Clustering and Classification in Fisheries

by

Yuki Fujita

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Statistics.

Victoria University of Wellington
2016

Abstract

This goal of this research is to investigate associations between presences of fish species, space, and time in a selected set of areas in New Zealand waters. In particular we use fish abundance indices on the Chatham Rise from scientific surveys in 2002, 2011, 2012, and 2013. The data are collected in annual bottom trawl surveys carried out by the National Institute of Water and Atmospheric Research (NIWA). This research applies clustering via finite mixture models that gives a likelihood-based foundation for the analysis. We use the methods developed by Pledger and Arnold (2014) to cluster species into common groups, conditional on the measured covariates (body size, depth, and water temperature). The project for the first time applies these methods incorporating covariates, and we use simple binary presence/absence data rather than abundances. The models are fitted using the Expectation-Maximization (EM) algorithm. The performance of the models is evaluated by a simulation study. We discuss the advantages and the disadvantages of the EM algorithm. We then introduce a newly developed function `clustglm` (Pledger et al., 2015) in **R**, which implements this clustering methodology, and perform our analysis using this function on the real-life presence/absence data. The results are analysed and interpreted from a biological point of view. We present a variety of visualisations of the models to assist in their interpretation. We found that depth is the most important factor to explain the data.

Acknowledgments

Firstly, I would like to thank my primary supervisor Associate professor Richard Arnold for the continuous support of my study, his patience, encouragement, and immense ideas for the research. His guidance helped me in all the time of my research and writing of this thesis. I could not have asked for a better supervisor.

I would also thank my secondary supervisor Dr. Daniel Fernádez for his passion and support. I am also grateful to the Marsden team members, Professor Shirley Pledger, Associate professor Ivy Ming, and Roy Costilla. I am thankful to them for sharing expertise, and discussions over the research topics.

I would like to show my deep appreciation to Associate professor Matthew Dunn. His wide knowledge in fishery and advice has helped with this research.

To Yuka, Sonia, Chiharu, and especially Tony - thank you for keeping me smiling and providing me your shoulders to cry on throughout this year. Your company is always warm and relaxing, and you were always welcoming whenever I needed comfort.

Special thanks to Mary, John, and David Johnson. You have been great part of my life here in New Zealand. Without your support, I would not

be studying in New Zealand. Thank you for sharing both good and bad times, and a big feed.

Finally, many thanks to my mum and dad for understanding me and supporting me in no matter what I decided to do. Although I am away from my home country, I always feel your support.

Contents

Chapter 1

Introduction

1.1 Aim of Research

In this research, we aim to investigate geographical and interspecies associations among fish species. We use likelihood-based cluster analysis on trawl survey data in New Zealand waters.

Ecological community datasets are often stored as a high dimension matrix of non-normal data. For example, species by site data recorded as a binary data (presence/absence), or counts (number of individuals of each species), or ordinal rankings (Braun-Banquet scale, Wikum and Shanholtzer, 1978). Detecting any patterns in high dimensional data can be quite difficult, and various methods have been implemented to simplify such datasets. A traditional approach when searching for any patterns of occurrence by reducing the dimensionality is cluster analysis. Cluster analysis is a statistical technique for grouping a set of objects where objects in the same group (called a cluster) are more similar to each other than to those in other groups. However, many methods (e.g. k-means cluster analysis) are based on mathematical techniques (Everitt et al., 2011). They do not have underlying probability distributions, and therefore it is impossible to make statistical inferences.

One type of analysis that overcomes this issue is cluster analysis based on mixture models. Mixture models assume that data is generated by a combination of two or more probability distributions, and were first applied by Pearson (1894). Clustering using mixture models gives a likelihood-based foundation for cluster analysis, and moreover, can model heterogeneity of the data. This approach can provide an effective way of clustering data under a variety of experimental designs (McLachlan and Peel, 2004).

The goal of this research is to investigate associations between presences of fish species, space, and time in a selected set of areas of New Zealand waters by using a mixture-model based clustering method. This research uses the methods developed by Pledger and Arnold (2014), and Fernández et al. (2014) to cluster species and/or sites into common groups. This clustering is conditional on measured covariates. We also aim to identify patterns of biological and ecological significance in the fisheries in these areas. To the best of our knowledge, this is the first time anyone has applied these methods incorporationing covariates.
More specific objectives of this research are

1. To detect clustering of species and geographical areas;

2. To summarize and display features of each cluster (What are the main characteristics of the clusters?, and are these characteristics able to be interpreted in a biologically/ecologically meaningful way?);

3. To extend the likelihood based clustering methods to cases where covariates (body size, depth, and water temperature) are included and compare the results with the clustering without covariates;

4. To compare cluster composition across time;

1.2 Background

1.2.1 Ecological Data

A variable with only two categories is called *binary*, *dichotomous* or *binomial* (Dobson and Barnett, 2008). Examples of binary data are presence/absence, dead/alive, male/female variables. Binary data are often expressed as dummy variables, and take 1 or 0 to indicate the outcomes (e.g. present $= 1$, absent $= 0$). Numerical data that can be measured without rounding are called *discrete* data. *Count*, the number of animals in a quadrat is a typical example of discrete variable. *Ordinal* data are categorical data where outcome categories have a natural order. Examples of ordinal responses and their ordered categorical scales are level of agreement with given statement (strongly agree, agree, neutral, disagree, strongly disagree), and any other variables that use a Likert scale. In biology, ordinal data arise in observations that are qualitative (e.g. absent/rare/plenty) as in the Braun-Blanquet scale.

Ecological data often consist of binary, count, and ordinal variables. For example, ecologists collect data on species composition as presence/absence or count. These data are often presented in a matrix format, for instance, the rows represent individual species and the columns represent sample locations. With possibly a large number of species or sites sampled, a dataset can be a very large matrix, and any overall patterns in such datasets are not obvious by inspection. Therefore, methods for reducing dimensions are necessary in order to detect any patterns of occurrence or abundance in the data.

1.2.2 Cluster Analysis Using Mixture Models

Cluster analysis can be applied to organize the observed data into meaningful structures to represent the large dataset in simpler and more efficient way. By doing so, ecologists can identify the relationships between the clusters and understand the information about the data more easily. They may want to have a better way of identifying diversity patterns associated with environmental and geographical conditions, spatial locations, and seasonal fluctuations. Cluster analysis is a useful tool to uncover true groups in the data, to help with discovering the pattern and structure of the data, and to investigate the association between the groups and also within the groups (Everitt et al., 2011). Another advantage of cluster analysis is data simplification (Gordon, 1999; Manly, 2004). If a large dataset can be condensed to a small number of clusters, these clusters can describe the patterns in the whole data set (Everitt et al., 2011). With the growing number of large datasets, cluster analysis is a useful method and has been applied in many domains.

A variety of cluster analysis techniques have been developed over the last three decades (Everitt et al., 2011). Approaches such as measurements of similarity/dissimilarity, partition optimization methods like k-means clustering, and hierarchical methods are commonly used. These approaches are also studied in many seminal papers. However, these approaches are based on mathematical techniques (Everitt et al., 2011) and are sample dependent. Moreover, they do not have underlying probability distribution (Fernández et al., 2014). Therefore, they do not fully exploit the nature of data, making it impossible to make formal statistical inferences or provide reliable model comparison methods. It is important to consider the nature of continuous, binary, count, ordinal, and nominal data and employ appropriate approaches accordingly.

Clustering using mixture models gives a likelihood-based foundation

for cluster analysis. This approach can overcome the issue described above and provide an effective way of clustering data under a variety of experimental designs (McLachlan and Peel, 2004). Finite mixture models in the context of clustering were first proposed by Pearson (1894) and have been studied in a number of papers (Govaert and Nadif, 2003; McLachlan and Chang, 2004; Pledger and Arnold, 2014). Mixture models assume that the data to be clustered are generated from a combination of two or more probability distributions. They enable us to estimate parameters in each distribution and posterior probability that each individual observation comes from a particular cluster. The likelihood is used to describe the observations. Models are often fitted using the Expectation-Maximization (EM) algorithm. It is an iterative method for finding maximum likelihood estimates of parameters in the model, and was first performed by Dempster et al. (1977). Mixture models are useful and being widely used to model the heterogeneity in many research areas (McLachlan and Chang, 2004). For example, McLachlan et al. (2002) demonstrated cluster analysis based on a normal mixture models for gene expression data. Another example is seen in Pledger (2000), where she performed cluster analysis using finite mixture models on binary data in an ecological capture-recapture study, and moreover, proposed a biclustering method. Biclustering, a relatively new development in cluster analysis, is a simultaneous procedure where the rows and columns are grouped into row clusters and column clusters. Govaert and Nadif (2003) and Pledger and Arnold (2014) developed biclustering via finite mixtures for binary and count data based on Bernoulli and Poisson likelihoods. Recently, Fernández et al. (2014) extended their approach to ordinal data.

Our challenge is to further extend the clustering method developed by Pledger and Arnold (2014) and Fernández et al. (2014) by adding covariates. We are particularly interested in seeing whether we get very different clusters when covariates are included or not. A covariate is an

observed variable that can possibly affect the outcome of an experiment. Alternatively, the terms "explanatory", "independent variable", or "predictor" are used. Covariates are incorporated into the models as terms for a linear regression. Generalized linear models (GLM) are commonly used families of statistical models incorporating covariates. For example, observed environmental conditions (e.g. temperature, soil moisture, vegetation coverage) are used as covariates to explain the abundance of species in ecological communities. An example of inclusion of covariates to cluster analysis is seen in a paper by Dunstan et al. (2011). Their aim was to group coral reef fish species that had similar ecological response to the environment from presence/absence data. They took a two-stage approach (called mixing GLM in their paper). They first carried out cluster analysis using mixture models to group species, then applied GLM using environmental measurements as covariates to each group, separately. Their approach did reduce the dimension of the original data and allowed them to describe each group's response to the environment. However, their approach lacks description of overall trend at a community level, and moreover, it did not consider the interactions between the species groups. Ignoring interaction among species opposes Francis et al. (2002) and Snelder et al. (2007)'s view that interaction between species is a major component of marine ecosystems. In addition, Bolck et al. (2004) raise their concern about multiple-stage analysis. They pointed out that analyses are carried out sequentially, so depending on the method used in the previous stage, the result from succeeding analysis changes leading to the final result being very different. Multiple-stage analysis has a risk of producing inconsistent results. In order to avoid this risk, we develop clustering analysis which incorporates covariates at the same time.

1.2.3 Previous Studies

Cluster analysis is a widely used method in ecology, in particular, in fisheries. Early examples of research into clustering in fisheries include He et al. (1997) and Francis et al. (2002). The literature from He et al. (1997) highlighted a classic characteristic of fishery-dependent data and risks involved in their analysis. Their aim was to classify fishing methods in relation to species composition of catches of sharks and tuna in Hawaiian waters. Their data were collected from local fishermen, so their data were fishery-dependent. Fishery-dependent data are defined as data from specific commercial and recreational fisheries as distinct from scientific ecological surveys. The characteristics of fishery-dependent data are highly diverse on fishing boats, gear, techniques, and target species. Furthermore, fishery-dependent data are self-reported data so some fishermen may not be honestly reporting, or discarding some catch before they land. Results from analyses using fishery-dependent data are likely to have unknown biases, leading to invalid conclusions. The authors suffered from a large variability of the fishery-dependent data and faced a great challenge in averaging the variations. Although He et al. (1997) adjusted for the variability as much as possible, they admitted that they could not adjust them fully. Therefore, it is uncertain whether their findings were conclusive.

On the other hand, a study by Francis et al. (2002) used fishery-independent data to demonstrate the cluster analysis for fish species in New Zealand waters. Their data came from research surveys, so the quality of the data is better than those from He et al. (1997). Unlike the paper from Dunstan et al. (2011), they considered the interaction between the species groups in their analysis. However, dissimilar to Dunstan et al. (2011), their clustering method was mathematical. They used binary data (presence/absence) and applied correspondence analysis, which does not have any underlying probability distribution of the data. As it was mentioned before, it is impossible to make any statistical inference from such analysis. These two

papers highlights the importance of the fishery-independent data, and approaches we need to use.

Cluster analysis is classified as an *unsupervised learning* technique, in which the task is to find hidden structure from data itself. However, many studies in the field of cluster analysis seem to have combined unsupervised learning approaches with *supervised learning*, where known information (prior information) is included in the analysis. Everitt et al. (2011) explains the basic idea of cluster analysis in his book. He states that cluster analysis is simply about discovering groups in the data. That is, cluster analysis reveals cluster structure but not the factors defining the cluster. He points out that many studies that performed cluster analysis seem to have predefined clustering factor prior to analysis, and interpreted the results according to them. He says that using distinct and natural clustering variables may lead to classification produced by artifact clustering, and may hinder researchers to discover unknown underlying structure. He is also concerned that using prior information to anticipate similarity measurements may not produce informative and interesting classifications. But Leathwick et al. (2003), who studied ecological clustering, argues that subjective choice of variables for clustering is required to increase robustness of the classification and have meaningful results. In fact, the studies of Francis et al. (2002) and Dunstan et al. (2011) predefined environmental measurements as covariates, and produced results that were interpreted in biologically and ecologically sensible way. In Snelder et al. (2007)'s paper, in which the importance of ecological classification for conservation management was highlighted, the authors developed a set of candidate variables that can explain the results from cluster analysis. It is true that we want to interpret the cluster structure in meaningful way, but it is also true that many previous studies paid little attention to Everitt et al. (2011)'s view. This research takes account of the concept of cluster analysis in Everitt et al. (2011)'s book. To evaluate whether the covariates are

in fact clustering factors, we compare clustering model with and without covariates. Information added as a covariate may make cluster structure be irrelevant, may change nothing in group membership as well as the numbers of clusters detected, or may produce different cluster structures. The grouping factor can be predefined covariates, or be a hidden variable called a latent variable. So, if the cluster structure is different between the simple cluster model and the cluster with covariate model, it means that the covariate explains the variations in the data and cluster effect will become weak. On the other hand, if cluster structure is unchanged or shows a little difference between the two models, then it means that effect of the covariate is small so we may need to investigate what might be contributing to clustering.

1.3 Mixture Models

In this section, we discuss the formulation of mixture models followed by the fitting of mixture models via maximum likelihood estimation. It also illustrates the usefulness of the EM algorithm.

Mixture models are probabilistic models that assume all the data are generated from a combination of two or more probability distributions. By combining individual probability distributions, mixture models provide a flexible and powerful framework for modeling heterogeneous data. Because of their usefulness, mixture models have been used in many applications, including cluster analysis, latent class analysis, discriminant analysis, and survival analysis (McLachlan and Peel, 2004). One of the first major analysis using mixture models was performed by Pearson (1894), where he fitted a normal mixture model using moments-based approach. Later, McLachlan and Basford (1988) have highlighted the advantage of using mixture models as an effective way of clustering various datasets.

McLachlan and Peel (2004) provide a guideline of the use of mixture models.

1.3.1 Finite Mixture Models

We consider application of finite mixture model in the context of cluster analysis. Let y be a realisation of random variable Y. Suppose we have an n dimensional vector $\mathbf{y} = (\mathbf{y_1}, \cdots, \mathbf{y_n})^{\mathbf{T}}$. With mixture model based clustering, we assume that $y_1, \cdots, y_n$ come from G distinct groups with some unknown proportions $\pi_1, \cdots, \pi_G$. The number of groups G is however fixed. In mixture models, the individual distributions that are combined to form a mixture distribution are called *mixture components*, and the probabilities π_g $(g = 1, \cdots, G)$ associated with each component are called *mixture weights*. Given a finite set of probability density functions $f_1(y_i), \cdots, f_G(y_i)$ $(i = 1, \cdots, n)$ with corresponding proportions $\pi_1, \cdots, \pi_G$, the full distribution of y_i is given by

$$f(y_i) = \sum_{g=1}^{G} \pi_g f_g(y_i) \tag{1.1}$$

where f_g is a mixture component density and π_g is its mixture weight. Note that $\pi_1, \cdots, \pi_g$ are non negative such that $\pi_g \geq 0, \forall g$, and $\sum_{g=1}^{G} \pi_g = 1$. A wide class of distributions can be approximated by equation (1.1). This type of mixture, where the number of components is fixed and finite, and a model probability density function is presented as a sum of parametrized functions, is called a *finite mixture*.

1.3.2 Infinite Mixture Models

In contrast, where the set of component distribution is uncountable (i.e. G is not a finite number), the sum of mixture components is replaced by an

integral over mixtures. Thus equation (1.1) becomes

$$f(y_i) = \int f(y_i|\theta_g)\pi_g(\theta_g)\mathrm{d}\theta_g$$

where θ_g is a parameter of the probability density function of $f(y_i|\theta_g)$ and in the continuous index of the mixture components rather than being discrete. This is called an *infinite mixture model*.

In this research, we use finite mixture models as a basis of cluster analysis. Once we define the underlying probability distribution, the next step is to find the parameter Θ that maximises the likelihood function. The typical approach is to use maximum likelihood estimation (MLE), which is explained in the next section.

1.3.3 Maximum Likelihood Fitting of Mixture Models

Here we review maximum likelihood estimation. For given data Y with n observations, the likelihood of the data, assuming that y_i are independent is

$$f(Y|\Theta, \underline{\pi}) = L(\Theta, \underline{\pi}|Y) = \prod_{i=1}^{n} f(y_i|\Theta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(y_i|\theta_g) \qquad (1.2)$$

where $Y = (y_1, \cdots, y_n)^T$, Θ is a set of parameters $(\theta_1, \cdots, \theta_G)$, and $\underline{\pi} = (\pi_1, \cdots, \pi_G)^T$. We wish to obtain the estimates for Θ that maximise the likelihood of equation (1.2);

$$\Theta^* = \arg\max_{\Theta} \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(y_i|\theta_g)$$

or, equivalently,

$$\Theta^* = \arg\max_{\Theta} \sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G} \pi_g f_g(y_i|\theta_g) \right\} \qquad (1.3)$$

However finding the MLE by solving equation (1.3) analytically in a finite mixture model is often difficult due to the following reasons. First, the number of G and π_g are unknown. Second, the summation over G that appears inside of the logarithm makes solving this equation complicated. Moreover this is further complicated due to several peaks in the likelihood, which are unique and unbounded. We have G number of the distributions and are required to find Θ that lead to the highest likelihood. A number of papers have studied the fitting of finite mixture models by maximum likelihood (McLachlan and Peel, 2004). Amongst these papers, the paper by Dempster et al. (1977) demonstrated fitting finite mixture distributions to modelling heterogeneous data using the Expectation-Maximisation (EM) algorithm. The use of the EM algorithm has been demonstrated for the analysis of heterogeneous data in a wide variety of fields (McLachlan and Peel, 2004). It is an iterative approach to solving the likelihood equations. The EM algorithm is also applicable in situations where the model has multiple parameters. Therefore, the EM algorithm is the primary tool in mixture model-based clustering. The EM algorithm is implemented by assuming that there are some missing observations, namely the group membership, which yield the complete data when combined with the observed data $\mathbf{y}$. We focus on the maximum likelihood fitting of mixture models via the EM algorithm in Chapter 3.

1.4 Outline of the Thesis

In this research, we aim to investigate the patterns of occurrence or abundance of fish species in relation to environmental factors by using cluster analysis based on mixture models. This method has not so far been applied with simultaneous inclusion of covariates. The data available in this research is fish abundance indices on the Chatham Rise from fishery-

independent scientific surveys from 1991 to 2013. The data are collected from annual bottom trawl surveys contracted by Ministry for Primary Industries (MPI) and carried out by National Institute of Water and Atmospheric Research (NIWA). We perform mixture-based clustering method including covariates. It is the first time that covariates have been included in mixture-based clustering analysis and also the first time this method has been applied in a fisheries context.

Chapter 2 introduces the annual bottom trawl survey data. We review the aim of this scientific survey, the data collection method, and the data management system. We also present a method to construct the data for the analysis from the survey data. Chapter 3 presents clustering methods used in this research including models with covariates. In order to evaluate our model performance, we set up simulation studies and their results are shown in this chapter. We then introduce a new **R** function `clustglm` and explain how this function is used for our analyses. Chapter 4 presents results from a selected fishery data. We illustrate the application of the `clustglm` function and present our results by using a number of visualisation tools. We conclude with final remarks and discussion in Chapter 5.

All the statistical programs throughout this research are written in **R** (R Core Team, 2015)

Chapter 2

Data

2.1 Description of Data

2.1.1 Overview

The data used for this research are fish abundance information of selected species from the hoki and middle depth trawl surveys of the Chatham Rise from 1992 to 2013.

This annual bottom trawl survey is contracted by Ministry for Primary Industries (MPI) and has been carried out by National Institute of Water and Atmospheric Research (NIWA) every summer (December to February) since 1992. The main purpose of this survey is to produce the relative biomass estimate of adult and juvenile New Zealand hoki (*Macruronus novaezelandiae*), which is New Zealand's largest fin-fish fishery. This trawl survey is the only survey that can produce fishery-independent estimates of abundance at these depths on the Chatham Rise (O'Driscoll et al., 2011). The methodological details of the survey are described in detail in Francis (1984); Hurst et al. (1992); O'Driscoll et al. (2011).

2.1.2 How the Survey Started

Hoki are widely distributed throughout New Zealand waters and support one of the most commercially valuable fisheries in New Zealand (Ministry of Fisheries, 2010). Hoki is a relatively fast growing species, approaching sexual maturity at three to five years of age. Hoki is assessed as two stocks; the eastern and western stocks (Figure not shown). Juvenile hoki from both stocks mix on the Chatham Rise, making the Chatham Rise a major nursery ground, especially on the western side (Horn, 1994), and therefore a principal area for recruitment (Stevens et al., 2013). Adult hoki also occur in deep water. So, the Chatham Rise is an important region for the derivation of estimates of hoki recruitment variability and biomass. Initially, several random trawl surveys were carried out, but they were not able to produce comparable estimates due to variability in boat size and fishing procedures (Horn, 1994). In order to provide a time series of comparable indices of abundance of adult hoki and to estimate future recruitment, this trawl survey on the Chatham Rise was commenced in 1992 (Horn, 1994).

2.2 Survey Methods

2.2.1 Survey Area and Design

The Chatham Rise is a broad area of ocean floor lying east of New Zealand (Figure not shown). It runs eastward from Banks Peninsula and extends over 150 km beyond the Chatham Islands. The convergence of southward and northward currents at the Chatham Rise bounds warm and cold waters, creating a subtropical front. This is a permanent oceanographic feature on the Chataham Rise (Dunn et al., 2010). The water of this convergence is nutrient rich, creating a region of high primary productivity that supports diverse and abundant animals. Therefore, the Chatham Rise is an important and valuable commercial fishing area.

All surveys since 1992 have been carried out in water depths of 200 to 800 m (O'Driscoll et al., 2011). In 2010, the survey was extended to deeper waters (to 1300 m). The reason for this was to provide fishery-independent relative biomass indices for orange roughy (*Hoplostethus atlanticus*) and to provide improved information for some species that are known to be observed in deeper waters on the Chatham Rise (O'Driscoll et al., 2011).

The survey follows a stratified random sampling design, where the Chatham Rise is stratified by depth and longitude (Horn, 1994). The stratification is a permanent setting for the survey but it has undergone several changes since 1992, in particular, a re-numbering of strata in 1996 and substratification of many strata in 2000 (O'Driscoll et al., 2011). These modifications were made to better define key species' (hoki and hake (*Merluccius australis*)) spawning area where high catch rates are common, thereby to have more precise biomass estimates. Therefore, the stratification is unique within the trips, although most strata remained the same. In 2010, deep strata from 800-1300 m on the northern and eastern Chatham Rise were introduced, making the total number of strata 34 since 2012 (Figure not shown). The stratification serves several purposes: It ensures that sampling locations occur in each stratum so that the survey covers all regions of interest, enables the creation of separate biomass estimates for each stratum, and most importantly, reduces the variance of the biomass estimates (Francis, 2006). The strata with 200-800 m water depths are referred as core strata, while the strata in deepwater (deeper than 800 m) are referred as non-core strata.

In addition to the stratification, the surveys follow a two-phase random design as a further refinement to allocate sampling locations (called trawl stations). A two-phase survey means that the survey is carried out in two parts. Francis (1984) proposed this strategy for the trawl surveys to

minimise the observed variance in catch rates, hence increasing precision of the biomass estimates. Allocation of trawl stations by stratum (starting locations of each trawl) in phase 1 is determined by computer simulations based on the catch rate of hoki and hake from all previous trawl surveys on the Chatham Rise. The simulation model takes into account stratum area and the distribution of key species. For the first survey in 1992, the phase 1 stations were allocated subjectively but with consideration of the distribution of juvenile and recruited hoki from past surveys (Horn, 1994). All surveys ensure that there are at least three phase 1 trawl stations in each stratum. Phase 2 trawl stations are additional sampling locations conducted during the survey. The aim is to improve the coefficients of variation (CV) for targeted species, particularly juvenile hoki. Allocation of phase 2 stations reflects the actual catch from phase 1 stations. After each tow at a phase 1 station, the computer simulation predicts possible phase 2 station locations that would best lower the overall biomass estimate CV. The allocations of phase 2 is done at the end of phase 1 tows. But where CVs in a stratum are known to be high, a phase 2 allocation may be done at the time in order to reduce streaming and survey time. This allows some degree of flexibility in achieving the desired CV. The proportion of trawls that are from phase 1 is fixed for all trips. The proportion of phase 2 tows (as a $\%$ of phase 1) should be kept as constant as possible. Because although phase 2 tows reduce the CV, they also introduce a bias. Keeping the proportion of phase 2 tows constant keeps this bias constant. Francis (1984) initially suggested that the optimal proportion of phase 1 trawl stations is 75$\%$, but given the budget for the survey and increasing historical data, the proportion of phase 1 stations has increased to 90$\%$ in recent years. Over the time of the survey, there have been about 25 to 34 strata, with around 100 trawl stations for the core strata of the surveys (Stevens and Livingstone, 2003; Stevens et al., 2012, 2013, 2014). The station allocation in phase 1 in non-core strata (strata in deeper water) is based on the biomass of orange roughy whose catch rates were obtained during the

trips in 2010 and 2011. There are no sampling stations allocated in phase 2 for these strata (O'Driscoll et al., 2011). All sampling stations in all strata are allocated at least 3 nautical miles away from each other. Allocations of strata for the trips in 2002 (tan0201), 2011 (tan1101), 2012 (tan1201), and 2013 (tan1301) are shown in Figure (not shown). As we can see the numbering of strata is slightly different from year to year. The trawl stations from the tan1201 trip are shown in Figure (not shown).

2.2.2 Vessel and Gear Specifications

The bottom trawl survey is carried out from *RV Tangaroa*, which is a purpose-built research vessel operated by NIWA. Trawling is a method of fishing that involves pulling a fishing net through the water behind one or more boats (Figure not shown). It is a commonly used method for commercial fishing as well as scientific research. During trawling, the net attached to the boat by warps is dragged along the ocean floor. The net narrows its shape towards the tail of the net (called the cod-end), where fish are caught. Floats, weights, and the trawl door otter board are used to keep the mouth of the net open. The speed and direction of trawling is chosen to maintain the geometry of the net so that the net won't collapse (Figure not shown). Trawling gear and procedures have been kept as consistent as possible over the trips in order to reduce sampling variability (O'Driscoll et al., 2011). The survey follows the guidelines of standard trawling procedures, described by Hurst et al. (1992). Towing speed and gear configuration are monitored regularly and maintained to be as constant as possible. At each trawling station the trawl is towed for 3 nautical miles (about 5.5 km) at a constant speed over the ground of 3.5 knots. Towing at a faster speed has a risk that the net is likely to be lifted from the sea floor. Towing at a slower speed will not be able to maintain the geometry of the net. Most of the tows at less than 800 m depth were carried out during daylight hours (as defined by Hurst et al., 1992). Trawling in deeper strata, where light level is low or absent, are carried out at night.

2.2.3 Measurements

Data collected by this research trawl surveys are stored in the database "**trawl**", managed for MPI by NIWA. This database contains several tables that represent trip (survey) information, strata information, trawl station information, catch information, and biological information. Each table consists of attributes that define properties of the table. The table **t_trip** holds the details for each trip. One of the attribute in this table is *trip_code*. It is a unique identifier for a each survey and coded by three letters and 4 digits describing the vessel name, trip year, and trip number. For example, TAN1201 indicates the survey that was carried out from NIWA's research vessel *RV Tangaroa* in 2012 and was the first trip of the year. Attribute *stratum* defines stratum number within each survey as in Figure (not shown).

Trawl stations (attribute: *station_no*) are labelled and the labels are stored with *trip_code* in table **t_station**. This table also includes data on location such as latitude, longitude, depth of both the gear and seabed at start of the tow as well as at the end the tow. Trawl survey performance can be measured by two attributes, *biomass_flag* and *gear_perf*, and they are also in **t_station**. The attribute *biomass_flag* indicates whether the trawl station is valid for biomass estimation. This flag takes three values: 0, 1, and 2 indicating "not valid ", "valid (for core strata)", and "valid (for non-core strata)", respectively. Trawls with *biomass_flag*$= 0$ ("not valid") could indicate that the towing was done during night when it was supposed to be carried out during day, towing was done for different purposes (e.g. non-survey), or towing was poor due to a gear fault. Gear performance during a tow is indicated by *gear_perf*. It is coded 1, 2, 3 or 4, indicating "Excellent", "Satisfactory", "Unsatisfactory probably due to malfunction or damage"and "Unsatisfactory due to malfunction or damage", respectively. This is subjectively measured by fishery scientists on board. Examples of unsatisfactory performance are that the net was torn during a tow, and trawl doors did not keep the net open. Gear used for the sur-

veys is coded in the attribute *gear_meth*. It is coded with two numbers and contains over 60 codes. Commonly used codes for trawl surveys are 01, 03, 05, 06, 51, 72, and 82, indicating bottom trawl, high operating bottom trawl, midwater trawl, pots, photography, and CTD (conductivity, temperature and Depth), respectively (Mackay, 2000). Overall survey performance is evaluated by the combination of *biomass_flag*, *gear_perf*, and *gear_meth*. These attributes are used to select appropriate data for this research.

The trawl database also indicates catch information for each trawl station on a trip. It is found in the table **t_catch**. Attributes in this table include *species*, *weight*, and *wt_meth*, and linked to station details via *trip_code* and *station_no*. At each station, all items in the catch are weighed and identified. Three-letter species code are assigned to all items caught (including non-animal species and rubbish) and recorded in the attribute *species*. There have been a few changes in the species codes over time. For example, the code for Tarakihi (*Nemadactylus macropteru*) was changed from TAR to NMP in 2010. These changes have been standardised to the most recent code (e.g. Tarakihi is coded as NMP throughout) in this research. Two closely related species are thought to be caught throughout the trips. One species from the genus *Brama* is southern Ray's bream (SRB, *Brama australis*), which is difficult to distinguish from Ray's bream (RBM, *Brama brama*) by external inspection. They have been reported as SRB and/or RBM in unknown ratios. So the codes for southern Ray's bream and Ray's bream is unified with RBM in this research.

Weight (in kilogram) of the item caught is recorded in attribute *weight*. All items caught are weighed at each station. While most items are weighed by scales, some items can be measured differently depending of the magnitude of the catch and the size of species. The attribute *wt_meth* indicates how the item was weighed with the code of 1 to 8. When the item is

weighed using scales, *wt_meth* is coded 1. Other definitions can be found in Mackay (2000). Weight information is recorded as the total weight of the item, and no counts of individuals are made.

Detailed biological data including length, weight, sex, age, and maturity stage of the individual fishes are measured for commercially important species. These date are found in table **t_fish_bio**. However, these species specific data are not required for this research. The tables and attributes referred in this research are listed in Table 2.1.

Table 2.1: Summary of the tables and the attributes used in this research

Table	Attribute	Information
t_trip	*trip_code*	unique identifier for a each survey
t_stratum	*stratum*	stratum code, unique within a trip
t_station	*station_number*	trawl stations, unique within a trip
	biomass_flag	evaluates whether the station is valid for biomass estimation
	gear_perf	evaluates the gear performance
	gear_meth	indicates the gear used for the survey
	max_gdepth	maximun depth (m) of lowest part of gear during the tow
	bot_temp	temperature at the bottom (°C)
t_catch	*species*	3 letters character indicating species
	weight	weight (kg)
	wt_meth	method of weighing

2.3 Data Used for This Research

In this research, the data records were throughly checked for quality before the analysis. One of the objectives of the research is to classify the fish species into common groups. Therefore species that are not classified as Teleostei (fin-fish) or Elasmobrachii (sharks, rays and skates) were excluded. Any species codes that do not identify any particular species were removed. For example, the code "FIS"and "SHA"were removed as they respectively indicate any fish and shark, otherwise unidentified. The trawls with $biomass\_flag = 0$ were also removed. This procedure left the data with $gear\_perf = 1$ or 2, indicating that the overall towing performance was satisfactory, hence catch information from these trawl stations are reliable. All of these data also had $gear\_meth = 01$ (bottom trawl). A further data quality check was performed by checking the trawl depth. The depth is measured at different positions (e.g. at the gear, at the vessel) but the depths recorded are similar. The minimum and maximum depth of the lowest part of gear (attribute:*min_gdepth*, *max_gdepth*) are checked whether they are within the depth range of core survey area (200 - 800 m as defined) or of deepwater strata (800 - 1300 m). Any records with inconsistent depths were removed. The labelling changes made for stratification were amended so that the location of the stratum is constant throughout the years.

The data of interest are: year/stratum/species/catch, for the analysis we record the catch as a presence/absence binary variable. That is, we record 1 if a species was caught at a particular stratum in a particular year, or 0 otherwise. With these records, a typical dataset has a matrix with rows representing species caught during the selected trip and the column representing the stratum in the trip. The dimension of the matrix differs year to year depending on the number of species caught and the number of strata. The number of strata increased after the deepwater strata were

introduced in 2010. We select the data from the survey in 2002 (tan0201), 2011 (tan1101), 2012 (tan1201), and 2013 (tan1301). We are interested in how the species groupings might have changed over time and want to have enough information for covariates. So three adjacent years from the recent trips and one earlier year were selected. Table 2.2 shows the number of Teleostei and Elasmobranchii species caught in these four trips. The number of strata increased after deep strata were introduced in 2010, and thereby the number of species caught also increased. Hoki is the most common species caught in all four trips.

Table 2.2: Summary of species caught and strata

Year	Number of species caught	Number of strata
2002	213	28
2011	286	33
2012	294	35
2013	323	35

2.3.1 Covariates

Some of our models require covariates at species and stratum level. For row (species) level cluster covariate, we use median body length (cm). We are interested in how much variation between clusters is actually mediated by species traits rather than environmental features. Initially, we considered the trophic level as a row level cluster covariate and attempted to assign the trophic levels to the species caught from the surveys. Trophic level is a representation of who eats whom in ecological communities. However, it is exceedingly difficult because it depends not only on body size, but also on diet, prey abundance, fishing activity, and animal behaviour (Dunn et al., 2013). The complexity of trophic levels is highlighted in Dunn et al. (2013). They studied 11 species of squaliforme shark species, and found that some species are opportunistic eaters and can shift between shallow and deep water levels to avoid competition and change their diet accordingly. They also concluded that the vulnerability to bottom trawl survey is likely to have an influence on species niches that may lead to alteration in the trophic level. Their study illustrates that the trophic level is unique to the region and changeable, also there is no global guide to define the trophic level for many of the species in our dataset.

