interval (confidence level <20%). In addition, distinct $\sigma_{j(2s)}$ are unnecessary for each metric, as there is significant overlap in their credible intervals within metric group, and could perhaps be replaced by distinct variances for each metric group. Thus, the model could have been reduced by addressing any

of these three points. However, we first removed λ_l as the previous chapter suggests that it was very unlikely to be significant. Thus, in the second baseline model, the equation for health became:

$$H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}_{i(l)}) + \varepsilon_{i(l)} \tag{4.14}$$

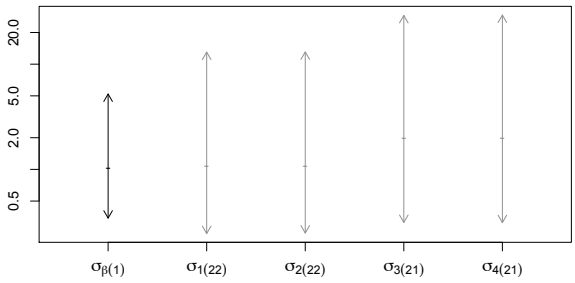


Figure 4.2: Model 4.4 estimates and 95% credible intervals for $\sigma_{\beta(1)}$ and $\sigma_{j(2s)}$ on a log-scale; '-' denotes posterior median

The results of implementing this model, denoted Model 4.14, were quite similar to those of the previous model. Health estimates, denoted LHFI(4.14) and their 95% credible intervals were practically unchanged (Fig. 4.1, bottom panel), although credible intervals for $\sigma_{H(l)}$ narrowed (Fig. 4.3). Estimates and credible intervals for $\sigma_{\beta(1)}$ and $\sigma_{j(2s)}$ were virtually unchanged (not plotted), and so were summary statistics for $\boldsymbol{\Omega}$ (Table 4.1). This indicates that this model, like the previous one, can be refined by employing distinct variances for each metric group instead of for each individual metric. As well, since ξ_{22} (and hence, ξ_{ms} for all m and s) is insignificant in this model (confidence level <20%), we could also refine the model additionally by removing ξ_{ms}. We chose to address the latter issue first in refining the baseline model, although addressing the former first would have been equally acceptable. Thus, in the third baseline model, the effects γ_m and θ_s are simply additive and this model adopts the following equation for metric effect:

$$\beta_{j(ms)} = \gamma_m + \theta_s + \omega_{j(ms)} \tag{4.15}$$

For this third baseline model, health estimates LHFI(4.15) and their 95% credible intervals were practically unchanged, though the intervals were perhaps slightly

Table 4.1: Summary statistics for LHFI-AMBI+ITI baseline models

		Mean	Median	2.5th %-ile	97.5th %-ile	MC Error	# Draws
Model 4.4	α_0	0.64	0.63	-0.77	2.05	0.01	52000
	γ_2	1.76	1.76	-0.37	3.86	0.01	
	λ_2	-0.13	-0.13	-0.98	0.71	0.01	
	θ_2	-2.11	-2.11	-4.24	0.02	0.01	
	ξ_{22}	-0.07	-0.06	-3.76	3.58	0.01	
	$\sigma_{\beta(1)}$	1.46	1.02	0.34	5.21	0.01	
	$\sigma_{1(22)}$	2.70	1.08	0.25	13.02	0.08	
	$\sigma_{2(22)}$	2.78	1.08	0.25	13.07	0.13	
	$\sigma_{3(21)}$	5.67	1.99	0.31	29.03	0.15	
	$\sigma_{4(21)}$	5.71	1.97	0.31	29.29	0.16	
	$\sigma_{H(1)}$	0.64	0.62	0.44	0.95	0.00	
	$\sigma_{H(2)}$	0.81	0.75	0.45	1.55	0.00	
Model 4.14	α_0	0.62	0.62	-0.77	2.03	0.01	50000
	γ_2	1.75	1.75	-0.35	3.85	0.01	
	θ_2	-2.11	-2.11	-4.28	0.07	0.01	
	ξ_{22}	-0.07	-0.05	-3.80	3.61	0.01	
	$\sigma_{\beta(1)}$	1.45	1.02	0.35	5.21	0.01	
	$\sigma_{1(22)}$	2.56	1.07	0.25	13.08	0.04	
	$\sigma_{2(22)}$	2.64	1.07	0.25	13.12	0.06	
	$\sigma_{3(21)}$	5.60	2.00	0.31	29.64	0.12	
	$\sigma_{4(21)}$	5.91	1.97	0.31	30.72	0.20	
	$\sigma_{H(1)}$	0.63	0.62	0.44	0.93	0.00	
	$\sigma_{H(2)}$	0.76	0.71	0.44	1.36	0.00	
Model 4.15	α_0	0.65	0.65	-0.59	1.90	0.01	60000
	γ_2	1.73	1.73	0.04	3.41	0.00	
	θ_2	-2.13	-2.13	-3.78	-0.51	0.00	
	$\sigma_{\beta(1)}$	1.32	0.97	0.34	4.38	0.01	
	$\sigma_{1(22)}$	2.43	1.03	0.25	11.71	0.07	
	$\sigma_{2(22)}$	2.53	1.04	0.25	12.07	0.06	
	$\sigma_{3(21)}$	4.91	1.88	0.32	24.58	0.13	
	$\sigma_{4(21)}$	4.80	1.78	0.32	23.90	0.15	
	$\sigma_{H(1)}$	0.64	0.62	0.44	0.94	0.00	
	$\sigma_{H(2)}$	0.76	0.71	0.44	1.36	0.00	
Model 4.16	α_0	0.63	0.63	-0.57	1.85	0.01	44000
	γ_2	1.73	1.73	0.12	3.34	0.00	
	θ_2	-2.14	-2.14	-3.69	-0.60	0.00	
	$\sigma_{\beta(1)}$	1.29	0.95	0.33	4.33	0.01	
	$\sigma_{\beta(21)}$	1.48	0.82	0.22	6.69	0.01	
	$\sigma_{\beta(22)}$	3.98	2.13	0.61	15.76	0.30	
	$\sigma_{H(1)}$	0.63	0.61	0.44	0.94	0.00	
	$\sigma_{H(2)}$	0.76	0.70	0.44	1.37	0.00	

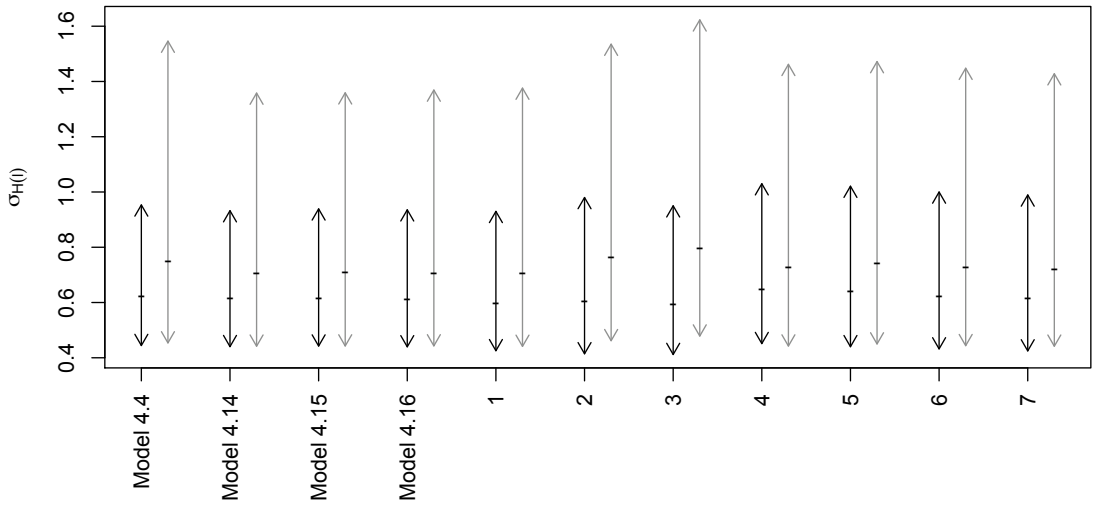


Figure 4.3: Estimates and 95% credible intervals for $\sigma_{H(1)}$ in black and $\sigma_{H(2)}$ in grey for all LHFI-AMBI+ITI models considered in this thesis; numbers 1 to 7 denote the 7 models run with covariates in Table 4.2 in that order; '-' denotes posterior median