The study by Jennings et al. (2001) also highlights the difficulty of assigning the trophic level to species. They also investigated the relationship between species' body size and the trophic level. Their study is motivated by Joel E. Cohen (1993), and found that body size could be used to predict the trophic level of fish in a community. Jennings et al. (2001)'s finding is supported by Rezende et al. (2009), in which they showed that body size and habitat affects the trophic level.

From these reasons above, we use only body length as a row level clustering covariate. We use median because the body size can vary largely among the individuals within the species depending on the sex and life

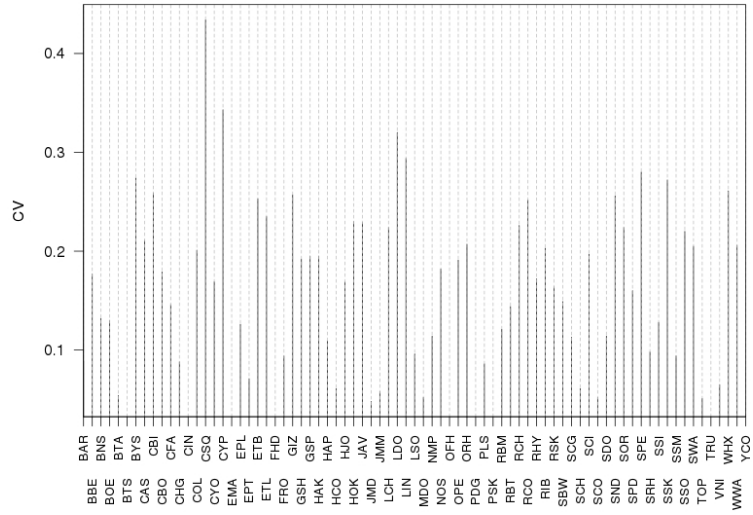
stage history (juvenile/adult), and often shows a multi-modal pattern. To note the body size variation within the species, we calculate the coefficient of variation (CV) of body length for each species. It is an indicator of the quality of the estimates and found by

$$CV = \frac{\text{standard deviation of body length}}{\text{mean of body length}}$$

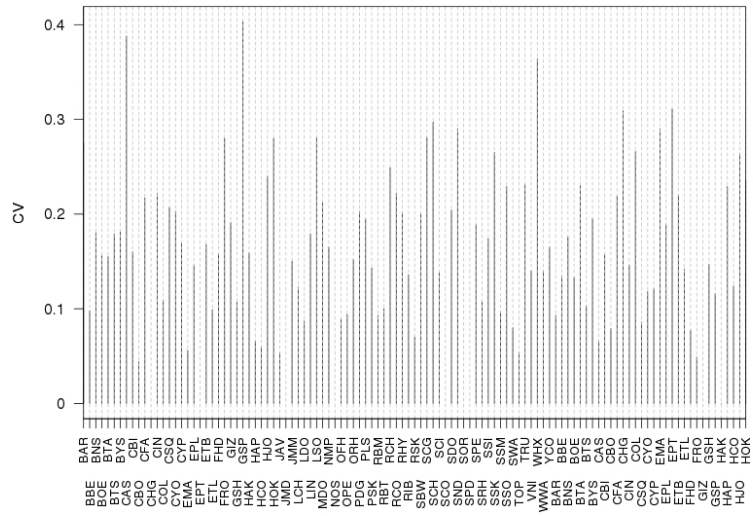
The values of CV for the species in this research vary between the species, and across the time within the species (Figure 2.1 and 2.2). As it can be seen, most of the CVs are under 20% which indicates the variation of body size is not too large. The species with the highest CV are the leafscale gulper shark (*Centrophorus squamosus*, CSQ) and the seal shark (*Dalatias licha*, BSH).

We use depth at the bottom at the towing gear (m), and sea floor temperature (°C), as column (stratum) level cluster covariates. We choose these two variables because they are thought to be drivers of biological patterns (Snelder et al., 2007), and have delivered ecologically and biological relevant results in clustering by Francis et al. (2002) and Dunstan et al. (2011). Moreover, environmental variables including depth and temperature are readily available measurements. They allow patterns of environmental variability to be more realistically portrayed and have been used as covariates in some studies of marine classification (Francis et al., 2002; Dunstan et al., 2011). These covariates are available in **trawl** database (Mackay, 2000). Depth is measured at various positions of the gear and the vessel. We use the attribute *max_gdepth*, which is the maximum depth of the lowest part of the gear during the tow. Water temperature (°C) at the bottom is recorded in the attribute *bot_temp*. There are some missing data for the temperature. We estimated the missing temperatures by taking the mean of the temperature in neighboring strata with similar depth from the same year (trip) and the temperature in the same strata from other

years. Figure 2.3 shows the temperature in the strata in the four trips. The records of sea floor temperature (°C) and depth at the bottom of the towing gear (m) in each stratum are presented in Figure 2.4 from the tan0201 data. Water in the shallow strata (core strata) are warmer, and temperature drops with depth. Figure 2.5, 2.6, and 2.7 are the visualisations of sea floor temperature (°C) and depth at the bottom of the towing gear (m) in each stratum from the tan1101, tan1201, and tan1301 data, respectively. There is no significant change in water temperature over time.

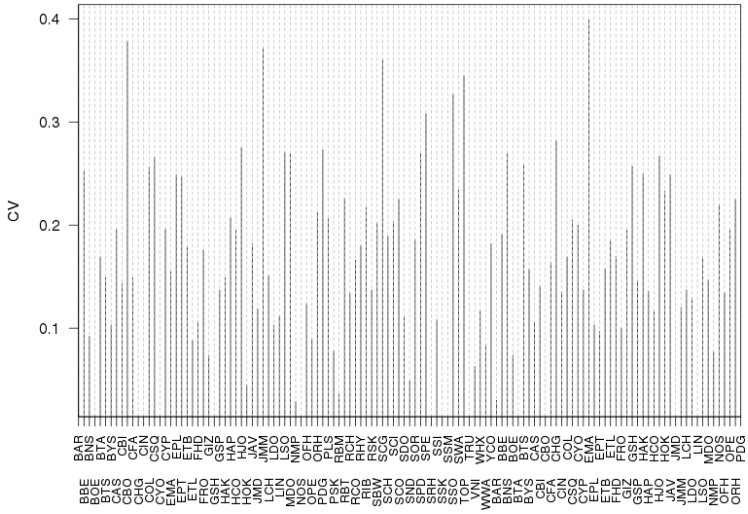


(a)

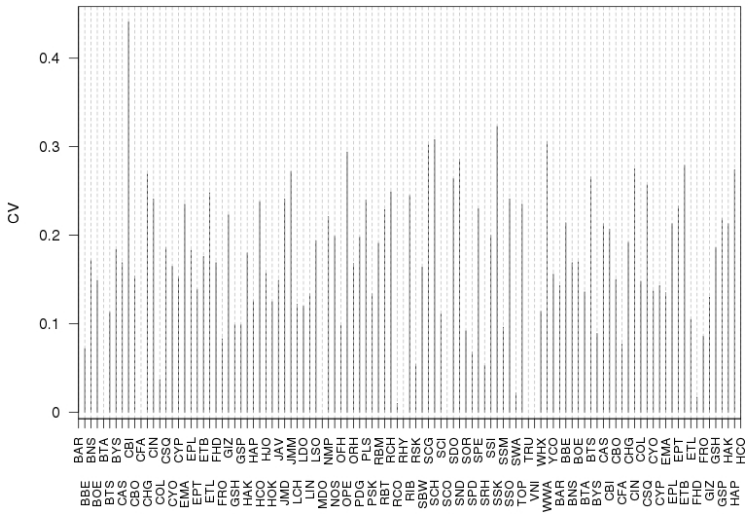


(b)

Figure 2.1: Coefficients of variation of the body length of species caught from tan0201 trip (a) and tan1101 trip (b).



(a)



(b)

Figure 2.2: Coefficients of variation of the body length of species caught from tan1201 trip (a) and tan1301 trip (b).

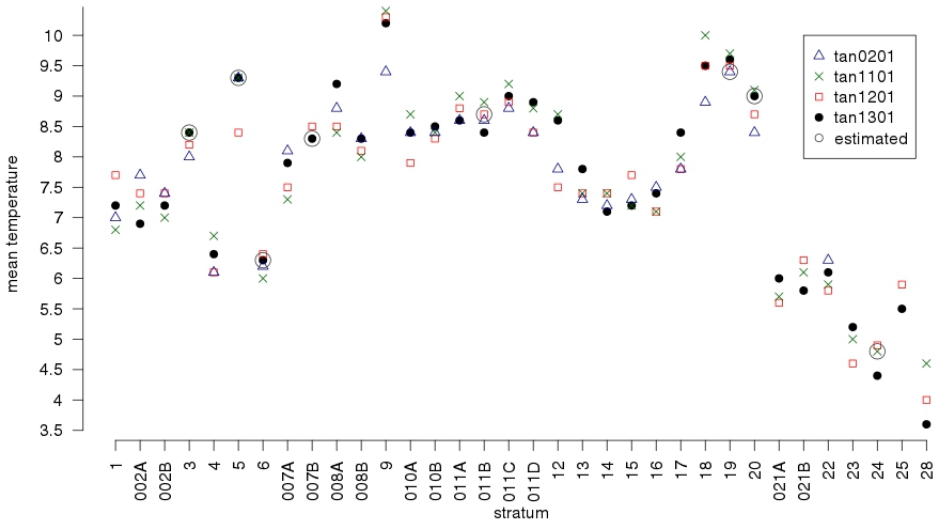
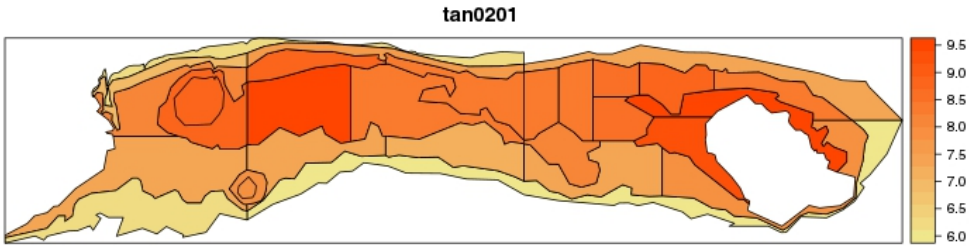
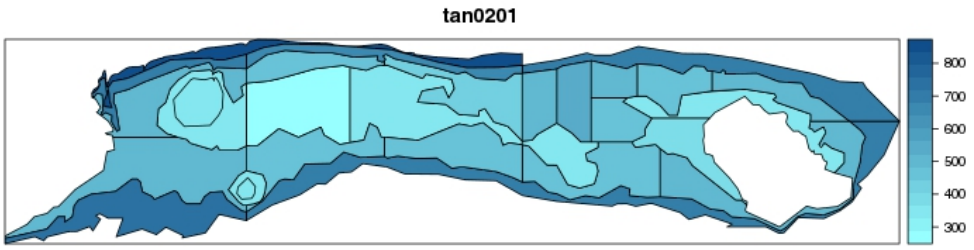


Figure 2.3: Bottom temperature recorded from the surveys in 2002 (tan0201), 2011 (tan1101), 2012 (tan1201), and 2013 (tan1301). The grey circle indicates the temperature that was estimated from the neighboring strata in the same year and the same strata from other years.

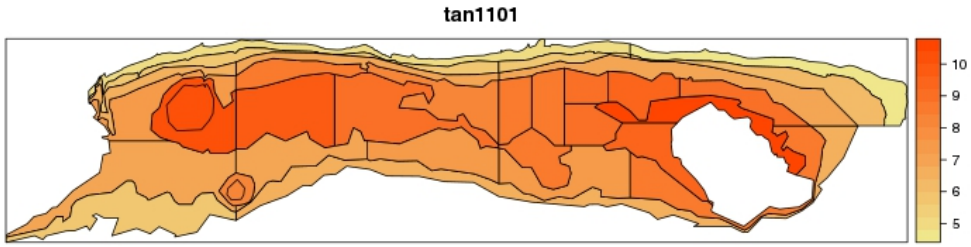


(a)

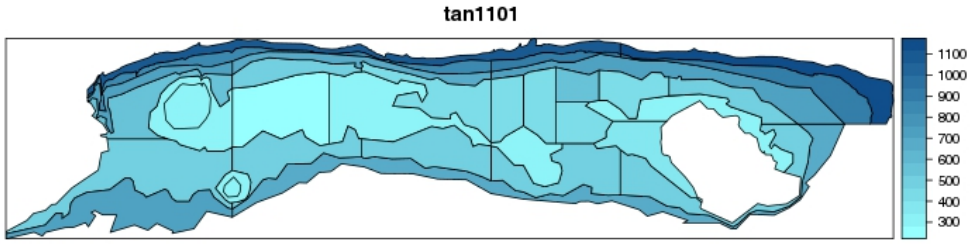


(b)

Figure 2.4: Map of the Chatham Rise displaying (a) seabed temperature (°C) and (b) depth (m) recorded for each stratum from the tan0201 data. The white area on the right includes the Chatham Islands therefore not surveyed.

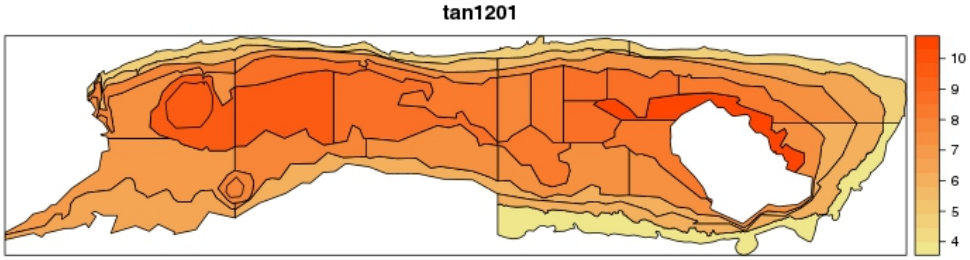


(a)

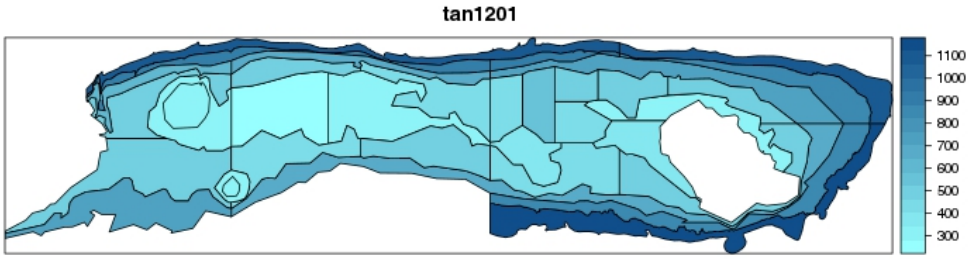


(b)

Figure 2.5: Map of the Chatham Rise displaying (a) seabed temperature (°C) and (b) depth (m) recorded for each stratum from tan1101 data. The white area on the right includes the Chatham Islands therefore not surveyed.

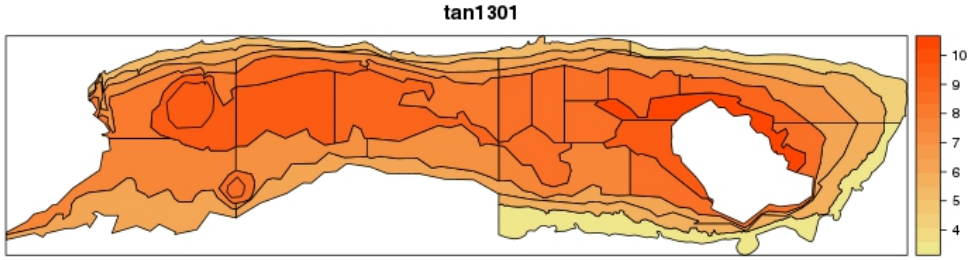


(a)

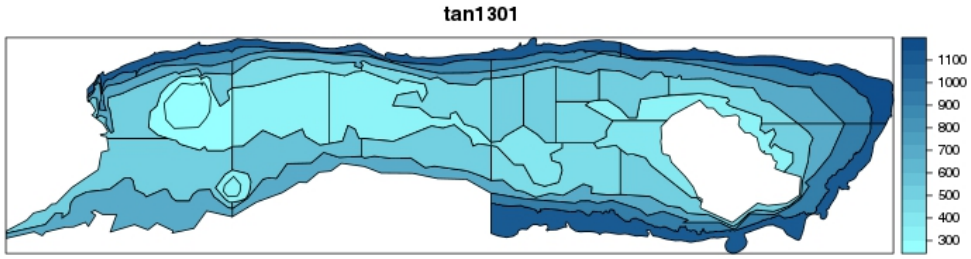


(b)

Figure 2.6: Map of the Chatham Rise displaying (a) seabed temperature (°C) and (b) depth (m) recorded for each stratum from tan1201 data. The white area on the right includes the Chatham Islands therefore not surveyed.



(a)



(b)

Figure 2.7: Map of the Chatham Rise displaying (a) seabed temperature (°C) and (b) depth (m) recorded for each stratum from tan1301 data. The white area on the right includes the Chatham Islands therefore not surveyed.

Chapter 3

Methods

3.1 Introduction

In this chapter, we first give the data, assumptions and likelihoods then present models used in this research. Some of the models have been previously developed, in particular, row and/or coulmn cluster models (Pledger and Arnold, 2014). We extend these existing models in the case where covariates are considered. In terms of model fitting, the Expectation-Maximisation (EM) algorithms is used. We explain the process of the EM algorithm with a particular model. We then demonstrate a simulation study in order to evaluate the performance of the maximum likelihood estimation models. Finally, we discuss the drawbacks of the EM algorithm and introduce a newly developed **R** function `clustglm` (Pledger et al., 2015), and explain how this function is used for this research.

3.2 Data, Assumptions and Likelihoods

Consider a binary event, called a "trial", with only two response outcomes: "success" or "failure". Independently repeated trials of an experiment with these two outcomes are called *Bernoulli* trials. We now present

the data being an $n \times p$ matrix $\mathbf{Y}$ of binary values with value y_{ij} a realization of random variable Y_{ij} $(i = 1, \cdots, n, j = 1, \cdots, p)$. The random variable Y_{ij} takes the value of 1 with the probability of success θ_{ij} or 0 with the probability of failure $1 - \theta_{ij}$. We denote $Y_{ij} \sim \text{Bernoulli}(\theta_{ij})$. Assuming independence in the rows and in the columns conditional on the rows, the likelihood function of the data is

$$L(\Theta|\mathbf{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{p} \theta_{ij}^{y_{ij}} (1 - \theta_{ij})^{1-y_{ij}} \tag{3.1}$$

The corresponding log likelihood is

$$\ell(\Theta|\mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{p} [y_{ij} \log \theta_{ij} + (1 - y_{ij}) \log(1 - \theta_{ij})] \tag{3.2}$$

The model specified in equation (3.2) has one parameter for every observation, and is *saturated*. Such a model has no predictive power, and simply identifies $\theta_{ij} = y_{ij}$. Of more interest is the appropriate reduction of the dimension of the parameters θ_{ij}. The simplest case is to have $\theta_{ij} = \theta$, a single parameter for all observations. Other alternatives are $\theta_{ij} = \theta_i$, or $\theta_{ij} = \theta_j$, one parameter for every row or column. In order to estimate the parameters Θ in these situations above, a generalised linear model (GLM) can be fitted. Of more interest are models that group the rows or columns into clusters, with the rows/columns being similar within clusters but different between the clusters. That is, $\theta_{ij} = \theta_{rj}$ for row clusters $(r = 1, \cdots, R)$ where row is a member of row cluster $r(i \in r)$. Likewise, $\theta_{ij} = \theta_{ic}$ for column clusters $(c = 1, \cdots, C)$ when $j \in c$. Furthermore, we can cluster rows and columns at the same time and have $\theta_{ij} = \theta_{rc}$. When clusters are made, the data no longer have a simple probability distribution but the mixture of R, C, or $R \times C$ probability distributions.

3.3 The Models and Model Fitting

In this section, we present likelihood-based clustering models for binary data. Some of these models have already been proposed by Pledger (2000), Govaert and Nadif (2003), and Pledger and Arnold (2014). In this research, we extend these existing methods by including covariates.

3.3.1 The Row-clustered Model

We start with the case of row clustering. Suppose that the rows come from a finite mixture with R groups with the rows being similar within groups but different between groups, yielding a clustering of the rows of the data matrix $\mathbf{Y}$. A prior probability that row i belongs to row group r is π_r. We assume that $1 \leq R < n$, $0 < \pi_r$, and $\sum_{r=1}^{R} \pi_r = 1$.

Let ϕ_{rj} be the probability that observation $y_{ij} = 1$ given that row i belongs to row group r (i.e. $\phi_{rj} = P(y_{ij} = 1 | i \in r)$). For Bernoulli distributions, the expected values of Y_{ij} for $i \in r$ is ϕ_{rj}, which is defined by

$$\operatorname{logit} \phi_{rj} = \log\left(\frac{\phi_{rj}}{1-\phi_{rj}}\right) = \eta_{rj} = \mu + \mathrm{a}_r + \mathrm{b}_j + \lambda_{rj} \tag{3.3}$$

The parameters that define ϕ_{rj} are μ, a_r, b_j, and λ_{rj}, where

- μ is the overall effect
- a_r is the effect of row group r ($r = 1, \cdots, R$)
- b_j is the effect of the jth column ($j = 1, \cdots, p$)
- λ_{rj} is the interaction between row cluster and columns

We temporarily assume that there is no association between row group and column ($\lambda_{rj} = 0$), and for avoiding identifiability problems we impose the constraints $\sum_{r=1}^{R} \mathrm{a}_r = \sum_{j=1}^{p} \mathrm{b}_j = 0$. We define Φ as the full parameter

set for this model that defines ϕ_{rj}, $\Phi = (\mu, \underline{\mathrm{a}}, \underline{\mathrm{b}})$, where $\underline{\mathrm{a}} = \mathrm{a}_1, \cdots, \mathrm{a}_r$, and $\underline{\mathrm{b}} = \mathrm{b}_1, \cdots, \mathrm{b}_j$. The likelihood function for the row-clustered model is

$$L(\Phi, \underline{\pi}|\mathbf{Y}) = \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}} (1 - \phi_{rj})^{1-y_{ij}} \right] \right\} \tag{3.4}$$

and thus the log likelihood function is

$$\ell(\Phi, \underline{\pi}|\mathbf{Y}) = \sum_{i=1}^{n} \log \left\{ \sum_{r=1}^{R} \pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}} (1 - \phi_{rj})^{1-y_{ij}} \right] \right\} \tag{3.5}$$

where $\underline{\pi} = (\pi_1, \cdots, \pi_r)$.

3.3.2 Fitting the Models

Once models are proposed, the next step is to fit the models and obtain parameter estimates. Here we show how the models are fitted by the EM algorithm.

The EM algorithm treats the observed data as being incomplete and considers the group membership as missing data. The missing information here is the actual group membership of the row groups. Following the procedure in McLachlan and Krishnan (2007), we express the missing data as the $n \times R$ matrix $\mathbf{Z}$ in the case of row clustering. Each element of $\mathbf{Z}$ is an indicator with $z_{ir} = 1$ if row i belongs to group r and $z_{ir} = 0$ otherwise. An example of $\mathbf{Z}$ when $n = 3$ and $R = 2$ is

$$\mathbf{Z}_{n \times r} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

This means the first row of $\mathbf{Y}$ is a member of group 1, while the second

and the third row of $\mathbf{Y}$ belong to group 2. We impose the constraint that each row i only belongs to one group, e.g. $\sum_{r=1}^{R} z_{ir} = 1, \forall i$. This implies that each row $\mathbf{Z}_i$ of $\mathbf{Z}$ independently follows a multinomial distribution (see Appendix A for more details).

$$\mathbf{Z}_i \sim \text{Multinomial}(1, \underline{\pi}) \qquad \text{for} \quad i = 1, \cdots, n$$

We temporarily assume that the membership $\mathbf{Z}$ is known and have

$$P(\underline{z}_{ir}|\underline{\pi}) = \prod_{i=1}^{n}\prod_{r=1}^{R} \pi_r^{z_{ir}} = \prod_{i=1}^{n}\sum_{r=1}^{R} \pi_r z_{ir}$$

Then the likelihood of the complete data is

$$\begin{aligned} L_c(\Phi, \underline{\pi}|\mathbf{Y}, \mathbf{Z}) &= P(\mathbf{Y}|\mathbf{Z}, \Phi, \underline{\pi}) \\ &= \prod_{i=1}^{n}\left\{\sum_{r=1}^{R} \pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}}(1-\phi_{rj})^{1-y_{ij}}\right]\right\} \times \prod_{i=1}^{n}\sum_{r=1}^{R} \pi_r z_{ir} \\ &= \prod_{i=1}^{n}\left\{\sum_{r=1}^{R} z_{ir}\pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}}(1-\phi_{rj})^{1-y_{ij}}\right]\right\} \end{aligned} \tag{3.6}$$

and the log likelihood of the complete data is defined as follows

$$\ell_c(\Phi, \underline{\pi}|\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} \log\left\{\sum_{r=1}^{R} z_{ir}\pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}}(1-\phi_{rj})^{1-y_{ij}}\right]\right\} \tag{3.7}$$

Since z_{ir} is either 0 or 1, we are able to bring z_{ir} and the sum over r outside of the log function. Hence equation (3.7) is further expanded as follows

$$
\begin{aligned}
\ell_c(\Phi, \underline{\pi}|\mathbf{Y}, \mathbf{Z}) &= \sum_{i=1}^{n} \log \left\{ \sum_{r=1}^{R} z_{ir} \pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}} (1-\phi_{rj})^{1-y_{ij}} \right] \right\} \\
&= \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \log \left\{ \pi_r \left[\prod_{j=1}^{p} \phi_{rj}^{y_{ij}} (1-\phi_{rj})^{1-y_{ij}} \right] \right\} \\
&= \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \left\{ \log \pi_r + \sum_{j=1}^{p} \log \left(\phi_{rj}^{y_{ij}} (1-\phi_{rj})^{1-y_{ij}} \right) \right\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} z_{ir} \left\{ y_{ij} \log \phi_{rj} + (1-y_{ij}) \log(1-\phi_{rj}) \right\} \\
&\qquad + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \log \pi_r
\end{aligned}
\tag{3.8}
$$

The EM algorithm is an iterative optimisation procedure for estimating the maximum likelihood estimates (MLE) of the parameter set Φ, $\underline{\pi}$, and $\mathbf{Z}$. The E-step of the algorithm estimates the $\mathbf{Z}$ matrix, conditional on a current parameter estimate $\widehat{\Phi}$ and $\widehat{\underline{\pi}}$. The M-step then re-estimates $\underline{\pi}$ and Φ conditional on $\mathbf{Y}$ and updated $\mathbf{Z}$, by maximising equation (3.8) over Φ and $\underline{\pi}$. These steps alternate until the estimates converge (Dempster et al., 1977).

E-step

In the E-step of the EM algorithm, we estimate the expected value of the latent variable $\widehat{z}_{ir}$, using the current estimates of the parameters (Φ and $\underline{\pi}$). We apply Bayes' theorem to obtain the expected values of z_{ir}. Applying Bayes' rule, we obtain

$$\begin{aligned}\widehat{z}_{ir} &= P(z_{ir}=1|y_{ij},\widehat{\Phi}) \\ &= \frac{P(y_{ij},\widehat{\Phi}|z_{ir}=1)P(z_{ir}=1)}{\sum_{u=1}^{R}P(y_{ij},\widehat{\Phi}|z_{iu}=1)P(z_{iu}=1)} \\ &= \frac{\widehat{\pi}_r\prod_{j=1}^{p}\widehat{\phi}_{rj}^{y_{ij}}(1-\widehat{\phi}_{rj})^{1-y_{ij}}}{\sum_{u=1}^{R}\{\widehat{\pi}_u\prod_{l=1}^{p}\widehat{\phi}_{ul}^{y_{il}}(1-\widehat{\phi}_{ul})^{1-y_{il}}\}}\end{aligned} \tag{3.9}$$

Equation (3.9) is the posterior probability that row i is in group r given that the data $\{y_{ij}\}_{j=1}^{p}$ have been observed, and conditioned on the current parameter estimates $\widehat{\Phi}$ and $\widehat{\underline{\pi}}$. The elements of $\widehat{\mathbf{Z}}$ is are not 0 or 1, but may take any value between 0 and 1. This is referred as "fuzzy" allocation of rows to a cluster because the estimated group membership is not definite, but expressed as probability.

M-step

In the M-step, we maximise the log likelihood function over the parameters Φ and $\underline{\pi}$, conditioned on the complete data $\mathbf{Y}$ and $\widehat{\mathbf{Z}}$. Using the updated $\{\widehat{z}_{ir}\}$, we obtain new values for μ, $\underline{\mathrm{a}}$, $\underline{\mathrm{b}}$ and $\underline{\pi}$. First, π_r are analytically estimated from the updated $\{\widehat{z}_{ir}\}$. Recall that we imposed the constraint on π. When we want to maximise the function subject to the constraint, we use the Lagrange multiplier method (see Appendix B for more details). The method of Lagrange multipliers is a strategy for finding the local maximum of a function subject to constraints. In our case, we want to maximise the function (3.8) subject to the constraint $\sum_{r=1}^{R}\pi_r = 1$. Applying the Lagrange multiplier method, we get

$$\widehat{\pi}_r = \frac{\sum_{i=1}^{n}\widehat{z}_{ir}}{\sum_{i=1}^{n}(\sum_{u=1}^{R}z_{iu})} = \frac{\sum_{i=1}^{n}\widehat{z}_{ir}}{\sum_{i=1}^{n}1} = \frac{\sum_{i=1}^{n}\widehat{z}_{ir}}{n} \tag{3.10}$$

This is simply the proportion of the n cases in group r given the estimated group memberships $\widehat{z}_{ir}$. The rest of parameter estimates do not have analytic solutions and must be estimated numerically. These estimates are obtained by using the numerical optimiser function `optim()` in **R** (R Core Team, 2015). These E- and M-steps are repeated in alternation until the estimates have converged. That is, the changes in the parameter estimates and/or the log likelihood between two successive iterations are less than a specified threshold (e.g. $\varepsilon = 1 \times 10^{-4}$). In this research the convergence of estimates is defined as

$$\frac{|\ell_c^t - \ell_c^{t-1}|}{|\ell_c^{t-1}|} < \varepsilon \qquad \text{and} \qquad \frac{|\Phi^t - \Phi^{t-1}|}{|\Phi^{t-1}|} < \varepsilon$$

where t indicates the tth iteration of the EM algorithm. We specify the threshold value to be $\varepsilon = 1 \times 10^{-6}$.

3.3.3 The Column-clustered Model

The model for column clustering is analogous to the row-clustered model, and is just the transpose of row clustering. Now we let ϕ_{ic} be the probability that observation $y_{ij} = 1$ given that column j belongs to column group c (i.e. $\phi_{ic} = P(y_{ij} = 1 | j \in c)$). The linear predictor for ϕ_{ij} is given by

$$\operatorname{logit} \phi_{ic} = \log\left(\frac{\phi_{ic}}{1-\phi_{ic}}\right) = \eta_{ic} = \mu + \mathrm{a}_i + \mathrm{b}_c + \lambda_{ic} \tag{3.11}$$

where μ is the overall effect, a_i is the effect of rows, b_c is the effect of the column group ($c = 1, \cdots, C$), and λ_{ic} is the interaction between rows and column clusters.

Model Fitting for Column-clustered Model

We use the same approach as described previously, but with $\mathbf{H}$, which is a $p \times C$ matrix that contains the information about the membership of column clusters, where $h_{jc} = 1$ if jth column is in column group c, or $h_{jc} = 0$ if not ($c = 1, \cdots, C$).
The likelihood of the complete data in case of column clustering is

$$L(\Phi, \underline{\kappa}|\mathbf{Y}, \mathbf{H}) = \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \left[h_{jc}\kappa_c \prod_{i=1}^{n} \phi_{ic}^{y_{ij}} (1-\phi_{ic})^{1-y_{ij}} \right] \right\}$$

where κ_c is the probability that jth column is in column group c. Constraints here are $\kappa_c > 0, \forall c$ and $\sum_{c=1}^{C} \kappa_c = 1$. The parameter Φ are the parameters that define ϕ_{ic}, which are $\underline{\mathrm{a}}, \underline{\mathrm{b}}$, and $\underline{\kappa}(\kappa_1, \cdots, \kappa_C)$.
The log likelihood function of the complete data is

$$\begin{aligned} \ell_c(\Phi, \underline{\kappa}|\mathbf{Y}, \mathbf{H}) &= \sum_{j=1}^{p} \log \left[\sum_{c=1}^{C} h_{jc}\kappa_c \prod_{i=1}^{n} \phi_{ic}^{y_{ij}} (1-\phi_{ic})^{1-y_{ij}} \right] \\ &= \sum_{j=1}^{p} \sum_{c=1}^{C} \sum_{i=1}^{n} h_{jc} \left[y_{ij} \log \phi_{ic} + (1-y_{ij}) \log(1-\phi_{ic}) \right] \\ &\quad + \sum_{j=1}^{p} \sum_{c=1}^{C} h_{jc} \log \kappa_c \end{aligned} \tag{3.12}$$

Following the same procedure as the row cluster case,

E step

$$\widehat{h}_{jc} = \frac{\widehat{\kappa}_c \prod_{i=1}^{n} \widehat{\phi}_{ic}^{y_{ij}} (1-\widehat{\phi}_{ic})^{1-y_{ij}}}{\sum_{v=1}^{C} \widehat{\kappa}_v \prod_{s=1}^{n} \hat{\phi}_{sv}^{y_{ij}} (1-\widehat{\phi}_{sv})^{1-y_{ij}}}$$

M Step

$$\widehat{\kappa}_c = \frac{\sum_{j=1}^{p} \widehat{h}_{jc}}{p}$$

The parameters μ, $\underline{\mathrm{a}}$ and $\underline{\mathrm{b}}$ are estimated by `optim` function by optimising equation (3.12).

3.3.4 Row Cluster Model with Row Level Covariates

We introduce a new term, the row covariate matrix $\mathbf{X}$ to equation (3.3). Let $\mathbf{X}$ be an $n \times D$ covariate matrix ($d = 1, \cdots, D$), and δ be the $D \times 1$ parameter vector for the covariates where D is the number of covariates. The covariate matrix may contain information which makes the clustering structure irrelevant, or allows a different clustering structure to emerge. We obtain the linear predictor for the model with covariates in the case of row clustering where $y_{ij}|i \in r, \underline{x}_i \sim \mathrm{Benoulli}(\phi_{rj}\,(\underline{x}_i))$ in the following form

$$\mathrm{logit}\,(\phi_{rj}(\underline{x}_i)) = \mathrm{logit}(\phi_{ijr}) = \eta_{ijr} = \mu + \mathrm{a}_r + \mathrm{b}_j + \underline{x}_i^T \underline{\delta} \tag{3.13}$$

where $\underline{x}_i$ is a $D \times 1$ vector of covariate, consisting the elements of the ith row of $\mathbf{X}$, $\sum_{r=1}^{R} \mathrm{a}_r = 0$, and $\sum_{j=1}^{p} \mathrm{b}_j = 0$. There is no constraint on δ. Note that $\phi_{ijr} = \frac{1}{1+\exp(-\eta_{ijr})}$. The log likelihood function of the complete data is

$$
\begin{aligned}
\ell_c(\Phi, \underline{\pi}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R} z_{ir}\left(y_{ij}\log\phi_{ijr} + \log(1-\phi_{ijr}) - y_{ij}\log(1-\phi_{ijr})\right) \\
&\quad + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R} z_{ir}\left(y_{ij}\log\left(\frac{\phi_{ijr}}{1-\phi_{ijr}}\right) + \log(1-\phi_{ijr})\right) \\
&\quad + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R} z_{ir}\left[y_{ij}\eta_{ijr} - \log\left(1+\exp(\eta_{ijr})\right)\right] \\
&\quad + \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log\pi_r
\end{aligned}
\tag{3.14}
$$

where $\Phi = \{\mu, \underline{\mathrm{a}}, \underline{\mathrm{b}}, \underline{\delta}\}$, and $\underline{\delta} = \delta_1, \cdots, \delta_D$. Note that the Φ becomes an $n \times p \times r$ array. That is, there is a different $n \times p$ matrix of probabilities ϕ_{ij} for each row group r.

3.3.5 Model Fitting for Covariate Model

The basic procedure is the same as the row-clustered model, however, further adjustment is required to allow for the altered dimension of Φ during computation.

E-step

$$
\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \widehat{\phi}_{ijr}^{y_{ij}}(1-\widehat{\phi}_{ijr})^{1-y_{ij}}}{\sum_{u=1}^{R}\{\widehat{\pi}_u \prod_{l=1}^{p} \widehat{\phi}_{ilu}^{y_{il}}(1-\widehat{\phi}_{ilu})^{1-y_{il}}\}} \tag{3.15}
$$

This equation is similar to that of the row-clustered model, but $\widehat{z}_{ir}$ is calculated in multiple stages. First, $\widehat{\phi}_{ijr}^{y_{ij}}$ is computed for the individual i conditional of being in group r and given its row covariate x_i, also conditional on the current parameters $\widehat{\mu}$, $\widehat{\underline{\mathrm{a}}}$, $\widehat{\underline{\mathrm{b}}}$ and $\widehat{\underline{\pi}}$. At this first stage, we have R matrices ($n \times p$) that contain ϕ_{ij} for each r. When $R = 3$, there will be three $\widehat{\phi}_{ijr}^{y_{ij}}(1 - \widehat{\phi}_{ijr})^{1-y_{ij}}$ matrices ($n \times p$). Next, each matrix is multiplied over the columns ($\prod_{j=1}^{p} \widehat{\phi}_{ij1}^{y_{ij}}(1 - \widehat{\phi}_{ij1})^{1-y_{ij}}$, for $r = 1$), making an $n \times 1$ column vector for each r. Then each vector is multiplied by the corresponding π_r, making the numerator of equation (3.15). Then z_{ir} is estimated by using Bayes' rule as before.

M step

The computation of π_r is the same as the row-clustered model.

$$\widehat{\pi}_r = \frac{\sum_{i=1}^{n} \widehat{z}_{ir}}{\sum_{i=1}^{n}(\sum_{u=1}^{R} z_{iu})} = \frac{\sum_{i=1}^{n} \widehat{z}_{ir}}{\sum_{i=1}^{n} 1} = \frac{\sum_{i=1}^{n} \widehat{z}_{ir}}{n}$$

The rest of the parameters Φ are also estimated as in the row-clustered model, but the calculation of log likelihood involves multiple steps because Φ is the array. The first term of the equation (3.14), $\Big[y_{ij}\eta_{ijr} - \log\left(1 + \exp(\eta_{ijr})\right)\Big]$ is calculated for each r, making r matrices ($n \times p$). Then, each matrix is summed up over columns, $\sum_{j=1}^{p}\Big[y_{ij}\eta_{ijr} - \log\left(1 + \exp(\eta_{ijr})\right)\Big]$, making an $n \times 1$ vector for each r. Next, the vectors are multiplied by corresponding $\widehat{z}_{ir}$ and summed up over rows and groups, resulting in a scalar value. The second term of the equation, $\sum_{i=1}^{n}\sum_{r=1}^{R} \widehat{z}_{ir} \log \pi_r$ is obtained by simple two dimensional matrix calculation, and also results in a scalar value. The sum of these two values is the complete data log likelihood, and we want to find the parameters $\underline{\mu}$, $\underline{\mathrm{a}}$, $\underline{\mathrm{b}}$, $\underline{\delta}$ that maximise the log likelihood. We repeat these two steps of the EM algorithm until estimates have converged.

3.3.6 Row Cluster Model with Column Level Covariates

Likewise, column level covariates can be included in the row-clustered model. Let $\mathbf{W}$ be a $p \times W$ matrix that contains column level covariates and W is the number of covariates. The linear predictor is

$$\text{logit}\left(\phi_{rj}(\underline{w}_j)\right) = \text{logit}(\phi_{ijr}) = \eta_{ijr} = \mu + \text{a}_r + \text{b}_j + \underline{w}_j^T \underline{\psi} \tag{3.16}$$

where w_j is a $W \times 1$ elements of $\mathbf{W}$ ($w = 1, \cdots, W$), and $\underline{\psi}$ is the vector of parameters for the column covariates $\underline{\psi} = (\psi_1, \cdots, \psi_W)$. The constrains are the same as before, $\sum_{r=1}^{R} \text{a}_r = 0$, and $\sum_{j=1}^{p} \text{b}_j = 0$. There is no constraint on ψ. The parameters are estimated the same way as the row cluster with row level covariates model in 3.3.5.

3.3.7 Inclusion of an Interaction Between Clusters and Covariates

A situation where there is an interaction between cluster membership and the covariates is further considered in this section. An interaction term can be introduced to equation (3.13) and (3.16). Taking the row-clustered model with column level covariates (3.16) for an example, let $\mathbf{T}$ be a $W \times R$ matrix consisting interaction parameters $\{\tau_{wr}\}$ between the row cluster and the column covariates. We impose the sum zero constraint on rows of the $\mathbf{T}$ matrix $\sum_{r=1}^{R} \tau_{wr} = 0$, $\forall w$). The linear predictor for the model with row cluster and column level covariates interaction term is

$$\text{logit}(\phi_{ijr}) = \mu + \text{a}_r + \text{b}_j + \underline{w}_j^T (\underline{\psi} + \underline{\tau}_r) \tag{3.17}$$

where $\underline{\tau}_r$ is the rth column ($w \times 1$) of the matrix $\mathbf{T}$. The complete data log likelihood is the same as equation (3.14), with Φ containing the extra parameters $\{\tau_{wr}\}$.

E-step

The computation of $\widehat{z}_{ir}$ takes the same process as the solving the equation (3.15). The only difference is that extra elements $\{\tau_{wr}\}$ are included to estimate $\{\phi_{ijr}\}$.

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \widehat{\phi}_{ijr}^{y_{ij}} (1 - \widehat{\phi}_{ijr})^{1-y_{ij}}}{\sum_{u=1}^{R} \{\widehat{\pi}_u \prod_{l=1}^{p} \widehat{\phi}_{ilu}^{y_{il}} (1 - \widehat{\phi}_{ilu})^{1-y_{il}}\}}$$

M step

As before we have the analytic result for $\widehat{\pi}_r$.

$$\widehat{\pi}_r = \frac{\sum_{i=1}^{n} \widehat{z}_{ir}}{n}$$

Again, the rest of parameters μ, $\underline{\mathrm{a}}$, $\underline{\mathrm{b}}$, $\underline{\psi}$, and $\{\tau_{wr}\}$ are obtained numerically by `optim()` function in **R** (R Core Team, 2015), and the EM algorithm is repeated until it satisfies the convergence rule. Note that we only estimate $W(R-1)$ elements of τ_{wr}. The rest of the elements of the matrix $\mathbf{T}$ are obtained by the constraints.

3.4 Row Standardised Model

Finally, we present the row cluster with row standardised model. This model has a new parameter α_i with $\sum_{i=1}^{n} \alpha_i = 0$. It can be used when the rows (species, for example) are assumed to differ in abundance. The parameter α_i is the deviation from the overall average of frequency of occurrence, and may also called the row main effect. The linear predictor for the row standardised model is

$$\mathrm{logit}(\phi_{ijr}) = \mu + \alpha_i + \mathrm{a}_r + \mathrm{b}_j \tag{3.18}$$

where $\sum_{r=1}^{R} \mathrm{a}_r = 0$, and $\sum_{j=1}^{p} \mathrm{b}_j = 0$. But this model will have $\mathrm{a}_r = 0$ for

all r, because α_i absorbs all variation among the rows, and therefore a has zero for all r. The only case where $\mathrm{a}_r \neq 0$ is when we have interaction term between row cluster and columns. Such situation is expressed as either

$$\mathrm{logit}(\phi_{ijr}) = \mu + \alpha_i + \mathrm{a}_r + \mathrm{b}_j + \lambda_{rj} \tag{3.19}$$

or with covariates

$$\mathrm{logit}(\phi_{ijr}) = \mu + \alpha_i + \mathrm{a}_r + \underline{w}_j^T(\underline{\psi} + \underline{\tau}_r) \tag{3.20}$$

If there are any clusters (i.e. $R \neq 1$), then $\sum_{r=1}^{R} \mathrm{a}_r = 0$ and $\sum_{i=1}^{n} \alpha_i z_{ir} = 0$, $\forall r$.

3.5 Model Options

We have shown some models for the linear predictor in the case of row cluster, column cluster, row cluster with row/column covariates, cluster with interaction between cluster and row/column covariates, and row standardisation. The data may be modeled by unclustered GLMs, clustering models with/without interactions, clustering models with covariates with/without interactions. Depending on the model we choose, the number of parameters varies. The list of models that may be fitted to the data is in Table 3.1.

Table 3.1: The list of the linear predictors.

Model equation for logitθ_{ij}	Number of parameters	Model description
GLM		
μ	1	Null
$\mu + \mathrm{a}_i$	n	Row effects
$\mu + \mathrm{b}_j$	p	Columns effects
$\mu + \mathrm{a}_i + \mathrm{b}_j$	$n + p - 1$	Main effects only
$\mu + \mathrm{a}_i + \mathrm{b}_j + \lambda_{ij}$	np	Saturated
Row Cluster models		
$\mu + \mathrm{a}_r$	$2R - 1$	All columns are alike
$\mu + \mathrm{a}_r + \mathrm{b}_j$	$2R + p - 2$	Column effects
$\mu + \mathrm{a}_r + \mathrm{b}_j + \lambda_{rj}$	$Rp + R - 1$	Saturated
$\mu + \mathrm{a}_r + \underline{x}_i^T \underline{\delta}$	$2R + D - 1$	Row covariates
$\mu + \mathrm{a}_r + \underline{x}_i^T \underline{\delta} + \mathrm{b}_j$	$D + 2R + p - 2$	Row covariates & column effects
$\mu + \mathrm{a}_r + \underline{x}_i^T (\underline{\delta} + \underline{\upsilon}_r) + \mathrm{b}_j$	$R + p + DR - 2$	Row covariates interaction, column effects
$\mu + \mathrm{a}_r + \underline{w}_j^T \underline{\psi}$	$2R + W - 1$	Column covariates
$\mu + \mathrm{a}_r + \underline{w}_j^T (\underline{\psi} + \underline{\tau}_r)$	$2R + WR - 1$	Column covariates interaction
$\mu + \underline{x}_i^T \underline{\delta}$	$D + 1$	Row covariates only (GLM)
$\mu + \underline{x}_i^T \underline{\delta} + \mathrm{b}_j$	$D + p$	Row covariates and column effects (GLM)
Row Standardised model		
$\mu + \alpha_i + \mathrm{a}_r + \mathrm{b}_j + \lambda_{rj}$	$n + R + Rp - 3$	Row standardisation with column effects
$\mu + \alpha_i + \mathrm{a}_r + \underline{w}_j^T (\underline{\psi} + \underline{\tau}_r)$	$n + 2R + WR - 2$	Row standardisation with column covariates

3.6 Random Starting Value

A note of caution is required when using the EM algorithm for mixture distributions. One of the disadvantages of mixture models is possible multimodal likelihood. Recall that mixture models assume that the data have more than two probability distributions (Section 1.2.2). This means that each mixture component has its own density function, $f_g(y)$ (see equation 1.1). There often exist many sub-optimal combinations of these distributions that are nevertheless a local maximum of the likelihood function. The EM algorithm may climb to one of these local maximum and terminate there. However, it will not necessarily have located the global maximum, and there is no diagnostic to indicate if an EM estimate is in fact the global maximum. But when the shape of mixture distribution is unknown and the EM algorithm is climbing-only iterative optimisation, how do we ensure that the EM algorithm has converged to the global maximum? McLachlan and Krishnan (2007) said that the EM algorithm does not guarantee the convergence at the global maximum. They reviewed a paper from Wu (1983) in which the properties of the EM algorithm were studied, and stressed that adjustments are required for convergence at the global maximum. One solution is to have several starting values for the EM iteration. However, it is still unknown how many sets of starting values are needed, how long it takes to converge, and more importantly, it still does not confirm that the EM algorithm reached at the global maximum. Karlis and Xekalaki (2003) compared several methods for selecting starting values in order to reach the global maximum in fewer iterations. They recommended that to use several sets of starting values and run only a few EM iterations without necessarily checking convergence first. Then, take the values with the largest likelihood from the initial step, and run the EM algorithm again using strict convergence rule. While this strategy seems to be sensible, their approach still does not suggest how to choose sets of random starting points. In a recent paper from Pledger et al. (2015),

the authors point out that having an uninformed set of random starting values is ineffective. Instead, they suggested to use k-means first to cluster the rows/columns, in order to have a realistic idea of group membership in the real data (called preliminary E-step). Then we can obtain estimates of suitable random starting points from this preliminary E-step. Once these informed random starting values are chosen, we run the EM algorithm to obtain the estimates. In our research, we combine the methods proposed by both papers. We run the preliminary EM algorithm with less strict convergence rule (less setting of the maximum number of iteration) in order to have good starting values followed by the concluding EM algorithm with strict convergence rule. In the concluding EM algorithm, the M-step is run twice to ensure that the EM algorithm has converged. The convergence rule for the second M-step is set to be more strict than the first M-step.

3.7 Model Selection Method

The next challenge is to find the optimal number of groups. We want the model which best presents the data including determining the number of groups. When the EM algorithm is applied, the different number of groups (R or C, for row-clustered and column-clustered, respectively) must be chosen first, and models are fitted accordingly. In order to determine the optimal number of groups, candidate models are separately fitted by the EM algorithm and compared. Commonly used approaches for the model comparison are the likelihood ratio test and information criteria. In this research, Akaike's Information Criterion (AIC, (Akaike, 1973)) is used for model selection. AIC is a relative measure of information loss, based on the likelihood function which gets penalized with the number of parameters in the model. The information loss is obtained by Kullback-Leibler distance (KLD) that measures the distance between true probability distribution and predicted probability distribution by a model (Richards, 2008). However, we can only approximate the unknown "truth", so the KLD is also approximate. Instead of KLD, AIC is used as a relative estimation of the KLD. The formula for AIC is

$$\mathrm{AIC} = -2\ell_c + 2K$$

where ℓ_c is the maximised log of the complete likelihood of the data (e.g. equation 3.8), K is the number of independent parameters in the model. AIC is a relative measure of information loss so we can only meaningfully interpret AIC value differences between the models. It is not an overall goodness of fit measure. The model with the smallest AIC is the model with the minimum expected KLD, thereby the best model. However, when the difference in AIC between the models is small, typically less than 3, the model with less parameters is chosen as the best model from the set of models that have been estimated.

3.8 Simulation Study

3.8.1 Introduction

In this section, we perform a simulation study on a set of models described in Section 3.3. We must be sure that we can reliably recover the parameters of a known model using the EM algorithm described above, in order to be confident that the estimation procedure will identify the correct model in a real dataset. We also need to be certain that our models satisfy underlying assumptions of the data distribution and the constraints, so that we can apply our models to the real-world data with confidence. But how do we achieve the evaluation of the models with the real data alone? It is not possible to evaluate our model performance with the data when the true parameters are unknown. For these reasons, we have carried out a set of simulation studies where we attempt to recover the correct parameter values for a set of simulated datasets when the true parameter values are known. Simulation is a numerical technique for conducting experiments and scenarios on the computer, and a rational way of investigating the model performance (Burton et al., 2006). The goal of the simulation study is to evaluate the reliability of the models and their robustness.

In this section, we conduct a simulation study on our likelihood-based clustering models. We first outline the process of the simulation study and then present our results.

3.8.2 General Procedure

All simulation studies involve generating independent simulated datasets. To do so we specify the true parameters first. A dataset is then generated using the true parameters and then a set of proposed models are fitted to the dataset. If the model is a good fit, we must be able to recover the true population parameters. In addition to creating true parameters, the specification of the simulation study includes the choice of the number of the

groups, the sample size, and the number of replicates for model runs. At each run, we fit the model with several random starting values that are chosen by the method described in Section 3.6. This procedure is repeated a number of times.

3.8.3 Outline of Simulation Study

Here we describe the process of the simulation study in detail. The method of the simulation study can be described in the following steps.

Notation

- n; sample size
- R; the number of clusters, $1 \leq R < n$
- p; the number of columns in a dataset, $p = 5$, $j = 1, \cdots, p$
- D; the number of covariates, $d = 1, \cdots, D$
- V; the number of replicates, $v = 1, \cdots, V$
- Q; the number of random starting points, $q = 1, \cdots, Q$

Step 1. Specify the model and true values for the parameters

- Select the model
- Select R from $\{2, 3, 4, 5\}$ and n from $\{100, 200, 500\}$
- Define the true parameters for the selected model.
  For example of simple row-clustered model (equation 3.3), we set μ, $\underline{\mathrm{a}}$, $\underline{\mathrm{b}}$. If we set $R = 2$ and $p = 5$, we need to specify $\phi = \mu, \mathrm{a}_1, \mathrm{b}_1, \mathrm{b}_2, \mathrm{b}_3, \mathrm{b}_4$. With the constraints, we generate a complete set of parameters; $\mathrm{a}_2 = -\mathrm{a}_1$, and $\mathrm{b}_5 = -\mathrm{b}_1 - \mathrm{b}_2 - \mathrm{b}_3 - \mathrm{b}_4$.
- Define the mixture weights, π_r. For each model we have two scenarios: Scenario 1 is the case when π_r are equal among R, and Scenario 2 is the situation where one π_r has a large value

(i.e. one group dominates in the population). When $R = 2$, we set $\pi_1 = \pi_2 = 0.5$ in Scenario 1, and $\pi_1 = 0.95, \pi_2 = 0.05$ in Scenario 2.

Step 2. Generate a dataset based on the true parameters

- The linear predictor for the row-clustered with column effects model without interaction (see equation 3.3) is

$$\text{logit}(\phi_{rj}) = \mu + \mathrm{a}_r + \mathrm{b}_j$$

For each $i = 1, \cdots, n$, we assign $r \sim \text{Discrete}(\pi_1, \cdots, \pi_r)$, then set $\theta_{ij} = \phi_{rj}$, and then for $j = 1, \cdots, p$, $y_{ij} \sim \text{Benoulli}(\theta_{ij})$ We then generate a $n \times p$ dataset based on $Y \sim \text{Bernoulli}(\Theta)$

Step 3. Fit the model using the EM algorithm with Q random starting points

For any values of R,

- Set Q. We set $Q = 5$. That means that we run the model for five times for one generated dataset.
- We obtain five sets of parameter estimates. When $R = 2$, we obtain five sets of $\{\mu, \mathrm{a}_1, \mathrm{b}_1, \mathrm{b}_2, \mathrm{b}_3, \mathrm{b}_4\}$ as well as the maximised log-likelihood value.
- We report the one with the highest log-likelihood as parameter estimates.

Step 4. Replicates

- Repeat step 2 and 3 for V times. We set $V = 100$. That is, we run the simulation for $Q \times V = 5 \times 100$ for each model in a particular situation.

Step 5. Report the results

- At the end of a simulation study, we have 100 sets of parameter estimates. We take the mean and standard deviation of these estimates and report these values and compare with the true parameters. We report the mean as the parameter estimate and the standard deviation as the standard error (S.E.).

3.8.4 Label Switching

One thing to note when reporting the results from the simulation study is label switching. Label switching is defined as re-labelling of the mixture components while the likelihood of the data is invariant (Stephens, 2000). More specifically, taking $R = 2$ row cluster model as a example, estimates from some EM algorithm during $V = 100$ runs may produce estimates with $\widehat{\mathrm{a}}_1 > \widehat{\mathrm{a}}_2$, while other EM algorithm runs produce with $\widehat{\mathrm{a}}_1 < \widehat{\mathrm{a}}_2$. Our true values always have $\mathrm{a}_1 > \mathrm{a}_2$, and we need to relabel the corresponding values so that we have $\widehat{\mathrm{a}}_1 > \widehat{\mathrm{a}}_2$. Label switching occurs naturally in mixture models where the likelihood can be multi-modal (McLachlan and Peel, 2004; Stephens, 2000). McLachlan and Peel (2004) states that switching of the component label is not a problem but it is important to consider the effect of the label switching in a simulation study. Since we know the true values for the parameters in the simulation, we were able to identify iterations with label switching and fix them accordingly. Identification of the label switching is impossible in the real-world data, but it is also unnecessary. In this research, we take the note of the occurrence of label switching.

3.8.5 Results

We carried out simulation studies for the row-clustered model without interaction (equation 3.3), and the row cluster with row level covariate model (equation 3.13). Results from the simulation study after adjusting for label switching for the row-clustered model ($R = 2$) are in Table 3.2

and 3.3 for Scenario 1 and 2, respectively. Results from the simulation study after adjusting for label switching for the row cluster with row level covariate model ($R = 3$) are presented in Table 3.4 and 3.5 respectively for Scenario 1 and 2. For the both models in Scenario 1, the estimates of the parameters became close to the true values, and as expected the variability (S.E.) decreases with increasing sample size n (Table 3.2, 3.4). The estimates for $\underline{\mathrm{b}}$ are very close to the true values with small amount of the variability, while the estimates for the μ and $\underline{\mathrm{a}}$ show little more variability. The ability to recover the true π_r values is different between the models in Scenario 1. The relative sampling error (RSE) is calculated for π_r to evaluate the quality of the estimate ($\mathrm{RSE}(\widehat{\pi}_r) = \frac{\mathrm{S.E.}(\widehat{\pi}_r)}{\widehat{\pi}_r}$)). For the row-clustered model when $n = 500$, $\mathrm{RSE}(\widehat{\pi}_2) = \frac{0.1421}{0.2012} = 0.71$. It is high (>70%), indicating our estimates for π_r are very poor. The RSEs for the row cluster with the covariate model are better than that of the row-clustered model. But the RSE for π_3 when $n = 500$ is 16%, again suggesting the estimate for π_r are poor (Table 3.4). Similarly, the estimates for $\underline{\mathrm{a}}$ have much larger variability in the row-clustered model ($\mathrm{S.E.} = 0.9737$, for both a_1 and a_2, Table 3.2), compared with that of in the row cluster with row level covariate model (S.E.= 0.1854, 0.2355, 0.1724 for a_1, a_2, a_3, respectively (Table 3.4). Therefore, in the situation when π_r has the equal value between the clusters, the row cluster with row level covariate model can perform better than the row-clustered model. It may be due to that the number of parameter is too few to explain the whole dataset, or the simpler model requires bigger sample size to evaluate its performance.

The row-clustered model appears to be performing poorly in Scenario 2. The estimates for $\underline{\mathrm{b}}$ are still close to the true values, but the estimate for the other parameters get further from the true values. Not only the estimates for μ and $\underline{\mathrm{a}}$ get further from the true values, but also the standard error of these parameter estimates showed small change with the size of n. The 95% confidence intervals when $n = 500$ for a_1 in Scenario 1 is (-1.0458,

2.7712), and (-1.6572, 3.6892) in Scenario 2. Both intervals have the true value of $\mathrm{a}_1 = 0.3749$, but the range of the interval in Scenario 2 is much larger than that in Scenario 1. For the row cluster with row level covariate model, the simulation results suggests there is no noticeable deterioration in the performance (Table 3.5). Although the estimates for a_r get far from the true values, the standard errors are similar to that of in Scenario 1, suggesting that the range of the 95% confidence intervals are relatively similar between the scenarios. The parameter δ is well estimated in the both scenarios ($\widehat{\delta} = 1.7949$ (S.E. $= 0.1929$), and 1.7832 (S.E. $= 0.1752$), for Scenario 1 and 2, respectively (Table 3.4, and 3.5).

The ability to estimate true π_r in Scenario 2 is similar to that of in Scenario 1. The RSEs of $\widehat{\pi}_r$ for the row-clustered model got worse, resulting in 92% when $n = 500$ (Table 3.3). Likewise, the RSEs for the row cluster model with row covariate model also became worse, about 20%. It can be still said that the ability to recover the true π_r is still better for the row cluster with row covariate model, but the quality of the estimates are poor for the both model. The parameter μ seems to be better estimated for the row cluster with row covariate model in the both scenarios.

Summary

In summary, the results from the simulation study are not really satisfactory. We did have improved estimation when the sample size is large (S.E.s are reduced), indicating having a large sample size is better for the overall estimation. The results also show that inclusion of the covariate is useful additional information that can improve the identification of groups. We also noted that the estimations for a_r and π_r are poor in most cases. It may be due to the particular data we generated. If the generated data for each run had many 0's or 1's, that means all species are either rare or very abundant in a ecological situation. It would make the simulation longer to converge and produce poor estimates when the data are dominated by

either 0 or 1. It is necessary to check generated dataset by taking the mean of the $\{y_{ij}\}$ is not close to 0 or 1, so that we could have better estimates from the simulation study.

Table 3.2: Simulation results for the row clustered model in Scenario 1 when $R = 2$

		$n = 100$		$n = 200$		$n = 500$	
Parameters	true	mean	S.E.	mean	S.E.	mean	S.E.
μ	0.1308	0.0444	1.6975	-0.0860	1.571	0.0120	1.2223
a_1	0.3749	1.0841	1.3933	1.017	1.2894	0.8627	0.9737
a_2	-0.3749	-1.0841	1.3933	-1.017	1.2894	-0.8627	0.9737
b_1	-0.0027	0.0250	0.1885	-0.0099	0.1379	-0.0085	0.0810
b_2	0.1248	0.1240	0.1826	0.1358	0.1254	0.1370	0.0843
b_3	0.4376	0.4534	0.2027	0.4475	0.1409	0.4375	0.0833
b_4	-0.9028	-0.9178	0.1813	-0.9136	0.1358	-0.9088	0.0812
π_1	0.5000	0.7572	0.1715	0.7658	0.1725	0.7988	0.1421
π_2	0.5000	0.2428	0.1715	0.2342	0.1725	0.2012	0.1421

Table 3.3: Simulation results for the row clustered model in Scenario 2 when $R = 2$

		$n = 100$		$n = 200$		$n = 500$	
Parameters	true	mean	S.E.	mean	S.E.	mean	S.E.
μ	0.1308	0.6170	1.6975	0.6726	2.0090	0.8183	1.6361
a_1	0.3749	1.2556	1.732	1.1925	1.6704	1.016	1.3639
a_2	-0.3749	-1.2556	1.732	-1.1925	1.6704	-1.016	1.3639
b_1	-0.0027	0.0234	0.1885	0.0120	0.1219	0.0018	0.0767
b_2	0.1248	0.1378	0.1826	0.1172	0.1350	0.1334	0.0782
b_3	0.4376	0.4677	0.2027	0.4242	0.1342	0.4416	0.0834
b_4	-0.9028	-0.9616	0.1813	-0.9068	0.1513	-0.9212	0.0743
π_1	0.9500	0.7510	0.19	0.7706	0.1944	0.7867	0.1955
π_2	0.0500	0.2490	0.19	0.2294	0.1944	0.2133	0.1955

Table 3.4: Simulation results for row clustered with a covariate model in Scenario 1 when $R = 3$

		$n = 100$		$n = 200$		$n = 500$	
Parameters	true	mean	S.E.	mean	S.E.	mean	S.E.
μ	0.1308	0.0734	0.2023	0.0768	0.1729	0.0612	0.1120
a_1	0.3749	0.4310	0.3531	0.3366	0.2237	0.3758	0.1854
a_2	0.1080	-0.0442	0.2738	0.0030	0.2544	0.0172	0.2355
a_3	-0.4829	-0.3868	0.3118	-0.3395	0.2642	-0.3930	0.1724
b_1	-0.0027	0.0290	0.2269	-0.0053	0.1527	-0.0027	0.0995
b_2	0.1248	0.1167	0.2408	0.1200	0.1431	0.1249	0.0856
b_3	0.4376	0.4393	0.2433	0.4321	0.1715	0.4626	0.0975
b_4	-0.9028	-0.9214	0.2183	-0.9073	0.1325	-0.9131	0.0872
δ	1.7846	1.8405	0.4027	1.8156	0.2880	1.7949	0.1929
π_1	0.3330	0.4342	0.0482	0.4461	0.0783	0.4440	0.0493
π_2	0.3330	0.3297	0.0383	0.3221	0.0783	0.3277	0.0355
π_3	0.3330	0.2360	0.0409	0.2319	0.0418	0.2283	0.0368

Table 3.5: Simulation results for row clustered with a covariate model in Scenario 2 when $R = 3$

		$n = 100$		$n = 200$		$n = 500$	
Parameters	true	mean	S.E.	mean	S.E.	mean	S.E.
μ	0.1308	0.4352	0.2473	0.4358	0.1696	0.4440	0.1128
a_1	0.3749	0.2035	0.2510	0.2184	0.2272	0.1798	0.1870
a_2	0.1080	0.0003	0.1935	-0.0054	0.1948	-0.0027	0.1597
a_3	-0.4829	-0.2037	0.2691	-0.2130	0.2311	-0.1772	0.1949
b_1	-0.0027	0.0279	0.2461	0.0145	0.1796	-0.0008	0.1115
b_2	0.1248	0.1034	0.2414	0.1320	0.1671	0.1269	0.0986
b_3	0.4376	0.4802	0.2979	0.4336	0.1816	0.4538	0.0928
b_4	-0.9028	-0.9271	0.2268	-0.9298	0.1534	-0.9042	0.0928
δ	1.7846	1.8525	0.4294	1.8156	0.2880	1.7832	0.1752
π_1	0.9500	0.4805	0.0524	0.5045	0.0528	0.4986	0.0502
π_2	0.0250	0.3217	0.0485	0.3155	0.0516	0.3157	0.0471
π_3	0.0250	0.1978	0.0402	0.1800	0.0398	0.1856	0.0367

3.9 The Use of the clustglm Function

The main disadvantage with the EM algorithm is computational time. As McLachlan and Krishnan (2007) said in their book, the EM algorithm can be slow to converge and moreover, it can fail to converge. Reasons why the EM algorithm takes so long to converge are that the function used in the M-step, optim, is a time consuming procedure, and the starting values can be far away from the global likelihood maximum. In order to reduce computational time, we use new computing package clustglm, developed by Pledger et al. (2015). This programme does simple row/column clustering and biclustering, using the glm function internally instead of optim. The switch from optim to glm immensely reduces computational time, because the glm function in **R** is fast and stable. In addition, the clustglm enable us to generate possible starting values that are close enough to a local peak of the likelihood by using three clustering methods that are also in **R**. They are the kmeans function that is used to generate the posterior probability matrix for row/column clustered models, the hclust function, which is a standard agglomerative hierarchical clustering, and lastly the diana function, a method of divisive hierarchical cluster analysis (Pledger et al., 2015). The values with the highest likelihood is chosen as a starting value to initialise the EM algorithm until convergence. Another big advantage is that the clustglm function can incorporate covariates. Because this package is based on the glm function, covariates can easily be included as predictor variables with clustering. This function can produce three information criteria; AIC, AICc (corrected AIC, Akaike, 1973), and BIC (Bayesian information criterion, Schwarz et al., 1978). The clustglm allow us to explore the data by fitting many models with a short computational time and compare between the models, therefore we use this function for the data analyses.

One minor hiccup is that the clustglm uses different model specifica-

tion from our specification in Section 3.3. In order to use this function, we need to arrange the dataset and models accordingly. The original dataset is the $n \times p$ matrix. This is arranged to a format so that `glm` function can be fitted. For example, if we have 4×2 matrix which consists binary response, the dataset to fit `glm` is

$$\underset{\text{original}}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}} \quad \rightarrow \quad \underset{\text{for } \texttt{clustglm}}{\begin{bmatrix} \mathcal{Y} & \mathcal{A} & \mathcal{B} \\ 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 3 & 1 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 2 \\ 1 & 3 & 2 \\ 1 & 4 & 2 \end{bmatrix}}$$

where $\mathcal{Y}$ column contains the response variable of the original data matrix, and $\mathcal{A}$ and $\mathcal{B}$ indicate row and column. Covariates can be easily added to this data frame by making additional columns. Our models also need to be arranged in order to accommodate this new data frame. As the `clustglm` is based on the `glm` function, we need to express the generalized linear regression formula. The model with main effects terms (using rows and/or columns as factors) is delivered via α_i and β_j, and cluster term is expressed in γ_{rc}. The general model equation in `clustglm` is

$$\text{logit}\, \phi_{rj} = \nu + \alpha_i + \beta_j + \gamma_{rc} \tag{3.21}$$

where

- ν is the intercept of the regression

- α_i are row main effects ($i = 1, \cdots, n$), $\quad \sum_{i=1}^{n} \alpha_i = 0$
- β_j are column main effects ($j = 1, \cdots, p$), $\quad \sum_{j=1}^{p} \beta_j = 0$
- γ_{rc} are the row/column cluster effects ($r = 1, \cdots, R$, $c = 1, \cdots, C$), $\sum_{r=1}^{R} \gamma_{rc} = 0, \forall c$, and $\sum_{c=1}^{C} \gamma_{rc} = 0, \forall r$

The row/column cluster term (e.g. a_r for the row clustering in equation 3.17) is expressed in γ_{rc}, and the number of row/column cluster is specified in r and c in γ. Covariates can be included in γ_{rc} term by replacing r or c with covariates.

3.9.1 Example

Here, we illustrate how our models are transformed to `clustglm` model formula.

Simple Row-clustered Model

Recall our model in simple row clustering case. Our model equation (3.3) is

$$\text{logit}\, \phi_{rj} = \mu + \mathrm{a}_r + \mathrm{b}_j + \lambda_{rj}$$

where we set $\lambda_{rj} = 0$. In order to use `clustglm`, the model above is transformed to the following equation

$$\text{logit}\, \phi_{rj} = \nu + \beta_j + \gamma_{r1} \tag{3.22}$$

where γ_{r1} is $R \times 1$ vector of row cluster effect. We set $\alpha_i = 0, \forall i$ and $C = 1$ in equation (3.21) as we don't cluster columns in this situation. The number of parameters is the same ($2R + p - 1$). The parameter estimates in our model (3.3) are computed by

$$\begin{aligned}\widehat{\mu} &= \nu + \overline{\gamma}_{..} \\ \widehat{\mathrm{a}}_r &= \gamma_{r.} - \overline{\gamma}_{..} \\ \widehat{\mathrm{b}}_j &= \beta_j\end{aligned}$$

Row Cluster with Column Covariates Model

Our model in the case of row cluster with column level covariates (3.16) is

$$\mathrm{logit}(\phi_{ijr}) = \mu + \mathrm{a}_r + \underline{w}_j^T \underline{\psi}$$

This is written in the following

$$\mathrm{logit}\,\phi_{ijr} = \nu + \underline{w}_j^T \underline{\psi} + \gamma_{r1}$$

The column effects b_j in the equation (3.22) is replaced by the column levels covariates, w_j^T. The parameter estimate for ψ is obtained from the model output. The number of parameters here is $2R + W - 1$.