narrower (Fig. 4.1, bottom panel in medium grey). Estimates for $\sigma_{H(l)}$ also remained the same. Estimates of the remainder of $\mathbf{\Omega}$ are also quite similar, though accompanying 95% credible intervals are somewhat narrower (Table 4.1); this corresponds to an increase in confidence levels of θ_2 and γ_2 from 90% to 95%. Note that credible intervals for $\sigma_{j(2s)}$ appeared quite similar on a log-scale to those for the two previous baseline models (and thus a figure was not included), although they were in fact slightly narrower as shown in Table 4.1). Since the overlap in credible intervals for $\sigma_{j(2s)}$ within metric group s remained quite substantial as in the previous two baseline models, the model was next refined by reducing $\sigma_{j(2s)}$ to $\sigma_{\beta(2s)}$. Thus, the fourth baseline model adopts the following change:

$$\left[\omega_{j(2s)}|\sigma_{\beta(2s)}\right] \overset{\text{ind}}{\sim} N(0, \sigma^2_{\beta(2s)}) \tag{4.16}$$

Unsurprisingly, 95% credible intervals for $\sigma_{\beta(2s)}$ are much narrower than for $\sigma_{j(2s)}$ in the three previous baseline models (compare Fig. 4.2 to Fig. 4.4). As well, the median estimates for $\sigma_{\beta(21)}$ and $\sigma_{\beta(22)}$ are more distinct from each other

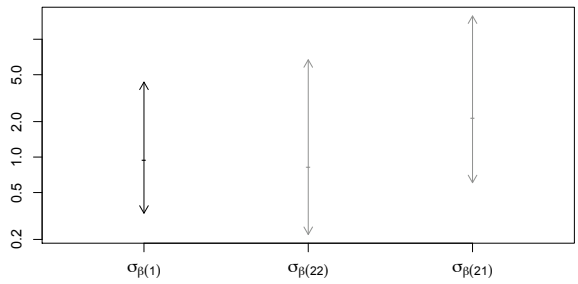


Figure 4.4: Model 4.16 estimates and 95% credible intervals for $\sigma_{\beta(1)}$ and $\sigma_{\beta(2s)}$ on a log-scale; '-' denotes posterior median

compared to $\sigma_{j(21)}$ and $\sigma_{j(22)}$ (Table 4.1). Otherwise, results for this model are similar to the previous baseline model: health estimates LHFI(4.16) and their 95% credible intervals are virtually unchanged (Fig. 4.1, bottom panel in light grey), as were those for $\sigma_{H(l)}$. As well, summary statistics for the remainder of $\mathbf{\Omega}$ are mostly unchanged (Table 4.1).

The four stages of the baseline model are quite similar. MCMC chains of posterior draws mixed fairly well, though not as well as those for LHFI-AMBI baseline models, as evidenced in the greater number of draws generally required for LHFI-AMBI+ITI models. As well, λ_2 and some of the p's experienced some mixing problems (high autocorrelation). The DIC was nearly the same for all models at either 8358 or 8359. Model 4.16 was employed from this point onwards as a base upon which covariates were added.

4.3 Incorporating Covariates

In the process of incorporating covariates into the LHFI-AMBI+ITI model, it was again important to consider the scientists' views regarding which covariates were important for determining health in Richibucto, as discussed in Section 3.3.

4.3.1 Preliminary Analysis

Preliminary analysis examining LHFI(4.16) and the covariates was conducted in much the same way as in Section 3.3 and with the same goal of finding covariates that potentially had relationships with health.

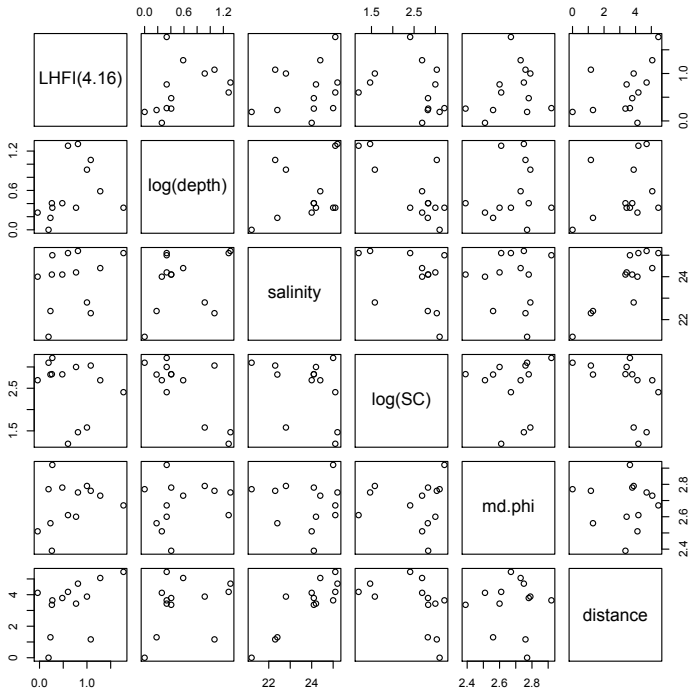


Figure 4.5: Matrix plot using the first temporal block of data of LHFI(4.16) and transformed covariates

A matrix plot of the covariates and LHFI(4.16) indicated that the same transformations should be taken as in Section 3.3. Thus, depth was log-transformed, and SC and OC were log-transformed since they are percentages. A matrix plot of the transformed covariates is provided in Fig. 4.5. (Recall that the trivariate components are highly correlated, and thus we only include log(SC) in this plot.) Comparing this to the corresponding plot for LHFI-AMBI (Fig. 3.6), it is clear that patterns between health and covariates are somewhat similar, though there appears to be a great deal more random scatter for LHFI-AMBI+ITI. As such, any relationships that may exist between health and covariates are no longer obvious. Recall

that there were strong linear relationships between LHFI(3.12) and salinity, distance, and log(SC) (somewhat strong); these no longer are evident for LHFI(4.16). Correlations between LHFI(4.16) and these three covariates are r =0.26, 0.45 and -0.27, respectively, as compared to r =0.64, 0.8 and -0.42 with LHFI(3.12). However, that is not to say that these relationships are not present, only that they are perhaps less easily detectable. A quadratic relationship between health and log(depth) still appears somewhat plausible; however, as explained previously we did not feel it was worthwhile to make use of this uninterpretable relationship.

Fitting linear regressions of LHFI values to various combinations of the five usable covariates provided some guidance as to which combinations of covariates should be implemented in a hierarchical LHFI model. They suggested that all five covariates could be useful in determining health, particularly salinity and distance, and so could the interactions log(depth)×log(SC), log(depth)×salinity, and log(SC)× distance.

4.3.2 Implementing the Models

Seven models with covariates were implemented with MCMC. The motivation behind each choice of model together with the confidence levels for each covariate in the model are described in Table 4.2. This table illustrates the evolution of the final LHFI-AMBI+ITI model with covariates. As with the LHFI-AMBI models, the estimates and 95% credible intervals for health, $\sigma_{\beta(1)}$, $\sigma_{\beta(2s)}$, γ_2 and θ_2, and the DIC values were similar for all the LHFI-AMBI+ITI models with covariates, and also very similar to those of the final baseline Model 4.16. As well, $\sigma_{H(l)}$ varied with the combination of covariates (Fig. 4.3, models with covariates are labelled 1 to 7 according to their order in Table 4.2).

Table 4.2: LHFI-AMBI+ITI models implemented with covariates

	Covariates	Motivation	Confidence levels of slopes associated with covariates
1	distance	Distance was very useful on its own for the LHFI-AMBI model, and in fact was one of the two proposed models for LHFI-AMBI. In the preliminary analysis on the AMBI+ITI baseline, distance still appeared to be a somewhat useful covariate, though its relationship with health is not as pronounced.	distance: 80-85%
2	log(depth), salinity, log(SC), log(depth)× log(SC)	These three covariates and interaction constituted the other proposed model for LHFI-AMBI. Salinity and log(depth)× log(SC) also appeared to be useful in the preliminary analysis on the AMBI+ITI baseline.	log(depth): 20-30%, salinity: 60-70%, log(SC): 70-80%, log(depth)× log(SC): 80-85%
3	log(depth), log(SC), log(depth)× log(SC)	Similar to the previous model but without salinity, as salinity was least significant in that model.	log(depth): 30-40%, log(SC): 70-80%, log(depth)× log(SC): 80-85%
4	log(depth), salinity, log(SC), mdϕ	The preliminary analysis with plots and regressions indicated that all of the five usable covariates could be useful in determining health. However, as distance and salinity are highly correlated, we exclude distance from this model.	log(depth): 30-40%, salinity: 30-40%, log(SC): <20%, mdϕ: 60-70%
5	log(depth), salinity, mdϕ, log(depth)× salinity	An adaptation of the model above: log(SC) was the least significant covariate, so it was removed. Log(depth)×Salinity was added since it appeared promising in the preliminary analysis.	log(depth): 50-60%, salinity: 30-40%, mdϕ: 50-60%, log(depth)× salinity: 50-60%
6	log(depth), log(SC), mdϕ, distance	Similar to the fourth model, but replace salinity with distance.	log(depth): 40-50%, log(SC): <20%, mdϕ: 60-70%, distance: 60-70%

Continued on Next Page...

Table 4.2 – Continued

	Covariates	Motivation	Confidence levels of slopes associated with covariates
7	log(depth), mdϕ, distance, log(depth)× distance	Similar to the two previous models above: log(SC) was removed as it was the least significant covariate in the previous model, and salinity was replace with distance in the interaction.	log(depth): 40-50%, mdϕ: 60-70%, distance: 60-70%, log(depth)× distance: 40-50%

MCMC chains of the posterior draws for all models mixed fairly well, though not as well as the LHFI-AMBI models. However, this is not a concern as they still mixed quite well. Models 5 and 7 required long burn ins (250,000), and models in general required more draws than for LHFI-AMBI model. In addition, autocorrelation was high for some of the p's.

Considering confidence levels as well as estimated precision of health ($\sigma_{H(l)}$), Models 1-3 in Table 4.2 seem to be the best of the models with covariates. Models 1-2 are the same the two proposed for LHFI-AMBI and Model 3 is simply a modification of Model 2. Denote these models respectively as Model 4.16a, 4.16b, and 4.16c. Some summary statistics for α_0 and $\boldsymbol{\alpha}$ for these models are in Table 4.3.

It is important to note that confidence levels of covariates were generally much lower; none of them came close to meeting the commonly accepted 95% level, but this result was not entirely unexpected given the results of the preliminary analysis. Two plausible explanations of the LHFI-AMBI+ITI model's inability to determine relationships between health and the covariates come to mind:

1. The LHFI construct is appropriate for describing health using AMBI and ITI metrics and the available covariates, but the ITI data are too noisy for the LHFI model to detect any patterns in the covariates.

2. AMBI and ITI are both good for describing health, and the data are not noisy, but our modelling technique is inappropriate for combining them with the available covariates.

Table 4.3: Summary Statistics for α_0 and $\boldsymbol{\alpha}$ for Models 4.16a, 4.16b, and 4.16c

	Covariate	Mean	99% Credible Interval	95% Credible Interval	90% Credible Interval	MC Error	# Draws
				Model 4.16a			
α_0	intercept	0.64	(-1.14, 2.44)	(-0.56, 1.88)	(-0.33, 1.63)	0.01	40000
α_5	distance	0.14	(-0.14, 0.43)	(-0.07, 0.35)	(-0.03, 0.31)	0.00	
				Model 4.16b			
α_0	intercept	0.78	(-1.03, 2.57)	(-0.45, 2.01)	(-0.22, 1.77)	0.01	60000
α_1	log(depth)	0.14	(-1.10, 1.38)	(-0.76, 1.05)	(-0.60, 0.88)	0.00	
α_2	salinity	0.12	(-0.24, 0.50)	(-0.14, 0.40)	(-0.10, 0.35)	0.00	
α_3	log(SC)	-0.52	(-1.88, 0.84)	(-1.50, 0.46)	(-1.32, 0.29)	0.01	
α^*_{13}	log(depth)× log(SC)	1.18	(-1.24, 3.49)	(-0.54, 2.85)	(-0.23, 2.55)	0.02	
				Model 4.16c			
α_0	intercept	0.79	(-1.02, 2.61)	(-0.44, 2.03)	(-0.20, 1.79)	0.01	40000
α_1	log(depth)	0.21	(-1.00, 1.42)	(-0.67, 1.10)	(-0.52, 0.94)	0.00	
α_3	log(SC)	-0.55	(-1.87, 0.82)	(-1.52, 0.45)	(-1.36, 0.26)	0.01	
α^*_{13}	log(depth)× log(SC)	1.11	(-1.27, 3.39)	(-0.64, 2.79)	(-0.34, 2.51)	0.03	

In light of the fact that field data have large measurement error by nature, the first explanation is certainly conceivable. If this is the case, perhaps the model cannot be fully remedied, and the models we have run are still useful. Then, these low confidence levels should not necessarily lead us to immediately dismiss our models as being unworthy.

The second explanation bears further investigation. We considered that AMBI and ITI might be too different to be combined into a single model, but ultimately rejected this viewpoint. The biologists who sampled the Richibucto data believe that AMBI and ITI can be modelled together. Additionally, there are no theoretical statistical objections to combining two different indices in a single LHFI model. Thus, while we believe that an AMBI and ITI model can work, some components of the model could perhaps be improved upon. For example, introducing a proper covariance matrix for the metric effects could improve the statistical significance of relationships between covariates and health. Recall that we assumed metric effects to be independent for simplicity as a preliminary tool, particularly since

implementing such a structured covariance matrix would be difficult in practice. Or, since AMBI and ITI focus on different aspects of health, it could be that the covariates measured are relevant to AMBI but less so to ITI. In consultation with biologists, we might find more appropriate covariates to be regressed upon both the AMBI and ITI metrics. However, further refinement of our current LHFI models in either manner is deferred as future work; please see Chapter 5.

4.4 Comparing LHFI-AMBI, LHFI-AMBI+ITI, AMBI and ITI

A matrix plot of AMBI, ITI, LHFI-AMBI (i.e. LHFI(3.12)), and LHFI-AMBI+ITI (i.e. LHFI(4.16)) is provided to illustrate the relationships between the two sets of model-based health indices and AMBI and ITI (Fig. 4.6).