Row Cluster with Row Standardisation

We presented row cluster with row standardisation models earlier (Section 3.4). These models are much more easily fitted in the `clustglm`. In the case of row clustering with cluster-column interaction and row standardisation, the model is

$$\mathrm{logit}(\phi_{ijr}) = \nu + \alpha_i + \beta_j + \gamma_{rj} \tag{3.23}$$

where $\sum_{i=1}^{n} \alpha_i = 0$, $\sum_{j=1}^{p} \beta_j = 0$, $\sum_{r=1}^{R} \gamma_{rj} = 0, \forall j$, and $\sum_{j=1}^{p} \gamma_{rj} = 0, \forall r$.

Chapter 4

Application to Trawl Survey Data

4.1 Introduction

In this chapter, we present the results from fitting models using `clustglm` to the fisheries trawl survey data. We have selected a set of presence/absence dataset from four survey trips, tan0201, tan1101, tan1201, and tan1301. These indicate an early year (2002) and three late years (2011, 2012, and 2013). The number of species caught varies year to year. Species considered for the analysis are the ones that were caught in all four trips, so we can minimise the risk of our analysis being affected by the presence of rare species, and can easily detect year to year changes. There are a total of 61 fish or shark species in the reduced datasets ($n = 61$). We first fit the simple row-clustered model, followed by the row/column level covariate models. The results from these models motivate us to propose two further models, which combine row and column level effects in one model. We select the best model using the information criterion AIC, interpret the results from the best model, and discuss possible biological/ecological explanations for our results. We use visualization techniques to present the results. Species cluster memberships, and the values of row cluster effects and the column effects are presented in graphical displays throughout this chapter. Geographical and depth distributions for the species studied in

our analysis were illustrated by Anderson et al. (1998).

4.2 Models Fitted

There are eleven models fitted for each of the four datasets. A suite of the models are listed in Table 4.1. We start with the simple row-clustered model (SRC, no predictor to explain the data), then add depth and bottom water temperature effect as main effect (Models 6, 7, and 8, Table 4.1), and the bottom temperature and depth interaction terms with row clusters (Models 1, 2, 3, 4, and 5, Table 4.1). We choose depth and bottom water temperature for the column covariates because there are thought to be important environmental factors to explain the species distribution in the New Zealand waters (Francis et al., 2002; Leathwick et al., 2003). We also fit the row standardised model (Model RS) and row covariate model (Model RC). For each model, $R = 2, \cdots, 10$ are investigated.

4.3 Model Selection

We have fitted a suite of models from the simple row-clustered model to the row cluster models with column/row covariates. For each model, the information criterion AIC is computed and the results from the tan0201 dataset are summarised in Table 4.2. An issue we encountered when selecting the best model according to the AIC is that the model selected may not necessarily give us results that are easily interpretable. For example, $R = 9$ is selected for Model 1. However, when the 61 species are clustered into nine groups, four groups had less than five species. For example, one group had only two species, ribaldo (*Mora Moro*, RIB) and shovelnose dogfish (*Deania calcea*, SND), which are common species on the Chatham Rise (Anderson et al., 1998). It is difficult and impractical to describe biological/ecological characteristics from such small groups, unless the species in that group are very different from other groups. For this reason, we set

Table 4.1: The model specifications for the trawl survey data. n = sample size ($n = 61$ for all datasets), p is the number of columns, R is the number of the groups, and W is the number of the column covariates.

Model Type	**Model Specification**	**npar**
Simple row-clustered (SRC)	$\mu + \mathrm{a}_r + \mathrm{b}_j$	$2R + p - 2$
Model 1	$\mu + \mathrm{a}_r + \text{depth}_j(\psi_1 + \tau_r) + \text{temp}_j(\psi_2 + \tau_r)$	$WR + 2R - 1$ ($W = 2$)
Model 2	$\mu + \mathrm{a}_r + \text{depth}_j\psi_1 + \text{temp}_j(\psi_2 + \tau_r)$	$WR + 2R - 1$ ($W = 1$)
Model 3	$\mu + \mathrm{a}_r + \text{depth}_j(\psi_1 + \tau_r) + \text{temp}_j\psi_2$	$WR + 2R - 1$ ($W = 1$)
Model 4	$\mu + \mathrm{a}_r + \text{temp}_j(\psi + \tau_r)$	$3R + W - 2$ ($W = 1$)
Model 5	$\mu + \mathrm{a}_r + \text{depth}_j(\psi + \tau_r)$	$3R + W - 2$ ($W = 1$)
Model 6	$\mu + \mathrm{a}_r + \text{depth}_j\psi_1 + \text{temp}_j\psi_2$	$2R + W - 1$ ($W = 2$)
Model 7	$\mu + \mathrm{a}_r + \text{temp}_j\psi$	$2R + W - 1$ ($W = 1$)
Model 8	$\mu + \mathrm{a}_r + \text{depth}_j\psi$	$2R + W - 1$ ($W = 1$)
Row standardised (RS)	$\mu + \alpha_i + \mathrm{a}_r + \mathrm{b}_j$	$2R + n + p - 3$
Row covariate (RC)	$\mu + \mathrm{a}_r + \text{body length}_i\delta + \mathrm{b}_j$	$2R + p - 1$

a cutoff minimum value for $\widehat{\pi}_r$. If there is any group with its $\widehat{\pi}_r < 0.08$, equivalent to fewer than five species, we check the species composition in that group. If it does not have any distinct characteristics (e.g. very rare deepwater sharks), we reduce the number of R until all $\widehat{\pi}_r$ values are greater than the cutoff value. Table 4.2 shows that the model we selected (highlighted in red) is not always the model with the minimum AIC. We show the final model selection results for the tan1101, tan1201, and tan1301 data in Table 4.3. For all datasets, the optimal model overall is highlighted in blue.

The optimal model overall is Model 1 for all datasets. The number of groups selected from this model is $R = 6$, 7, 5 and 7 for the tan0201, tan1101, tan1201, and tan1301 data, respectively. This model has both depth and bottom temperature covariates interacting with the row clusters, suggesting that they explain the data differently. For all datasets, Models 3 and 5 have similar AIC (for example, AIC $= 1644.10$, 1645.23 for Models 3 and 5, for the tan1101 data, Table 4.3), and so does Models 2 and 4 (AIC $= 1756.30$, 1757.60 for Models 2 and 4, for the tan1101 data, Table 4.3). Models 3 and 5 both include depth interaction term, and Model 3 has an extra covariate, bottom temperature. Likewise, both Models 2 and 4 have the bottom temperature interaction term, with Model 2 having an extra covariate, depth. The AIC's are very similar for Models 3 and 5, also for Models 2 and 4, indicating that fitting one interaction term might be enough to explain the data. However, when we compare Model 1 with Models 4 and 5, the AIC for Model 5 is close to that for Model 1. This indicates that addition of bottom temperature terms improved the model only a small amount. Putting together, having depth and bottom temperature with interaction effects best presents the data, but depth seems to be the strongest explanatory variable.

Table 4.2: A suite of models fitted for the tan0201 data. The minimum AIC in each model is shown in bold, and the model selected in each model option is shown in red. The overall best model is highlighted in blue. Note when the difference of the AIC is less than 3 between the two adjacent models, we take the simpler model.

Model	R	npar	AIC	the minimum π_r
Simple row-clustered model	3	31	1710.70	0.16
	4	33	**1708.93**	0.12
	5	35	1712.89	0.10
Model 1	6	23	1343.77	0.13
	7	27	1341.55	0.06
	8	31	1318.31	0.05
	9	35	**1312.83**	0.05
	10	39	1323.75	0.01
Model 2	5	15	1539.43	0.13
	6	18	**1533.12**	0.05
	7	21	1572.39	0.05
Model 3	6	18	1372.31	0.09
	7	21	1354.59	0.07
	8	24	**1347.78**	0.05
	9	27	1353.08	0.03
Model 4	5	14	1537.47	0.13
	6	17	1531.14	0.05
	7	20	**1526.08**	0.05
	8	23	1525.81	0.01
Model 5	7	20	1358.85	0.08
	8	23	**1352.87**	0.05
	9	26	1358.21	0.03
Model 6	3	7	**1732.60**	0.16
	4	9	1731.73	0.12
	5	11	1735.73	0.01
Model 7	3	6	**1730.88**	0.16
	4	8	1730.02	0.12
	5	10	1734.02	0.01
Model 8	3	6	**1733.99**	0.16
	4	8	1733.16	0.12
	5	10	1737.164	0.01
Model RS	2	89	**1616.35**	0.04
	3	91	1620.35	0.04
	4	93	1624.35	0.04
Model RC	2	30	1777.59	0.49
	3	32	**1712.26**	0.16
	4	34	1709.37	0.11

Table 4.3: A suite of best models from each model option for the tan1101, tan1201, tan1301 data. The overall best model in each year is highlighted with blue.

Dataset	Model	R	npar	AIC	minimum π_r
tan1101	SRC	4	38	2223.52	0.16
	Model 1	7	27	1629.07	0.08
	Model 2	6	18	1756.30	0.11
	Model 3	7	21	1644.10	0.08
	Model 4	6	17	1757.60	0.11
	Model 5	7	20	1645.23	0.08
	Model 6	4	9	2240.35	0.16
	Model 7	4	8	2238.49	0.16
	Model 8	4	8	2242.38	0.16
	Model RS	2	94	2135.91	0.04
	Model RC	4	39	2221.30	0.15
tan1201	SRC	4	40	2358.55	0.17
	Model 1	5	19	1869.45	0.16
	Model 2	5	15	1958.88	0.13
	Model 3	5	15	1883.84	0.15
	Model 4	5	14	1957.68	0.13
	Model 5	5	14	1887.25	0.15
	Model 6	4	9	2365.25	0.17
	Model 7	4	8	2364.60	0.17
	Model 8	4	8	2367.68	0.17
	Model RS	2	96	2255.22	0.01
	Model RC	4	41	2361.10	0.16
tan1301	SRC	3	38	2386.23	0.18
	Model 1	7	27	1786.66	0.11
	Model 2	6	18	1878.70	0.10
	Model 3	7	21	1808.31	0.08
	Model 4	6	17	1881.63	0.10
	Model 5	7	20	1810.29	0.08
	Model 6	3	7	2402.00	0.18
	Model 7	3	7	2400.06	0.18
	Model 8	4	7	2403.21	0.18
	Model RS	2	96	2299.77	0.01
	Model RC	3	39	2386.52	0.17

4.4 Species Membership Results

We use visualization techniques to present a posteriori membership $\{\widehat{z}_{ir}\}$. This is the probability that species i is in row cluster r. The numerical expression is

$$\widehat{z}_{ir} = P(i \in r|\mathbf{Y}) \tag{4.1}$$

In the case of row cluster with covariates model, $\widehat{z}_{ir}$ is calculated by equation (3.15).

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \widehat{\phi}_{ijr}^{y_{ij}}(1-\widehat{\phi}_{ijr})^{1-y_{ij}}}{\sum_{u=1}^{R}\{\widehat{\pi}_u \prod_{l=1}^{p} \widehat{\phi}_{ilu}^{y_{il}}(1-\widehat{\phi}_{ilu})^{1-y_{il}}\}}$$

Note that $\sum_{r=1}^{R} \widehat{z}_{ir} = 1$ for all i. In these visualisations we assign each species to a single row group for simplicity. For the each row of $\mathbf{Z}$, the highest z_r values is selected for the membership. Each species i is assigned uniquely to a row group with the highest posterior probability $\widehat{z}_{ir}$. Figure 4.1 is the visualisation of species group memberships for the tan0201 data, ordered by the result from SRC model. The best R selected from each model is $R = 3, 6, 5, 6, 5, 7, 3, 3, 3$ respectively for SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. Group 1 of SRC model in Figure 4.1 (the bottom group, shown in pale blue) contains hoki (HOK) and ling (LIN), which are the main target species for the trawl survey. The species in this group were caught at almost all strata and occur most frequently around 500 m depth. Their distributions are concentrated on a broad area of the Chatham Rise, and are also common in the other areas of the New Zealand's EEZ, such as the sub-antarctic region or west coast region (Anderson et al., 1998). Species in Group 2 of SRC model are also frequently observed, but they are less abundant species on the Chatham Rise (Anderson et al., 1998). Group 3 of SRC model in Figure 4.1 is the least frequently observed species group. The clusters from SRC and

the models with covariates as the main effects only (Models 6, 7, and 8) are identical, indicating that including the covariates only as main effects in the model is not useful (M6, M7, and M8, Figure 4.1)

Figure 4.2 shows exactly the same group membership results for the tan0201 data but they are ordered by the membership result from Model 1 (M1). The membership results from Model 3 (M3) and Model 5 (M5) look very consistent with each other. Model 5 has seven groups, but one group (Group 7, in dark blue) has only five species. Group 7 in Model 5 is a subset of Group 6 in Model 3. They both have main effect of depth and interaction with row groups. The only difference between them is that Model 3 includes bottom water temperature as a main effect. It indicates that temperature as a main effect alone (without interaction) does not have power as an explanatory variable when depth (with interaction) is already in the model. The membership in Models 2 (M2) and 4 (M4) are also identical to each other. They both have main effect of bottom water temperature and interaction with row groups. The only difference between them is that Model 2 includes depth as a main effect. It suggests that depth as a main effect alone does not have power as an explanatory variable when bottom temperature (with interaction) is already in the model. The membership results from Model 4 and Model 5 are different, suggesting depth and bottom temperature explain the data differently. The memberships from Model 1 (M1) are consistent with Models 3 and 5. This agrees with the results from the model selection (Section 4.3) that depth is the most powerful predictor to explain the data.

The membership result for the tan1101 data is shown in Figure 4.3 (memberships ordered by SRC), and 4.4 (ordered by Model 1). The number of groups R selected for each model is $R = 4, 7, 6, 7, 6, 7, 4, 4, 4$ respectively for SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The value of R increased for all models com-

pared with the results from the tan0201 data. Similar to the tan0201 results, species are clustered by its frequency of occurrence by SRC model. The difference from SRC for the tan0201 data is that species are clustered into four groups for the tan1101 data by SRC. Group 1 species are most frequently occurred species and Group 4 species are least frequently observed species. Species in Groups 2 and 3 also frequently occur, however the difference between them are not clear. Figure 4.4 shows the same group membership results in the tan1101 data ordered by the membership result from Model 1 (M1). Again the membership results from Model 3 (M3) and Model 5 (M5) look very consistent with each other, and so does Models 2 and 4. The memberships from Model 1 is slightly different from that of Models 3 and 5 but still very similar. Results for the tan1101 data again suggest that depth is the strongest predictor.

The membership results from the tan1201 data (Figure 4.5 (ordered by SRC) and 4.6 (ordered by Model 1)), and from the tan1301 data (Figure 4.7 (ordered by SRC) and 4.8 (ordered by Model 1)) also show similar results from the tan0201 data and the tan1101 data. For the tan1201 data, the value of R selected for each model is $R = 4$, 5, 5, 5, 5, 5, 4, 4, 4 from SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8, respectively. This is the only year that species are clustered into five groups by all models with column covariates (Models 1, 2, 3, 4, and 5). For the tan1301 data, the value of R selected for each model is $R = 3$, 7, 6, 7, 6, 7, 3, 3, 3. The models with depth and interaction between depth and row groups (Models 1, 3, and 5) resulted in having more groups, indicating that depth can explain the data better.

Overall, the more information we add in the model, the larger R is selected. These figures show the cluster structure changes after we add covariates in the model, confirming that the covariates we use (depth and bottom water temperature) explain variations in the data. The member-

ship visualisations reflect what we found from the model comparison; Depth is the strongest predictor to explain the data. The number of groups selected for each model is different across the years, which may reflect changes in frequency of occurrence, or other environmental factors affecting the preferred location for each species.

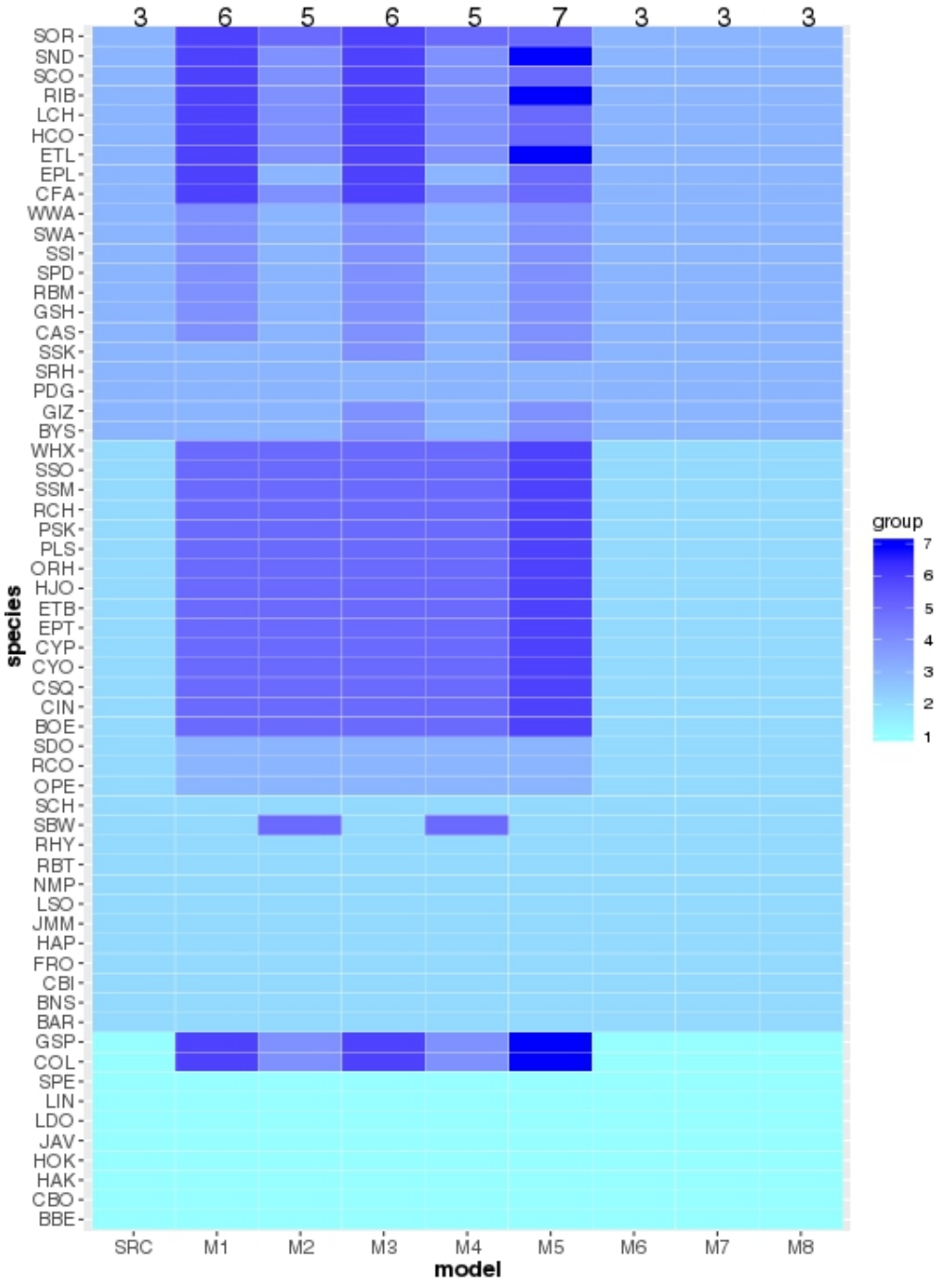


Figure 4.1: Plot showing species group memberships for the tan0201 data from a set of models shown in Table 4.1. The membership is ordered by result from SRC. From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

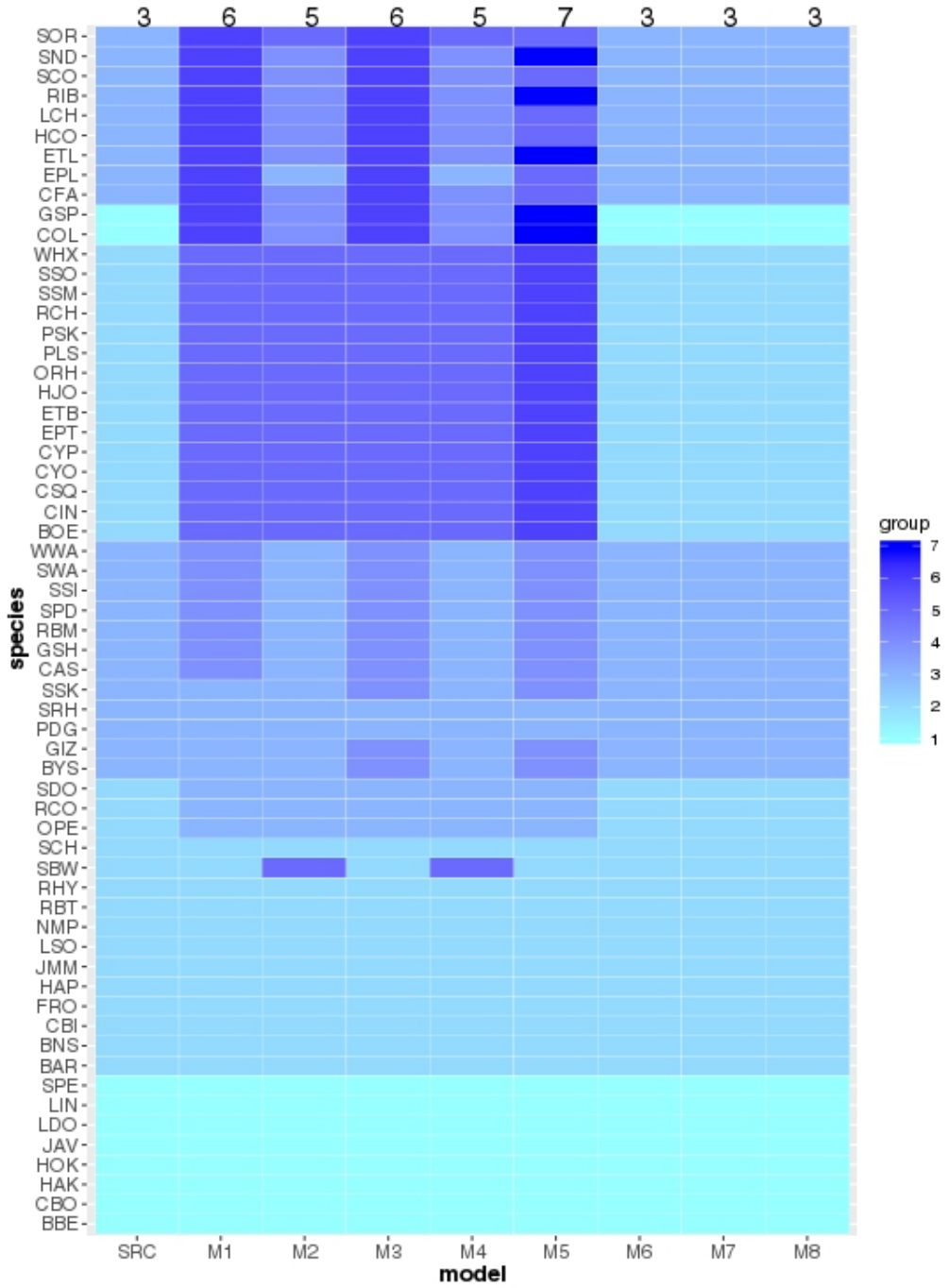


Figure 4.2: Plot showing species group memberships for the tan0201 data from a set of models shown in Table 4.1. The membership is ordered by result from Model 1 (M1). From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

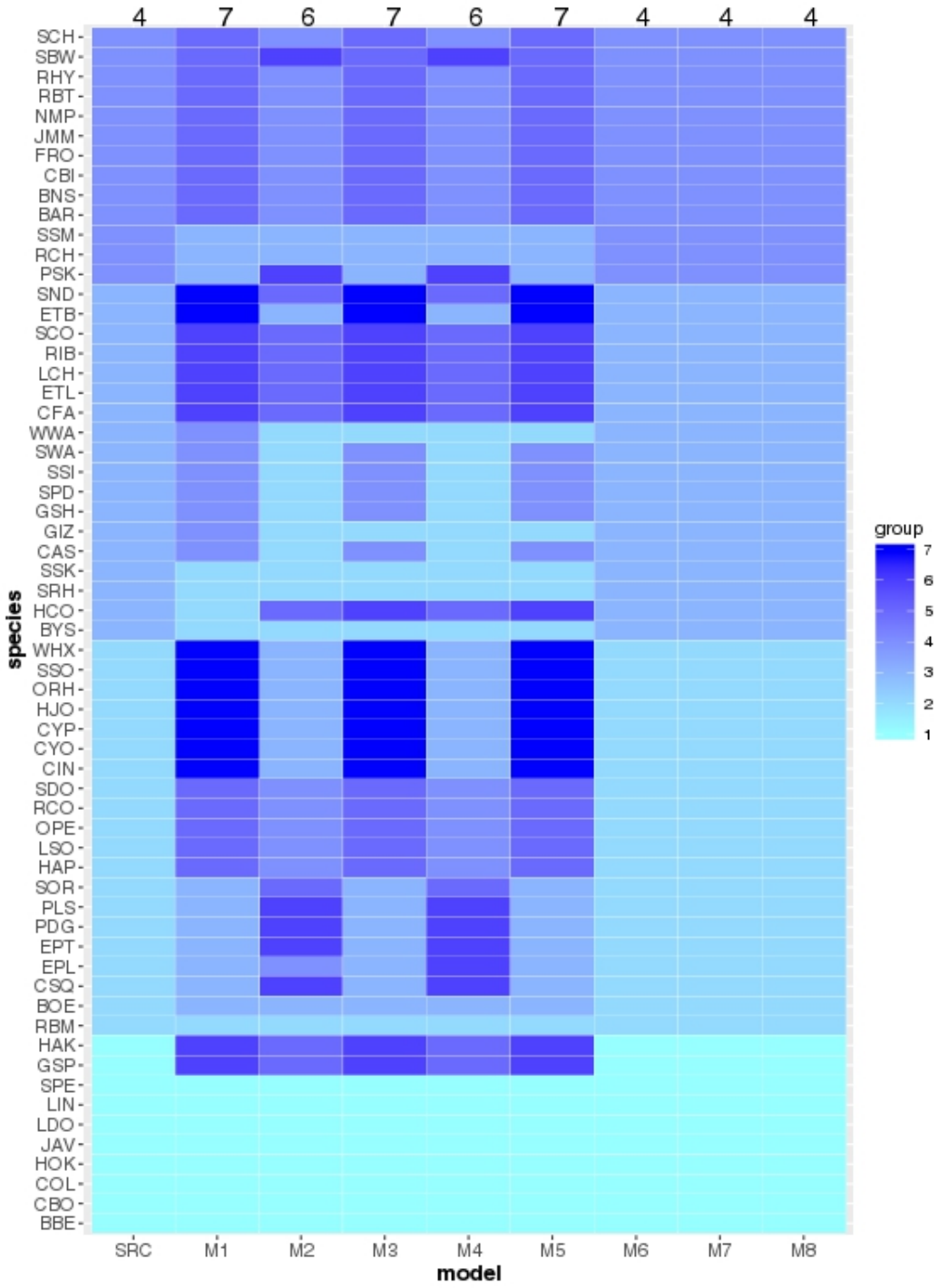


Figure 4.3: Plot showing species group memberships for the tan1101 data from a set of models shown in Table 4.1. The membership is ordered by result from SRC. From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

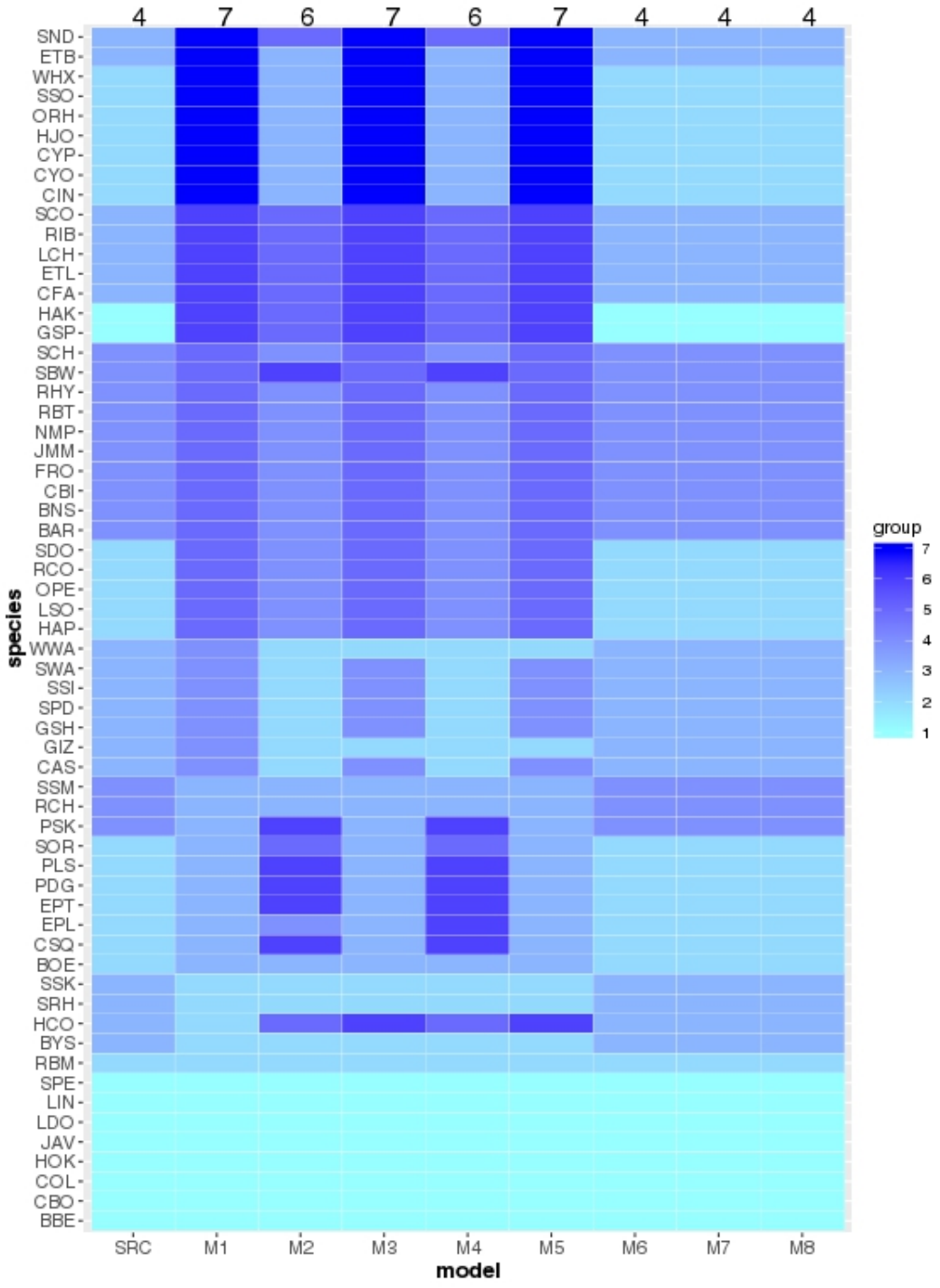


Figure 4.4: Plot showing species group memberships for the tan1101 data from a set of models shown in Table 4.1. The membership is ordered by result from Model 1 (M1). From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

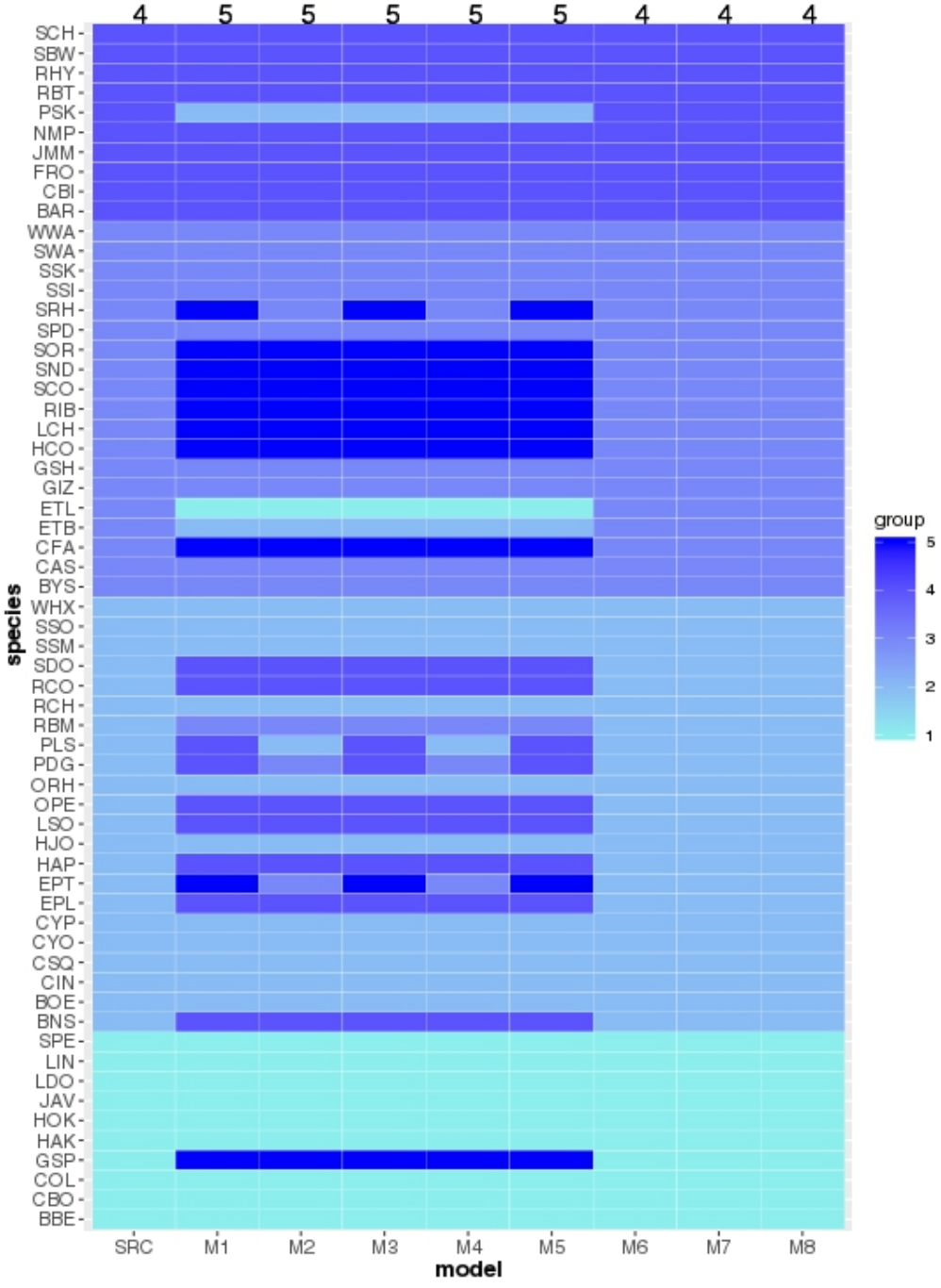


Figure 4.5: Plot showing species group memberships for the tan1201 data from a set of models shown in Table 4.1. The membership is ordered by result from SRC. From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