LHFI-AMBI and LHFI-AMBI+ITI are strongly positively correlated with each other ($r = 0.85$), and there is a clear linear trend between the two. Both LHFI models are strongly negatively correlated with AMBI itself ($r = -0.81$ for both) though it appears that adding ITI degrades the relationship between LHFI and AMBI, as the linear relationship with LHFI-AMBI+ITI contains more scatter. While ITI is essentially uncorrelated with LHFI-AMBI ($r = 0.01$), it is somewhat correlated with LHFI-AMBI+ITI ($r = 0.44$), and an upward trend between ITI and LHFI-AMBI+ITI is visible. Thus, it appears that the LHFI-AMBI+ITI model may have managed to incorporate both aspects of health from ITI and AMBI.

In this chapter and the previous, we discussed in detail and at great length the design and implementation of LHFI models for Richibucto using the AMBI and ITI metrics. The next and last chapter presents an overall summary of the findings of this thesis, focussing on general conclusions for the LHFI for estuarine systems, particularly for Richibucto, and also discusses future work.

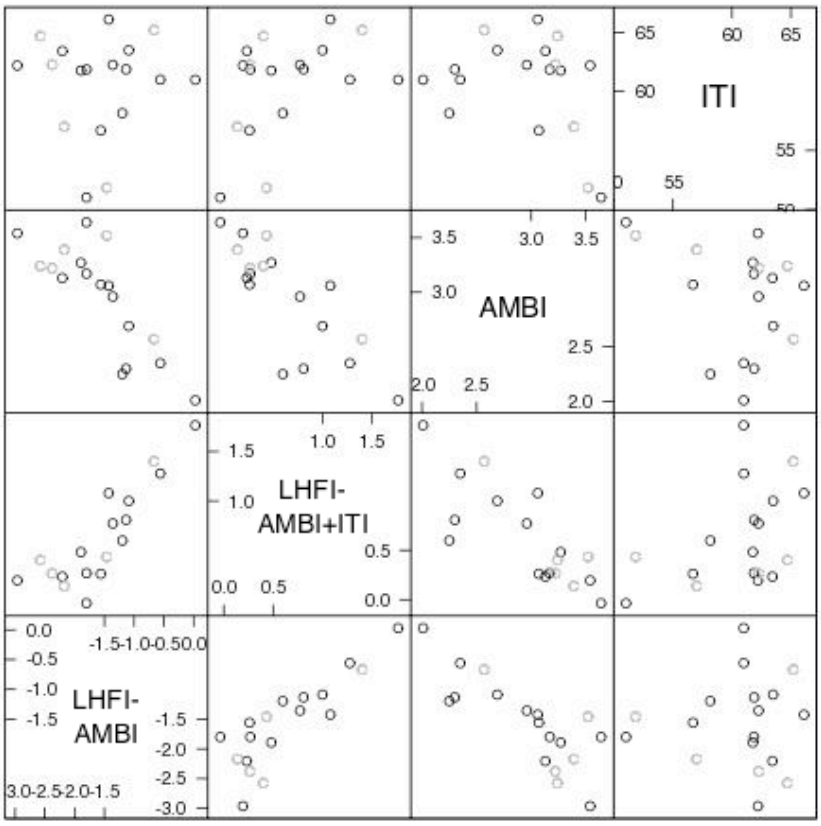


Figure 4.6: Matrix plot of LHFI(3.12), LHFI(4.16), AMBI and ITI; a black '∘' denotes a point in the first block and a grey '∘' denotes a point in the second block

Chapter 5

Conclusion

5.1 Restatement of Objectives

The objectives of this thesis were to investigate whether the LHFI is applicable to estuarine systems, since it has to date only been applied to freshwater ecosystems; to determine the form of the LHFI model if it is applicable; and to assess the health of the previously unassessed Richibucto system. Accordingly, this thesis described the method of and motivation behind constructing a latent health factor model for measuring ecosystem health, and applied this methodology to the estuarine system in Richibucto, NB using metrics from AMBI and ITI, which are two indices commonly used by biologists for measuring different aspects of estuarine health.

Using standard statistical modelling methods, the methodology for constructing a latent health factor model is in many ways an improvement upon the conventional methods used by biologists for estimating health. Standard statistical practices allow the LHFI to avoid a great deal of the arbitrariness typically involved in constructing conventional indices, and also allow for straightforward model inference and prediction. As well, the added statistical features of the LHFI do not necessarily detract from its biological worthiness, since the LHFI involves the same data as the indices and also can require input from scientists in model specification.

5.2 Conclusions

The models resulting from applying this methodology to Richibucto performed similarly to each other. Both LHFI-AMBI and LHFI-AMBI+ITI models were able to make distinctions between health levels at different sites, and also provided somewhat similar site rankings. While credible intervals for health were narrower for LHFI-AMBI+ITI than for LHFI-AMBI, the former did not appear to have a greater ability to distinguish between sites. Thus it was unclear whether an LHFI based upon both AMBI and ITI is superior to an LHFI based upon only AMBI. In addition, including covariates in both models through $f()$ appeared to be advantageous since estimates of the precision of health $\sigma_{H(l)}$ were lower for certain combinations of covariates, although estimates and credible intervals of health itself were not affected. The best combinations of covariates for LHFI-AMBI and LHFI-AMBI+ITI were quite similar: for LHFI-AMBI, we found two models involving covariates that we felt would be of good use to ecologists: one containing distance, and the other containing log(depth), salinity, log(SC), log(depth)×log(SC). For LHFI-AMBI+ITI, the three best models were the same two as for LHFI-AMBI and a model containing log(depth), log(SC), log(depth)×log(SC).

Overall, we can conclude that the LHFI modelling process is indeed applicable to estuarine systems and that LHFI-AMBI and LHFI-AMBI+ITI are both appropriate models for Richibucto which are useful for estimating ecosystem health. Since LHFI-AMBI and LHFI-AMBI+ITI are both highly correlated with AMBI and LHFI-AMBI+ITI is somewhat correlated with ITI, it appears that both models are still effective in comparison with the original indices, although we cannot state whether LHFI is better at estimating health than AMBI and ITI on the strength of this. Overall, the thesis objectives have largely been achieved.

5.3 Suggestions for Future Work

Some more specific results of the analyses follow, leading into suggestions for further work. Confidence levels for coefficients for the LHFI-AMBI+ITI models were quite low (60-85% for the "best" models), especially when comparing them to levels for LHFI-AMBI. This indicates that the extra data from ITI metrics helped to improve the precision of health estimates but weakened the overall relationship between health and covariates. Since confidence levels for the LHFI-AMBI+ITI models were low, we suggested several options for improving the models in Chapter 4, with the hope that these changes could clarify relationships between health and covariates. Firstly, the model could potentially be improved by including a structured covariance matrix for the metric effects, instead of independent metric effects as we assumed to keep the preliminary model uncomplicated. However, this option may not be viable without the addition of more data. As well, implementing a structured covariance matrix could prove quite difficult in practice. An inverse Wishart prior is commonly used for unstructured covariance matrices, but imposing a structure upon the covariance matrix rules out this option. Determining a prior that will retain the structure of the posterior covariance matrix and that is also positive definite seems to be a current popular topic of research, and such problems appear to be addressed on a case-by-case basis. Another suggestion for improving the LHFI-AMBI+ITI model stems from the possible explanation that the covariates measured are relevant to AMBI but not necessarily ITI. Consultation with biologists might lead to finding additional covariates that are more appropriate for ITI, which could then be modelled along with the original AMBI covariates.

Although the latent health factor model has already provided useful information on the health of Richibucto, it can be further developed in other ways as well. In the preceding paragraph, we discussed introducing a structured covariance matrix to metric effects for LHFI-AMBI+ITI as a potential means of improving relationships with covariates; even though the LHFI-AMBI model is not prone to the same issues, we could implement a similar change to it, with the hope of improving an already

promising model[1].

5.3.1 Introducing Additional Regression Layers

As well, since scientists believe in strong associations between some of the covariates, additional levels of regression could be introduced. Given the results of the analysis on covariates, we could regress salinity upon distance, or regress the variables of the trivariate (silt clay, sorting, and organic content) upon each other. Recognizing these relationships between the covariates in the LHFI model could perhaps improve parameter estimation, and the scientific meaning of the resulting index.

As an illustration of how such a change would affect the model, consider introducing a regression of salinity upon distance. This would change the relationships between health and covariates from as visualized in Fig. 1.1(b) to the more complex schematic visualized in Fig. 5.1. The equation for health and the form of $f()$ would accordingly be altered, to perhaps the following:

$$H_{i(l)} = \alpha_0 + f(\boldsymbol{\alpha}, \boldsymbol{x}^*_{i(l)}) + \lambda_l + \varepsilon_{i(l)} \tag{5.1}$$

$$x_{1i(l)} = g(\boldsymbol{\alpha}^*, x_{5i(l)}) + \delta_{1i(l)} \tag{5.2}$$

where $\boldsymbol{x}^*_{i(l)}$ represents $\boldsymbol{x}_{i(l)}$ without salinity, $f()$ is an appropriately chosen regression of all covariates but salinity with corresponding intercept and regression coefficients α_0 and $\boldsymbol{\alpha}^*$, $g()$ is an appropriately chosen regression of salinity ($x_{1i(l)}$) on distance ($x_{5i(l)}$) with corresponding regression coefficents $\boldsymbol{\alpha}^*$, and $\delta_{1i(l)}$'s are independent and identically distributed errors with mean zero[2].

[1]Note that Chiu et al. (2008) fit LHFI models with unstructured covariance matrices, but these were poorly estimated, quite likely because there were few sites (18) and many parameters.

[2]Note that in (5.2) there is no intercept term since inestimability issues could arise between an intercept here and α_0.

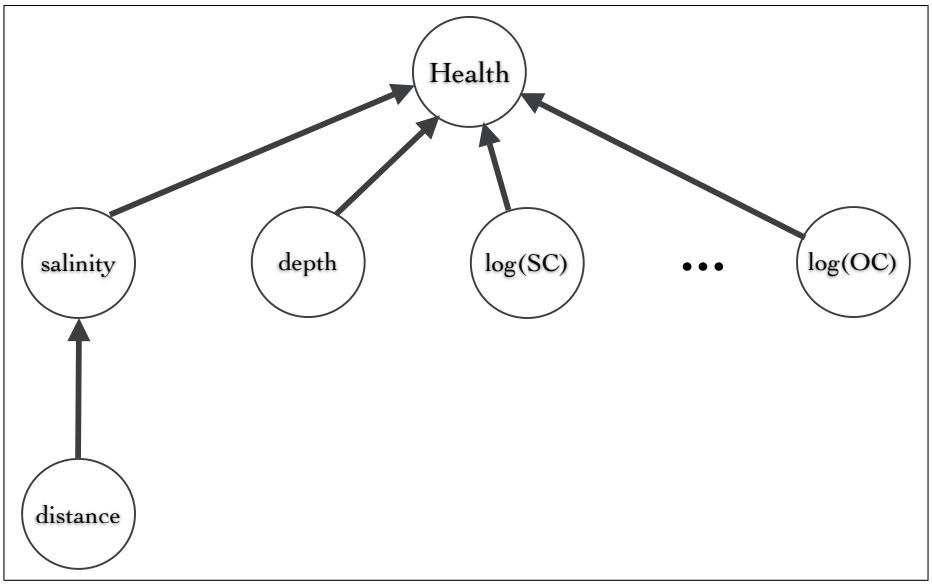


Figure 5.1: Graphical depiction of relationships between health and covariates if an additional regression of salinity upon distance is introduced, as described in equations (5.1)-(5.2)

5.4 Prediction of Health for a New Site

Finally, we return to the discussion of site prediction under the LHFI framework which was first mentioned in Chapter 2. As stated previously, the posterior distribution provides a simple means of predicting a new site's health once a model has been specified and given that measurements of covariates are available. This capability is one of the major advantages of the LHFI over conventional indices, as gathering benthic data is time-consuming and expensive. However, it was not demonstrated with the Richibucto data, as benthic data had been gathered for all sites under consideration. However, we outline the method of prediction here which was originally described in Chiu et al. (2008) should the reader consider doing so in practice.

Let "$*$" denote a new site. Then, the predicted health of the new site is calculated as

$$\widehat{H}^* = E(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*) \tag{5.3}$$

A single draw of H^* from the predictive posterior $P(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*)$ can be simulated as follows: generate a Monte Carlo draw of $\boldsymbol{\Omega}$ using (2.7), and substitute the relevant parts of this draw along with $\boldsymbol{x}^*$ into the equation for health in (2.2). The predicted health $\widehat{H}^*$ can be estimated as the mean of a collection of draws simulated in this manner. Credible intervals can also be estimated using quantiles of the same collection of draws.

In summary, this thesis focussed on constructing latent health factor models for estuarine systems, specifically for the Richibucto estuarine system using metrics from the two conventional indices AMBI and ITI. Meaningful LHFI models were found using the AMBI metrics alone, and using the AMBI and ITI metrics in conjunction. These models were able to distinguish health levels of different sites at Richibucto, and in fact both sets of models produced somewhat similar rankings of sites by health level. Several environmental covariates were found to have a significant impact on health as well, though such relationships were less clear under the LHFI-AMBI+ITI model than under the LHFI-AMBI model. This matter led to some suggestions for future development of the LHFI model for estuarine systems, which were also discussed in detail.

APPENDICES

Appendix A

The Richibucto Data

Table A.1: Replicates and Total Organism Counts

Site	Number of Replicates	Total Organism Counts		
		First Replicate	*Second Replicate*	*Third Replicate*
1	3	136	194	140
2	3	389	449	337
3	2	167	183	n/a
4	3	230	208	205
5	3	198	277	96
6	2	269	336	n/a
7	2	264	391	n/a
8	2	285	357	n/a
9	3	224	210	238
10	3	189	125	103
11	3	223	212	170
12	2	467	214	n/a
13	3	324	337	273
14	3	216	173	193
15	3	454	249	551
16	3	456	430	532
17	3	195	245	288
18	3	89	128	94

Table A.2: AMBI Group Counts

Site	Group Counts														
	First Replicate					*Second replicate*					*Third replicate*				
	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
1	2	11	55	61	7	0	17	104	65	8	1	13	69	53	4
2	3	121	109	152	4	8	139	146	156	0	8	108	81	140	0
3	20	9	76	56	6	17	18	88	57	3			*n/a*		
4	9	29	112	79	1	8	23	118	57	2	8	25	96	72	4
5	9	34	27	53	75	8	96	47	86	40	4	33	29	26	4
6	8	39	134	85	3	3	31	213	81	8			*n/a*		
7	15	121	77	50	1	37	163	112	74	5			*n/a*		
8	13	39	125	81	27	32	37	141	98	49			*n/a*		
9	23	63	50	78	10	23	62	62	63	0	16	45	69	102	6
10	19	30	76	58	6	9	21	25	61	9	10	9	23	57	4
11	21	46	68	83	5	17	49	57	77	12	26	23	50	62	9
12	53	116	52	246	0	79	30	82	22	1			*n/a*		
13	31	60	59	126	48	37	65	67	83	85	16	27	58	105	67
14	20	28	82	64	22	20	23	66	54	10	22	34	54	72	11
15	120	45	223	66	0	42	32	142	33	0	146	94	254	55	2
16	91	73	243	46	3	74	70	220	65	1	109	146	217	60	0
17	35	70	36	43	11	56	92	56	38	3	44	96	72	45	31
18	20	25	13	24	7	42	58	4	19	5	25	42	8	11	8