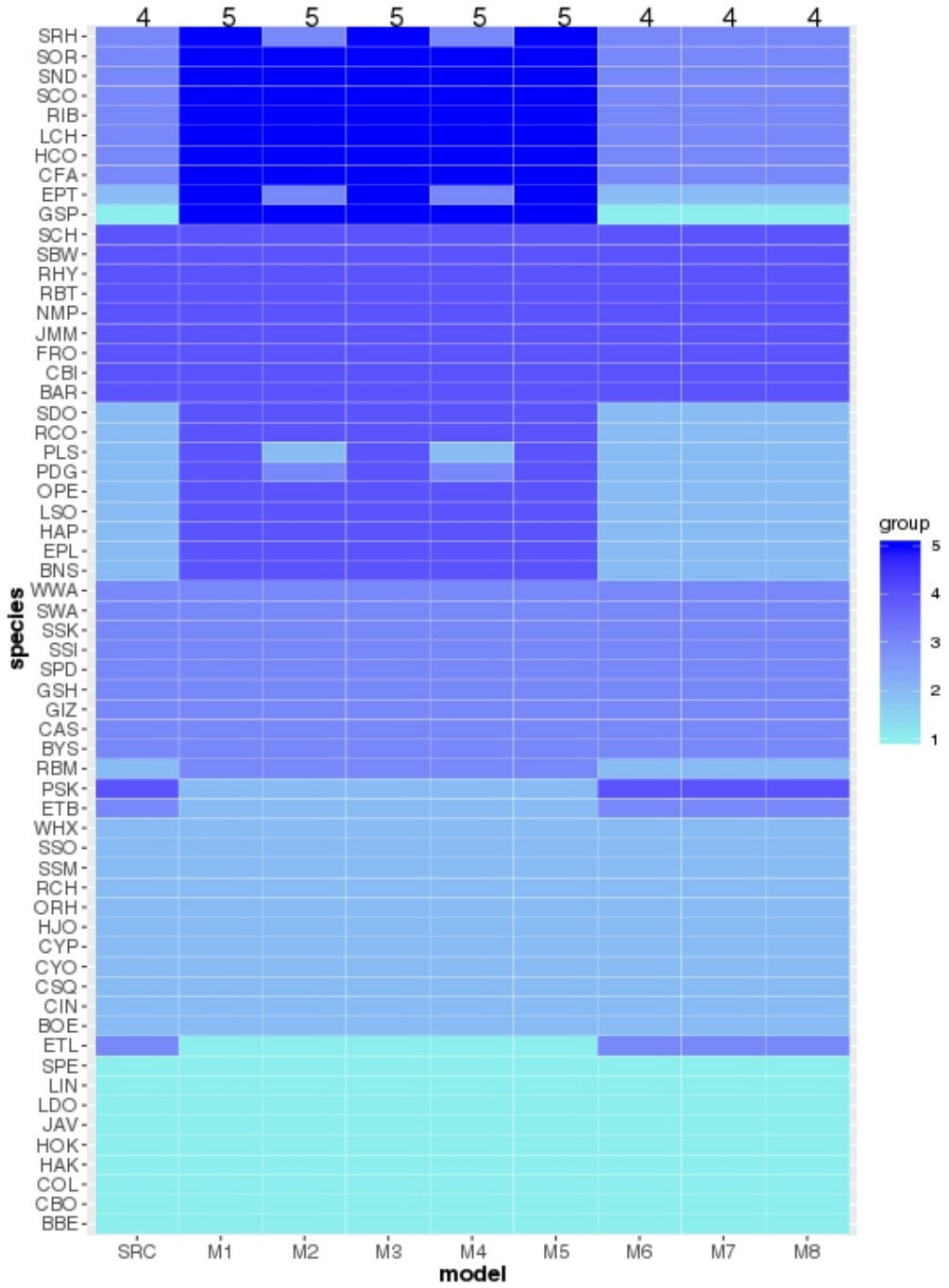


Figure 4.6: Plot showing species group memberships for the tan1201 data from a set of models shown in Table 4.1. The membership is ordered by result from Model 1. From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

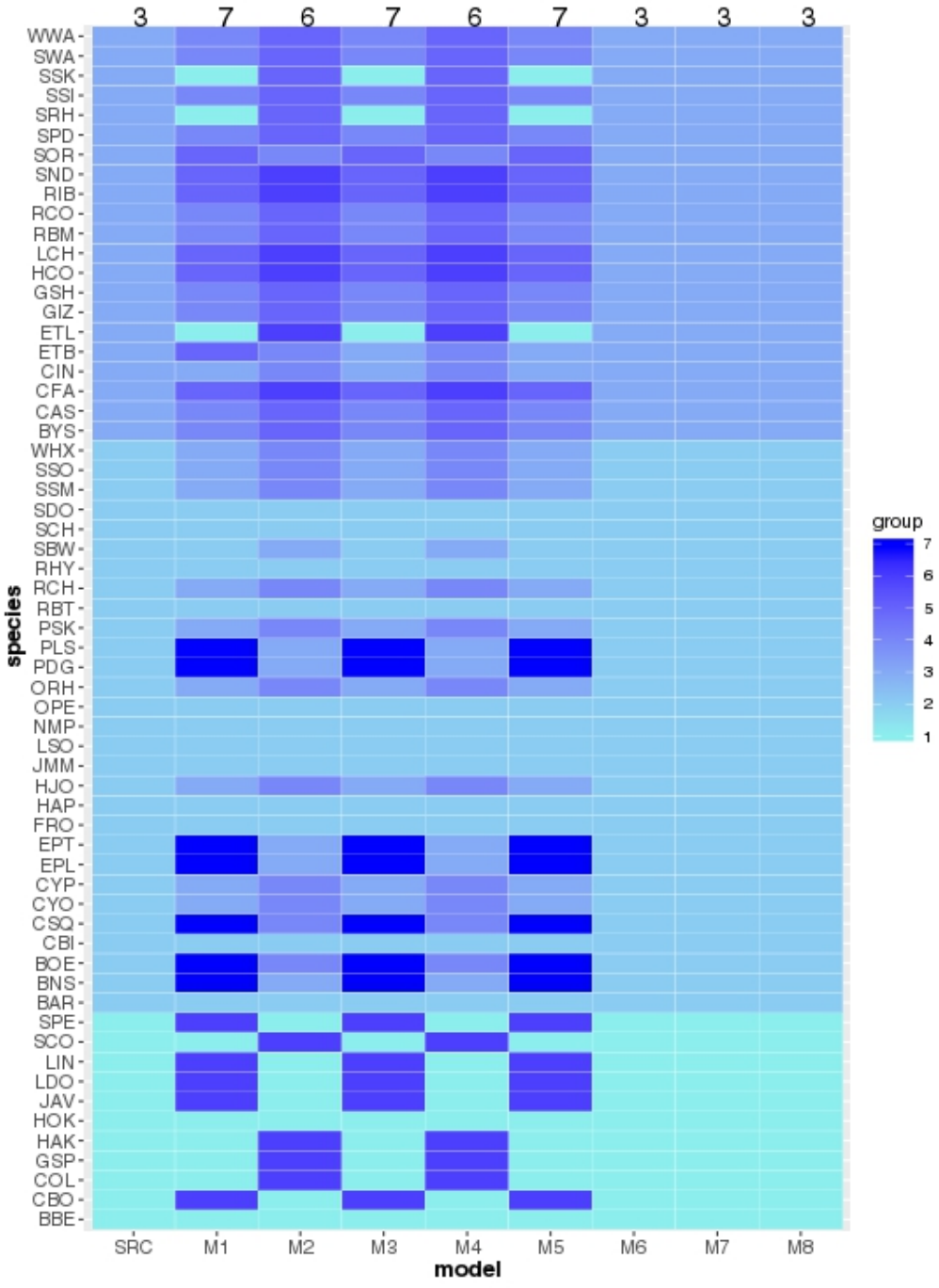


Figure 4.7: Plot showing species group memberships for the tan1301 data from a set of models shown in Table 4.1. The membership is ordered by result from SRC. From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

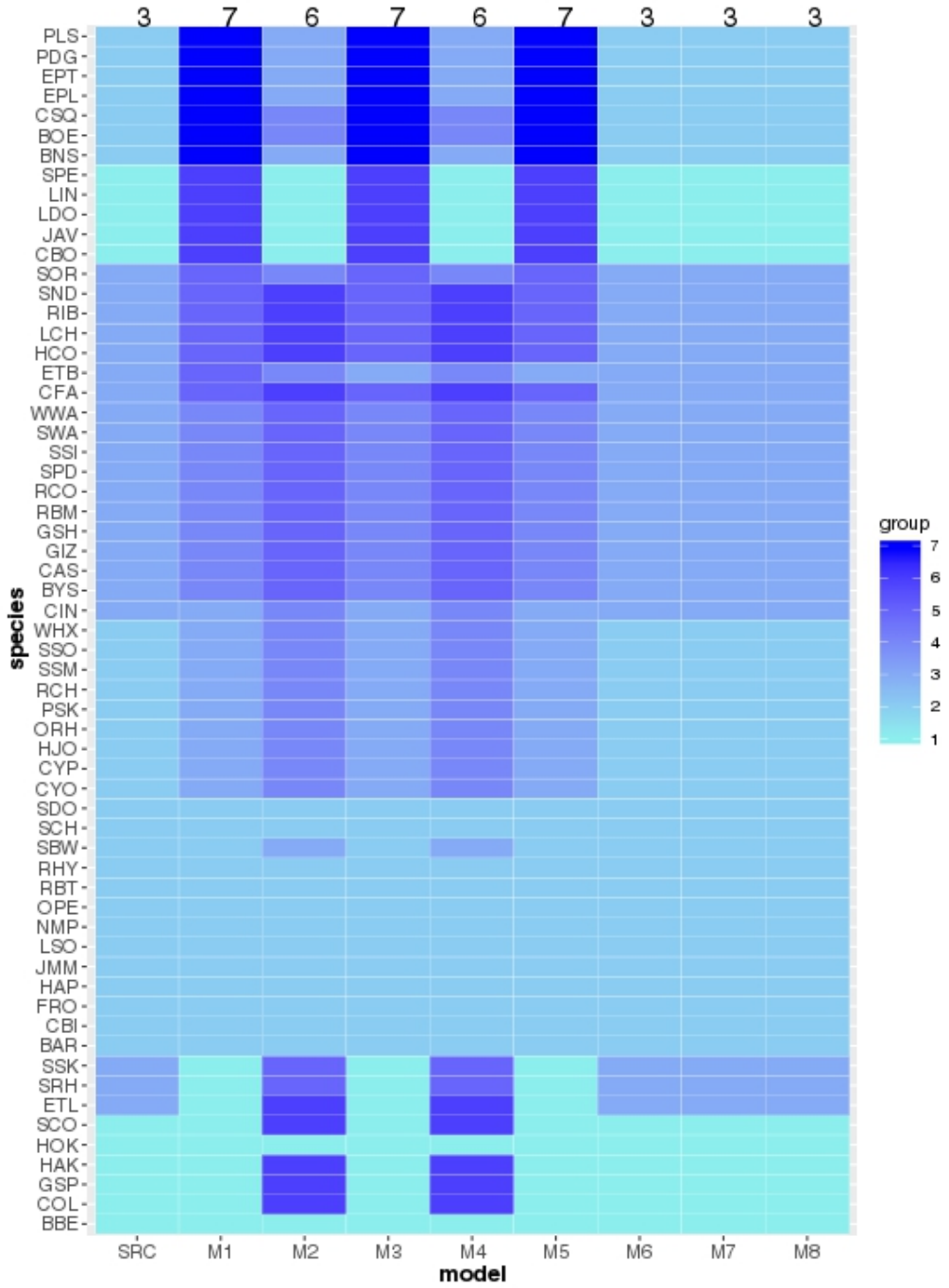


Figure 4.8: Plot showing species group memberships for the tan1301 data from a set of models shown in Table 4.1. The membership is ordered by result from Model 1. From left, SRC, Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7, and Model 8. The number of clusters R in the fitted model is shown on the top.

4.5 Column Effects

So far we have found that the depth is the most powerful covariate explaining the data, but bottom water temperature is also important. In this section, we further investigate how similar or different depth and bottom temperature effects are. To do this, we calculate the column effect values ($\widehat{\mathrm{b}}_j$) from SRC model, Model 1, Model 4 and Model 5 (Table 4.1). We select these four models so that we can see column effect when there is no predictor in the model (SRC), when depth is the only predictor (Model 5), when bottom water temperature is the only predictor (Model 4), and when depth and temperature are both in the model (Model 1). The $\widehat{\mathrm{b}}_j$ term is not included in Models 1, 4, or 5 (Table 4.1), so we need to compute $\widehat{\mathrm{b}}_j$ values. The $\widehat{\mathrm{b}}_j$ are computed as follows.
For SRC model, we use the estimates of b_j, the independent column effect. For Models 1, 4, and 5, the column effects are determined by the covariates:

$$\begin{aligned}
\mathrm{logit}(\phi_{ijr}) &= \eta_{ijr} \\
&= \mu + \mathrm{a}_r + w_j^T(\psi + \tau_r) && (4.2) \\
\widehat{\eta}_{j\cdot} &= \sum_{r=1}^{R} \widehat{\pi}_r \eta_{ijr} \\
&= \sum_{r=1}^{R} \widehat{\pi}_r [\mu + \mathrm{a}_r + w_j^T(\psi + \tau_r)] \\
&= \mu + \sum_{r=1}^{R} \widehat{\pi}_r \mathrm{a}_r + \sum_{r=1}^{R} \widehat{\pi}_r w_j^T(\psi + \tau_r) \\
&= C + w_j^T \widehat{\pi}_r(\psi + \tau_r) && (4.3) \\
\widehat{\mathrm{b}}_j &= w_j^T \widehat{\pi}_r(\psi + \tau_r) && (4.4)
\end{aligned}$$

where

- ψ is the $W \times 1$ of coefficient for the covariates

- τ_r is the rth column of the row cluster - covariates interaction matrix $\mathbf{T}$ coefficients
- C is a constant ($C = \mu + \sum_{r=1}^{R} \widehat{\pi}_r \mathrm{a}_r$) across all columns

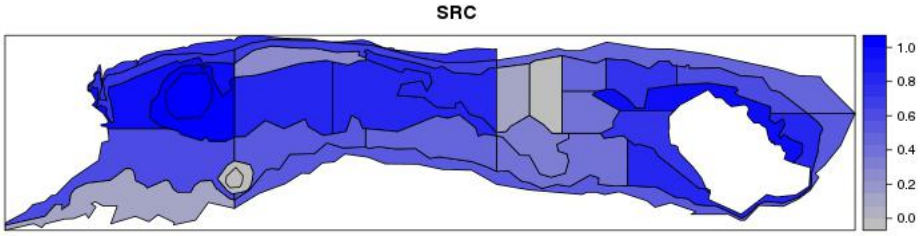
These values are further standardised by

$$\widehat{\mathrm{b}}_{j,std.} = \frac{\widehat{\mathrm{b}}_j - \min\widehat{\mathrm{b}}_j}{\max\widehat{\mathrm{b}}_j - \min\widehat{\mathrm{b}}_j} \tag{4.5}$$

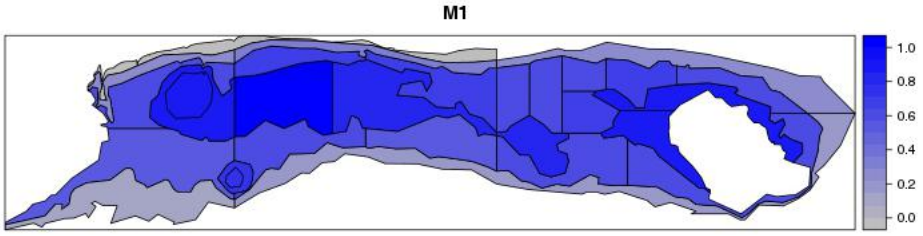
The value of $\widehat{\mathrm{b}}_{j,std.}$ for the tan0201 data are visualised in Figure 4.9. A darker colour means that the chance of observing species is high, and a lighter, greyer colour means the species are rarely seen from that stratum. The area with white colour was not surveyed as this area includes the Chatham Islands (Figure **??**). The visualisations from SRC model (Figure 4.9a) simply show where species were mostly caught, and the area called Mernoo Bank (stratum 18) and Reserve Bank (stratum 19 and 20, Figure **??**) seem to be the most popular locations for many fish. SRC model has an independent parameter for each stratum, showing the strongest variation between areas, as expected. Figure 4.9c shows stronger peaking along a narrow ridge of strata. Model 4 has bottom water temperature and temperature-row group interaction term in the model. Figure 4.9c is very similar to Figure 2.4a, indicating that species are more likely to be caught in warmer water. In contrast, dark area in Figure 4.9d is more spread out, indicating species also occur in deeper strata. The colour tones in Figure 4.9d is opposite of the colour tones in Figure 2.4b. Model 5 is able to show high probability of catching species in wider areas than Model 4, but these two figure suggests that species are likely to be caught in shallow water strata (200-400 m). The column effects from Model 1 seems to be taking a balance between Model 4 and Model 5 (Figure 4.9b).

Similar patterns are seen for the tan1101, tan1201, tan1301 column effects plots (Figure 4.10a, 4.11a, 4.12a). The visualizations of the b_j val-

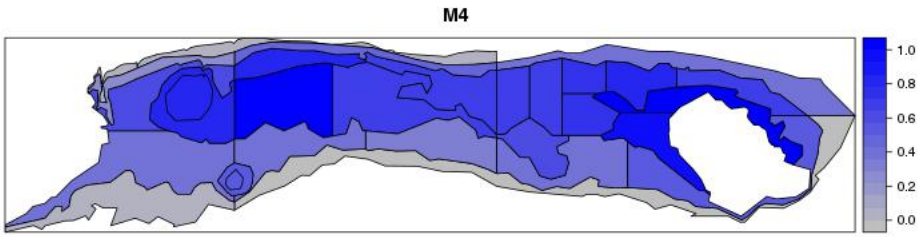
ues for SRC show notable difference between the strata for all trips (Figure 4.10a, 4.11a, 4.12a). Areas with high probability of catch in Model 4 are narrower than Model 5, and Model 1 shows counterbalanced effects of depth and water temperature. There is year to year variation of catch probability, as the visualisations from SRC models changes with time (Figure 4.10a, 4.11a, 4.12a). The visualisations of column effects from Model 4 for the tan1101, tan1201, and tan1301 data (Figure 4.10c, 4.11c, 4.12c) show similar colour gradations as Figure 2.5a, 2.6a, and 2.7a, respectively. As before, Model 4 indicates that more species are caught in warm water areas, and this pattern do not change with time. The column effects for Model 5 seems to have changed with time as more deeper areas are showing darker colour. (Figure 4.10d, 4.11d, 4.12d). This is probably because of additional tows in deep strata in the recent years, so more species were caught in these areas. Similar to Figure 4.9d, these figures show high probability of catching fish in wider areas, but the probability is still small in the strata that are deeper than 800 m (Figure 2.5b, 2.6b, and 2.7b)



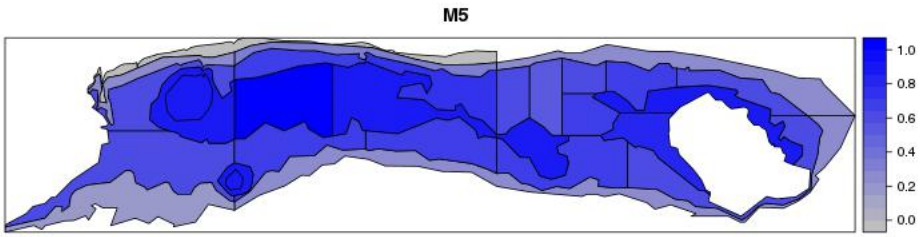
(a)



(b)

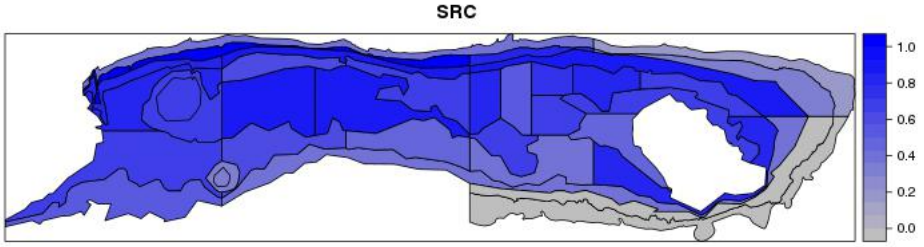


(c)

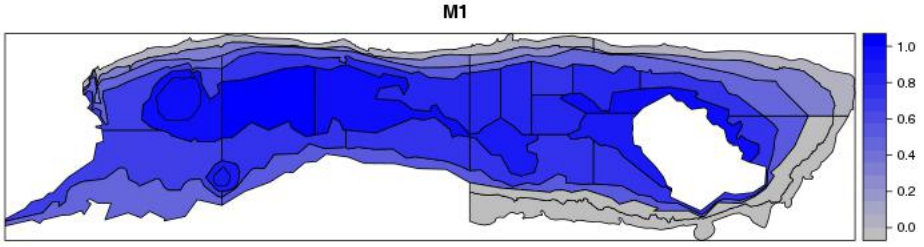


(d)

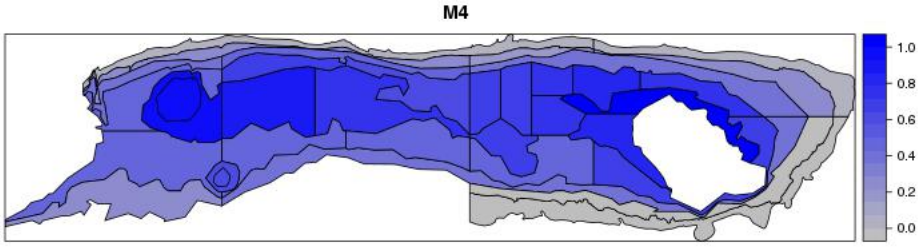
Figure 4.9: Map of the Chatham Rise displaying the logit-scaled probability of catching species for each stratum for the tan0201 data, calculated by equation 4.4 and 4.5 for four different models. From the top, the simple row clustered model (SRC), Model 1, Model 4, and Model 5. The white area on the right includes the Chatham Islands therefore not surveyed.



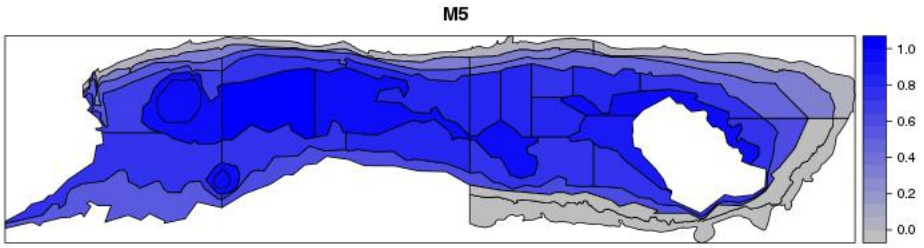
(a)



(b)

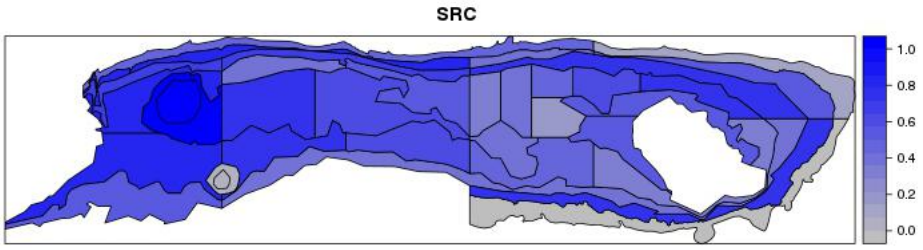


(c)

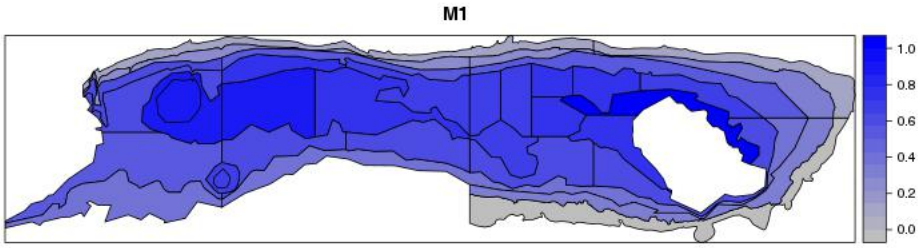


(d)

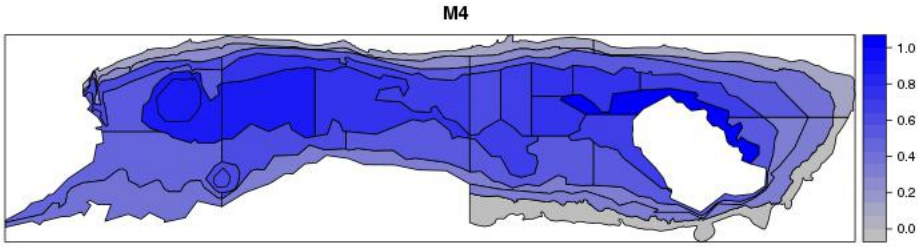
Figure 4.10: Map of the Chatham Rise displaying the logit-scaled probability of catching species for each stratum for the tan1101 data, calculated by equation 4.4 and 4.5 for four different models. From the top, the simple row clustered model (SRC), Model 1, Model 4, and Model 5. The white area on the right includes the Chatham Islands therefore not surveyed.



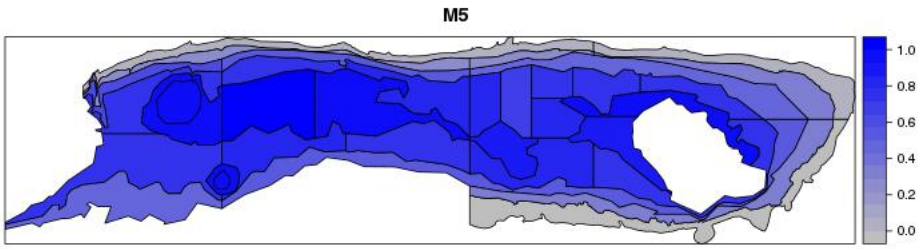
(a)



(b)

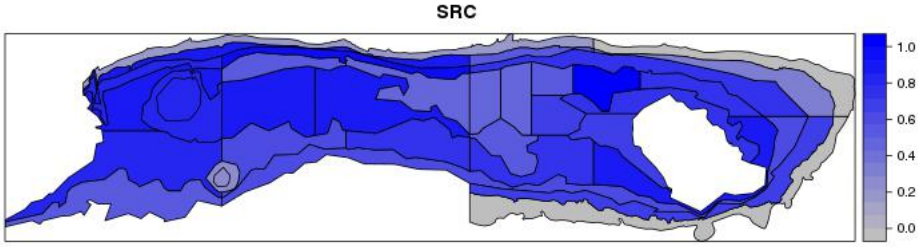


(c)

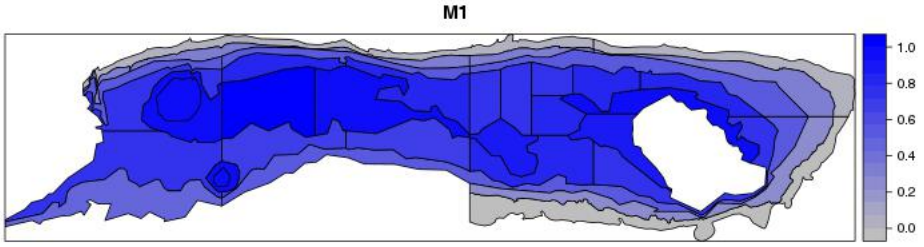


(d)

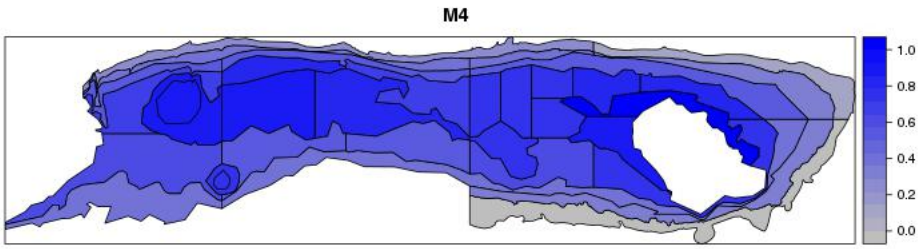
Figure 4.11: Map of the Chatham Rise displaying the logit-scaled probability of catching species for each stratum for the tan1201 data, calculated by equation 4.4 and 4.5 for four different models. From the top, the simple row clustered model (SRC), Model 1, Model 4, and Model 5. The white area on the right includes the Chatham Islands therefore not surveyed.



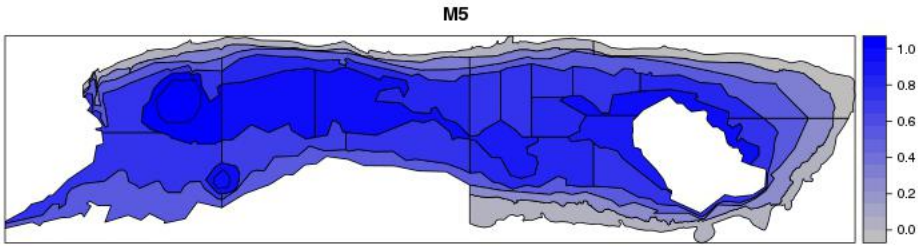
(a)



(b)



(c)



(d)

Figure 4.12: Map of the Chatham Rise displaying the logit-scaled probability of catching species for each stratum for the tan1301 data, calculated by equation 4.4 and 4.5 for four different models. From the top, the simple row clustered model (SRC), Model 1, Model 4, and Model 5. The white area on the right includes the Chatham Islands therefore not surveyed.

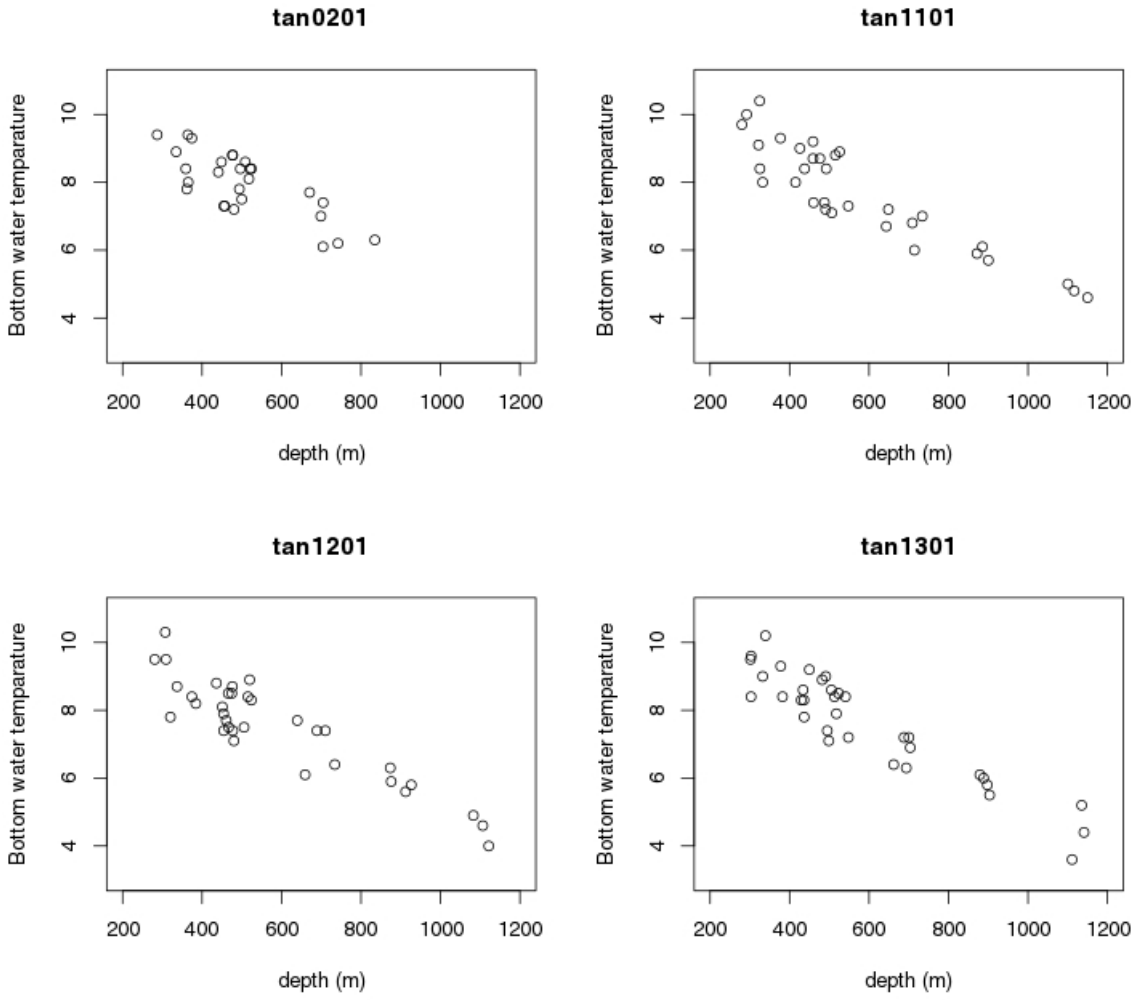


Figure 4.13: Plots showing the relationship between depth (m) and bottom temperature (°C) for the for the tan0201, tan1101, tan1201, and tan1301 data.

4.6 Summary

We found that depth and bottom temperature are both associated with species occurrences, and modify the clustering when included in the model. We also found that species are also clustered by frequency of occurrence. Comparing depth effect with bottom temperature effect, we found that depth is the strongest predictor to explain the data. Our findings so far are supported by previous studies, such as the paper from Francis et al. (2002), in which they clustered 123 fish and squid species into four groups. They found depth is the most important predictor to explain species assemblages. They also found that latitude and longitude are the important factors to explain species distribution, and mentioned a possibility that the species distributions could be explained by bottom temperature. We have detected the effect by bottom temperature but our results do not provide convincing explanations whether it has a significant effect. The biggest reason is that temperature is correlated with depth (Figure 4.13). It also may be due to the way we selected the species for the analysis. We selected the species that were caught in all four trips. Recall that the deep-water strata were introduced in 2010, therefore the most species caught in the deep strata are likely to have been excluded, as they were probably not caught in tan0201 trip. Another reason is that the species are also clustered by its frequency of occurrence. To see the effect of the bottom temperature more clearly, we need to control the difference in frequency of occurrence for the all species.

4.7 The Row Standardised and the Row Covariate Models

In this section, we present the results from the row standardised model (Model RS) and the row covariate model (Model RC). Both models are presented in Table 4.1 earlier, here we revisit these models. The row standardised model is expressed in equation (3.18)

$$\text{logit}(\phi_{ijr}) = \mu + \alpha_i + \text{a}_r + \text{b}_j$$

where α_i is row standardisation term for each row i with constraint $\sum_{i=1}^{n} \alpha_i = 0$, a_r is the cluster effect for each row cluster r with constraint $\sum_{r=1}^{R} \text{a}_r = 0$, and b_j is the effect of the column j with $\sum_{j=1}^{p} \text{b}_j = 0$ constraint.

The row covariate model is written in equation (3.13)

$$\text{logit}(\phi_{ijr}) = \mu + \text{a}_r + \text{b}_j + \underline{x}_i^T \underline{\delta}$$

where δ is the coefficients of the row covariate (body length), x_i is the ith element of row covariate matrix $\mathbf{X}$ (e.g. x_i is median body length for each species i), $\sum_{r=1}^{R} \text{a}_r = 0$, and $\sum_{j=1}^{p} \text{b}_j = 0$.

In terms of the model selection, Model RS and Model RC both have larger value of the AIC than Model 1, therefore we still take Model 1 as the optimal model for all datasets (Table 4.2, 4.3).

To investigate the effects of row level terms, we display the values of $\widehat{\text{a}}_r$ from SRC model, $\widehat{\alpha}_i$ and $\widehat{\text{a}}_r$ from Model RS, and $\widehat{\text{a}}_r$, $x_i^T\widehat{\delta}$, and $\widehat{\text{a}}_r + x_i^T\widehat{\delta}$ from Model RC. Figure 4.14 illustrates the values described above for the tan0201 data. Values with small numbers are shown in red tones, and they shift to yellow tones as the value become larger. The column in far left

(SRC, a_r) are the value of row cluster effect $\widehat{\mathrm{a}}_r$, shown in three different tones as there are three clusters made from this model (Table 4.2). The second column from left in this figure shows the value of row cluster effect $\widehat{\mathrm{a}}_r$ from Model RS. There is no colour difference (i.e. monotonic) indicating that a_r values are not identifiable. This is expected because the variations among the species are explained by $\widehat{\alpha}_i$, and therefore, $\widehat{\mathrm{a}}_r = 0$ for the simple row standardised model. Then Model RS drops a_r term from the model, becoming the GLM where $y_{ij} = \theta_{ij}$, with $\mathrm{logit}(\phi_{ij}) = \alpha_i + \mathrm{b}_j$. This model does not reduces the dimension of the original data, therefore this model cannot explain the species groupings. The row standardisation value $\widehat{\alpha}_i$ from Model RS is displayed in the third column (Row Std, α_i). It is difficult to see the difference in abundance among the species as there is not much colour difference. But there are three species in yellow colour. They are hoki (HOK), javelin fish (JAV), and bollons rattail (CBO), which are most frequently occurred species in this trip. The three right hand columns in Figure 4.14 illustrate the row level effects value from Model RC. The row cluster effect (shown as Row Cov a_r) has three colour bands with the same clusters as SRC, indicating that body length does not explain the data. This is confirmed by the column showing $x_i^T\widehat{\delta}$, as it shows no colour variation, indicating that $\widehat{\delta}$ does not have significant effect, and thereby the data is explained by the row clusters. Therefore, the far right column, which is a visualisation of $\widehat{\mathrm{a}}_r + x_i^T\widehat{\delta}$, shows the same cluster structure as SRC.

Similar patterns are seen for the tan1201 and tan1301 data (Figure 4.16 and 4.17). But Figure 4.15, the results for the tan1101 data, show very different colour display from others. It is clearer to see the variation among the species explained by α_i (third column from left, Figure 4.15). There is no obvious colour variation within the species in Group 4 (shown in red), indicating that frequency of occurrence is similar between these species. It is also clearer to see that there is not much variation explained by $x_i^T\widehat{\delta}$ (second column from right, Figure 4.15), indicating that the body length

is not important predictor for the data. The reason why Figure 4.15 is because the row standardisation value $\widehat{\alpha}_i$ for hoki (HOK) from tan1101 trip is lower than other three trips. The value of $\widehat{\alpha}_i$ for hoki is 17.178 (tan0201), 3.8295 (tan1101), 16.7592 (tan1201), and 16.7706 (tan1301), showing the value from tan1101 trip is less than a quarter of the values from the other years. This can be explained by the observation frequencies. Hoki is the dominant species caught on the Chatham Rise and it was caught in almost all strata in every year. In fact, hoki was present in the all strata in tan0201, tan1201, and tan1301 trip (e.g. recorded 1 in the all strata). But it was not observed in all strata in the tan1101 trip, and where it was was stratum 28. So the observation frequencies for hoki in the tan1101 data was treated the same as other species that were also not observed in one stratum, regardless of the location. This is probably why hoki did not stand out in the tan1101 data like other datasets. However, the survey in stratum 28 was not completed in this trip due to the lack of time (Stevens et al., 2012). Because this stratum is deep, it is likely that the survey there was carried out targeting only deepwater species, not hoki, so surveying in stratum 28 was considered a low priority. This is responsible for the dramatically different appearance in Figure 4.15: The value of $\widehat{\alpha}_i$ and $\widehat{\mathrm{a}}_r$ are much lower, and so the heat map covers a smaller range of values, showing structure not easily visible in the heat maps for the other three years. This uncompleted survey significantly affected on our analysis, therefore stratum 28 should have excluded from the tan1101 data.

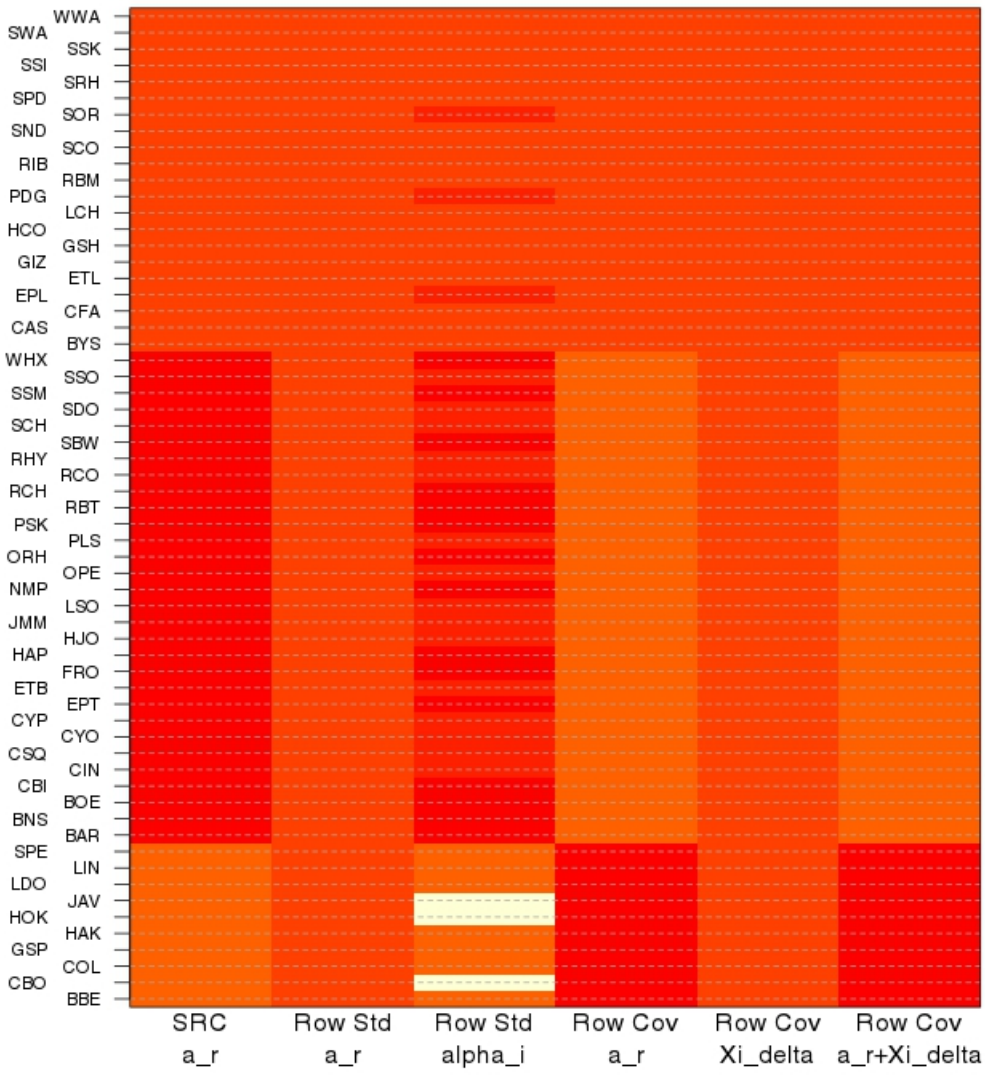


Figure 4.14: Heat map of the row cluster effect value ($\widehat{\mathrm{a}}_r$) of SRC model, the row cluster effect value ($\widehat{\mathrm{a}}_r$) and the row standardisation value ($\widehat{\alpha}_i$) of Model RS, and fitted values of their effects of Model RC for the tan0201 data.

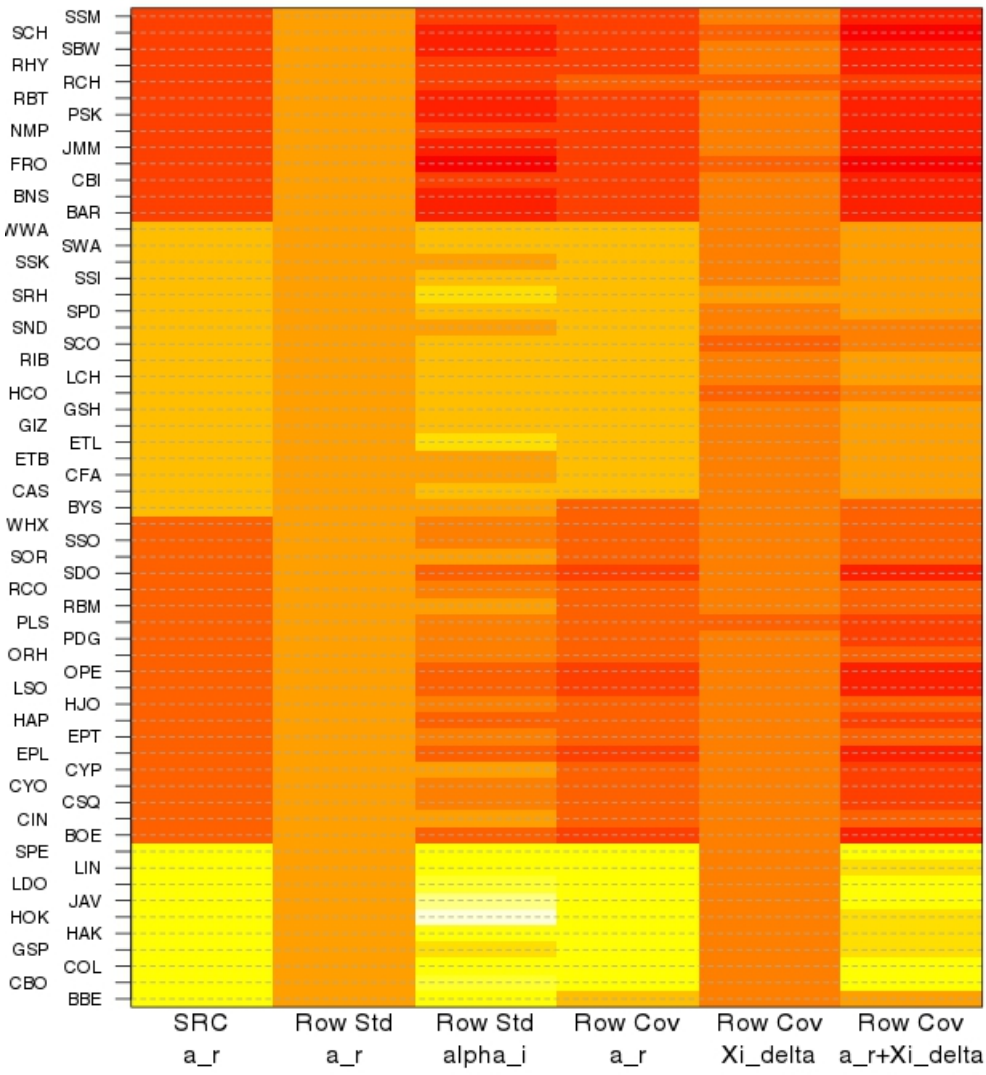


Figure 4.15: Heat map of the row cluster effect value ($\widehat{\mathrm{a}}_r$) of SRC model, the row cluster effect value ($\widehat{\mathrm{a}}_r$) and the row standardisation value ($\widehat{\alpha}_i$) of Model RS, and fitted values of their effects of Model RC for the tan1101 data.

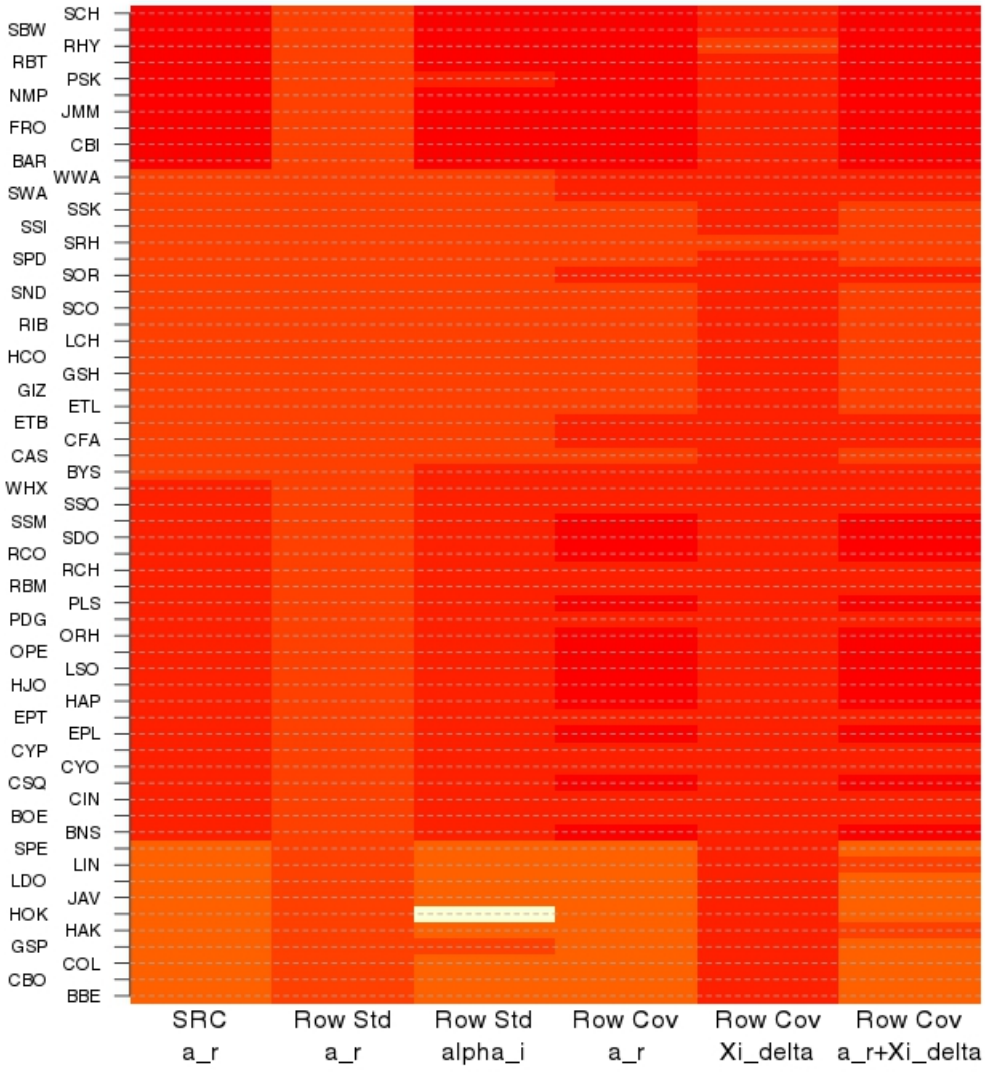


Figure 4.16: Heat map of the row cluster effect value ($\widehat{\mathrm{a}}_r$) of SRC model, the row cluster effect value ($\widehat{\mathrm{a}}_r$) and the row standardisation value ($\widehat{\alpha}_i$) of Model RS, and fitted values of their effects of Model RC for the tan1201 data.

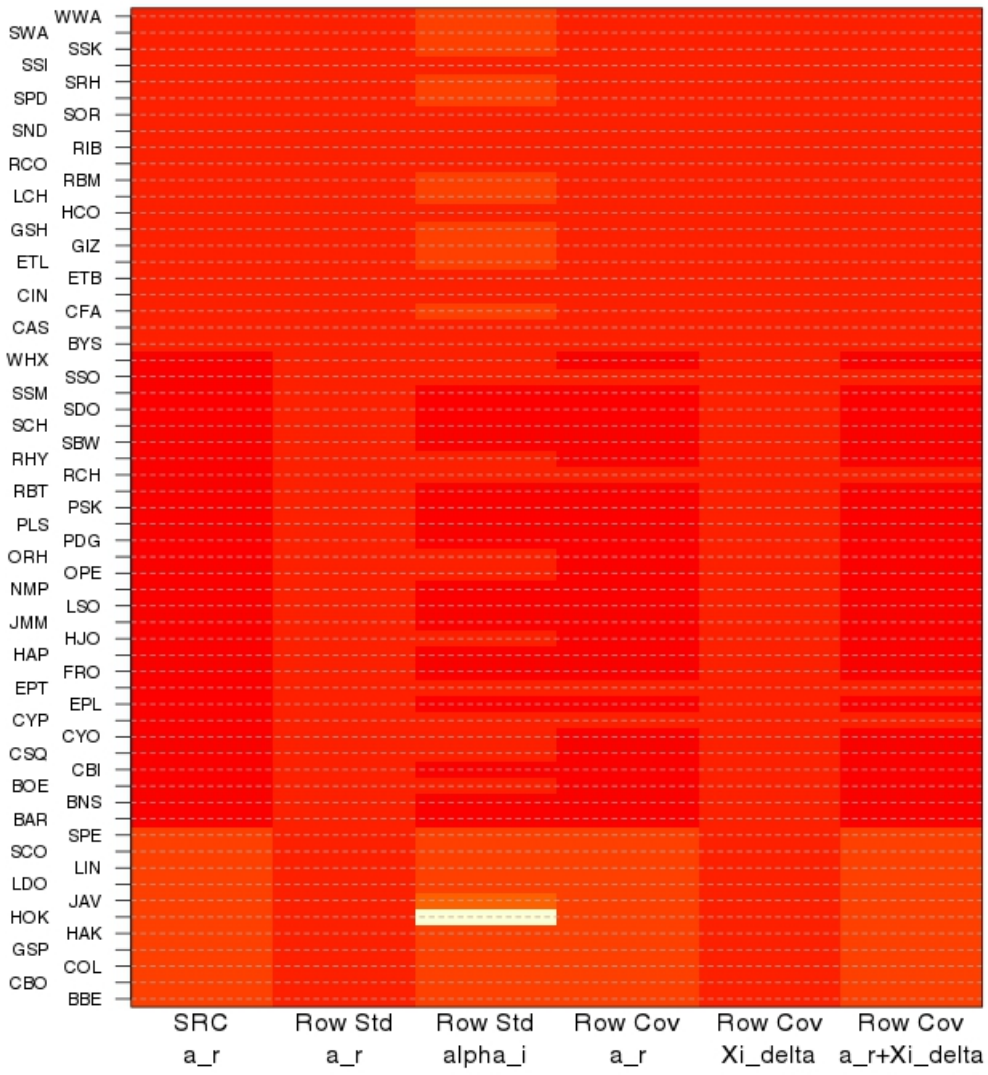


Figure 4.17: Heat map of the row cluster effect value ($\widehat{\mathrm{a}}_r$) of SRC model, the row cluster effect value ($\widehat{\mathrm{a}}_r$) and the row standardisation value ($\widehat{\alpha}_i$) of Model RS, and fitted values of their effects of Model RC for the tan1301 data.

4.8 Further Models

The results from previous sections have motivated us to investigate two further models. We combine the row level information and the column level information to a single model and further investigate the species clustering. The models we propose here are: Row cluster with row standardisation and the column covariates (Model 10, M10); and row cluster with row and column covariates model (Model 11, M11). More specifically,

- Model 10: Row cluster with row standardisation and the column covariates (equation 3.20 in Section 3.4)

$$\begin{aligned} \text{logit}(\phi_{ijr}) &= \mu + \alpha_i + \text{a}_r + \underline{w}_j(\underline{\psi} + \underline{\tau}_r) \\ &= \mu + \alpha_i + \text{a}_r + \text{depth}_j(\psi_1 + \tau_r) + \text{temp}_j(\psi_2 + \tau_r) \end{aligned} \tag{4.6}$$

- Model 11: Row cluster with the row covariates and the column covariates

$$\begin{aligned} \text{logit}(\phi_{ijr}) &= \mu + \underline{x}_j^T \underline{\delta} + \underline{w}_j(\underline{\psi} + \underline{\tau}_r) \\ &= \mu + \text{body length}_i\delta + \text{a}_r + \text{depth}_j(\psi_1 + \tau_r) + \text{temp}_j(\psi_2 + \tau_r) \end{aligned} \tag{4.7}$$

4.8.1 Model Selection

Table 4.4 compares the AIC values from Model 1, Model 10, and Model 11. The same cutoff value ($\widehat{\pi}_r > 0.08$) is applied. For the all datasets, the AIC value for Model 10 is the smallest of all. As Model 1 was the best model with the minimum value of AIC (Table 4.2, 4.3), Model 10 best presents the data for all trips.

4.8.2 Membership Results

We use the same visualization techniques as in Section 4.3 to present the membership $\{\widehat{z}_{ir}\}$. Figures 4.19 compared membership results from SRC, Model 1, and Model 10 for the tan0201 data. It shows that the memberships have changed dramatically by Model 10. The biggest change from the earlier results is that the species are no longer clustered by its frequency of occurrence, as the species memberships obtained by SRC model are fairly scattered. This is expected because α_i term in Model 10 absorbs the effect of frequency of occurrence. The same results are seen for the tan1101 data (Figure 4.20), tan1201 data (Figure 4.21), and tan1301 data (Figure 4.22). The species are also clustered differently between Model 1 and Model 10 for all datasets. But it can be seen that the membership results from Model 1 are retained in Model 10 to some degree. The most obvious example is seen from the tan1101 data in Figure 4.20. The species in Group 7 in Model 1 (the darkest blue group in M1), which is a deepwater species group, are seen in the same group (Group 5) with other deep-

Table 4.4: Comparison of Model 1, Model 10, and Model 11. The overall best model in each year is highlighted with blue.

Dataset	Model	R	npar	AIC	minimum π_r
tan0201	Model 1	6	23	1343.77	0.13
	Model 10	5	79	1208.19	0.18
	Model 11	6	24	1340.524	0.11
tan1101	Model 1	7	27	1629.07	0.08
	Model 10	5	79	1536.28	0.14
	Model 11	6	24	1675.110	0.10
tan1201	Model 1	5	19	1869.45	0.16
	Model 10	6	83	1641.42	0.08
	Model 11	6	24	1823.60	0.09
tan1301	Model 1	7	27	1786.66	0.11
	Model 10	5	79	1656.50	0.14
	Model 11	6	24	1813.14	0.09

water species in Model 10. The species are not clustered in the same way by Model 10 across the years (Figure 4.18). However, it seems that there are two big groups that do not change over time. One group is a group of deepwater species, and another is a group of shallow water species. The species membership changes across Groups 1, 2, 3, and 4 across time for all trips, but they were hardly clustered to Groups 5 or 6. Likewise, most species in Groups 5 and 6 did not join to Groups 1, 2, 3, or 4 for all trips.

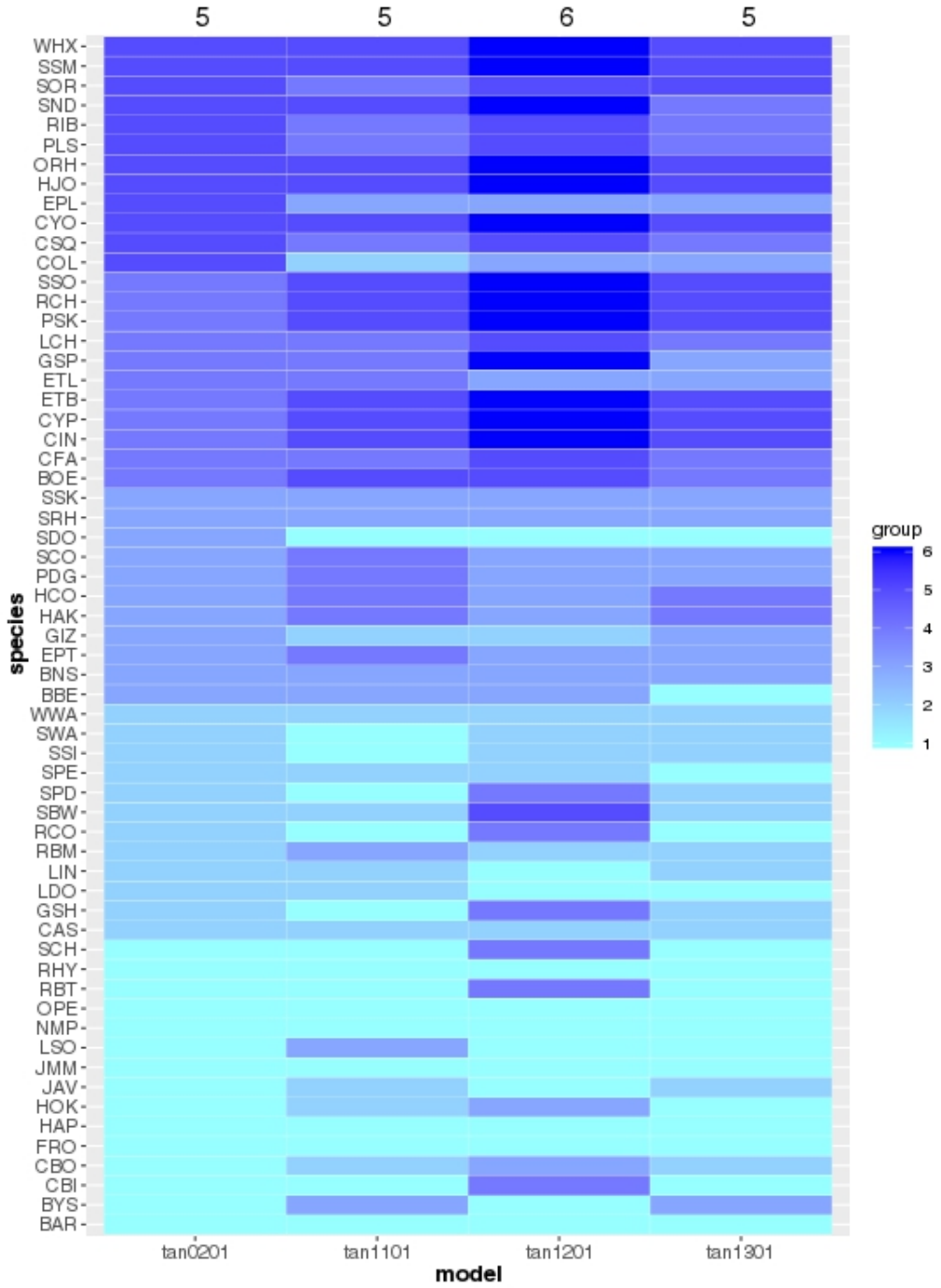


Figure 4.18: The membership results from Model 10 for the tan0201, tan1101, tan1201, and tan1301 data. The number of clusters made is shown on top.

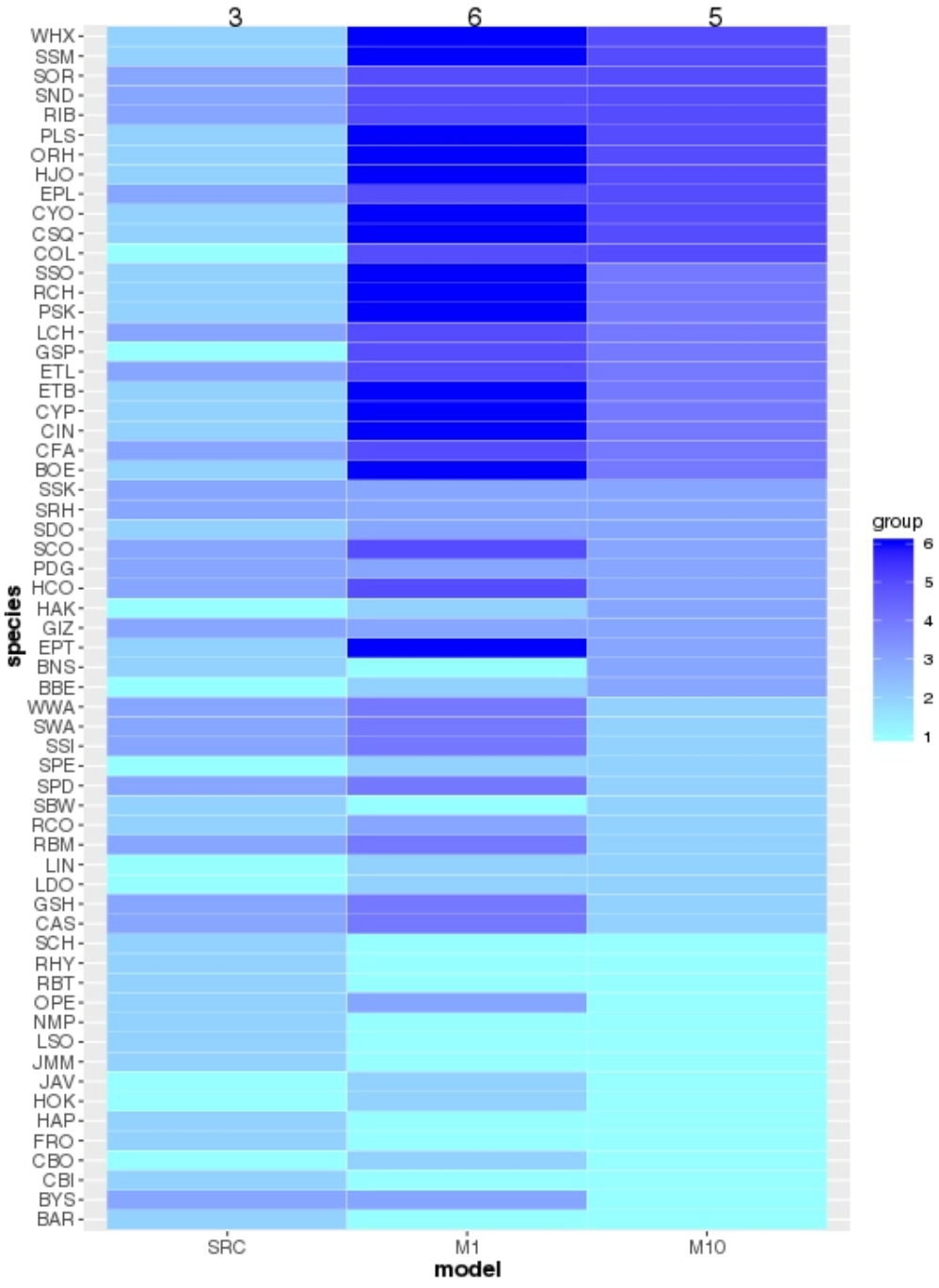


Figure 4.19: Plot showing species group memberships for the tan0201 data from SRC, Model 1, and Model 10. The membership is ordered by result from Model 10. The number of the best cluster R for the model is shown on the top.

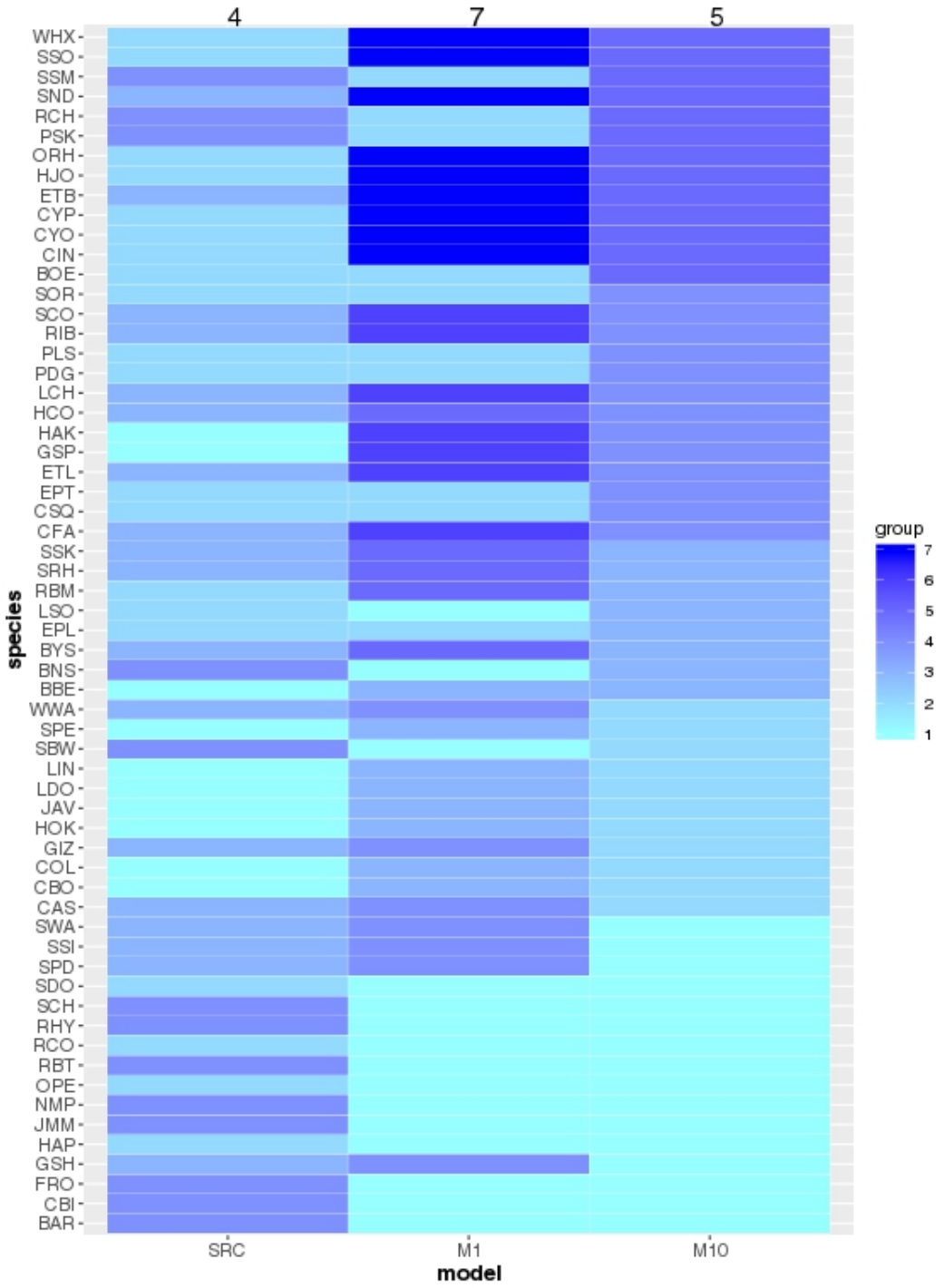


Figure 4.20: Plot showing species group memberships for the tan1101 data from SRC, Model 1, and Model 10. The membership is ordered by result from Model 10. The number of the best cluster R for the model is shown on the top.