Table A.3: ITI Group Counts

Site	Group Counts											
	First Replicate				*Second replicate*				*Third replicate*			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
1	0	121	8	7	0	177	9	8	0	128	8	4
2	0	378	7	4	3	441	5	0	2	330	5	0
3	19	127	15	6	15	130	35	3		*n/a*		
4	5	191	33	1	1	184	21	2	1	176	24	4
5	0	117	6	75	0	230	7	40	0	89	3	4
6	6	244	16	3	5	314	9	8		*n/a*		
7	1	236	26	1	2	355	29	5		*n/a*		
8	7	225	26	27	16	261	31	49		*n/a*		
9	14	174	26	10	1	185	24	0	7	197	28	6
10	2	161	20	6	8	102	6	9	3	86	10	4
11	6	155	57	5	3	158	39	12	3	113	45	9
12	9	431	27	0	11	146	56	1		*n/a*		
13	32	194	49	49	24	195	33	85	18	160	28	67
14	34	123	37	22	32	108	23	10	24	130	28	11
15	16	325	60	53	27	174	27	21	46	368	100	37
16	48	331	61	16	39	317	44	30	43	386	69	34
17	7	156	21	11	19	191	31	4	14	224	18	32
18	2	72	7	8	7	105	11	5	4	76	6	8

Table A.4: Environmental Covariates

Site	Depth (m)	Temp (C°)	Salinity (ppt)	Silt-clay (%)	Mdϕ	Sorting	Organic Content (%)	Distance (km)
1	1	15.4	21.2	22.18	2.77	1.13	7.94	0
2	2.9	15.1	22.3	20.76	2.76	1.22	11.66	1.164
3	1.2	15.5	22.4	16.81	2.56	1.23	6.29	1.298
4	1.3	12.7	21.3	13.25	2.53	1.44	5.22	1.731
5	4.5	12.5	23	23.7	2.62	1.41	12.72	2.179
6	1.8	12.5	21.9	13.71	2.39	1.16	5.55	1.686
7	2.9	12.2	23.5	23.21	3.03	1.25	10.09	2.463
8	1.8	12.4	23.6	18.77	2.84	1.12	8.47	2.970
9	1.4	14.7	24.2	20.07	2.6	1.36	6.74	3.433
10	1.5	14.7	24.1	16.89	2.78	1.14	4.65	3.790
11	1.5	15	24.1	16.96	2.39	1.27	7.59	3.358
12	2.5	14.1	22.8	4.85	2.79	0.76	1.92	3.880
13	1.3	14.2	24	14.72	2.51	1.15	6.62	4.119
14	1.4	14.7	25	24.83	2.92	1.25	6.41	3.642
15	3.6	14.7	25.1	3.31	2.61	0.69	1.85	4.179
16	3.7	15.2	25.2	4.34	2.75	0.71	1.92	4.701
17	1.8	14.6	24.4	14.7	2.73	1.11	7.22	5.060
18	1.4	14.7	25.1	11.11	2.67	1.04	6.25	5.448

Appendix B

Convergence Checks

This appendix discusses steps taken with the LHFI model to ensure convergence inasmuch as possible. Many of the following methods are popular tools and as such are facilitated by or built into OpenBUGS.

1. For every LHFI model, multiple chains were run concurrently, each with a different set of initial values. This often facilitates the exploration of a greater part of the parameter space, which lessens the chance of converging to a local solution instead of the target distribution. In addition, the two chains can be compared to each other; if the chains have converged, they should be quite similar in behaviour, or provide estimates that differ at the most by a slight constant offset. This situation occurred when the LHFI was fit to the Puget Sound Lowlands data (Chiu et al., 2008, 2007): the relative rankings of health remained the same, though the absolute estimates of health differed by a small offset between two chains.

2. A visual inspection of the complete trace plot of all variables, that is, a plot of the sample values versus the iteration number, is often used to give one a general idea of if and when convergence has been reached, and the length of an appropriate burn-in. An option to view the trace is available in OpenBUGS.

3. An inspection of summary statistics, e.g. mean, median, quartiles, etc., for each variable for each chain is another informal check of convergence, and also available in OpenBUGS. One can compare the statistics for individual chains; if the chains have converged, the means of individual chains should be similar. (Comparing medians may be more appropriate for variables whose distributions are bound at one end and tend to produce extreme values.) It is also a good idea to look at the estimate of the Monte Carlo standard error of the mean, MC error for short. This estimate is an indication of how well the simulation has run. If the chains have converged, the MC error should be similar between multiple chains. If the MC error is different among different chains, it means that the simulation is different to chain, and could indicate that the chains have not been run long enough.

4. A modified version of the Gelman-Rubin statistic is another commonly used diagnostic (Spiegelhalter et al., 2007). This statistic determines whether inferences from m chains are similar enough be believed to have converged by comparing the pooled variance of all m chains to the within-chain variance of individual chains. OpenBUGS plots the width of the central 80% interval of the pooled runs, the average width of the 80% intervals stabilize within the individual runs, and their ratio R. If the chains have converged, the pooled variance and within-chain variance should both stabilize, and the ratio R should approach 1.

5. Examining smoothed kernel density estimate plots (also provided by OpenBUGS) for continuous variables can sometimes reveal non-convergence through multi-modal posterior densities, or other irregular shapes.

6. Autocorrelation of Markov chains is often of interest, though high autocorrelation is not necessarily an indication of non-convergence; it only indicates that a model may be slow to converge. High autocorrelation can also make it difficult to determine whether a chain has converged. OpenBUGS provides autocorrelation plots up to lag 100. One way to lower the autocorrelation

without losing much information is to take only every kth observation for analysis, as several highly correlated values do not provide much more information than one value alone, and this will decrease the time and storage space required as well.

Appendix C

Hierarchical Centring

With complex hierarchical models involving many parameters, such as our LHFI models in the Bayesian framework, there frequently exist strong correlations between parameters, which make MCMC methods slow to converge. MCMC methods that update the parameter space one parameter at a time, such as the Gibbs Sampler, are often the most practical option for models with high dimensional parameter spaces (Dey et al., 2000). With these methods, high correlation between parameters implies that each successive step of the MCMC is unlikely to be far from the previous, and hence a long time is required to explore the parameter space fully. The parameterisation of a complex model therefore can greatly affect the correlation between parameters and accordingly the convergence time.

Gelfand et al. (1995) propose a technique called hierarchical centring to reparameterise models and mitigate mixing problems. This technique roughly entails splitting the model into as many levels as possible where each level has as few parameters as possible. As well, our experience is such that linearity of parameter at each level of the hierarchy would be ideal; such a parameter can thus be typically modelled with a normal distribution or an approximately normal distribution. Gelfand et al. (1995) argue that hierarchical centring will often result in "better behaviour of the Markov chain Monte Carlo algorithm". Hierarchical centring does not necessarily provide an optimal parameterisation, and there are no set rules for

carrying out this technique. Often, experimenting with several parameterisations is necessary before a useful one is found (Dey et al., 2000).