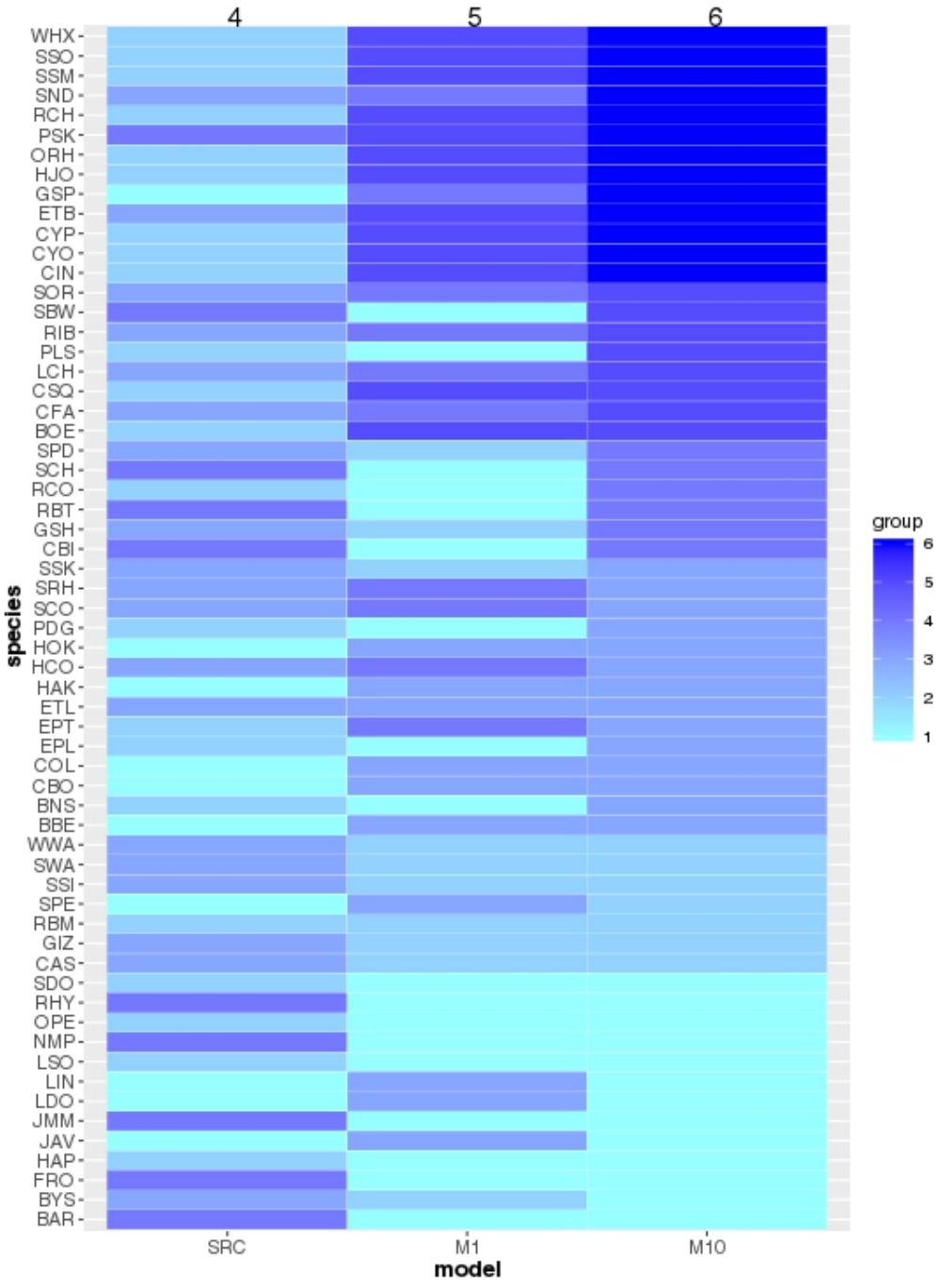


Figure 4.21: Plot showing species group memberships for the tan1201 data from SRC, Model 1, and Model 10. The membership is ordered by result from Model 10. The number of the best cluster R for the model is shown on the top.

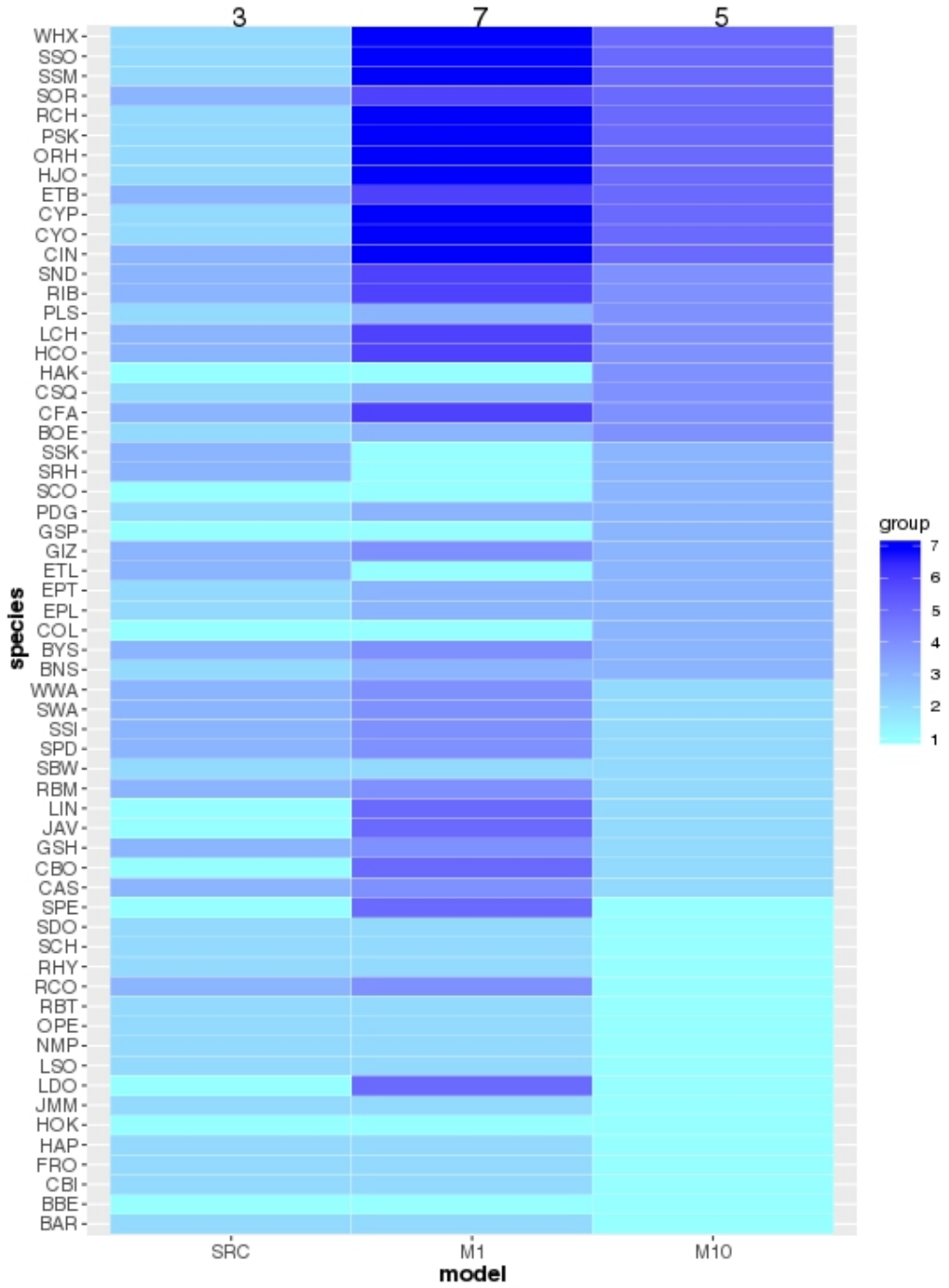


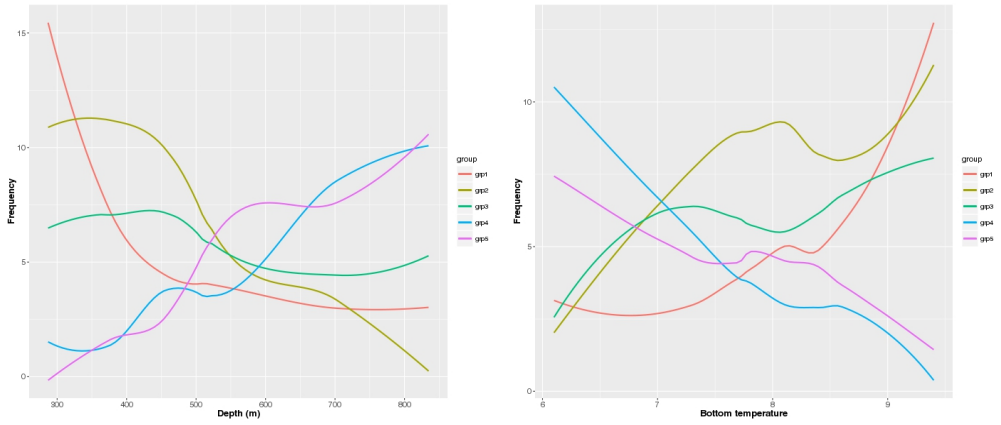
Figure 4.22: Plot showing species group memberships for the tan1301 data from SRC, Model 1, and Model 10. The membership is ordered by result from Model 10. The number of the best cluster R for the model is shown on the top.

4.9 Description of Group Characteristics

In this section, we present species frequency against depth and bottom temperature. We also give a brief description for each group in a selected trip. Figure 4.23, 4.24, 4.25, 4.26 are smoothed frequency polygons of the depth for each group. We first find the frequencies of presence in each strata for a selected group species. So a larger number means more species were caught in a particular stratum. The frequency for each group is given by

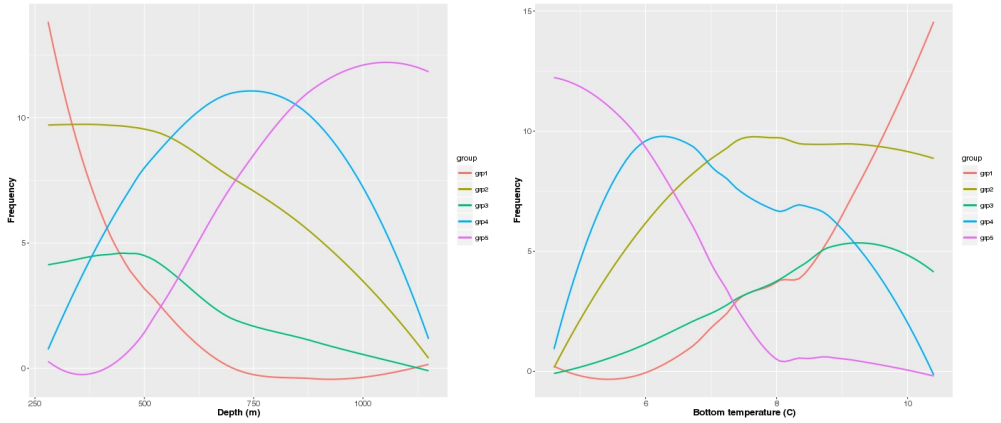
$$\sum_{i=1}^{n} I(i \in r) y_{ij} = f_{rj} \tag{4.8}$$

where f_{rj} is the frequency of the species in group r observed at stratum j, $I(i \in r)$ is the indicator function selects the species in the group r (i.e. the group for which z_{ir} is the highest). Once we found f_{rj}, we refer to our dataset to obtain the value of the depth (denoted by d_j) and the bottom temperature (denoted t_j) for in each group. The value of $\{d_j\}$ and $\{t_j\}$ are plotted first, and a kernel of density function is drawn as a smooth line, coloured by the group. The smoothed frequency polygons of bottom temperature are just mirror images of depth, because they are correlated (Figure 4.13).



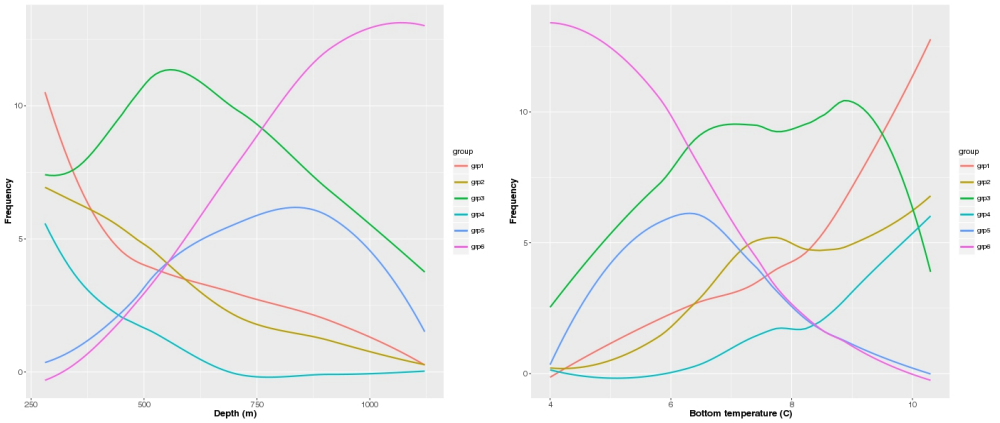
(a) (b)

Figure 4.23: The smoothed frequency polygons of depth (a) and bottom temperature (in °C, (b)) for each species group, clustered by Model 10 for the tan0201 data.



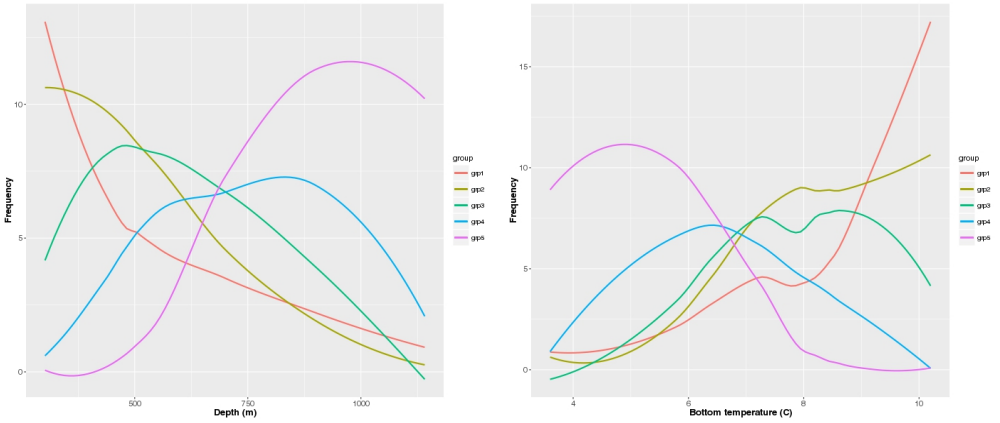
(a) (b)

Figure 4.24: The smoothed frequency polygons of depth (a) and bottom temperature (in °C, (b)) for each species group, clustered by Model 10 for the tan1101 data.



(a) (b)

Figure 4.25: The smoothed frequency polygons of depth (a) and bottom temperature (in °C, (b)) for each species group, clustered by Model 10 for the tan1201 data.



(a) (b)

Figure 4.26: The smoothed frequency polygons of depth (a) and bottom temperature (in °C, (b)) for each species group, clustered by Model 10 for the tan1301 data.

4.9.1 Group Characteristics for the tan1301 data

Here we show the features of groups clustered by Model 10 for the tan1301 data. The species are clustered to five groups by Model 10. There are two shallow water (200-400 m) species groups, two middle water species group (400-800 m), and one deepwater species group (>800 m). The depth preference is the lower quartile and the upper quartile of depth records (d_j) for each group r. Trawl survey area for the tan1301 trip is shown in Figure **??**.

Group 1: Shallow water species

The species in this group can be classified as shallow water species. Their distribution is concentrated at Mernoo Bank (stratum 18) and Reserve Bank (stratum 19 and 20). The depth preference ranges from 360 to 550 m.

Group 2: Shallow water species

Group 2 species were caught in almost every stratum except the deep strata. Similar to Group 1, they show concentrated distribution at Mernoo Bank (stratum 18), but they cover the south side of shallow strata (stratum 12, 13, 14, 15, 16, and 17). They have similar depth preference as Group 1 with the range of 380-530 m.

Group 3: Middle water species

The depth preference of this group is at 430- 630m, but their distributions are concentrated in shallow water strata. The species in this group were most frequently caught in the west side of shallow strata (stratum 10A, 10B, 11A, 11B, 11C, and 11D).

Group 4: Middle water species

The distributions of Group 4 species are similar to Group 3 but less dense in the shallow water strata. So this group prefers slightly deeper water at 500 -800 m.

Group 5: Deepwater species

Species in this group are deepwater species with the depth preference of 800-1000 m. They were seen in the all deep strata, but quite a few species were also seen in the stratum 1, which is neighboring stratum of the stratum 22.

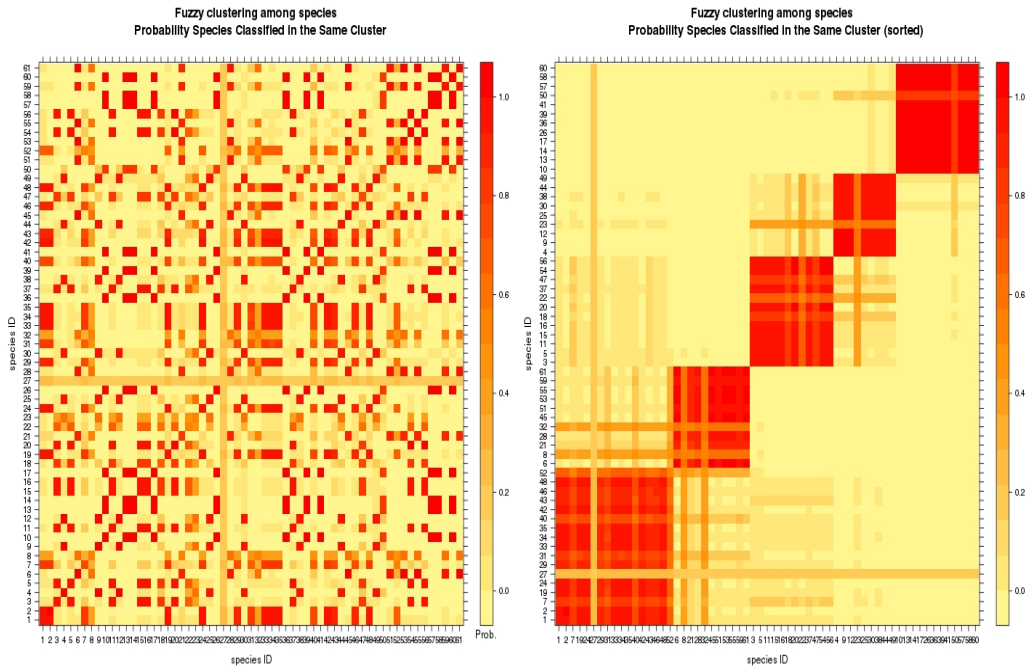
4.9.2 Fuzziness of Clustering

Our approach applies fuzzy clustering via finite mixtures, so any fuzziness in the cluster structure should appear in any visualisation tools. Figure 4.27 shows two heat maps showing the probability $C_{ii'}$ of any pair of species i and i' $(i, i' = 1, \cdots, n)$ of being allocated to the same cluster for the tan1301 dataset. The displayed probability $C_{ii'}$ in both heat maps is calculated as follows:

$$\begin{aligned} C_{ii'} &= \sum_{r=1}^{R} P\left(z_{ir}=1, z_{i'r}=1|\mathbf{Y}, \Phi, \underline{\pi_r}\right) \\ &= \sum_{r=1}^{R} P\left(z_{ir}=1|z_{i'r}=1\mathbf{Y}, \Phi, \underline{\pi_r}\right) P\left(z_{i'r}=1|\mathbf{Y}, \Phi, \underline{\pi_r}\right) \\ &= \sum_{r=1}^{R} P\left(z_{ir}=1|\mathbf{Y}, \Phi, \underline{\pi_r}\right) P\left(z_{i'r}=1|\mathbf{Y}, \Phi, \underline{\pi_r}\right) \\ &= \sum_{r=1}^{R} z_{ir} z_{i'r} \end{aligned}$$

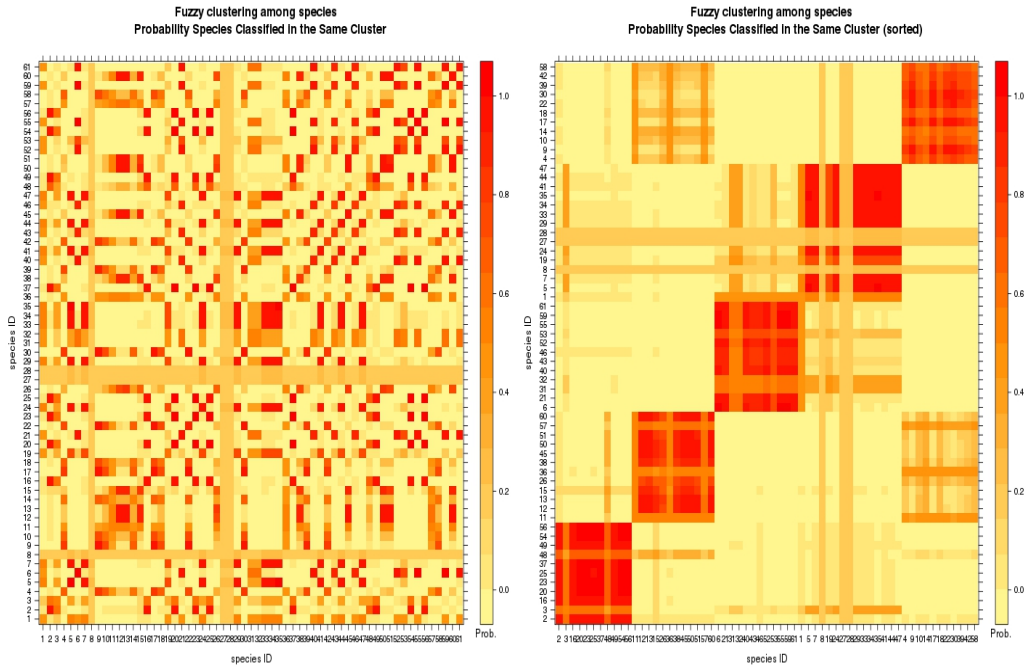
where z_{ir} and $z_{i'r}$ are the posterior probabilities that row i and i' respectively belong to row group r as defined in equation (3.15). The equation above is valid because we assume the independence over the rows conditional on Φ. It is difficult to detect any pattern from Figure 4.27a, because the species are presented by species ID (Table C.1). On the other hand, it is much easier to see that there are five clusters in the tan1301 data in Figure 4.27b, because the species are sorted by the clusters. A strong red

colour indicates that the two species are highly likely to be allocated to the same cluster. Otherwise, an orange colour are the species with a moderate probability and a yellow colour occurs for those species with lower probability of being allocated to the same cluster. It seems that the posterior probabilities that row i and i' belong to the same row group r is sharp for most species, as the graph is mostly dark red or yellow. The species with ID$= 27$ shows moderate probability of being in the same group for all species. This is hoki (Table C.1). Because hoki was caught in all strata in every year (except one stratum in tan1101 trip), hoki has the same probability of being allocated to the same row group for any other species. Fuzziness of clustering from the other datasets are presented in Figure 4.28, 4.29, and 4.30 for tan0201, tan1101, and tan1201 data, respectively. They all show similar patters as Figure 4.27. There are three species that show moderate probability of being in the same group for all species for the tan0201 data (Figure 4.28). They are hoki (ID=27), javelin fish (ID=28), and bollons rattail (ID=8). The same pattern is seen for hoki for the tan1201 data (Figure 4.30). No species has constant probability of being in the same group in Figure 4.29. As it was mentioned before, this is because hoki was not caught in stratum 28 due to the limited time for the survey in tan1101 trip. These graphs have an advantage of showing the fuzzy assignment of rows to clusters based on the posterior probabilities z_{ir}. However, they all show sharp boundaries between the clusters, indicating that most species are assigned to a group with high probability.



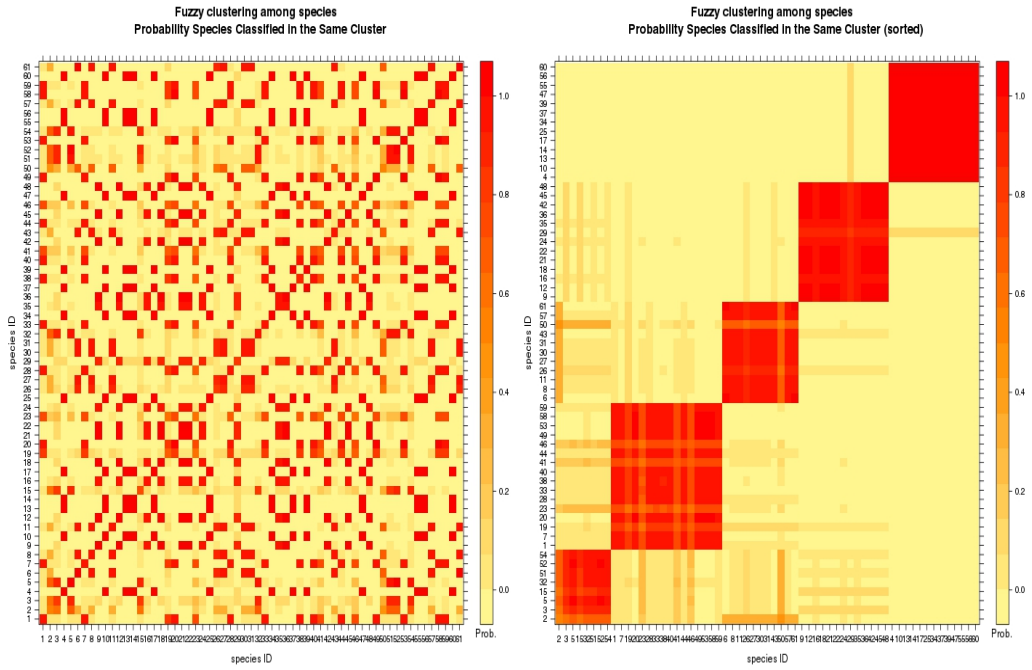
(a) (b)

Figure 4.27: Heat map plot depicting the fuzzy cluster structure for the tan1301 data with $R = 5$ groups. Both axes identify the species (rows). The figure (a) shows the species without any sorting (i.e. as they appear in the original data set). The figure (b) is sorted by the row cluster structure given by Model 10.



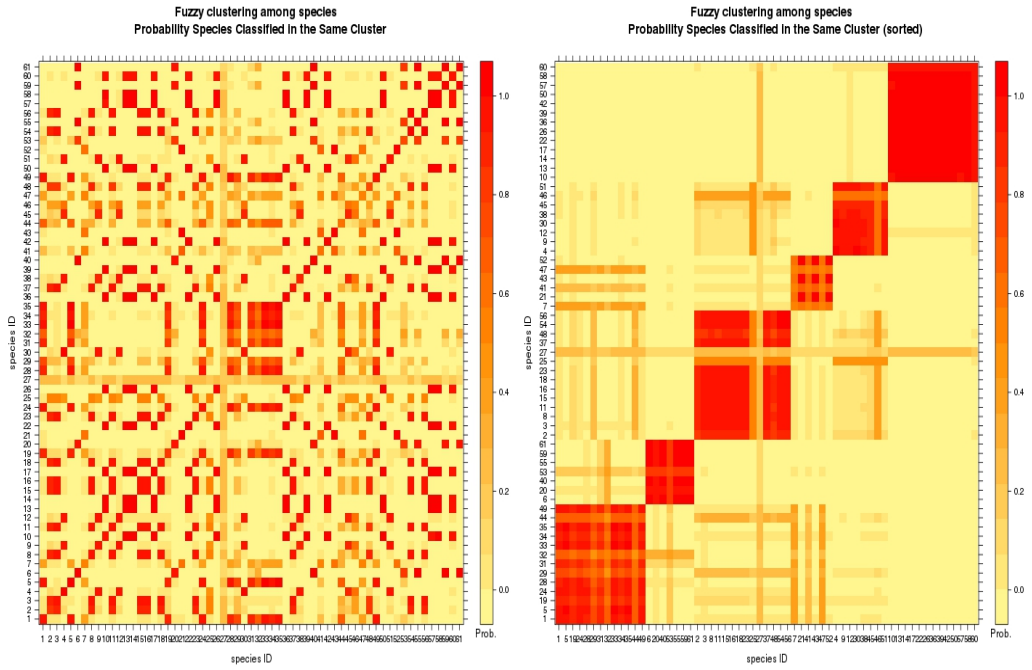
(a) (b)

Figure 4.28: Heat map plot depicting the fuzzy cluster structure for the tan0201 data with $R = 5$ groups. Both axes identify the species (rows). The figure (a) shows the species without any sorting (i.e. as they appear in the original data set). The figure (b) is sorted by the row cluster structure given by Model 10.



(a) (b)

Figure 4.29: Heat map plot depicting the fuzzy cluster structure for the tan1101 data with $R = 5$ groups. Both axes identify the species (rows). The figure (a) shows the species without any sorting (i.e. as they appear in the original data set). The figure (b) is sorted by the row cluster structure given by Model 10



(a) (b)

Figure 4.30: Heat map plot depicting the fuzzy cluster structure for the tan1201 data with $R = 5$ groups. Both axes identify the species (rows). The figure (a) shows the species without any sorting (i.e. as they appear in the original data set). The figure (b) is sorted by the row cluster structure given by Model 10

4.9.3 Any Other Candidate Factors?

In this section, we fitted two further models and found that the species clusters are best presented when we control for frequency of occurrence. We also confirmed that depth is the important factor to explain the data. We presented the results using visualisation tools for easier interpretation. The membership visalisations (Figure 4.19, 4.20, 4.21, 4.22) show that when row standardisation is included the species are no longer clustered by its frequency of occurrence, but only by depth. There are year to year changes in clustering, but the main groups (shallow water, deepwater species) remained constant across the years (Figure 4.18). Figure 4.23, 4.24, 4.23 and 4.26 confirm that we cannot separate the effect of bottom temperature from depth as they are inverse of each other. The descriptions of the species features in each group from the tan1301 data suggest that each group has different depth preference but there is an overlap, especially in the shallow water groups. The results also suggest that there might be other factors contributing to the clusters. Possible factors are spatial location (e.g. latitude, longitude), because we can see the spatial difference between Groups 1 and 2. Another reason is that spatial location has been used to explain species assemblages in previous studies (for example, Francis et al., 2002).

In order to explore the characteristics of the groups in more details, we present the distributions of the frequencies of observation by depth, bottom temperature, latitude, and longitude across all of the strata, but separately by group. That is, f_{rj} from equation (4.8) is referred to the trawl survey database (Mackay, 2000) and we obtain raw data of depth (d_j^*), bottom temperature (t_j^*), latitude (denoted by s_j), and longitude (denoted by e_j) for each species. Figure 4.31 shows the species frequencies according to these four environmental variables for the tan0201 data. It shows species frequencies against depth (first column from left), bottom water temperature (second column), latitude (third column) and longitude (forth col-

umn) for each group (from the top row, Group 1, 2, 3, 4, 5). Overall, species are more likely to be present on the north side of the Chatham Rise (centered at 44°S). This pattern is particularly strong for deep water species (Group 5). This is because the most deep strata are on the north side of the Chatham Rise. There are only two deep water strata, stratum 25 and 28 on the south while there are five deep strata on the north. In addition, there is only one deep strata for tan0201 trip (stratum 22, Figure **??**) which sits on the north side of the Chatham Rise. Group 1, 2, and 3 are shallow water species groups with no obvious difference in depth histograms, with the mean around 500 m for all groups. There is no difference in temperature either, reaching peak at 8 °C. The histograms of latitude overlap each other, suggesting there is no latitudinal difference between these three groups. All three groups show flat frequencies on longitude. Groups 4 and 5 both show deeper depth range than Groups 1, 2, and 3, and are seen in colder water, about 7 °C. The histogram of latitude for Group 4 is similar to the three shallow water groups. The histograms of longitude for Groups 4 and 5 are similar, showing a peak on the west side of the Chatham Rise.

Similar patterns are seen for the rest of the data (Figure 4.32, 4.33, and 4.34). The tan1201 data is the only data that species are clustered to six groups. It seems that there are two middle water groups (Groups 3 and 6). Species in these groups share similar depth and spatial location range, but Group 3 species seems to occur more frequently in warmer water than Group 6.

There is no obvious difference in the latitude range between the shallow and middle water species groups, so we doubt whether latitude could be an important variable to explain species clusters. It is difficult to say whether there is any pattern in the histograms of longitude. They show similar pattern across the groups in each dataset, and there is no significant change from year to year. Our visual inspections suggest that the

spatial information may not be influential to explain the data. Therefore we conclude that spatial location is not effective predictor and the depth is still the most powerful predictor.

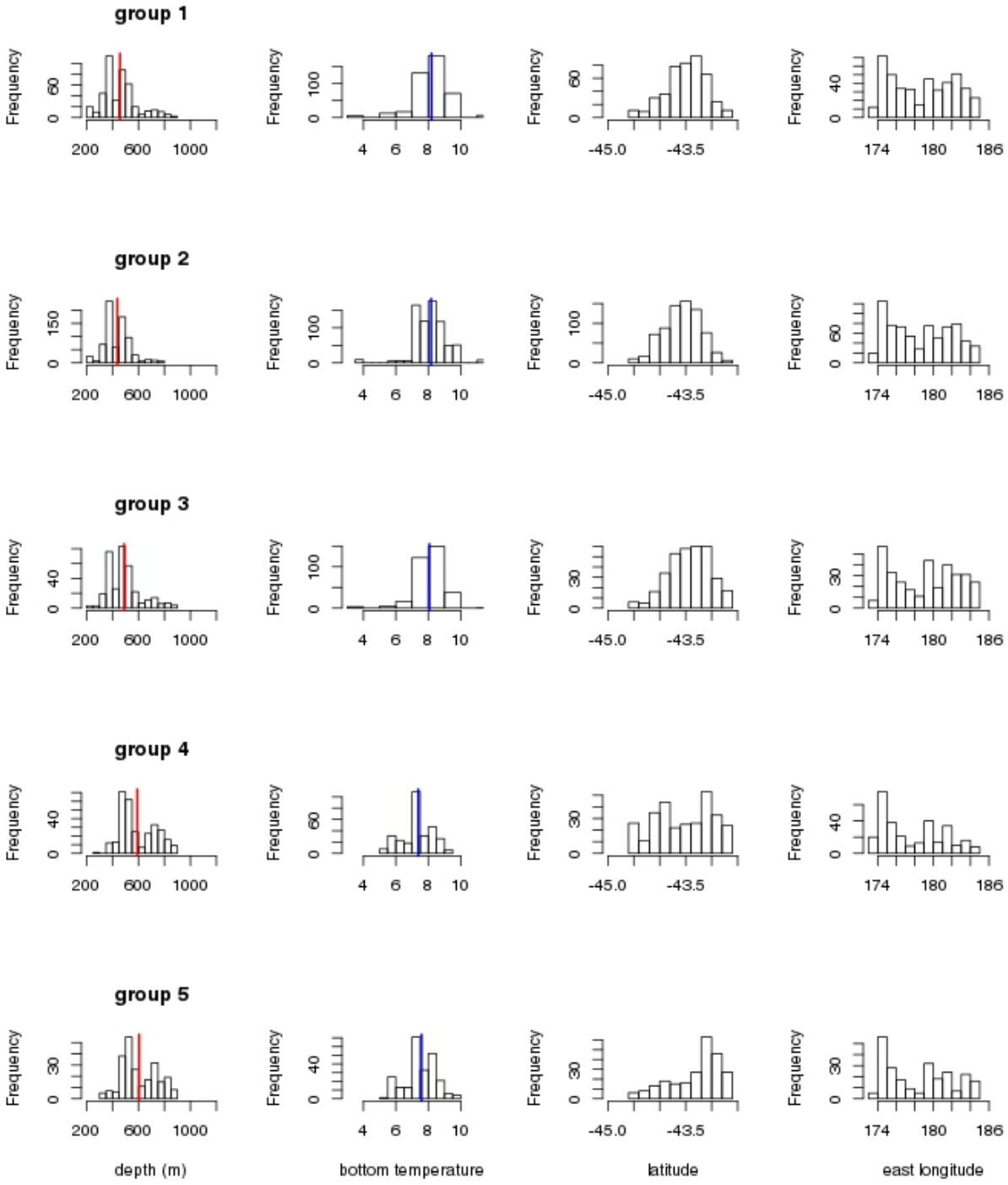


Figure 4.31: Histogram of depth, bottom temperature, latitude, and longitude record obtained from the database for each group selected by Model 10 for the tan0201 data. From left, depth (m), bottom temperature (°C), latitude (Negative sign means south), and east longitude. The vertical lines in the histograms of depth and bottom water temperature are the mean.