As a simple example of hierarchical centring, consider the model $Y_{ij} = \mu+\alpha_i+\varepsilon_{ij}$ where α_i and ε_{ij} are random effects such that $\alpha_i \overset{\text{iid}}{\sim} N(0, \sigma^2_\alpha)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. A fully hierarchically centred parameterisation of this model can be constructed by defining $\mu_i = \mu + \alpha_i$, i.e. $Y_{ij} = \mu_i + \varepsilon_{ij}$ where $\mu_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2_\alpha)$. The centring of the model is evident in diagram representations of these parameterisations (Fig. C.1).

The LHFI-AMBI models were similarly hierarchically centred, since Markov chains from the original parameterisations in Chapter 3 mixed extremely slowly. Several partial hierarchical centring parameterisations were attempted as full centring was difficult with such a complex model; the final parameterisation chosen was the one that mixed the quickest. For Model 3.8 (the version with covariates) this parameterisation was achieved by defining new parameters:

$$\begin{cases} \widetilde{H}_{i(l)} = H_{i(l)} - \alpha_0 \\ \psi_s = \alpha_0 + \theta_s \\ b_{j(s)} = \beta_{j(s)} + \psi_s \end{cases} \tag{C.1}$$

The parameterisation was thus:

$$\begin{aligned} \nu_{ijls} &= \widetilde{H}_{i(l)} + b_{j(s)} \\ \left[\widetilde{H}_{i(l)} \,|\lambda_l, \boldsymbol{\alpha}, \boldsymbol{x_{(l)}}, \sigma^2_H\right] &\overset{\text{ind}}{\sim} N(f(\boldsymbol{\alpha}, \boldsymbol{x_{(l)}}) + \lambda_l, \sigma^2_{H(l)}) \\ \left[b_{j(s)}|\psi_s, \sigma^2_{j(s)}\right] &\overset{\text{ind}}{\sim} N(\psi_s, \sigma^2_{j(s)}) \end{aligned} \tag{C.2}$$

Note that α_0 is no longer relevant by employing the above definition of $\widetilde{H}_{i(l)}$. Diagram representations of the original and centred parameterisations are in Fig. C.2, and reveal the centredness of the new parameterisation. The other LHFI-AMBI models were centred in the same manner.

The LHFI-AMBI+ITI models were also partially hierarchically centred in much the same way. For Model 4.4, we defined:

$$\begin{cases} \widetilde{H}_{i(l)} = H_{i(l)} - \alpha_0 \\ \phi_{ms} = \alpha_0 + \gamma_m + \theta_s + \xi_{ms} \\ b_{j(ms)} = \beta_{j(ms)} + \alpha_0 = \phi_{ms} + \omega_{j(ms)} \end{cases} \tag{C.3}$$

The centred parameterisation was thus:

$$\nu_{ijlms} = \widetilde{H}_{i(l)} + b_{j(ms)} \tag{C.4}$$

$$\left[\widetilde{H}_{i(l)} \,|\boldsymbol{\alpha}, \boldsymbol{x_{(l)}}, \lambda_l, \sigma_H^2\right] \overset{\text{ind}}{\sim} N(f(\boldsymbol{\alpha}, \boldsymbol{x_{(l)}}) + \lambda_l, \sigma^2_{H(l)}) \tag{C.5}$$

$$\left[b_{j(1s)}|\phi_{1s}, \sigma^2_{\beta(1)}\right] \overset{\text{ind}}{\sim} N(\phi_{1s}, \sigma^2_{\beta(1)}) \tag{C.6}$$

$$\left[b_{j(2s)}|\phi_{2s}, \sigma^2_{j(2s)}\right] \overset{\text{ind}}{\sim} N(\phi_{2s}, \sigma^2_{j(2s)}) \tag{C.7}$$

Diagram representations of the original and centred parameterisations are in Fig. C.3.

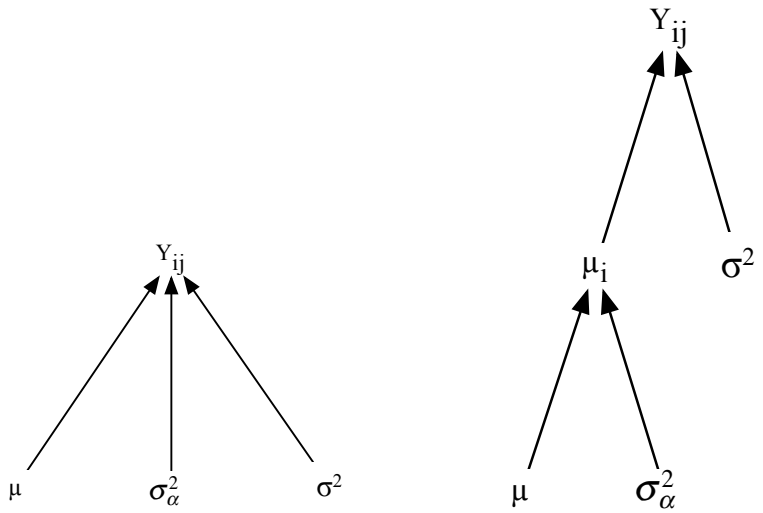
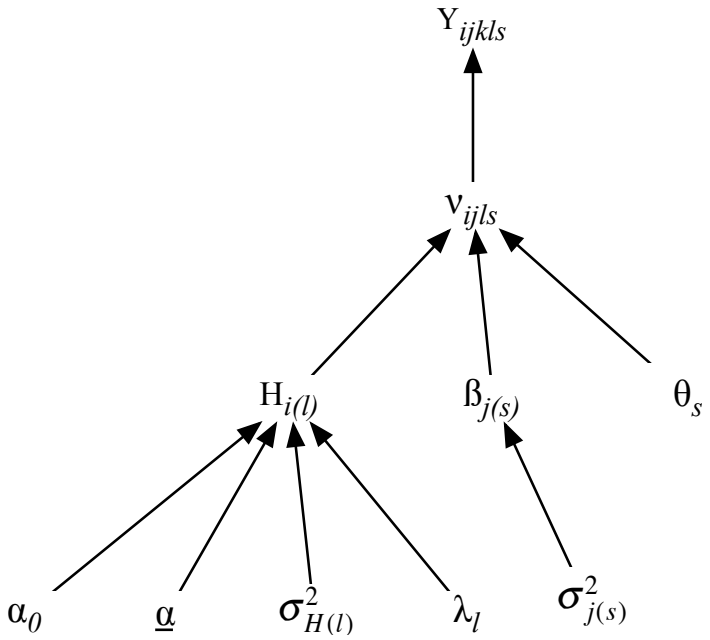
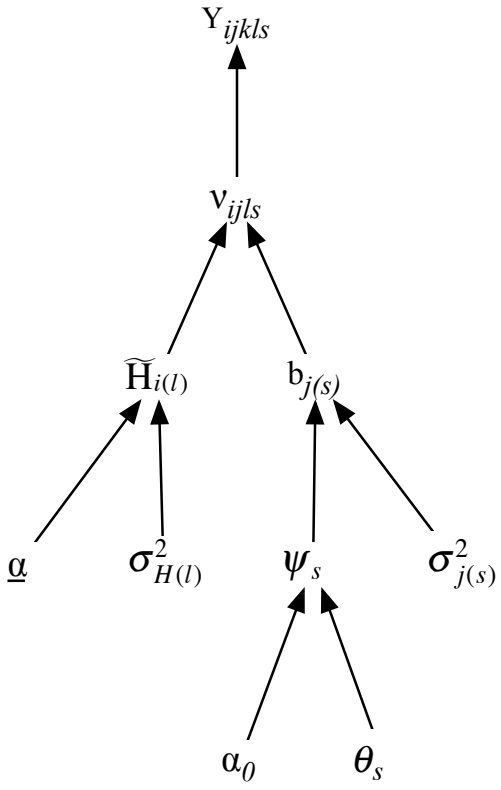