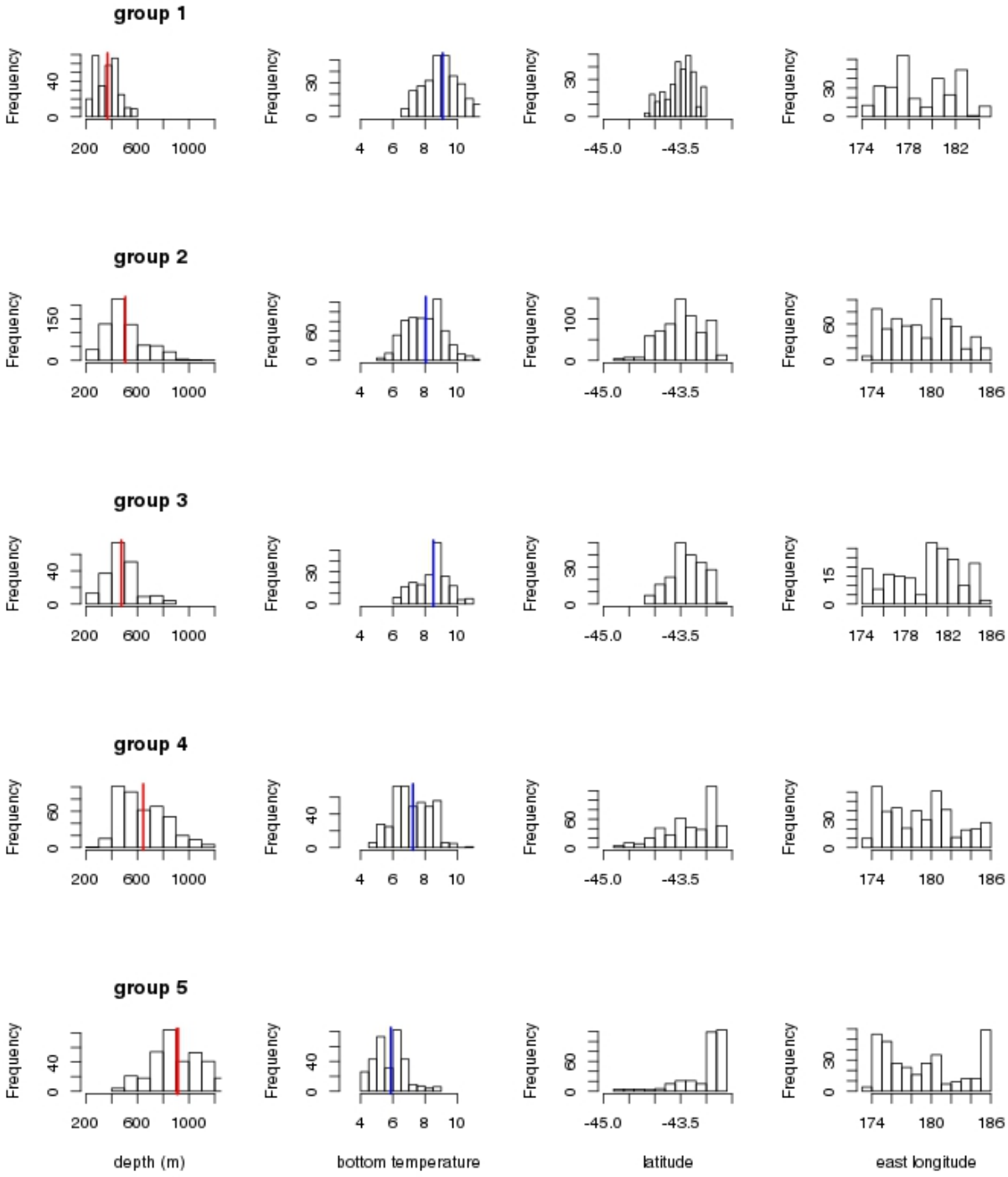


Figure 4.32: Histogram of depth, bottom temperature, latitude, and longitude record obtained from the database for each group selected by Model 10 for the tan1101 data. From left, depth (m), bottom temperature (°C), latitude (Negative sign means south), and east longitude. The vertical lines in the histograms of depth and bottom water temperature are the mean.

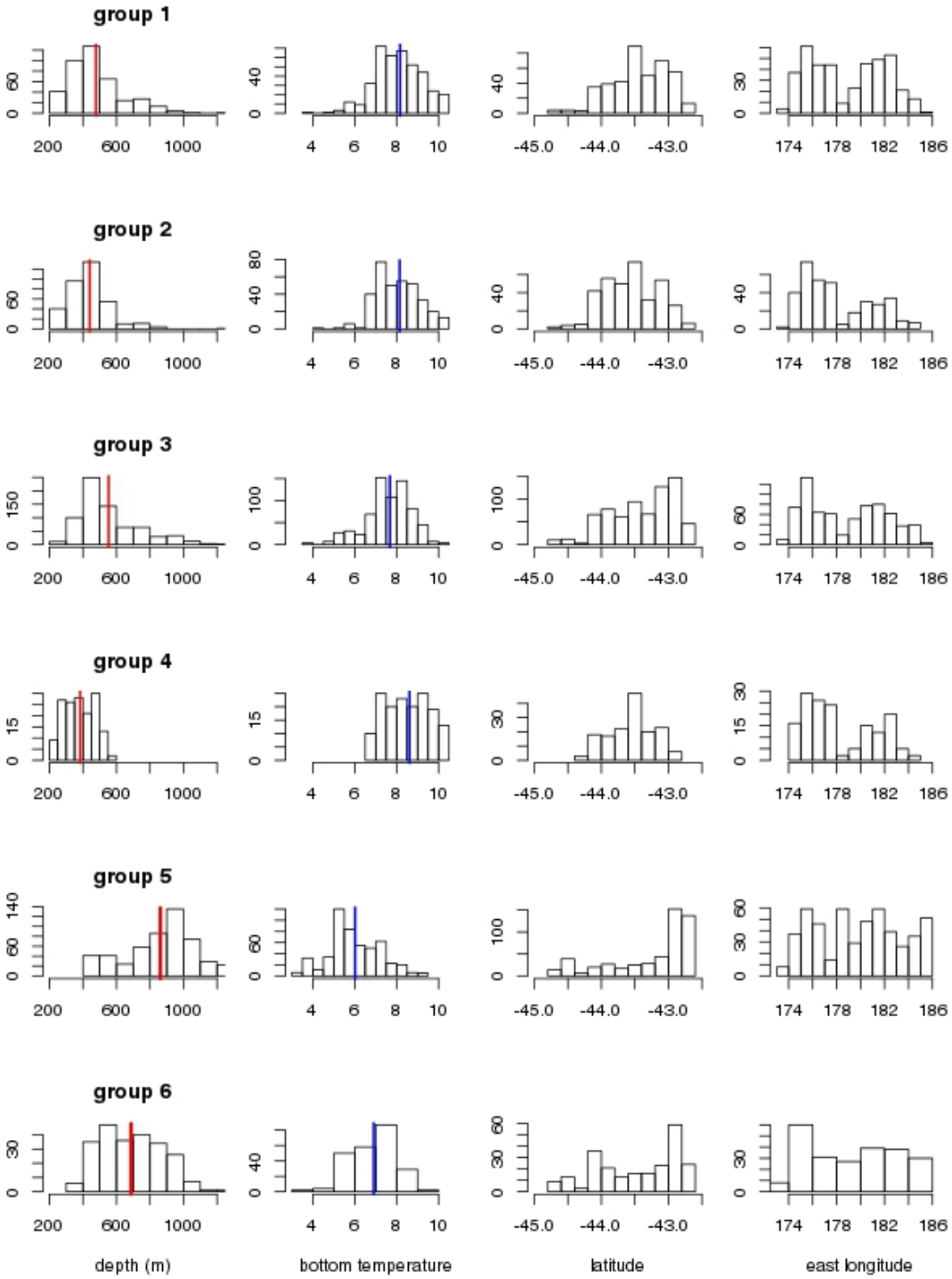


Figure 4.33: Histogram of depth, bottom temperature, latitude, and longitude record obtained from the database for each group selected by Model 10 for the tan1201 data. From left, depth (m), bottom temperature (°C), latitude (Negative sign means south), and east longitude. The vertical lines in the histograms of depth and bottom water temperature are the mean.

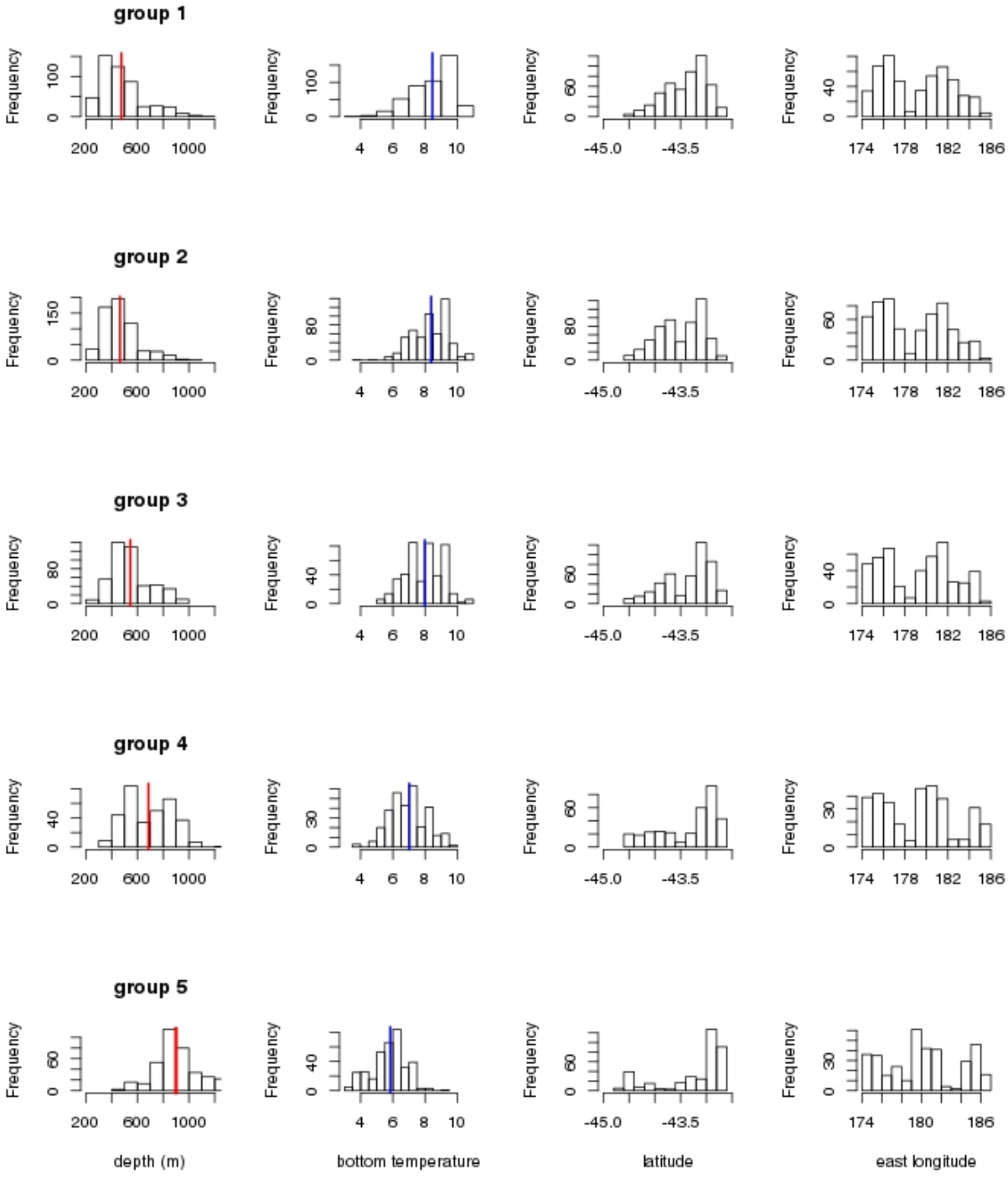


Figure 4.34: Histogram of depth, bottom temperature, latitude, and longitude record obtained from the database for each group selected by Model 10 for the tan1301 data. From left, depth (m), bottom temperature (°C), latitude (Negative sign means south), and east longitude. The vertical lines in the histograms of depth and bottom water temperature are the mean.

Chapter 5

Discussion

In this project we conducted cluster analysis for fish and shark species on the Chatham Rise, using a likelihood based clustering methods via finite mixture models. We implemented a clustering method that for the first time includes covariates and compared them with the existing models. We first programmed from scratch our own implementation of finite mixture fitting using standard **R** functions. The simulation study revealed that our new models might not be so robust for very rare or very common species, and required a long computational time. We discussed the advantages and the disadvantages of the EM algorithm, and introduced a newly developed **R** package, `clustglm` function (Pledger et al., 2015), in order to improve our analysis. We then performed cluster analysis with inclusion of covariates using the `clustglm` routine, and used several visualisation tools to present our results. Our initial analysis, which indicated species were clustered by the environmental variables and frequency of occurrence, prompted us to improve our models by combining the row and column level information. Inclusion of a term that controls the species' frequency of occurrence gave us improved model performance and results that are easier to understand.

5.1 Links between Species Clusters and the Environment

Our results indicate that there is a strong relationship between the species clusters and ocean depth. We also found that sea floor temperature is a significant predictor of species presence/absence, but depth is the strongest predictor. Our main findings support results from Francis et al. (2002) and Leathwick et al. (2003). They studied the relationship between the species richness and the environment in the New Zealand waters, and also concluded that depth is the most important environmental predictor. Leathwick et al. (2003) give us possible explanations why the bottom temperature effect is weak. They said water temperature is highly correlated with depth, which we also found, but they also said that the relationship is not perfectly linear. This is because water temperature becomes more stable with depth due to decreasing influence of other environmental variations (e.g. sunlight). Then temperature becomes more nearly constant, and is no longer related to depth. Another reason is that temperature in shallow water varies with other environmental factors, such as the weather, season, and time of the day (morning, afternoon, and night). But this applies only to regions that are not affected by a subtropical front (Leathwick et al., 2003). A subtropical front is an oceanographic feature creating by warm water currents and cold water currents, and it occurs on the Chatham Rise (Section 2.2). In addition, it is unlikely that the environmental factors can change water temperature deeper than 200 m, because sea water temperature is relatively stable. Having considered that temperature is highly related with depth, temperature in shallow water strata (200-400 m) are too deep to be affected by the environmental factors, and the effect of the subtropical front, water temperature is a weak predictor to explain our data. However, water temperature is time variant whereas depth is not, so water temperature might become more powerful predictor in longer term analysis, or when surveys are carried out in each season.

The Chatham Rise subtropical front creates nutrient rich water that affects fish distribution, so including variables that capture the patterns of the front is important. Dunn et al. (2010) explained that the subtropical front had a wide latitudinal range, up to 100 km. They also said that the surface water temperature gradient could be explained by latitude. We however argue whether latitude is a important factor to explain the data. Our visual inspections (for example, Figure 4.31) show that latitude distribution is different only in deep species groups. But the deepwater species group is clearly separated from other groups by depth. It is uncertain whether latitude is a good predictor for the data, we therefore agree with Leathwick et al. (2003)'s view of the importance of latitude. In their paper they compared several studies on species richness in the New Zealand waters. They stated that spatial information could be used, along with a range of environmental variables, but only in a large scale study. So the inclusion of spatial variables in a regional scale study like ours is unlikely to be important.

Our initial analysis included body length information. It was the only biological information used in the analysis. Our focus on use of the median body length as a predictor was that the body length might explain the trophic level of species on the Chatham Rise, and thereby explains the data. But we found the median body length covariate was not an important predictor. This suggests that the median body length does not explain the trophic levels, or the trophic level on the Chatham Rise is too complex. A study from Dunn et al. (2010) separated chimaera species niche by their diet, the depth and spatial distributions. So where a range of the biological information is available, it may be worthwhile to carry out a cluster analysis including biological information and depth.

5.2 Analytical Considerations

Our ability to analyse and interpret fish species clusters is limited by the use of presence/absence data, rather than abundance. Both common species and rare species are simply recorded as present when they were observed, irrespective of their abundance. Francis et al. (2002) and Dunstan et al. (2011) also raise their concerns about that cluster analysis for presence/absence data may not fully explain the species relationship at a community level. The original data we used were recorded as weights (kg), so some might say we could have used their weights instead. However, using the catch weight as a response variables also hinders species abundance. One big species might weight as much as thousands of a small fish. A possible alternative to our application is to transform the catch weights to a rank (e.g. very abundant, abundant, rare), with a consultation from a fishery expert. Such an approach (though without covariates) was studied by Fernández et al. (2014). Species clustering may occur differently in this method.

We may have found the importance of spatial location for the data if we had included column standardisation term. Our best model, Model 10 (3.20) can be written as $\text{logit}(\phi_{ijr}) = \mu + \alpha_i + \beta_j + \text{depth}_j(\psi_1 + \tau_r) + \text{temp}_j(\psi_2 + \tau_r)$, with β_j controlling variation in the columns. Clustering after controlling variation over rows and columns should be undertaken to fit clustering models to complex ecological data. It is also important to have sensible cutoff value for π_r when a cluster contains only few species. Setting the range of number of clusters before analysis interfere model performance and may result in failing to cluster rare species. When a cluster has few species, it is important to investigate species biological/ecological features, and also to set sensible cutoff value for π_r to provide results that make sense in biological/ecological way.

Other further research directions are

- Investigation of cluster structure of all species observed rather than 61 species. Inclusion of all fish and shark species for analysis may give a different clusters.

- Considering biclustering models that reduce column dimensions. This is already implemented in the `clustglm` function, so the analysis should be straight forward.

- Analysis of spatial effects for clustering. Our visual inspections only indicated that they may not be important. Thereby spatial information needs to be included in models. However, care should be exercised in fitting the models for this datasets, because the strata on the Chatham Rise are defined by depth and longitude (Stevens and Livingstone, 2003). It may be necessary to carry out analysis at the sampling station level (location within each stratum), not at the stratum level.

- Performing cluster analysis on a subset selected species, for example, rattail species, chimaeras, and hoki, whose biological information is more comprehensively available. It may give us deeper insights into how the species distribution changes with their life stage.

Appendix A

Multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution. In the binomial distribution, a random variable resluts from an experiment consisting of n independent trials with two possible outcomes (often called success/failure). Each outcome has a fixed probability; the probability of success is expressed as π and the probability of failure is $1 - \pi$.

We may wish to model trials with more than two possible outcomes. In a multinomial distribution, a random variable from n independent and identical trials can have more than two outcomes, say, K possible outcomes ($K \geq 2$), with each outcome having a fixed probability $\pi_k \quad (k = 1, \cdots, K)$ and $\sum_{k=1}^{K} \pi_k = 1$. Suppose we have n independent and identical trials that can have an outcome in any of K categories. Let y_i be a realisation of random variable Y_i. $y_{ik} = 1$ if trial i has outcome k, and $y_{ik} = 0$ otherwise ($i = 1, \cdots, n$, $k = 1, \cdots, K$). Such dataset can be seen as

$$Y = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & K \end{array} \\ \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0\cdots & 0 \\ 1 & 1 & 1\cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} \end{array}$$

π_k is the probability of that ith trial falls into category k, where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

If we define $y_k = \sum_{i=1}^{n} y_{ik}$, with $\sum_{k=1}^{K} y_k = n$, then the multinomial probability mass function (pmf) is

$$f(Y|n) = P(y_1, \cdots, y_k|\pi) = \frac{n!}{y_1!y_2!\cdots y_K!}\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_K^{y_K}$$

is denoted by

$$Y \sim \text{Multinomial}(n, \pi) \qquad \text{for } n \times k \text{ matrix } \mathbf{Y}$$

For the ith row,

$$Y_i \sim \text{Multinomial}(1, \pi) \qquad \text{for } k \times 1 \text{ vector } \underline{Y}_i$$

Appendix B

Maximun Likelihood Estimation of the Mixing Proportion π_r

In this section, we show how to find the MLE for the parameter π_r analytically. Recall that we impose a constraint on π, $\sum_{r=1}^{R} \pi_r = 1$. When we want to maximise function subject to a constraint, we can use the Lagrange multiplier method. The method of Lagrange multipliers is a strategy for finding the local maximum of a function subject to constraints. In the case of row clustering, we want to maximise the function (3.8) subject to the constraint $\sum_{r=1}^{R} \pi_r = 1$.

We introduce a new variable λ, called *Lagrange multiplier*, to the equation (3.8), and define a new function Q to be

$$\begin{aligned} Q(\Theta, \pi_r, \lambda | y_{ij}, z_{ir}) &= l_c(\Theta, \pi_r | y_{ij}, z_{ir}) + \lambda(\sum_{r=1}^{R} \pi_r - 1) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} z_{ir} \left[y_{ij} \log \theta_{rj} + (1 - y_{ij}) \log(1 - \theta_{rj}) \right] \\ &\quad + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \log \pi_r + \lambda(\sum_{r=1}^{R} \pi_r - 1) \end{aligned} \tag{B.1}$$

To maximise Q, we take partial derivatives with respect to θ_{rj}, π_r and λ, and set them simuteneously to zero. θ_{rj} depends on a set of parameters,

denoted by Ω. We can maximise $l_c(\Theta, \pi_r|y_{ij}, z_{ir})$ numerically over Ω.

However, to maximise over π_r, we have an analytical solution. We take the derivative of the equation (B.1) with respect to π_r to get the MLE:

$$\frac{\partial Q}{\partial \pi_r} = \frac{1}{\pi_r}\sum_{i=1}^{n} z_{ir} + \lambda$$

Set the equation above to be zero,

$$\begin{aligned}
\frac{1}{\pi_r}\sum_{i=1}^{n} z_{ir} + \lambda &= 0 \\
\frac{1}{\pi_r} d_r &= -\lambda \quad (\text{Set} \sum_{i=1}^{n} z_{ir} = d_r) \\
\pi_r &= -\frac{d_r}{\lambda} \qquad \text{(B.2)}
\end{aligned}$$

Since $\sum_{r=1}^{R} \pi_r = 1$,

$$\sum_{r=1}^{R} \pi_r = -\sum_{r=1}^{R} \frac{d_r}{\lambda} = 1$$

We then get the estimate for the $\lambda = -\sum_{r=1}^{R} d_r$, which we put back into (B.2)

$$\begin{aligned}
\widehat{\pi}_r &= -\frac{d_r}{-\sum_{r=1}^{R} d_r} \\
&= \frac{\sum_{i=1}^{n} z_{ir}}{\sum_{r=1}^{R}\sum_{i=1}^{n} z_{ir}}
\end{aligned}$$

Since $\sum_{r=1}^{R} z_{ir} = 1, \forall i$

$$\begin{aligned} \widehat{\pi}_r &= \frac{\sum_{i=1}^{n} z_{ir}}{\sum_{i=1}^{n} 1} \\ &= \frac{\sum_{i=1}^{n} z_{ir}}{n} \end{aligned} \tag{B.3}$$

Lastly, differentiating (B.1) with respect to λ ensures the constraint

$$\begin{aligned} \frac{\partial Q}{\partial \lambda} &= \sum_{r=1}^{R} \pi_r - 1 = 0 \\ \sum_{r=1}^{R} \pi_r &= 1 \end{aligned}$$

Appendix C

List of the Species

Table C.1: Common and Scientific names of species used for the analysis. These species were caught fron all tan0201, tan1101, tan1201, and tan1301 survey. The species are listed in alphabetical order of the species code.

ID	Code	Common name	Scientific name
1	**BAR**	Barracouta	*Thyrsites atun*
2	**BBE**	Banded bellowsfish	*Centriscops humerosus*
3	**BNS**	Bluenose	*Hyperoglyphe antarctica*
4	**BOE**	Black oreo	*Allocyttus niger*
5	**BYS**	Alfonsino	*Beryx splendens*
6	**CAS**	Oblique banded rattail	*Caelorinchus aspercephalus*
7	**CBI**	Two saddle rattail	*Caelorinchus biclinozonalis*
8	**CBO**	Bollons rattail	*Caelorinchus bollonsi*
9	**CFA**	Banded rattail	*Caelorinchus fasciatus*
10	**CIN**	Notable rattail	*Caelorinchus innotabilis*
11	**COL**	Olivers rattail	*Caelorinchus oliverianus*
12	**CSQ**	Leafscale gluper shark	*Centrophorus squamosus*

Continued on next page

Table C.1 – *Continued from previous page*

ID	Code	Common name	Scientific name
13	**CYO**	Smooth skin dogfish	*Centroscymnus owstoni*
14	**CYP**	Longnose velvet dogfish	*Centroscymnus crepidater*
15	**EPL**	Bigeye cardinalfish	*Epigonus lenimen*
16	**EPT**	Deepsea cardinalfish	*Epigonus telescopus*
17	**ETB**	Baxter's dogfish	*Etmopterus baxteri*
18	**ETL**	Lucifer dogfish	*Etmopterus lucifer*
19	**FRO**	Frostfish	*Lepidopus caudatus*
20	**GIZ**	Giant stargazer	*Kathetostoma giganteum*
21	**GSH**	Dark ghost shark	*Hydrolagus novaezealandiae*
22	**GSP**	Pale ghost shark	*Hydrolagus bemisi*
23	**HAK**	Hake	*Merluccius australis*
24	**HAP**	Hapuku	*Polyprion oxygeneios*
25	**HCO**	Hairy conger	*Bassanago hirsutus*
26	**HJO**	Johnson's cod	*Halargyreus johnsonii*
27	**HOK**	Hoki	*Macruronus novaezelandiae*
28	**JAV**	Javelin fish	*Lepidorhynchus denticulatus*
29	**JMM**	Slender jack mackerel	*Trachurus murphyi*
30	**LCH**	Longnose spookfish	*Harriotta raleighana*
31	**LDO**	Lookdown dory	*Cyttus traversi*
32	**LIN**	Ling	*Genypterus blacodes*
33	**LSO**	Lemon sole	*Pelotretis flavilatus*
34	**NMP**	Tarakihi	*Nemadactylus macropterus*
35	**OPE**	Orange perch	*Lepidoperca aurantia*

Continued on next page

Table C.1 – *Continued from previous page*

ID	Code	Common name	Scientific name
36	**ORH**	Orange roughy	*Hoplostethus atlanticus*
37	**PDG**	Prickly dogfish	*Oxynotus bruniensis*
38	**PLS**	Plunkets shark	*Centroscymnus plunketi*
39	**PSK**	Longnosed deepsea skate	*Bathyraja shuntovi*
40	**RBM**	Rays' bream	*Brama brama*
41	**RBT**	Redbait	*Emmelichthys nitidus*
42	**RCH**	Pacific spookfish	*Rhinochimaera pacifica*
43	**RCO**	Red cod	*Pseudophycis bachus*
44	**RHY**	Common roughy	*Paratrachichthys trailli*
45	**RIB**	Ribaldo	*Mora moro*
46	**SBW**	Southern blue whiting	*Micromesistius australis*
47	**SCH**	School shark	*Galeorhinus galeus*
48	**SCO**	Swollenhead conger	*Bassanago bulbiceps*
49	**SDO**	Silver dory	*Cyttus novaezealandiae*
50	**SND**	Shovelnose dogfish	*Deania calcea*
51	**SOR**	Spiky oreo	*Neocyttus rhomboidalis*
52	**SPD**	Spiny dogfish	*Squalus acanthias*
53	**SPE**	Sea perch	*Helicolenus spp.*
54	**SRH**	Silver roughy	*Hoplostethus mediterraneus*
55	**SSI**	Silverside	*Argentina elongata*
56	**SSK**	Smooth skate	*Dipturus innominatus*
57	**SSM**	Smallscaled brown slickhead	*Alepocephalus antipodianus*
58	**SSO**	Smooth oreo	*Pseudocyttus maculatus*

Continued on next page

Table C.1 – *Continued from previous page*

ID	Code	Common name	Scientific name
59	**SWA**	Silver warehou	*Seriolella punctata*
60	**WHX**	White rattail	*Trachyrincus aphyodes*
61	**WWA**	White warehou	*Seriolella caerulea*

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International*.

Anderson, O. F., Bagley, N., Hurst, R., Francis, M., Clark, M., and McMillan, P. (1998). *Atlas of New Zealand fish and squid distributions from research bottom trawls*. NIWA.

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1):3–27.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Dunn, M. (2013). Lecture notes in applied marine biology. University Lecture.

Dunn, M. R., Griggs, L., Forman, J., and Horn, P. (2010). Feeding habits and niche separation among the deep-sea chimaeroid fishes harriotta

raleighana, hydrolagus bemisi and hydrolagus novaezealandiae. *Mar Ecol Prog Ser*, 407:209–225.

Dunn, M. R., Stevens, D. W., Forman, J. S., and Connell, A. (2013). Trophic interactions and distribution of some squaliforme sharks, including new diet descriptions for deania calcea and squalus acanthias. *PloS one*, 8(3):e59938.

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, 222(4):955–963.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley.

Fernández, D., Arnold, R., and Pledger, S. (2014). Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*.

Francis, M. P., Hurst, R. J., McArdle, B. H., Bagley, N. W., and Anderson, O. F. (2002). New Zealand demersal fish assemblages. *Environmental Biology of Fishes*, 65(2):215–234.

Francis, R. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand Journal of Marine and Freshwater Research*, 18(1):59–71.

Francis, R. (2006). Optimum allocation of stations to strata in trawl surveys. *New Zealand Fisheries Assessment Report*, 23:50.

Gordon, A. (1999). *Classification, 2nd Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

He, X., Bigelow, K. A., and Boggs, C. H. (1997). Cluster analysis of long-line sets and fishing strategies within the Hawaii-based fishery. *Fisheries Research*, 31(1–2):147–158.

Horn, P. L. (1994). *Trawl survey of hoki and middle depth species on the Chatham Rise, December 1991-January 1992 (TAN9106)*. MAF Fisheries.

Hurst, R., Bagley, N., Chatterton, T., Hanchet, S., Schofield, K., and Vignaux, M. (1992). Standardisation of hoki/middle depth time series trawl surveys. maf fisheries greta point internal report no. 194. 89 p. *Unpublished report held in NIWA library, Wellington*.

Jennings, S., Pinnegar, J. K., Polunin, N. V., and Boon, T. W. (2001). Weak cross-species relationships between body size and trophic level belie powerful size-based trophic structuring in fish communities. *Journal of Animal Ecology*, 70(6):934–944.

Joel E. Cohen, Stuart L. Pimm, P. Y. J. S. (1993). Body sizes of animal predators and animal prey in food webs. *Journal of Animal Ecology*, 62(1):67–78.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590.

Leathwick, J., Overton, J., and McLeod, M. (2003). An environmental domain classification of new zealand and its use as a tool for biodiversity management. *Conservation biology*, 17(6):1612–1623.

Mackay, K. N. W. (2000). *Database documentation: trawl*. NIWA internal report. NIWA, Wellington. Unpublished report. Not to be cited without permission of the author.

Manly, B. F. (2004). *Multivariate statistical methods: a primer*. CRC Press.

McLachlan, G. and Chang, S. (2004). Mixture modelling for cluster analysis. *Statistical methods in medical research*, 13(5):347–361.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J. and Basford, K. E. (1988). Mixture models: Inference and applications to clustering. *Applied Statistics*.

McLachlan, G. J., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422.

Ministry of Fisheries (2010). National fisheries plan for deepwater and middle-depth fisheries part1b hoki. [Online; accessed 14-July-2015].

O'Driscoll, R., MacGibbon, D., Fu, D., Lyon, W., and Stevens, D. (2011). A review of hoki and middle-depth trawl surveys of the chatham rise, January 1992-2010. *New Zealand Fisheries Assessment Report*, 47:814.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:pp. 71–110.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, 56(2):434–442.

Pledger, S. and Arnold, R. (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, 71:241–261.

Pledger, S., Pledger, L., and Arnold, R. (2015). Linking finite-mixture bi-clustering with generalized linear models. Manuscript submitted for publication.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rezende, E. L., Albert, E. M., Fortuna, M. A., and Bascompte, J. (2009). Compartments in a marine food web associated with phylogeny, body mass, and habitat structure. *Ecology Letters*, 12(8):779–788.

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1):218–227.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Snelder, T. H., Leathwick, J. R., Dey, K. L., Rowden, A. A., Weatherhead, M. A., Fenwick, G. D., Francis, M. P., Gorman, R. M., Grieve, J. M., Hadfield, M. G., et al. (2007). Development of an ecologic marine classification in the new zealand region. *Environmental Management*, 39(1):12–29.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Stevens, D. and Livingstone, M. (2003). Trawl survey of hoki and middle-depth species on the chatham rise, January 2002 (tan0201). *New Zealand Fisheries Assessment Report*, 19:57.

Stevens, D. W., O'Driscoll, R. L., Dunn, M. R., Ballara, S. L., and Horn, P. L. (2012). Trawl survey of hoki and middle-depth species on the chatham rise, January 2011 (tan1101). *New Zealand Fisheries Assessment Report*, 10:98.

Stevens, D. W., O'Driscoll, R. L., Dunn, M. R., Ballara, S. L., and Horn, P. L. (2013). Trawl survey of hoki and middle-depth species on the chatham rise, January 2012 (tan1201). Available at `http://www.mpi.govt.nz/news-and-resources/publications/` (2015/07/14).

Stevens, D. W., O'Driscoll, R. L., Oeffner, L., Ballara, S. L., and Horn, P. L. (2014). Trawl survey of hoki and middle-depth species on the chatham

rise, January 2013 (tan1301). Available at http://www.mpi.govt.nz/news-and-resources/publications/ (2015/07/14).

Wikum, D. A. and Shanholtzer, G. F. (1978). Application of the braun-blanquet cover-abundance scale for vegetation analysis in land development studies. *Environmental management*, 2(4):323–329.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.