Figure C.1: Diagram representations of (a) original and (b) hierarchically centred parameterisations of $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

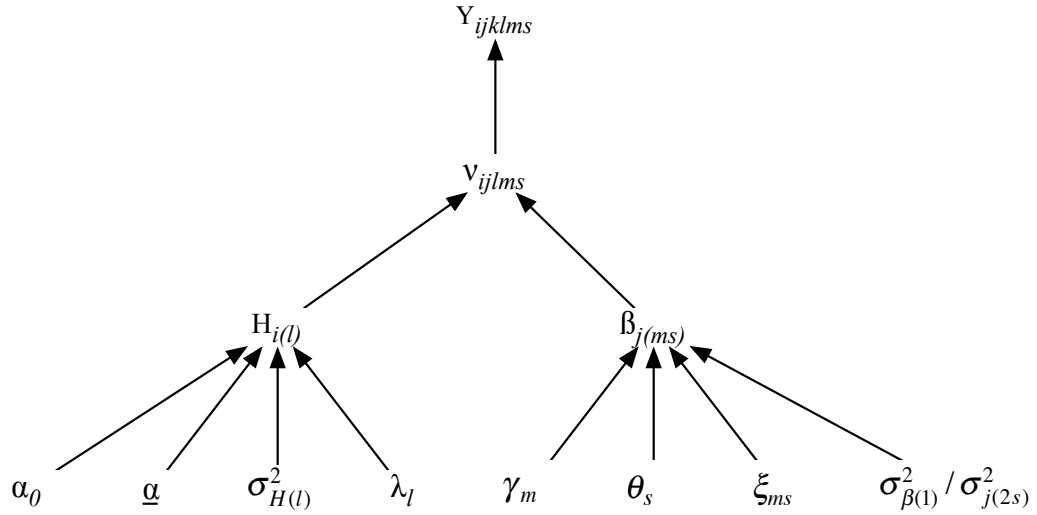


(a)

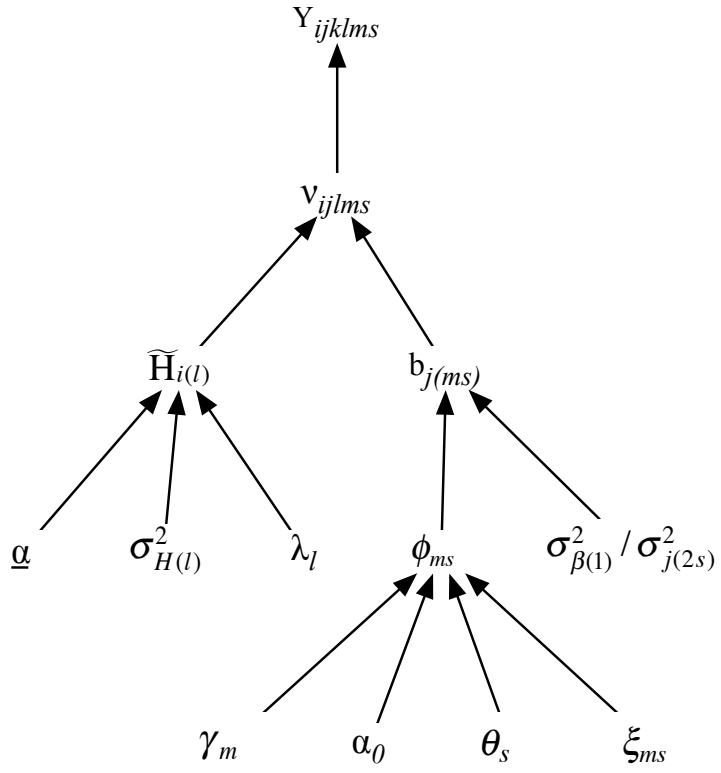


(b)

Figure C.2: Diagram representations of (a) original and (b) hierarchically centred parameterisations of Model 3.8



(a)



(b)

Figure C.3: Diagram representations of (a) original and (b) hierarchically centred parameterisations of Model 4.4

References

A. Borja, J. F. and Erez, V. P. (2000). A marine biotic index to establish the ecological quality of soft-bottom benthos within european estuarine and coastal environments. *Marine Pollution Bulletin*, 40(12):1100–1114. 2, 16, 42

Bilyard, G. R. (1987). The value of benthic infauna in marine pollution monitoring studies. *Marine Pollution Bulletin*, 18(11):581–585. 1

Chiu, G. S. (2008). On identiability of covariance components in hierarchical generalized analysis of covariance models. submitted, available at `http://www.stats.uwaterloo.ca/stats_navigation/techreports/08WorkingPapers/08-09.pdf`. 21

Chiu, G. S., Guttorp, P., Khan, S. A., Liang, J., and Westveld, A. H. (2007). An ecological latent health factor index via a random-effects model for taxa richness and composition. Technical report, Department of Statistics and Actuarial Science, University of Waterloo. Working Paper Series No. 2006-02, available at `http://www.stats.uwaterloo.ca/stats_navigation/techreports/06WorkingPapers/2006-02.pdf`. 73

Chiu, G. S., Guttorp, P., Westveld, A. H., Khan, S. A., and Liang, J. (2008). A latent health factor index modelling approach via generalized linear mixed models, with application to ecological health assessment. preprint, available at `http://www.stats.uwaterloo.ca/stats_navigation/techreports/08WorkingPapers/08-08.pdf`. 2, 3, 4, 6, 8, 9, 10, 11, 12, 41, 66, 67, 73

Cromey, C. J., Nickell, T. D., and Black, K. D. (2002). Depomodmodelling the deposition and biological effects of waste solids from marine cage farms. *Aquaculture*, 214(1-4):211–239. 44

Dauer, D. M. (1993). Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin*, 16(5):249–257. 1

Dey, D., Ghosh, S. K., and Mallick, B. K., editors (2000). *Generalized Linear Models: A Bayesian Perspective*, chapter 2. CRC Press. 76, 77

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488. 76

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). Efficient parametrisations for generalized linear mixed models. *Bayesian Statistics*, 5:165–180.

Gentle, J. E. (2002). *Elements of Computational Statistics*. Springer, New York. 25

Hogg, R. V., McKean, J. W., and Craig, A. T. (2005). *Introduction to Mathematical Statistics*. Pearson Education, Upper Saddle River, sixth edition. 12, 26

Lu, L., Grant, J., and Barrell, J. (2008). Macrofaunal spatial patterns in relationship to environmental variables in the richibucto estuary, new brunswick, canada. *Estuaries and Coasts*, 31:994–1005. 32

Montgomery, D. C. (1997). *Design and Analysis of Experiments*. Wiley, New York, fifth edition.

National Oceanic and Atmospheric Administration Coastal Services Center, U. S. D. o. C. (2009, accessed). Benthic habitat mapping - grab sampling. available at `http://www.csc.noaa.gov/benthic/mapping/techniques/sensors/grab.htm`. 16

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007). *OpenBUGS User Manual*, 3.0.2 edition. available at `http://www.mrc-bsu.cam.ac.uk/bugs`. 74

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*, 64(4):583–639. 30

Word, J. Q. (1978). *The Infaunal Trophic Index*, pages 19–39. Coastal Water Research Project, California, U.S.A. 2